An Annotation Language for Semantic Search of Legal Sources

Adeline Nazarenko[1], **François Levy**[1], **Adam Wyner**[2]
[1]LIPN, Paris 13 University – Sorbonne Paris Cité & CNRS
{Adeline.Nazarenko, Francois.Levy}@lipn.univ-paris13.fr
[2]Department of Computing Science, University of Aberdeen
azwyner@abdn.ac.uk

Abstract

While formalizing legal sources is an important challenge, the generation of a formal representation from legal texts has been far less considered and requires considerable expertise. In order to improve the uniformity, richness, and efficiency of legal annotation, it is necessary to experiment with annotations and the annotation process. This paper reports on a first experiment, which was a campaign to annotate legal instruments provided by the Scottish Government's Parliamentary Counsel Office and bearing on Scottish smoking legislation and regulation. A small set of elements related to LegalRuleML was used. An initial guideline manual was produced to annotate the text using annotations related to these elements. The resulting annotated corpus is converted into a LegalRuleML XML compliant document, then made available via an online visualisation and query tool. In the course of annotating the documents, a range of important interpretive and practical issues arose, highlighting the value of a focused study on legal text annotation.

Keywords: annotation, semantic search, legal information, methodology

1. Introduction

Formalizing legal sources has been identified for years as an important challenge for the development of legal content management in order to increase interoperability or support dialogue between various legal systems and actors. The generation of the envisaged formal representation from the legal texts has been far less considered and requires considerable expertise. Nevertheless, enriching the texts with a layer of semantic annotations and opening new access routes to textual content represent a critical issue for all the legal work that rests essentially on the analysis and interrogation of sources.

Legal annotation requires legal skills to understand the text and the significance of legal statements as well as logical skills to understand what conclusions can be drawn with respect to the annotations. In order to improve the uniformity and the cost of legal annotation, it is necessary to experiment with annotations which allow us to query, analyse, and interrogate legal sources.

Our research was driven by requirements set by the parliamentary counsel of the Scottish Government's Parliamentary Counsel Office, which aims to improve internal legislative drafting. As part of this, it is useful to provide a corpus of law in electronically readable form which can be queried to address the following competence questions:

1. What are all the offences and associated penalties or defences?
2. What prohibitions apply to tobacco products?
3. What obligations have been placed on which entities, *e.g.* shop owners?
4. What permissions are given to Scottish Ministers?
5. Given a provision, what are related overriding or reparation provisions?

For parliamentary counsels, who draft the law, the answers to these questions ought to be all relevant provisions so that they can compare formulations or check interactions between provisions. For the technical solution to extracting such questions, semantic meta-data must be added to the text of the corpus. To realise such a corpus, we require an annotation language along with a sound methodology of annotation. The main aim of this paper is to provide initial elements of such a methodology, given some elements of LegalRuleML as an annotation language. In the conclusion, we briefly discuss the extent to which the competence questions were addressed.

As a first experiment, a campaign has been organized to annotate 10 legal instruments provided by the Scottish Government's Parliamentary Counsel Office (41,859 words, $\sim$ 140 pages). All bear on Scottish smoking legislation and regulation. An initial set of guidelines was produced to annotate the text using annotations related to LegalRuleML elements. There was a team of 6 student annotators - 3 from the law school, 2 from the computer science department, and 1 from the linguistics department. This paper reports on this experiment and the guidelines. Relatedly, it presents the interpretative issues that annotators raised during the first annotation phase. It shows how difficult it is to comply with an annotation language and to LegalRuleML semantics. We present solutions to make legal rule annotation feasible on a large scale and with high inter-annotator agreement.

The paper is organized as follows. Section 2. discusses existing work and motivates our choice of annotation. Section 3. sketches the LegalRuleML formalism, which is the target of the annotation process. Section 4. describes the simplified annotation language, which has been designed for legal annotators. In Section 5., we outline our annotation manual, which gives our solutions to how we resolve ambiguities and favor agreement between annotators.

2. Annotation of Legal Texts

Formalization of legal texts has since long been considered for automating legal reasoning (Bench-Capon et al., 1987). Several ontologies have been built for the representation of

legal concepts (Hoekstra et al., 2009; Wyner and Hoekstra, 2012), and there are formal languages for the representation of the content of legal documents, such as SBVR (OMG, 2008) and LegalRuleML (Athan et al., 2015). However, a direct translation from natural language, particularly legal language, into a given formal language is extremely difficult to accomplish. Several transformation steps are required to cover the "formalization continuum" (Baumeister et al., 2011; Lévy and Nazarenko, 2013); controlled languages have been proposed to bridge the gap between the proper languages of legal actors and a logical language supporting formal reasoning (Hoefler, 2012; Feuto Njonko et al., 2014; Lévy et al., 2015; Wyner et al., 2016; Wyner et al., 2017).

Based on the advances of the semantic web, another approach consists in enriching documents with annotations so as to ease the content management which is beyond keyword search, for example, encoding in XML the document structure of legal sources (e.g. Akoma Ntoso[1]) (Casanovas et al., 2016). The identification and resolution of cross references has also been recognized as a key element for the exploitation of legal sources, e.g. Google Scholar case law searches. Following this track, we have laid the foundations for an annotation language that is compatible with LegalRuleML (Nazarenko et al., 2016). In this paper we focus on the semantics of rules rather than on the text structure (Akoma Ntoso), but both types of annotation are complementary and are expected to work together in the end.

Many annotation tasks have been launched for the last decades. Among methodological issues, the guidelines definition and training phase (the annotators learn the task, the target annotation language, and application requirements) are essential (Fort, 2016). The present paper reports on this critical preparatory phase.

3. A Glimpse of LegalRuleML

LegalRuleML has been chosen as the target formal language, since it has been specifically developed to suit legal texts. It is at the crossroad of two sources – RuleML (Boley et al., 2010), which encodes logical rules in a portable way, and Akoma Ntoso, which annotates meta-properties of legal documents.

In Akoma Ntoso, meta-properties include the name of the document, its date, version, its jurisdiction (area where applicable), its classification among legal documents (is it an Act, a Decree, a Statutory Instrument, a decision, etc.), the authorship (Minister, Council, Judge, etc.), the status (in course of elaboration, promoted, ammended, cancelled, etc.)... They also include information about the layout, headings, and provisions as they are numbered in the text. All this information is essential for lawyers who need to know the structure and authority of the text as well as its relation to other texts. We take as given that the document structure is annotated in Akoma Ntoso. RuleML facilitates representation of Propositional and Predicate Logic, as well as negation-as-failure, modalities, and other logical features. RuleML encodes each level of syntactic analysis of a formal rule: e.g. antecedent and consequent parts, the atoms and arguments which build each part, logical coordinators, quantifiers, ...

LegalRuleML adopts much of RuleML and extends it with elements for legal classification of statements (see also Section 4.). A *prescriptive* statement expresses a deontic conclusion such as *permission*, *prohibition*, or *obligation*. A constitutive statement defines the conceptual meaning of the terms, in order to help the interpretation. *Penalty* statements list and describe penalties which can be imposed by an authority, while *reparation* statements make the link between a prescriptive statement and the penalties that apply where the prescription has been violated. LegalRuleML has also a mechanism named *association* to link one or several rules to one or several sources, provided the sources have an IRI. The *context* annotation adds one more level so that several concurrent interpretations can be recorded. Figure 1 is an example of LegalRuleML representation of a legal rule in its XML encoding.

4. A Simple Annotation Language

Encoding legal documents in LegalRuleML to take advantage of standard Semantic Web technologies is difficult to achieve in one step. On the one hand, legal professionals cannot reasonably be expected to engage with the complexity of LegalRuleML conformant encoding of legal documents. On the other hand, it is important to keep track of the text which has been translated. Moreover, legal professionals rather than computer scientists are best placed to understand the legal sources and how they should be annotated. To accommodate our requirements as well as to track the steps and quality of the translations, we only worked with a small palette of LegalRuleML elements as text annotations which classify legal statements and their relations. The selected elements are key parts of the description of rules in LegalRuleML. These annotations leave room for further refinement in subsequent steps:

- Permission: the bearer is allowed to do something or be in a state.
- Obligation: the bearer is bound to do something or be in a state, for otherwise, the bearer is in violation.
- Prohibition: the bearer is bound not to do something or be in a state, for otherwise, the bearer is in violation.
- Constitutive: a legal definition of a concept.
- Override: an indication that one legal rule takes precedence over another.
- Penalty: a description of a sanction.
- Reparation: an indication of a link between a prescriptive norm and a penalty to be applied in case of violation.

Any legal rule relies on one of the prescriptive statement types (permission, obligation or prohibition). Constitutive statements and rule relationships (override and reparation) are essential to the interpretation of legal rules. In the texts that we have considered, nearly all of it falls under one or

[1] http://www.akomantoso.org/

```
<lrml:LegalRuleML ...>
 <lrml:LegalSources memberType="TBD">
    <lrml:LegalSource key="ref1"
      sameAs="http://www.legislation.gov.uk/ukpga/1998/42/contents" />
  <lrml:Associations key="sourceBlock1">
    <lrml:Association>
      <lrml:appliesSource keyref="#ref1"/>
      <lrml:toTarget keyref="#rule_1a"/>
    </lrml:Association>
  </lrml:Associations>
  <lrml:Context key="ruleInfo1">
    <lrml:appliesAssociations keyref="#sourceBlock1"/>
    <lrml:inScope keyref="#stmts_1a"/>
  </lrml:Context>
<lrml:Statements key="stmts_1a">
    <lrml:ConstitutiveStatement>
      <ruleml:Rule key=":rule_1a">
        <ruleml:if>
          <ruleml:Atom>
            <ruleml:Rel>P</ruleml:Rel>
          </ruleml:Atom>
        </ruleml:if>
        <ruleml:then>
          <ruleml:Atom>
            <ruleml:Rel>Q</ruleml:Rel>
          </ruleml:Atom>
        </ruleml:then>
      </ruleml:Rule>
    </lrml:ConstitutiveStatement>
  </lrml:Statements>
</lrml:LegalRuleML>
```

Figure 1: A Conditional Rule and its Source in LegalRuleML

the other of the chosen categories, which we take as indicative of a good selection of annotations.
However, annotators met more ambiguity than expected (see section 5.). In concrete terms, the annotators have to select the relevant statements and to annotate them as in

```
<ConstitutiveStatement id=c20>
In this schedule, ``fixed
penalty notice'' means a
notice offering a person the
opportunity of discharging any
liability to conviction for
an offence under section 1 by
payment of a fixed penalty.
<ConstitutiveStatement>
```

Identifiers are facultative and are used for reference, *e.g.* in an `Override` or `Reparation` statement. For instance, Statement c22 overrides c20 is written `<override over c22 under c20>`.
Whatever their background, annotators had no problem to learn and use this simple language. They could follow the definitions and the explanations given in the manual.
Presently, 558 statements are annotated, the annotation manual has been updated after discussions among annotators, and the corpus can be viewed and queried by a search tool available at `http://tal.lipn.univ-paris13.fr/LexEx`.

5. Annotation Instructions

Annotation instructions must be clear for annotators, but nevertheless allow them to analyse complex cases. In addition, several issues arose. The more important points of discussion are considered here.

Annotation goal Annotation guidelines must explain what is the goal of the task and what kinds of annotations are expected. In our case, annotators were told to concentrate on statements expressing definitions and rules, and thus to skip facts and factual statements since our texts are provisions rather than cases. This means that some fragments of text were left unannotated.

List and nested annotations It was difficult for annotators to identify the borders of the statements and their parts, especially for lists which are numerous and often nested in legal documents. Unless the list items stand for autonomous statements, the annotators were instructed to annotate a whole list as a single statement. This gives a coarse-grained annotation, but directly relates to the source complexity.

Modals Recognizing the deontic status of a statement appeared to be challenging for annotators. They tried to rely on the presence of a modal verb but this is far from obvious. Modal verbs may be ambiguous: *may* or *must* can have an epistemic as well as a deontic meaning – the latter related to permission and prohibition, the former to the degree of certainty. For instance, in

> An 'age verification policy' is a policy that steps are to be taken to establish the age of a person attempting to buy a tobacco product [...] if it appears to the person selling the tobacco product that the customer may be under the age of 25 [...].

may has an epistemic rather than a permission value.
Deontic values can be expressed otherwise than by modals:

- A permission can be considered as an absence of prohibition (the answer to *Is it allowed to... ?*). In such weak cases (von Wright, 1963), there is often no explicit statement, as opposed to strong permissions which are usually explicitly stated (they create exceptions to a prohibition which would otherwise apply).
- Prohibitions are often introduced by *It is an offence to...*. They were often wrongly annotated as constitutive statements (offence definitions). After discussion, we considered that those statements had a performative character and should be tagged as prohibitions.
- Sentences introduced by *It is a defence with respect to [subsection xyz] to...* raised problems for annotators. We classified exceptions to offences as strong permissions. The solution was the same for *A person does not commit an offence...* or *No offence is commited with respect to...*. As this interpretation was not obvious for annotators, dedicated explanations and examples were added to the annotation guide.

Surprisingly, *may* also happens to express an obligation as in *the sheriff may do A, B or C*, which states that the sheriff has to take one action, be it A, B or C.
Another difficulty arose with a distinction between permissions or powers, as in:

> The Scottish Ministers may by regulations amend the age specified in subsection (3).

> A constable making a requirement under subsection (1) may also require the person to supply the constable with the person's name and address....

The preceding examples could be defined as *powers*, which are permissions given to some officials to modify or adapt obligations or prohibitions on other parties. For instance, violation of a power generally entails that some action is ineffective. A similar difficulty arises with *must* and could be solved with *duties*.

Exceptions Exceptions are frequent in legal sources: they describe a general case and state that a different conclusion holds in some specific sub-conditions. Even if the generic and specific cases can be difficult to correlate for the annotators, annotating exceptions in legal sources is of prime importance. However, the wording can be quite different from one case to another. In the following example, the exception immediately follows the general case and is marked by *but* and *only*, but was missed by the annotators:

> (1) An enforcement officer may give a person a fixed penalty notice if the officer has reason to believe that the person has committed an offence under Chapter 1 or 2.
>
> (2) But a person may be given a fixed penalty notice only if the person is aged 16 or over.

When marked by *except that*, *at the exception of*, etc., exceptions are easy to detect, but exceptions are not always semantically self-explaining:

> "care service" has the meaning given by section 47(1) of the Public Services Reform (Scotland) Act 2010, except that it does not include a service mentioned in paragraph (k) of that section (child minding).

Reparations and penalties The majority of reparation statements have a simple form:

> A person who commits an offence under subsection (2) is liable on summary conviction to a fine not exceeding level 3 on the standard scale.

but roughly 40% of these sentences where classified as penalties. The distinction is subtle. The following case

> The fixed penalty for an offence under section 1 is £100.

which looks like a penalty definition, but actually states an offence reparation and describes the related penalty. To clarify, we give priority to reparation and only ask annotators to mark up penalties when they are stated separately. Once the text has been annotated and accepted as correct (peer-reviewed, then adjudicated), it is passed to LegalRuleML annotators, who transform the annotated text into a LegalRuleML compliant XML document. These documents can then be queried on the search tool.

6. Conclusion

In this paper, we motivated and applied the annotation of legal texts using elements related to LegalRuleML, an XML markup language for legal texts. We proposed using a simplified palette of elements from LegalRuleML so as to focus attention on the statement classification and relations. As anticipated, this raised a range of important and interesting interpretive issues, which had to be addressed in the annotation guidelines.
In (Wyner et al., 2017), the corpus was evaluated with respect to the competency questions introduced in Section 1. and the web search application, which retrieves the annotated statements based on their types as well as on the keywords or text patterns they contain. An example of competency questions is searching all the definitions of offences, which could be done to check that something is not prohibited (weak permission). All these definitions involve the

word *offence*. Searching this word yields 70 statements of different kinds. To focus on definitions, we require also that the statement be a `Prohibition`, which reduces to 26 answers. All the erroneously recovered statements specify the procedure which applies in case of offence and not the offence itself.

An other example is searching provisions stating obligations placed on shop owners. They are of course `Obligations`, but *Shop* appears only once in these texts and *owner* never. Checking by hand, *business* is the more usual term, but also *management*, *control*, and *responsible person*. Some further semantic annotation is needed in order to recognize the contextual synonymy of these terms.

The results are high quality. For example, for definitions of offences the recall is 1 and precision is .84; associated defenses, obtained by searching `Permission` elements which contain any of *defence* or *offence*, reach a recall 1 and precision .60; querying *Scottish Ministers* in `Permission` elements yields precision .952 and recall .875. The results also highlighted areas for further refinement. For parliamentary counselors, such results are attractive, producing meaningful and accessible results quickly.

While we have applied a relatively small set of annotations to a modest textual corpus, several valuable lessons were learned that can be taken forward. Working to formalise natural language, it was very helpful to work with a highly scoped annotation language, challenging us understand the interpretation of both the annotation and the language. Addressing competency questions provided a clear and useful goal towards demonstrable solutions. Yet, despite the scoped annotation language and modest corpus, the annotation task proved to be more complex than anticipated, partly due to the complexities of the source language in structure and sense. The observations will be useful in ongoing work. In future work, we plan to extend the corpus based on the revised manual, apply machine learning to the resulting gold standard corpus, and enrich LegalRuleML with elements such as *right*, *duty*, *power*, and *defence*.

7. Acknowledgements

We thank the University of Aberdeen's Impact, Knowledge Exchange, and Commercialisation Award for this 10 week study. The French National Research Agency (ANR-10-LABX-0083) supported this work in the context of the Labex EFL. We also thank the student staff: A. Andonov, A. Faulds, E. Onwa, L. Schelling, R. Stoyanov, and O. Toloch.

8. Bibliographical References

Athan, T., Governatori, G., Palmirani, M., Paschke, A., and Wyner, A. (2015). LegalRuleML: Design principles and foundations. In *Reasoning Web International Summer School*, pages 151–188. Springer.

Baumeister, J., Reutelshoefer, J., and Puppe, F. (2011). Engineering intelligent systems on the knowledge formalization continuum. *International Journal of Applied Mathematics and Computer Science (AMCS)*, 21(1).

Bench-Capon, T., Robinson, G., Routen, T., and Sergot, M. (1987). Logic programming for large scale applications in law: A formalisation of supplementary benefit legislation. In *International Conference on Artificial Intelligence and Law*, pages 190–198.

Boley, H., Paschke, A., and Shafiq, M. O. (2010). RuleML 1.0: The overarching specification of web rules. In Mike Dean, et al., editors, *Proceedings of Semantic Web Rules*, pages 162–178. Springer.

Casanovas, P., Palmirani, M., Peroni, S., van Engers, T. M., and Vitali, F. (2016). Semantic web for the legal domain: The next step. *Semantic Web*, 7(3):213–227.

Feuto Njonko, P., Cardey, S., Greenfield, P., and El Abed, W. (2014). RuleCNL: A controlled natural language for business rule specifications. In Brian Davis, et al., editors, *Controlled Natural Language*, pages 66–77. Springer International Publishing.

Fort, K. (2016). *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. Wiley-ISTE, July.

Hoefler, S. (2012). Legislative drafting guidelines: How different are they from controlled language rules for technical writing? In Tobias Kuhn et al., editors, *Proceedings of CNL 2012*, pages 138–151.

Hoekstra, R., Breuker, J., Di Bello, M., and Boer, A. (2009). LKIF core: Principled ontology development for the legal domain. In *Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood*, pages 21–52.

Lévy, F. and Nazarenko, A. (2013). Formalization of natural language regulations through SBVR structured english. In Leora Morgenstern, et al., editors, *Proceedings of RuleML2013*, pages 19–33, Seattle, WA, USA, July.

Lévy, F., Nazarenko, A., and Wyner, A. (2015). Towards a high-level controlled language for legal sources on the semantic web. In *Workshop on Legal Domain And Semantic Web Applications (LeDA-SWAn 2015)*.

Nazarenko, A., Lévy, F., and Wyner, A. (2016). Towards a methodology for formalizing legal texts in LegalRuleML. In Floris Bex et al., editors, *Proceedings of JURIX2016*, pages 149–154. IOS Press.

OMG. (2008). Semantics of business vocabulary and business rules SBVR. formal specification, v1.0. Technical report, The Object Management Group.

von Wright, G. (1963). *Norm and action: A logical inquiry*. Routledge and Kegan Paul.

Wyner, A. and Hoekstra, R. (2012). A legal case OWL ontology with an instantiation of *Popov v. Hayashi*. *Artificial Intelligence and Law*, 20(1):83–107, March.

Wyner, A., Nazarenko, A., and Lévy, F. (2016). Towards a high-level controlled language for legal sources on the semantic web. In Brian Davis, et al., editors, *Proceedings of CNL2016*, pages 92–101, Aberdeen, UK, July. Springer.

Wyner, A. Z., Gough, F., Lévy, F., Lynch, M., and Nazarenko, A. (2017). On annotation of the textual contents of scottish legal instruments. In Adam Z. Wyner et al., editors, *Proceedings of JURIX 2017*, pages 101–106. IOS Press.