# Semantic Annotations for Legal Text Processing using GATE Teamware

**Adam Wyner, Wim Peters**

Department of Computer Science; Department of Computer Science
University of Liverpool; University of Sheffield
Liverpool, UK; Sheffield, UK
adam@wyner.info; W.Peters@dcs.shef.ac.uk

**Abstract**

Large corpora of legal texts are increasing available in the public domain. To make them amenable for automated text processing, various sorts of annotations must be added. We consider semantic annotations bearing on the content of the texts - legal rules, case factors, and case decision elements. Adding annotations and developing gold standard corpora (to verify rule-based or machine learning algorithms) is costly in terms of time, expertise, and cost. To make the processes efficient, we propose several instances of GATE's Teamware to support annotation tasks for legal rules, case factors, and case decision elements. We engage annotation volunteers (law school students and legal professionals). The reports on the tasks are to be presented at the workshop.

**Keywords:** Text processing, Legal corpora, Web-based annotation

## 1. Introduction

Large, public domain corpora of legal texts are increasing available and searchable. Advanced Scholar Search in Google Scholar makes patents, legal opinions, and journals searchable according to: keywords, author, publication, and collections. The searches can be refined by subject areas, court hierarchy, states, and decision date. Each decision is annotated with respect to decisions cited in it, enabling the presentation of a web of citations to be presented[1] WorldLLI, the global, free, independent, and nonprofit organisation of the Legal Information Institutes, offers search in its database of legal texts, generally by keyword and selected database. Similarly, the United Kingdom offers legislation online The National Archives - legislation, where each act contains links to related acts.

There is, plainly, an enormous volume of textual legal material available. To search for complex information or to make use of it in automated processing, the unstructured textual information of sentences, references, and textual presentation must be structured and made machine readable. To do so, we must annotate corpora of texts with semantic annotations (among other potential annotations, e.g. syntactic) and create Gold Standard corpora, which support the development of rule-based or machine learning algorithms that can be used to annotate large volumes of textual data. In some currently available corpora (those mentioned above), such tasks have been carried out, and we find documents with *linked data*, e.g. references within a text are associated with a *URL*, as well as some *metadata*, e.g. data, location, and judicial context. Yet, clearly, there is a great range of information that can be annotated and used.

Recent work by (Maynard and Greenwood, 2012) shows just how far such an approach - semantic annotation of text, creation of a Gold Standard, and development of automated annotation tools - can go. In this study, 42 terabytes of data from the electronic archives of the UK's National Archives were annotated and indexed with respect to a range of elements: dates, government departments and agencies, measurements, a large knowledge base and associated ontology, and so on. Central to the effort was semantic annotation and the creation of a Gold Standard corpus to evaluate the performance of the system; this was created by four domain experts who manually annotated 13 documents from the source corpus using GATE's Teamware to enter and analyse the annotations (Bontcheva et al., 2010).

Related efforts in the legal domain have created annotation tools for smaller corpora evaluated against relatively constrained Gold Standards for arguments (Moens et al., 2007; Mochales-Palau and Moens, 2008; Wyner et al., 2010), elements of legal cases (Francesconi and Pratelli, 2011; Wyner, 2010), rules and norms (de Maat and Winkels, 2010; Wyner and Peters, 2011), and case factors (Ashley and Brüninghaus, 2009; Wyner and Peters, 2010a). Yet, semantic annotation and the creation of Gold Standards is not, in and of itself, straightforward and unproblematic. Generally, a small number of annotators are deployed on a fragment of the corpus due to the cost and complexity of the task. Moreover, annotation guidance and adjudication are significant issues (Maeda et al., 2008).

In view of the problems and limitations of current annotation campaigns, we suggest a means to broaden participation of annotators. This will allow us to annotate more text with more semantic annotations, leading to higher quality, richer Gold Standard corpora. To do this, we use GATE's Teamware to support annotation tasks for legal rules, case factors, and case decision elements. We engage annotation volunteers (law school students and legal professionals), who are domain specialists. We exercise the tool on a corpus of texts appropriate to the domain - regulations and intellectual property decisions. In the following, we outline GATE's Teamware. In section 3., we mention the approach to annotators, guidelines, and evaulation. The annotations

[1] Accessed April 2, 2012. Search for exact phrase *intellectual property* among legal opinions in California Advanced Scholar Search returns, among others, *Moore v Regents of University of California*. The corpus is based on law.resource.org, which offers bulk access to primary legal materials.

and corpora we work with are discussed in section 4. . We close with a sample screenshot from a previous online annotation exercise. We report on the results of the current compaign at the SPLeT 2012 workshop.

## 2. Teamware

To create high quality, annotated corpora, we need a clear methodology, guidelines for annotators, a means to serve text and annotation tools to annotators, storage of the annotated texts, measures for inter-annotator agreement, and adjudication of annotations (Maynard and Greenwood, 2012). Teamware provides a unified enviroment to carry out these various tasks. The tool is web-based, so no local installation of software is required, and the data is stored in a central repository. The tool supports a range of roles (e.g., annotators, editors, managers) appropriate to different actors and phases of the annotation process, allowing non-specialists to participate in the annotation task. On the other hand, expert curators can then adjudicate the gathered annotations. In addition to supporting users in annotating text, Teamware uses GATE components to *preannotate* the text for a range of annotations, which relieves the annotator of some aspects of the annotation task. Business process statistics are kept on the tasks, representing time each annotator spends per document, percentage of completed documents, and other measures.

## 3. Annotators, Guidelines, and Evaluation

For annotators, we propose to work with contacts in law schools and legal societies to engage them voluntarily in a collaborative task that is similar to the annotating task they already individually engage in to *brief cases*, but using an online tool to annotate, compare, and evaluate corpora created by the users. In future work, we look forward to tying together more closely the annotation tasks with learning objectives in law schools, for example, by using the tool to support legal case analysis and comparison as a basis for student discussion.

To support the annotators in their task, they must be receive guidance. In a small pilot study for online annotation of legal case factors, we provided instructions on how to access and use the tool itself as well as information on the annotations to be identified in the text, e.g. Legal Case Annotation. We expect to extend and expand these instructions for more widespread use.

The tool supports multiple annotators who are annotating the same text with the same annotation set. Thus, a text is multiply annotated and can be compared for *interannotator agreement*, the extent to which annotators agree not just on the selection of annotations on the text, but the exact textual span covered by the annotation. GATE Teamware provides tools for measuring interannotator agreement. In addition, there is an *adjudication* tool so that differences between annotators can be decided in favour of the *correct* or *consistent* annotation. In this way, GATE Teamware supports the development of *Gold Standards*.

## 4. Annotations and Corpora

The target annotations are based on prior work for each of the following topic areas. As each of the sub-topics may have a large range of possible annotations, we make a selection of relevant annotations as a basis for further systematic and controlled development of use of the tool.

### 4.1. Legal Case Factors

To facilitate legal case-based reasoning, the legal case factors must be analysed. We focus on cases concerning *intellectual property* and factors discussed in the cases. For a corpus, we have 140 cases that have been used in the CATO analysis of legal cases (Aleven, 1997). The factors are expressed in (Wyner and Peters, 2010b), which are then further decomposed in (Wyner and Peters, 2010a). From (Aleven, 1997), we have 27 base level factors such as follows, where we have an index F1, a label *Plaintiff-disclosed-information-in-negotiations*, the side of the dispute that the factor favours, a description of the factor, and comments on when the factor does and does not apply. The latter three elements can be used to aid the annotator.

- F1 Plaintiff-disclosed-information-in-negotiations
- Favours defendant.
- Plaintiff disclosed information during negotiations with defendant. The defendant fairly obtained the information and the plaintiff was not interested to maintain the information as a secret.
- Applies if the plaintiff disclosed the information to defendant during negotiations for a joint venture, licensing agreement, sale of a business, etc..
- Does not apply if the defendant learned the information while employed by plaintiff.

Among others, we have the following factors:

- F6 Plaintiff-adopted-security-measures
- F7 Defendant-hired-plaintiff-employee
- F10 Plaintiff-disclosed-information-to-outsiders
- F21 Defendant-knew-information-confidential
- F27 Plaintiff-disclosed-information-in-public-forum

In this task, the objective is for the annotator to annotate the sentence or sentences that indicate the relevant factor in the case decision.

### 4.2. Rules

For the analysis of regulations and legislation, it would be very useful to identify, extract, and process *legal rules*. This part of the task is based on (Wyner and Peters, 2011). As an initial basis, we use a corpus of passages from US Code of Federal Regulations for blood banks on testing requirements for communicable disease agents in human blood. This is a four page document of 1,777 words. The model of analysis proposed includes annotation for:

- Agent and theme, which are semantic roles that must be associated with noun phrases in grammatical (subject or object) roles in the sentence. These are used to account for active-passive alternations and identify the individual's relationship to the deontic concept.
- Deontic modals and verbs.
- Main verbs.
- Exception clauses, which may appear in lists.
- Conditional sentences along with their antecedents and consequences. Antecedents may appear in lists.

### 4.3. Case Elements

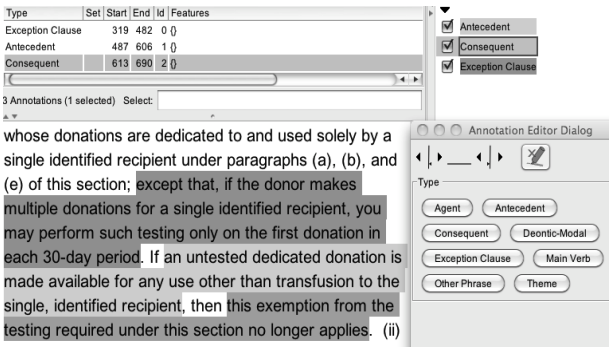
In addition to factors relevant to legal case-based reasoning, we are interested to identify and extract for further processing a range of elements that appear in legal cases (Wyner, 2010; Francesconi and Pratelli, 2011). We use some of the cases for the CATO case base. Among the elements of interest are:

- Case citation, cases cited, precedential relationships.
- Names of parties, judges, attorneys, court sort....
- Roles of parties, meaning plaintiff or defendant, and attorneys, meaning the side they represent.
- Final decision.
- Case structural features such as sections.
- Causes of action.

### 4.4. A Sample

In Figure 1, we have a screen shot of Teamware after a user has annotated parts of the document she has been served. The annotator receives a document in the online tool, highlights a passage, selects an annotation from the *Annotation Editor Dialog*, then moves on to annotate the next passage. In the sample, we have annotations (in colour in the original, but in greyscale in this paper) for an exception clause *except that...30-day period*, an antecedent of a conditional *an untested...recipient*, and the consequent of a conditional *this exemption...applies.* Once annotated by several annotators, we can evaluate interannotator agreement, export the annotated information in XML, and further process it.

Figure 1: Sample of Teamware to Annotate Rules




## 5. Conclusion

In this paper, we have briefly outlined motivation, background, tool, corpora, and target annotations that we study in the annotation exercise. The results of the exercise are to be reported at the SPLeT 2012 meeting, which is part of LREC 2012. We expect that this will be the start of a broader movement to *crowdsource* legal text analytics and semantic analysis on a larger scale, which will yield greater understanding of and use for legal information.

## 6. Acknowledgements

The first author was supported by the FP7-ICT-2009-4 Programme, IMPACT Project, Grant Agreement Number 247228. The views expressed are the author and should not be taken as representative of the project.

## 7. References

Vincent Aleven. 1997. *Teaching case-based argumentation through a model and examples*. Ph.D. thesis, University of Pittsburgh.

Kevin D. Ashley and Stefanie Brüninghaus. 2009. Automatically classifying case texts and predicting outcomes. *Artificial Intellence and Law*, 17(2):125–165.

Kalina Bontcheva, Hamish Cunningham, Ian Roberts, and Valentin Tablan. 2010. Web-based collaborative corpus annotation: Requirements and a framework implementation. In *Proceedings of New Challenges for NLP Frameworks*, Malta, May.

Emile de Maat and Radboud Winkels. 2010. Automated classification of norms in sources of law. In Enrico Francesconi, et al., editors, *Proceedings of SPLeT '10*, pages 170–191. Springer.

Enrico Francesconi and Tommaso Pratelli. 2011. A twofold parsing strategy for italian court decisions. In Katie Atkinson, editor, *Proceedings of JURIX '11*, pages 125–129. IOS Press.

Kazuaki Maeda, Haejoong Lee, Shawn Medero, Julie Medero, Robert Parker, and Stephanie Strassel. 2008. Annotation tool development for large-scale corpus creation projects at the linguistic data consortium. In Nicoletta Calzolari, et al., editors, *Proceedings of LREC '08*. European Language Resources Association.

Diana Maynard and Mark Greenwood. 2012. Large scale semantic annotation, indexing and search at the national archives. In *Proceedings of LREC '12*. European Language Resources Association.

Raquel Mochales-Palau and Marie-Francine Moens. 2008. Study on the structure of argumentation in case law. In Enrico Francesconi, et al., editors, *Proceedings of JURIX '08*, pages 11–20. IOS Press.

Marie-Francine Moens, Erik Boiy, Raquel Mochales-Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of ICAIL '07*, pages 225–230. ACM Press.

Adam Wyner and Wim Peters. 2010a. Lexical semantics and expert legal knowledge towards the identification of legal case factors. In Radboud Winkels, editor, *Proceedings of JURIX '10*, pages 127–136. IOS Press.

Adam Wyner and Wim Peters. 2010b. Towards annotating and extracting textual legal case factors. In *Proceedings of SPLeT '10*, Malta. To appear.

Adam Wyner and Wim Peters. 2011. On rule extraction from regulations. In Katie Atkinson, editor, *Proceedings of JURIX '11*, pages 113–122. IOS Press.

Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. 2010. Approaches to text mining arguments from legal cases. In Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors, *Proceedings of SPLeT '09*, pages 60–79. Springer.

Adam Wyner. 2010. Towards annotating and extracting textual legal case elements. *Informatica e Diritto: Special Issue on Legal Ontologies and Artificial Intelligent Techniques*, 19(1-2):9–18.