For the Automated Mark-Up of Italian Legislative Texts in XML

Andrea Bolioli[1] Luca Dini[1] Pietro Mercatali[2] Francesco Romano[2]
[1] *CELI, C.so Moncalieri 21,10131 Torino, Italy*
[2] *ITTIG, CNR, Via Panciatichi 56/16, 50127 Firenze, Italy*

Abstract. In this paper we will present a method for mining information within legal texts, in particular in regards to corpora of statutes. Text mining, or more in general Information Extraction, can provide a valuable help to people involved in research about the linguistic structure of statutes, and, as a side effect can be the seed for a new generation of applications for the validation and conversion in the legislative domain.

1 Introduction

In order to improve the retrieval of legislative sources on the Internet, parliamentary bodies and public administrations in many countries have begun a process of converting their legislative "deposits" into a standard format to facilitate searching for and displaying texts.

The XML mark-up language seems to be the tool that is proving most successful for this purpose (see, for example, the MetaLex project [1]). This language, in fact, combining its dual nature as a mark-up language and a Web standard, is able to form the common ground for both action "at the source", namely, legislative drafting, and that "downstream" relating to the publication of the texts and the identification of the tools for accessing legislative information [2].

In Italy, the introduction of the XML language for processing legislative texts[1] was proposed and experimented in the "*Norme in rete*" [Law on the Net] Project, solicited by the Ministry of Justice, financed by AIPA (Agenzia per l'informatica nella pubblica amministrazione) and developed under the guidance of the Istituto di Teoria e Tecniche dell'Informazione Giuridica (ITTIG) of the Italian National Research Council. The project has produced, amongst other things, the Document Type Definition (DTD) rules adopted as a standard by AIPA[2] for the on-line publication of Italian legislative instruments.

In order to adopt this language as a standard and, above all, for the conversion of the legislative texts in force into the format provided for by the DTD rules, in our opinion, two factors must interact.

A) Definition and promotion of a "controlled" legislative language

Legistics or the techniques for legislative drafting has introduced unambiguous and recurrent elements into legislative texts, whereby it is possible to identify a more controlled

[1]We prefer using the word 'text' instead of 'instrument' or 'measure', especially when we want to underline that we are referring to co-ordinated lists of signs expressing the legislative instrument rather than simply the instrument itself.

[2]AIPA Circular, 22 April 2002, "Formato per la rappresentazione elettronica dei provvedimenti normativi tramite il linguaggio di marcatura XML" [Format for the Electronic Representation of Legislative Provisions by Means of the XLM Mark-up Language] . The text can be consulted at : http://www.normeinrete.it/standard/circolare-xml.htm.

Andrea Bolioli, Luca Dini, Pietro Mercatali and Francesco Romano, 'For the Automated Mark-Up of Italian Legislative Texts in XML' in T.J.M. Bench-Capon, A. Daskalopulu and R.G.F. Winkels (eds.), *Legal Knowledge and Information Systems. Jurix 2002: The Fifteenth Annual Conference*. Amsterdam: IOS Press, 2002, pp. 21-30.

language in legislative language compared to natural language. In fact, specific rules of orthography, lexicon, syntax, style and structure for the drafting of legislative texts have been adopted. A collection of these rules is to be found in a circular[3] issued in 2001 by the President of the Council of Ministers, and the Speakers of the Chamber of Deputies and the Senate and adopted by the Government and Parliament. The circular updates an earlier one of 1986[4]. For drafting their legislative measures, almost all the Regions in Italy have adopted the "Rules and Suggestions for Drafting Legislative Texts" Manual, a set of rules that are almost the same as the state rules, compiled in 1991 and updated in 2002.

These rules have been applied and complied with in the drafting of legislative texts enacted by the State and Regions since the end of the 1980's. Leafing through the legislative documents, it cannot be said that these rules have, up until now, been strictly and uniformly applied by all law-makers. However, some analyses of sample texts have shown that use of the legislative drafting rules is spreading [5].

The drafting of other normative legislative documents (such as the regulations of local authorities, collective contracts, etc.) is not bound by these rules. It can, however, be said that it is widespread, in practice, to make reference to these drafting rules, even if their application depends on the sensitivity and knowledge of the drafter. On the other hand, many initiatives are underway for the formal and binding adoption of the State/Regional drafting rules by all those persons who produce legislative documents. The "*Norme in rete*" [Law on the Net] Project has contributed to accelerating the process for spreading and receiving the standards for legislative drafting, drawing attention to their utility for electronic processing, whilst still holding that the main purpose of these rules is to guarantee greater clarity in and ability to understand legislative texts [6].

At the same time, research into legal theory, legal language, and legal artificial intelligence have contributed to the definition of the syntactical and semantic structures and to morphological and lexical behaviour peculiar to the legal discourse. Among the authors that have looked at this matter from different points of view see: [7], [8], [9], [10] and, furthermore, reference can be made to numerous studies on normative theory, in particular in the analytically oriented Italian legal philosophers, such as Guastini, Tarello, Conte.

B) Use of tools for natural language recognition

It is evident that the presence of common rules consolidate the definition of text models, which the interdisciplinary studies we have just mentioned describe with ever increasing exactitude. It is also evident that this modelling assists in the automated recognition of the structures of legislative texts and their tagging according to the XML standard. In fact, this tagging will be difficult for the law-maker to obtain as it is extraneous to the tasks and objectives involved in his normal activities. If other professionals do it later, it may provoke an often unsustainable increase in the time needed and the costs involved in building and managing the legislative knowledge base structured according to XML standards.

It is precisely from the perspective of implementing an efficient parsing tool for the automated recognition of the linguistic and structural elements of legislative texts and their subsequent tagging and the conversion of these texts into the XML format, that the research we present here began. The project has, however, an immediate objective, which we can define as recursive: to verify and determine, with the analysis of a gradually widening legislative

[3]Circular 20 April 2001, no. 10888 of the Presidency of the Council of Ministers, "Regole e raccomandazioni per la formulazione tecnica dei testi legislativi" [Rules and Recommendations the Technical Formulation of Legislative Texts], published in the Gazette Ufficiale No. 97 of 27 April 2001. The same rules have also be adopted by the Chamber of Deputies and the Senate with identical circulars by their relative Speakers.

[4]Circulars of the Speaker of the Senate, the Speaker of the Chamber of Deputies and the Presidency of the Council of Ministers of 24 February 1986 (G.U. No. 123 of 29 May 1986, Supplemento ordinario No. 40). For an in-depth illustration of the rules for legislative drafting in Italy and in Europe see: [3], [4].

corpus, how legislative drafting apply and interpret the models inferred from the legislative drafting rules and described by the theoretical research we have just mentioned.

2 Finite State Transducer Methodology and the Extraction of Legislative Knowledge

In the partial analysis of the text through finite state automata and the finite state transducer [11], the appropriate methodology was identified for recognising and tagging the relevant portions of the text.

For this purpose, the project has used a flexible and configurable technical tool: the Sophia 2.1 [12] information extraction system, which enables rules and specific models (already defined or in the course of definition) to be formalised.

In particular, we are working with this software on analysing and tagging the first sample of legislative texts in the following phases:

- normalisation of the text in input, properly tagging all those structures and textual segments that can be recognised on the basis of characters or, in other words, without resort to consultation of the lexicon-dictionary;
- lexical (syntactical category) and morphological analysis of the text in input (inflectional features);
- disambiguation of the syntactical category of the words (*Part of Speech Tagging*);
- partial syntactical analysis (called *chunking*), aimed at identifying the minimum syntactical groups present in the text in input and at grouping them in constituents;
- semantic analysis and identification of the relevant conceptual structures in the text in input;
- conversion of the analysed document from the Microsoft? Word (HTML, RTF, txt, etc.) format into the XML format, according to the established DTD.

3 The Legislative Structures that are the Object of Analysis

The initial phase of the project was focused on identifying and formalising the two conceptual structures of the legislative instrument: the provision for explicit textual amendment or the *novella* and the explicit referral citation, that we have considered:

- to have a high level of formalisation, strictly dictated by the drafting rules and a function circumscribed and sufficiently defined by the legislative drafting technique;
- to carry out an important role in organising and co-ordinating the text and the legislative system, since they are vehicles of supplementary or modifying links[5] between legislative provisions in the same instrument or among different instruments.

The methodology for analysis adopted by the project, both for amendment provisions and for citations is the following:

[5]For normative link we mean “every possible relationship between two (or more) legislative provisions” [13].

- on the basis of the structures laid down by the rules or defined by theory or practice, the creation of one or more models for describing the micro-text to be analysed, and its relationship within the entire structure of the text;
- re-writing of these models into production rules;
- choice of the sample;
- analysis of the sample, examination of the results and definition of the new models described, starting from the identified results;
- further verification of the models obtained in this way on a wider corpus and the transfer of the models into an automated language recognition system.

3.1 Explicit Textual Amendment

Amendment provisions, according to Sartor, fall within the main types of legislative links, classified on the basis of the impact of the legislative link on the legal provision involved. Amendments (or modifications) distinguished from the other large branch of referrals or references, are legislative links characterised by the fact that the active provision affects the passive provision, eliminating it, changing the text or changing the legal significance (whilst leaving the text unchanged). This effect is, instead, lacking in the referral, where the active provision avails itself of the passive provision to complete its meaning, without influencing the latter [13].

In relation to the nature of the impact of the amendment of the provision on the passive provision, we distinguish between textual amendments, time-based amendments (that influence the period of time of the applicability of the passive provision), material amendments, (that amend the legislative content of the passive provision without affecting the text). We shall only look at the first type, the explicit amendments of the text which, traditionally lawyers in Italy call *novelle*.

Indeed, it is perhaps more correct to say that the function of the explicit legislative amendment is expressed through three aspects:

- the structure of the *novella,* which is made up of an introductory part, called *subsection* (in italian *alinea*) [6] and a part that contains the *explicit textual amendment;*
- the characteristics of the amending legislative act and the amended act: indispensable for subsequently being able to reconstruct the amending links between the different legislative sources;
- the citation with which the document to be modified is cited, that expresses the legislative reference (also a textual reference), a fundamental element of the amending provisions.

On the basis of the three aspects mentioned here, we have attempted to define and describe the qualifying elements of the amendment provision. This description, which is set out here, is derived from the rules for legislative drafting and from the analysis of a sample of approximately 100 amending provisions found in 8 State legislative instruments (the four so-called Bassanini Laws and other legislative instruments related to them), enacted between 1968 and 1999.

[6] Understood as the 'part of the provision that introduces the amendment': it contains the purview aimed at specifying the relationship (substitution or integration or abrogation) between the provision in force previously and that provided by the textual amendment. The new sub-section generally ends with a colon, followed by the textual amendment placed between inverted commas.

Type of amending act: indicates the type (law, decree-law, decree of the President of the Republic, legislative decree, etc.) of the legislative act in which the amendment is found; it serves to quickly reconstruct the links between the provisions when there are amendments and it is composed of:

Name of the act, Date, Number: they indicate the essential elements of the amending legislative act, both in the form of the full citation and in its simplified form;

Position of the *novella*: this is the position, within the amending text, where the amendment provision is found, in such a way as to identify the amendment formula with precision, and also to immediately highlight at what level of the structure it is present.

Object of the amendment: this indicates the object of the amendment in the strict sense (or, in other words, whether the amendment affects the entire act, or a part of it, which section, subsection, etc.). This element is also important from the point of view of the structure, because when a part is modified, the effect of the amendment also reflects on that of a directly superior level, in particular, on additions or repeals.

Type act to be amended, composed of:

Name, Date, Number of the act to be amended: with these elements, the characteristics of the legislative act to be amended, the type of act and the essential elements of the document are indicated.

Action: this element describes the action of amendment; it should only take on standard values, sometimes in combination: *repeal, substitution, insertion, addition*, but may take on other values (for example: *replacement*).

Expression: it is the linguistic form with which the amendment is provided for, enclosed by inverted commas or other orthographic signs (colon, brackets, etc.), that delimit the amendments. The expression contains the enunciation that provides for the action, up until the colon that introduces the new text.

The text of the amendment , which, on the basis of the drafting rules, is enclosed within inverted commas and proceeded by a colon.

Furthermore, some textual elements (prepositions, adverbs, conjunctions, etc.) have been identified, which act as connectors and qualifiers of the various elements of the amendment provision [14].

3.2 Legislative Citation

We have already mentioned the fact that the citation expresses the explicit legislative reference or referral, the constituent element of the amendment provision, but it is also an independent element with a linking function that integrates different provisions, whilst not changing the text. Its strict formalisation is, therefore, indispensable for both the recognition of the explicit amendment, and for all automated activities of co-ordination among the legislative instruments.

We have identified what we consider are the constituent and distinctive elements of the legislative citation:

- the **Part** which indicates any part explicitly marked by a particular graphic expression into which the Act is subdivided (for example article, paragraph, letter, chapter, title, etc.) with the relative graphic sign of numeration (example 3, c), 8 A, etc.);
- the **Act (name+date+number):** for Act, we mean the unambiguous identifier of the legislative act referred to, which, usually, is expressed by the official name of the act (for example, law, decree-law, ministerial decree, etc.), the date of promulgation or enactment, and the progressive number assigned to the act.

Two comments need to be made with regard to the name:

- to be unambiguous, the name often needs some specifications to be added to the proper name of the act; for example, to cite a ministerial decree, the Ministry that has enacted it must be indicated, or suffer the consequences of the non unambiguous identification of the cited act;
- usually when a legislative act is cited, found in another act, the name of the container is cited. For example, a regulation is not cited as such, but as a Decree of the President of the Republic or a Decree of the Minister. However, this rule is not always applied (see, for example, citation of the Codes or of consolidations).

Furthermore, legislative citations may be integrated with several elements that link the parts of the citation (in particular, prepositions and punctuation). It is, for example, a frequent practice to use a preposition to link the citation of the parts to that of the act (for example: Article 20 *of the* Statute. . .).

On the basis of the different writing of the element Act, we have classified the various citation formats provided for by the drafting rules, into three categories, called: *normal, simplified, non paradigmatic citations*. We have also described the structure of the Part element which, for reasons of brevity, we shall not illustrate here.

3.2.1 Model of a Normal Citation

Above all, the citation must be unambiguous for identifying that, and only that, specific act to which it refers. Its unambiguous nature almost always is obtained by recording three pieces of information:

the name of the act (which indicates the category to which the act belongs);

the date, usually that of promulgation or enactment (the standard format is: dd (in numbers)/month (in letters)/yyyy (in numbers));

the number that identifies the legislative act (it has a maximum of four numbers).

As an example, let us look at how the model of normal citation is expressed in the module of the formalisation of the semantic structure, in which well-defined elements of the production rules of the lower levels (lexical, morphological and syntactical) are used.

This formalization allows the normal citations to be identified and the template to be generated in output (figure 1).

The elements that make up the rule outlined above need some explanation.

Firstly, it should be noted that this rule has been formalised by accepting all the elements of the citation as obligatory: the name of the act (M-NAME-ACT), the date (DATE), the comma (PUNCTX) between the date and the number of the act and the same number of the

Table 1:

Citation_normal	[M-NAME-ACT]+ DATE:vdate+ PUNCTX+ NUM-ACT:vnum	{actionNEW(vltype:Reference:"Citation_normal", vldefinition:definition_type_act:alllist, vltype:type:alllist, vlauthority:authority:alllist, vdate:date:all, vnum:number:all)}

act (NUM-ACT). We have, that is, written the Citation_normal rule so that it fully corresponds with the Italian legislative drafting rules.

During this phase of the project, we decided to give a single semantic category (NAME-ACT) to all types of legislative instruments. It seemed to us, however, advisable to pose the problem of a semantic categorisation that takes into account the different effectiveness and origin of legislative sources. This further categorisation may be dictated by the need for substantive verifications in the use of the reference from one source to another; for example can a state law cite a regional law and in what way,? Or can the distinction between the sources cited permit or assist in processing the different sources in different ways; for example, can the automated recognition of a regional source assist in selecting the Web site and file where the source is found?

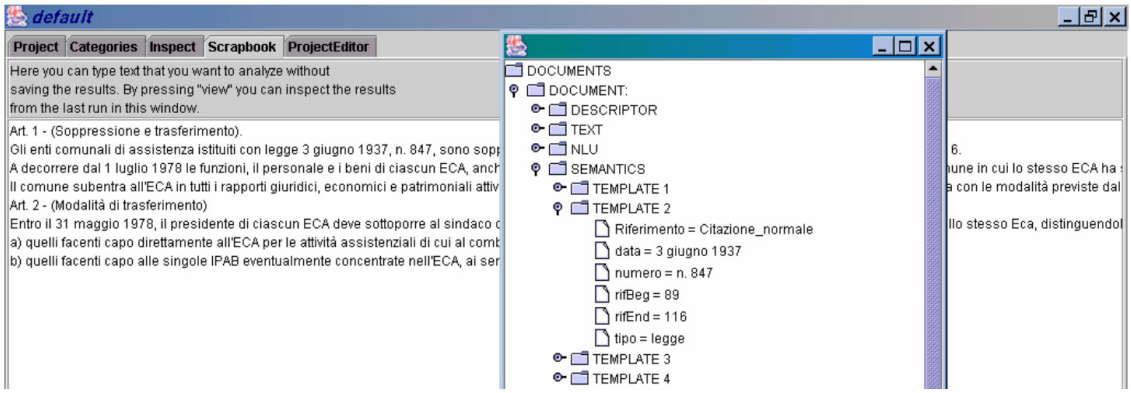

Figure 1:

3.2.2 Model of Simplified Citation

The legislative drafting rules provide for the use of abbreviated or simplified citations within the corpus of the text of the legislative act.

Two types of *simplified* citation are provided for which can be formalised in the following way:

> Act name (full or abbreviated) + Act number/Year enactment-promulgation (example: L. 400/1988 or Law 400/1988);
>
> Act name (full) + **n.** + Number act + **of** + Year enactment-promulgation (example: Law No. 400 of 1988).

3.2.3 Model of non Paradigmatic Citation

The citation of some legislative acts, including the Codes and the Constitution, do not follow the normal model (name act+date+number), but, due to consolidated practice accepted by the

drafting rules, it is expressed by indicating the part, followed only by the full name of the act. In these cases, the format is, therefore, the following:

part+preposition (optional) + name act.
Examples: *Article 345 of Civil Code; Article 3 Constitution.*

The three models we have just described and translated into the formalism of the information extraction system, permits citations that can be called well-formed citations to be recognised; in other words, those written in compliance with the legislative drafting rules.

As we already know (see. parag. 1), the formal rules for writing citations in legislative texts are not always respected. For this reason, based on the results of the analysis in course of the selected legislative corpus, we shall formalise rules for recognising those citations which may be defined as irregular forms of citation.

3.2.4 Irregular Citations

It is evident that the formalisation of irregular citations, uncertain by definition, presents considerable difficulties. In particular, it may be considered extremely difficult to recognise those citations that do not even meet the minimum conditions for lack of ambiguity of the legislative reference, making their specific function fail.

With regard to these forms of citation, we intend to proceed in the following way:

1. to define *a priori* a series of cases of the citations with minimum "deviation" from the normal model of citation like:

   legge 25 aprile 2001, n. 57 articolo 2

2. where the citation is still unambiguous, but the deviation from the normal model is just the elements' order;

3. to try to define *a posteriori* (on the basis of the analysis of a sample and statistical surveys) recurrent irregular citations with greater or lesser deviation from the normal model like:

   articolo 20 della legge 675

4. where the citation is ambiguous and the only possibility to reconstruct the exact meaning of it is to analyse the context.

A second phase of the research begins here which should lead to the extraction of the *template* of irregular or incorrect citations to be verified and, possibly, returned to normality.

The first analyses carried out on the laws enacted in the 1990s making up part of the selected legislative corpus, confirm that, up until now, thanks to the regular citation models we have just described, it is possible to identify and extract more than 95% of the explicit textual legislative references, conforming to the legislative drafting rules.

4 Conversion into XML

Once the elements making up the citations and the amending provisions have been identified in the text, the tagging of these elements in XML is almost immediate. The system for analysing the document, in fact, saves the positions of the elements recognised within the text [15].

It is therefore easy to generate a new document containing the marked-up text, as exemplified in the fragment that follows here:

```
<documento>
  <intestazione start="12" end="73">
    Dlgs 498/97 Decreto Legislativo 30 dicembre 1997, n. 498
  </intestazione>
    "Modifiche alla normativa concernente la posizione di ..."
     pubblicato nella Gazzetta Ufficiale n. 21 del 27 gennaio 1998
     IL PRESIDENTE DELLA REPUBBLICA
     Visti gli articoli 76 e 87 della Costituzione;
     ...Emana

     il seguente decreto legislativo:

     Art. 1.
     ... 1.<modifica azione-tipo="SOSTITUZIONE-COMMA" start="3999" end="3999">
             <riferimento start="3999" end="4065">
               Il primo comma dell'articolo 45 della legge 10 maggio 1983, n.212
             </riferimento>, e' sostituito dal seguente:
             <novella start="4096" end="4691" tipo="struttura">
               "La categoria dell'ausiliaria comprende il personale ..."
             </novella>
           </modifica>.
   ...
</documento>
```

This example was automatically generated starting from the original legislative text and from the output of Sophia 2.1 containing the information about the identified elements.

These elements, namely, the references and the amending provisions, are marked up with the predefined tags, and the attributes of the element are filled in (for example, the ”action-type” attribute where the type of amendment is indicated).

5 Future Prospects

We have already said that the research proposes to build models of the typical structures of legislative texts and to verify, through textual analysis on a sample corpus, their correctness and utility for the implementation of information extraction systems; however, the research also aims at being preparatory for the development of applications that could deal with:

1. integration of Sophia’s linguistic engine with the Lexedit [16], developed at ITTIG. The new system will support structural analysis of legal documents, spelling, syntactic and stylistic errors correction, semantic inconsistencies detection (a similar approach is described in [17]);

2. on the basis of the model implemented for legislative amendments, the possibility of obtaining an application capable of automating the h*yperlinking* mechanism among legislative instruments, aiding, in this way, the creation of a structured network of legislative references and the co-ordination of the texts;

3. on the basis of the parallel development of the “Access to Law on the Net” Project, the possibility will be evaluated of using the procedure created for automating the XML conversion of existing legislative deposits (or those being formed) in those cases where the formal structures are not sufficient for full conversion, but require linguistic and semantic comprehension of the text.

References

[1] A. Boer, R. Hoekstra, R. Winkels, T. van Engers, F. Willaert, Proposal for a Dutch Legal XML Standard. (PDF), in Proceedings of the EGOV 2002 Conference (DEXA 2002).

[2] A. Marchetti, F. Megale, E. Seta, F. Vitali, Marcatura XML degli atti normativi italiani. I DTD di Norma in rete, in *Informatica e diritto*, 1, 2001, pp. 123-148.

[3] R. Pagano, Le direttive di tecnica legislativa in Europa, Camera dei deputati, Rome, 1997.

[4] R. Pagano, Introduzione alla legistica. L'arte di fare leggi, Giuffrè, Milan, 2001.

[5] S. Baroncelli, S. Faro, Tecnica legislativa e legislazione regionale: l'esperienza delle regioni Toscana, Emilia Romagna e Lombardia, in *Iter Legis*, 1998, pp. 173-213.

[6] C. Biagioli, E. Marinai, P. Mercatali, Documento normativo: note esplicative del DTD per i documenti normativi, in *Informatica e diritto*, 1, 2000, pp. 55-106.

[7] C. Biagioli, Ipotesi di modello descrittivo del testo legislativo per l'accesso in rete a informazioni giuridiche, contribution to the Feasibility Study for the Implementation of the "Access to the Law on the Net", 31 gennaio 2000, pp. 1-90, www.ittig.cnr.it;

[8] B. Mortara Garavelli, Le parole della giustizia, Einaudi, Turin, 2001;

[9] G.U. Rescigno, L'atto normativo, Giappichelli, Torino, 1998;

[10] F. Sabatini, Analisi del linguaggio giuridico, in D'Antonio M., Corso di studi superiori legislativi 1988-1989, CEDAM, Padua, 1990.

[11] E. Roche, Y. Schabes (eds.), Finite-State Language Processing, The MIT Press, 1997.

[12] http://www.celi.it/english/sophia.htm

[13] G. Sartor, Riferimenti normativi e dinamica dei nessi normativi, in Il procedimento normativo regionale, Cedam, Padua, 1996.

[14] M.C. De Lorenzo, Modelli di novelle, in *Informatica e diritto*, 1, 2002.

[15] http://www.celi.it/conversione.htm

[16] P. Mercatali, Produzione legislativa, informatica e intelligenza artificiale. Sistema sperimentale per il drafting legislativo, in Mercatali P., Soda G., Tiscornia D., Progetti di intelligenza artificiale per la pubblica amministrazione, Franco Angeli, Milan, 1996, pp. 84-98.

[17] T.M. van Engers, R.A.W., The POWER light-version: Improving Legal Quality under Time Pressure, in http://lri.jur.uva.nl/∼epower (2002).