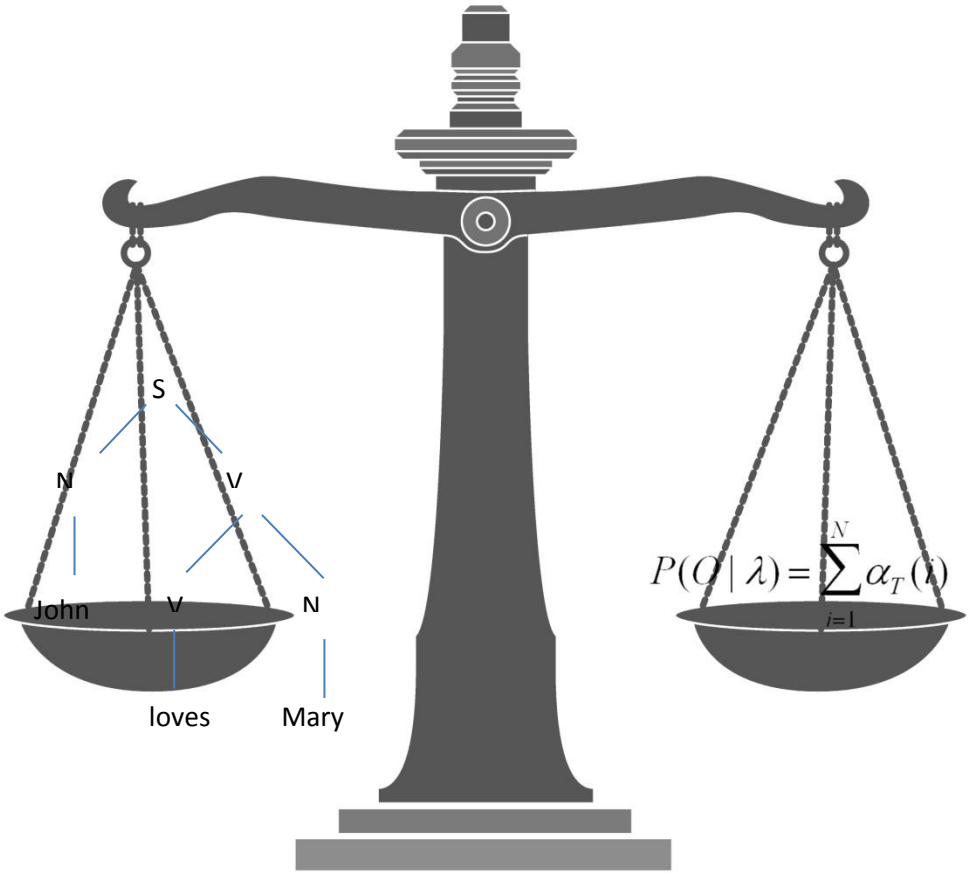
Proceedings of the ICAIL 2011 Workshop

Applying Human Language Technology to the Law



University of Pittsburgh School of Law

June 10, 2011

Preface from the Workshop Co-Chairs

Over the last decade, there have been dramatic improvements in the effectiveness and accuracy of Human Language Technology (HLT), accompanied by a significant expansion of the HLT community itself. Over the same period, there have been widespread developments in web-based distribution and processing of legal textual information, e.g. cases, legislation, citizen information sources, etc. More recently, a growing body of research and practice has addressed a range of topics common to both the HLT and Artificial Intelligence and Law communities, including automated legal reasoning and argumentation, semantic information retrieval, cross and multi-lingual information retrieval, document classification, logical representations of legal language, dialogue systems, legal drafting, legal knowledge discovery and extraction, linguistically based legal ontologies, among others. Central to these shared topics is use of HLT techniques and tools for automating knowledge extraction from legal texts and for processing legal language.

The goal of the workshop is to foster interdisciplinary collaboration between researchers in HLT and those in AI and Law. The workshop is intended to provide a forum for researchers already in this collaborative field. In addition, we hope that the workshop will both introduce HLT researchers to the materials and problems of processing legal language and inform AI and Law researchers about current theories, techniques, and tools from HLT that may be applied to legal language. We hope that interactions among these researchers will promote research and applications in both disciplines.

Adam Wyner, University of Liverpool
Karl Branting. MITRE Corporation

Workshop Organization

Program Co-Chairs:

Adam Wyner (University of Liverpool, UK)
Karl Branting (The MITRE Corporation, USA)

Program Committee:

Kevin Ashley (University of Pittsburgh, USA)
Johan Bos (University of Rome, Italy)
Sherri Condon (The MITRE Corporation, USA)
Jack Conrad (Thomson Reuters, USA)
Enrico Francesconi (ITTIG-CNR, Florence, Italy)
Ben Hachey (Macquarie University, Australia)
Alessandro Lenci (Università di Pisa, Italy)
Leonardo Lesmo (Università di Torino, Italy)
Emile de Maat (University of Amsterdam, Netherlands)
Thorne McCarty (Rutgers University, USA)
Marie-Francine Moens (Catholic University of Leuven, Belgium)
Simonetta Montemagni (ILC-CNR, Italy)
Raquel Mochales Palau (Catholic University of Leuven, Belgium)
Craig Pfeifer (The MITRE Corporation, USA)
Wim Peters (University of Sheffield, United Kingdom)
Paulo Quaresma (Universidade de Évora, Portugal)
Mike Rosner (University of Malta, Malta)
Tony Russell-Rose (Endeca, United Kingdom)
Erich Schweighofer (Universität Wien, Austria)
Rolf Schwitter (Macquarie University, Australia)
Manfred Stede (University of Potsdam, Germany)
Mihai Surdeanu (Stanford University, USA)
Daniela Tiscornia (ITTIG-CNR, Italy)
Radboud Winkels (University of Amsterdam, Netherlands)
Jonathan Zeleznikow (Victoria University, Australia)

Support From:

The University of Pittsburgh School of Law

Contents

The Role of HLT in High-End Search and the Persistent Need for Advanced NLP Technologies

Jack Conrad
Senior Research Scientist
Research & Development, Thomson Resuters

In this talk, I will first discuss the multiple views exploited by a high-end legal search engine like WestlawNext. These dimensions may include the traditional document view (e.g., modified tf.idf scoring of a document relative to a query), a taxonomic view (the classification of a candidate document using an expansive legal taxonomy such as the Key Number System), the network view (documents that are both cited by the instant document and that cite the instant document), the user view (thousands of user interactions with candidate documents including views, prints, finds, etc.). This isn't our father or mother's retrieval system, nor does it closely resemble early versions of Westlaw. It is a powerful and modern search capability that leverages multiple document perspectives. Yet with all this search capacity, there remains an essential role for HLT technologies. Together, we will examine some illustrations of user-driven searches and the imperfect nature of some of the search dimensions presented above. We will then discuss the remaining challenges that present themselves to researchers working in our domain.

Lexical vs. Surface Features in Deceptive Language Analysis

Tommaso Fornaciari
Center for Mind/Brain Sciences
Corso Bettini 31, Rovereto
Università di Trento
tommaso.fornaciari@unitn.it

Massimo Poesio
Center for Mind/Brain Sciences
Università di Trento
and Language and Computation Group
University of Essex

ABSTRACT

Methods for identifying deceptive statements in language could be of great practical utility in court and in other legal situations. Among the best known proposals in this direction are methods proposed by Pennebaker and colleagues relying on the Linguistic Inquiry and Word Count, and tested with language representing a good sample of situations in which deception may be used, but collected in artificial situations. We analyze the performance of these techniques to identify deceptions in genuine court testimonies from criminal proceedings for calumny and false testimony, in which deceptive statements are precisely identified by court judgments, and compare it with that of methods relying exclusively on surface information.

1. INTRODUCTION

Methods for identifying deceptive statements in language could be of great practical utility in court and in other legal situations, e.g., to help the work of the Police Forces, which faces every day situations where they have to evaluate questionable testimonies. Detecting deception isn't easy—humans finds this task difficult, and their performances recognizing deception is not much better than chance [2]. Worse, it seems that specific trainings don't improve their skills [5]. But fortunately stylometric techniques have often been shown to be effective at picking up clues identifying aspects of a text or its author that humans can't spot, for example the authors of anonymous text [6] or particular dimensions of personality [14]. In the case of detecting deception, the hope is to find cues in communication not under conscious control of the person producing the language, that might reveal the deceptive character of a statement. The idea that "statements that are the product of experience will contain characteristics that are generally absent from statements that are the product of imagination" is historically known as Undeutsch Hypothesis [15]. In more formal terms, it could be asserted that, from a cognitive point of view, the elaboration of a false narrative is different from a simple memory recovery, so that some evidences of this difference could be found in the communicative outputs.

The major stumbling block in testing the Undeutsch hypothesis is the availability of texts in which deceptive statements have been annotated. Such texts are not easy to come by, and as a result, most studies of deception study artificially produced language [8, 13]. One of the key characteristics of the work discussed here is that we rely instead on real life data—the (Italian) Corpus of DEception in COURt (DECOUR), currently under construction and consisting of transcripts of criminal proceedings for calumny and false testimony in which the defendant was found guilty. In the sentences issued by the judge for these trials, the defendant's deceptive statements are explicitly listed, often verbatim. This makes it possible to collect deception data with a great deal of reliability.

The work described in this paper had two objectives. First, we intended to evaluate the effectiveness with these real-life data about deception, and for Italian, of lexically-based techniques for deception detection—and in particular, of the methods proposed in [8]—so far only evaluated for English, and with artificially produced data. Second, we intended to compare these techniques with methods relying purely on surface features of the text. The structure of the paper is as follows. We first discuss the lexical-based approach we investigated. We then discuss our methods, and the experimental setting we used to compare techniques; in this section we also discuss out datasets. Results and a Discussion follow next.

2. BACK GROUND

2.1 Stylometry

Stylometry is the study of linguistic style in text, typically through statistical techniques. In forensic linguistics, typical stylometric tasks include author profiling [3, 11], author attribution [7, 6] and plagiarism analysis [12]; another well-established type of stylometric analysis is deducing age and sex of authors of written texts [4].

As Koppel *et al.* (*op.cit.*) point out, the features used in stylometric analysis belong to two main families: surface-related and content-related features. The first type of features includes the frequency and use of function words or of certain n-grams of words or part-of-speech. Such features have been shown to be surprisingly effective in work, e.g., by Daelemans and his lab [6]. The second kind of features specifies information about the semantic content of words,

accessed from dictionaries and lexical resources. Perhaps the best-known lexical resource for deception detection is the Linguistic Inquiry and Word Count (LIWC), created by Pennebaker [9] and used by his group for a number of studies of deceptive language [8]. In addition LIWC has been employed in studies of deceptive language carried out by other groups, such as the work by Strapparava and Mihalcea [13], who obtained good results at classifying into "sincere" or "deceptive" texts collected with the Amazon Mechanical Turk service. Strapparava and Mihalcea used the LIWC for post-hoc analysis only, measuring several language dimensions, as positive or negative emotions, self-references, and so on. So they were able to identify some distinctive characteristics of deceptive texts, but only in descriptive terms: they did'nt make use of the LIWC outputs to distinguish the deceptive texts from sincere ones. Newman *et al.* [8], by contrast, used LIWC to carry out the classification itself.

LIWC includes also dictionaries of languages other than English, among which Italian. We were therefore able to employ the categories of the Italian LIWC dictionary [1] as features to train models aimed to estimate if the statements of our Italian corpus are deceptive or sincere. Our corpus and our analysis units are different from the work of Newman *et al.*, but we followed an analogous methodological path.

2.2 Newman et al.

Newman *et al.* collected a corpus of sincere and deceptive texts through five different studies. In three of them, the subjects had both to describe their true opinions about abortion, and also to try to support the opposite point of view. The opinions were videotaped, typed and handwritten, respectively. The fourth study was videotaped, and the subjects had to express true and false feelings about people they liked or they disliked. Finally, the fifth study, also videotaped, consisted in a mock crime, in which the subjects were accused, rightly or not, of a little theft by an experimenter, and they had to reject any responsibility.

As a result, Newman *et al.* obtained ten groups of texts, five sincere and five deceptive. These texts were preliminarily analyzed using the LIWC. Of the 72 linguistic dimensions considered by the program, the authors selected the 29 variables considered more promising to detect deception. In particular, they excluded the categories that could reflect the content of the texts, those used less frequently in the texts, and those specific of one form of communication (for example the nonfluencies, that are specific of spoken language). At the end, they considered the following list of variables:

- Standard linguistic dimensions:
  1. Word Count;
  2. % words captured by the dictionary;
  3. % words longer than six letters;
  4. Total pronouns;
  5. First-person singular;
  6. Total first person;
  7. Total third person;
  8. Negations;
  9. Articles;
  10. Prepositions;
- Psychological processes:
  11. Affective or emotional processes;
  12. Positive emotions;
  13. Negative emotions;
  14. Cognitive processes;
  15. Causation;
  16. Insight;
  17. Discrepancy;
  18. Tentative;
  19. Certainty;
  20. Sensory and perceptual processes;
  21. Social processes;
- Relativity:
  22. Space;
  23. Inclusive;
  24. Exclusive;
  25. Motion verbs;
  26. Time;
  27. Past tense verb;
  28. Present tense verb;
  29. Future tense verb.

For the analyses, first, the values of the 29 variables were standardized by conversion of the percentages - that are the output of the LIWC - to z scores. Then a 5-fold cross validation was performed, training a logistic regression on the texts of four studies and testing on the fifth. Whereas chance performance was 50% of correct classifications, the authors reached an accuracy of about 60% (with a peak of 67%) in three of the five studies. In the remaining two studies, the performances was not better than chance.

To evaluate simultaneously the five studies, from the 29 LIWC categories, the following five were selected:

1. First-person singular pronouns;
2. Third person pronouns;
3. Negative emotions words;
4. Exclusive words;
5. Motion verbs.

They were the variables that were significant predictors in at least two studies, and also in this case the accuracy of the previsions was about 60%.

3. METHODS

In this work we aimed, first of all, to adapt to Italian the deception detection methods proposed by Newman *et al.*; and secondly, to compare the results obtained this way with those obtained using only surface features. We discuss each method in turn in this Section, and present the results in the next.

3.1 Adapting Newman et al.'s Techniques to Italian

In order to use the LIWC for deception detection, we collect for each utterance features vectors based on the categories of the Italian LIWC dictionary (*op.cit.*). We did not directly employ the LIWC software for tokenization, preferring to make use instead of our tokenization rules. We simply counted out the correspondences in our corpus with the items of the Italian LIWC dictionary, incrementing the scale of the corresponding categories and then normalizing the frequencies so obtained.

We built five kinds of vectors, with the following features:

"Newman 29" First, for uniformity with the work of Newman *et al.*, we selected the features of the Italian LIWC dictionary, corresponding to the categories of the English dictionary employed in the cited work. Due to the fact that the Italian categories for pronouns are larger than the English ones, the 29 categories of Newman *et al.* became 35. These categories are listed in Table 1.

"All" A second model employed as features all the 85 categories of the Italian LIWC dictionary, the words counted and the percentage of the words longer than six letters and captured by the dictionary.

"Our 29" Third, we selected the best 29 features on the basis of the *beta* weights of all variables, as obtained by the models trained with the "All" set of features. These were the LIWC variables with $beta > 1$. Table 2 shows the features and their weight.

"Newman 5" Then, a vector was built reproducing the 5 categories which Newman *et al.* employed to evaluate all their corpus simultaneously. Also in this case, to pass to the Italian categories implied to collect more categories, that is 10. The variables are shown in table 3.

"Our 5" Last, we collected our five features with highest *beta* weights, that is:

English categories	Italian categories
Feeling	Sentim
You	Tu
Sleep	Dormire
Metaphysics	Metafis
Anxiety	Ansia

3.2 Surface strings

The surface features were extracted from a training set, better described below, of 623 utterances. First, we lemmatized and part-of-speech tagged these utterances, using a version of TreeTagger[1] [10] trained for Italian. Then we considered the "true" and the "false" utterances separately, as two independent *corpora*. For each set of utterances, we built six frequency lists, selecting their most frequent items, as follows:

[1] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

Frequency list	Selected
Lemmas	first 200
Bigrams of lemmas	first 200
Trigrams of lemmas	first 200
POS	first 25
Bigrams of POS	first 25
Trigrams of POS	first 25
Total	675

So we collected 675 surface features for each class of utterances. Afterwards we merged the features of both sets. Therefore, theoretically, we could have had a vector of whom the length could vary from 675 features, in case of perfect identity of the features of the two sets of utterances, until 1350 features, in case of no overlap. At the end, we obtained a vector of 1021 features, including two features not related to the frequency lists: the length of the utterances themselves, with or without punctuation. Just the fact that there was not a lot of overlap between the most frequent surface features of "true" and "false" utterances, seemed promising about the possibility to distinguish the two classes.

Table 1: The features of the "Newman 29" vector

English categories	Italian categories
Word Count	Word Count
% words captured by dic.	% words captured by dic.
% words > six letters	% words > six letters
Total pronouns	Pronomi
First-person singular	Io
	Io Ver
Total first person	Noi
	Noi Verb
Total third person	Lui lei
	Loro
	Se
	Lui Verb
	Loro Ver
Negations	Negazio
Articles	Articol
Prepositions	Prepos
Affective/emotional proc.	Affett
Positive emotions	Emo Pos
Negative emotions	Emo Neg
Cognitive processes	Mec Cog
Causation	Causa
Insight	Intros
Discrepancy	Discrep
Tentative	Inibiz
Certainty	Certez
Sensory/perceptual proc.	Proc Sen
Social processes	Social
Space	Spazio
Inclusive	Inclusi
Exclusive	Esclusi
Motion verbs	Movimen
Time	Tempo
Past tense verb	Passato
Present tense verb	Present
Future tense verb	Futuro

Table 2: The features of the "Our 29" vector

English features	Italian features	*beta* weights
Feeling	Sentim	1820.3733
You	Tu	1031.9776
Sleep	Dormire	741.7434
Metaphysics	Metafis	674.1313
Anxiety	Ansia	195.5052
Leisure	Svago	64.6542
School	Scuola	30.2712
Affect	Affett	11.3263
He/She	Lui lei	10.8196
Body	Corpo	10.5046
Humans	Umano	10.1749
Down	Sotto	6.0316
Transitive	Transiti	5.8079
Achieve	Raggiun	5.3449
Conditional	Condizio	4.7417
Anger	Rabbia	4.0103
To be	Essere	3.6371
Space	Spazio	3.3979
You verb	Voi Verb	3.36
To have	Avere	2.8534
Senses	Proc Sen	2.4546
Dictionary	Dic	1.9339
Discrepancy	Discrep	1.7454
Social	Social	1.7434
Number	Numero	1.4363
We verb	Noi Verb	1.4317
Negate	Negazio	1.2563
Certainty	Certez	1.0186
Pronouns	pronomi	1.0133

4. EXPERIMENTS

4.1 The Data

The data used in this work is the (Italian) Corpus of Deception in Court (DECOUR), a collection under construction of transcripts of criminal proceedings for "calumny" and "false testimony", in which the truthfulness or deceptiveness of testimonies is certain and easily verifiable, because when the defendant is found guilty, the trial ends with a sentence which explains the facts and points out the lies told by the subject, often *verbatim*.

At present DECOUR is constituted by the transcripts of 18 testimonies interrogating a total of 17 subjects and collected in the Italian Courts of Trento, Bolzano and Prato. The average age of the subjects is about 36; 14 of our subjects are male, 2 females, and 1 transgender; 8 subjects are from the North of Italy, 2 from the Center, 3 from the South, and 4 from abroad. Finally, we only know the educational level of five subjects: in four cases this is high school qualification, in the last case Italian middle school.

Unlike the study of Newman *et al.*, our analysis units are not whole documents, but the single utterances issued by the subjects. We have 1437 utterances issued by the heard subjects, which appears in the hearings as defendant, witness or expert witness. The utterances of other figures of the hearings, typically the judge, the prosecutor and the lawyer, are by default assumed as true and not considered in this work.

Table 3: The features of the "Newman 5" vector

English categories	Italian categories
First-person singular	Io
	Io Ver
Total third person	Lui lei
	Loro
	Se
	Lui Verb
	Loro Ver
Negative emotions	Emo Neg
Exclusive	Esclusi
Motion verbs	Movimen

Each utterance of the subject being questioned receives a label as regards the truthfulness or less of the utterance itself on the basis of the information found in the sentence issued by the judge. Obviously, between the white of the truth and the black of the falsity, there are wide gradations of gray, and the sentence, that describes the fact and points out the lies of the defendant, can't be used to label each statement issued in the courtroom; we developed therefore a coding scheme taking these issues into account.

The labels used to mark utterances are chosen among these categories:

"False" The utterance is clearly identified in the sentence as false, or its falsity is a logic consequence of some ascertained lie.

"True" The utterances that are consistent with the reconstruction of the facts contained in the sentence, are considered true. Also the utterances that explain something not considered in the sentence, because uninfluential in respect of the investigated facts, are generally considered true.

"Not reliable" An utterance is considered not reliable if it is related to the facts under investigation, but the sentence does not prove its deceptiveness.

"True or not reliable" Like the "not reliable" utterances, the "true or not reliable" ones are related to the topic of investigation, and the sentence nothing demonstrates about them. The only, sometimes slippery difference is that, according to the event and to other statements certainly true or false, and/or on the basis of a weak connection with the interests that the subject tries to defend, it is logical to suppose that they are probably true. In brief, according to the common sense those utterances should be true, but the fact is not demonstrated, and ultimately questionable.

"False or not reliable" This is the specular situation in respect of the previous point.

"Undecidable" The utterances that, from a logical point of view, cannot be neither true neither false, are considered undecidable. This is the case of lot of questions

(like "Excuse me, can you repeat?"), but also of several utterances stopped in mid-sentence, that haven't a complete sense. This is also the case of the utterances that have a meta-communicative function, and regulate the relations between actors, like "Now I explain." or "If you think so..." and so on.

The amount of the labeled utterances and of their tokens (with and without punctuation) is shown in the following table.

Label	Utterances	Tokens	
		with punct.	without punct.
False	333	5778	4802
True	537	7908	6628
Not reliable	225	3351	2746
True or not reliable	83	1758	1452
False or not reliable	78	1648	1360
Undecidable	181	1146	886
Total	1437	21589	17874

Only the utterances labeled as "true" and "false" were used in our study, and the other ones are discarded. We obtained therefore a corpus of 870 utterances, of whom about 61.7% are "true" and 38.3% are "false".

4.2 The logistic regression

To carry out the analyses, the corpus of 870 "true" or "false" utterances was split in this way:

- 10 hearings were used as training set, for an amount of 623 utterances: it means about 72% of the corpus, in terms of utterances. It is also the part of corpus from which we collected the features of the surface vectors;
- 4 hearings were used as test set, for an amount of 148 utterances, equal to 17% of the corpus.
- 4 hearings were used as development set, for error analysis and so on.

Using the training set above mentioned, we built models performing logistic regression in the Weka package[2]. We employed separately the vectors made by the content features of the Italian LIWC dictionary (*op.cit.*), and the vectors of surface features collected from the training set. The test set was employed for the classification task.

4.3 Chance levels

To evaluate the results of the analyses, we defined our chance level through a Monte Carlo simulation. The test set had 81 "true" utterances and 67 "false", it means respectively 54.73% and 45.27%. 10000 times, a random simulator simply produced 148 previsions, obtaining the result "true" with p = .5473.
Comparing the simulated results with the test set, we found that less than 1% of simulations exceeded the 60% of "correct answers". So we assumed the 60% of correct classifications as threshold for our test set.

[2]`http://www.cs.waikato.ac.nz/ml/weka/`

5. RESULTS

5.1 The content features vectors

The results of the experiments with content features vectors are shown in Tables 4, 5, 6, 7 and 8. The performances of "Newman 29" and "All" vectors are similar and clearly higher than chance level. The "Our 29" features also did better than chance level, but the results are less good. "Newman 5" and "Our 5" vectors, instead, did not exceed the chance level. In other words, the feature selection techniques we used do not seem to be very useful—in general, the more features are employed in the vectors, the better the results.

Always, the fluctuations in performance are due to different levels of effectiveness in detecting deceptive utterances. "Newman 29", "All" and "Our 29" vectors, indeed, have exactly the same accuracy detecting "true" utterances. But the worst models are progressively blind to deceptiveness, and tend to evaluate all utterances as "true": the "Our 5" vector, for example, judges "true" 146 of the 148 utterances of the test set! Also for this reason, the recall of "true" utterances is always high. The crucial challenge, therefore, is to discover the "false" utterances: the recall of the best vectors is few less than .5, until almost 0 of the worst ones. However, the best vectors reach high levels of precision in detecting deception, close to .9. This means that, if on the one hand it is not simple to recognize deceptive utterances, on the other, when models judge an utterance as deceptive they are unlikely to be wrong. The same precision is not found regarding the "true" utterances: it is due to the tendency of the models to "see" "true" utterances, with advantage for the recall, and disadvantage for the precision.

5.2 The surface features vectors

The results of the experiments with surface features vectors are shown in Table 9. The model trained with surface features also achieves results well above chance level—indeed, almost as good as those with the best content features vectors. This difference is mainly due to "false positive" errors: more utterances are classified as "false" even if they are not. In fact, in this model the precision in detecting "false" utterances is lower (although consequently the recall is slightly better).

6. DISCUSSION

Even if the 29 features of Newman *et al.* were selected for English texts, they are very effective with Italian testimonies, as well. The "Newman 29" vector is the best, but performs better than "all" only because it well classified a single "false" utterance more than the other one: so their results are substantially equivalent. This confirms the reasonable supposition of Newman *et al.*, that to exclude from the vectors the features related to the topic of the texts does not result in worse performance. Moreover, the "Our 29" vector, which collects the most weighty features of the "All" vector, contains also content-related features, and their performances are inferior to "All" and "Newman 29" vectors. It seems that, if typical features of deceptive language exist, they should not be found in the topic of the speech. It is also possible that it could be simply damaging to exclude selectively only some content-related features, creating imbalances in evaluating specific topics.

Table 4: Logistic regression - "Newman 29" vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	33	34	0.868	0.493	0.629
True utterances	76	5	0.691	0.938	0.796
Total	109	39			
Total per cent	**73.65**%	**26.35**%			

Table 5: Logistic regression - "All" vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	32	35	0.865	0.478	0.615
True utterances	76	5	0.685	0.938	0.792
Total	108	40			
Total per cent	**72.97**%	**27.03**%			

Table 6: Logistic regression - "Our 29" vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	21	46	0.808	0.313	0.452
True utterances	76	5	0.623	0.938	0.749
Total	97	51			
Total per cent	**65.54**%	**34.46**%			

Table 7: Logistic regression - "Newman 5" vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	4	63	0.5	0.06	0.107
True utterances	77	4	0.55	0.951	0.697
Total	81	67			
Total per cent	**54.73**%	**45.27**%			

Table 8: Logistic regression - "Our 5" vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	1	66	0.5	0.015	0.029
True utterances	80	1	0.548	0.988	0.705
Total	81	67			
Total per cent	**54.73**%	**45.27**%			

Table 9: Logistic regression - Surface features vectors

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-Measure
False utterances	35	32	0.729	0.522	0.609
True utterances	68	13	0.68	0.84	0.751
Total	103	45			
Total per cent	**69.59**%	**30.41**%			

Differently from Newman *et al.*, instead, the smaller feature sets do not perform well in our corpus. This is probably due to the fact that our analysis units - the utterances - are considerably shorter than the texts of their study, and therefore they need to be defined by a lot of features, to be adequately identified.

Our results show that using LIWC does in fact result in slightly better performance than using surface features alone, but not by much, which suggests that reasonable results at deception detection could be obtained with resource-poor languages as well. On the other hand, experiments in progress combining both content and surface features suggest that this combination may result in improved performance.

Our results also show that our subjects did spend some effort to conceal their lies. In the Monte Carlo simulation, in fact, less than 1% of the simulation had a recall of "true" utterances better than 63%. Our models based on the Italian LIWC dictionaire categories, instead, show a clear bias, so that they tend to judge as "true" a lot of utterances, and their recall is never lower than 93.8%... at the expense of the recall of "false" utterances. Therefore, to detect deception is just a task which consists in finding out hidden items in a multitude. The good news is that, when an utterance is recognized as "false", the models trained with content features are probably right. It would be crucial in a real life scenario, where it would be very important to be confident about the previsions carried out. This could be a practical reason why the content features seem to be better than the surface ones, regardless of their overall accuracy.

The moral could be that, in the context of the hearings in front of the judge, there are "false" utterances that are linguistically similar - or identical - to the "true" ones. Maybe, they can't be recognized with tools of textual statistics. But also there is a portion of "false" utterances - maybe about 50%, like our results suggest? - which is different in the style from the "true" ones. We hope that this portion could be used to support and to orientate police investigations and judges's decisions, especially in cases in which other kind of evidence are scarce or absent.

7. ACKNOWLEDGEMENTS

The data collection for this work is really complex, and it couldn't have been realized without the help of a lot of people. Many thanks to Dr. Heinrich Zanon, President of the Court of Bolzano, to Dr. Sabino Giarrusso, President of the Court of Trento, and to Dr. Francesco Antonio Genovese, President of the Court of Prato. Many thanks also to Dr. Piero Tony, Chief Prosecutor of the Public Prosecutor's Office of Prato, to Dr. Biagio Mazzeo, Prosecutor in the Public Prosecutor's Office of Genova, and to Rita Fava of the Public Prosecutor's Office of Prato.

8. REFERENCES

[1] A. Agosti and A. Rellini. The italian liwc dictionary. Technical report, LIWC.net, Austin, TX, 2007.

[2] C. F. Bond and B. M. De Paulo. Accuracy of Deception Judgments. *Personality and Social Psychology Review*, 10(3):214–234, 2006.

[3] M. Coulthard. Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25(4):431–447, 2004.

[4] M. Koppel, J. Schler, S. Argamon, and J. Pennebaker. Effects of age and gender on blogging. In *In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[5] T. R. Levine, T. H. Feeley, S. A. McCornack, M. Hughes, and C. M. Harms. Testing the Effects of Nonverbal Behavior Training on Accuracy in Deception Detection with the Inclusion of a Bogus Training Control Group. *Western Journal of Communication*, 69(3):203–217, 2005.

[6] K. Luyckx and W. Daelemans. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 513–520, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[7] F. Mosteller and D. Wallace. *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, 1964.

[8] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards. Lying Words: Predicting Deception From Linguistic Styles. *Personality and Social Psychology Bulletin*, 29(5):665–675, 2003.

[9] J. W. Pennebaker, M. E. Francis, and R. J. Booth. *Linguistic Inquiry and Word Count (LIWC): LIWC2001.* Lawrence Erlbaum Associates, Mahwah, 2001.

[10] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, September 1994.

[11] L. M. Solan and P. M. Tiersma. Author identification in american courts. *Applied Linguistics*, 25(4):448–465, 2004.

[12] B. Stein, M. Koppel, and E. Stamatatos. Plagiarism analysis, authorship identification, and near-duplicate detection pan'07. *SIGIR Forum*, 41:68–71, December 2007.

[13] C. Strapparava and R. Mihalcea. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In *Proceeding ACLShort '09 - Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009.

[14] S. A. Sushant, S. Argamon, S. Dhawle, and J. W. Pennebaker. Lexical predictors of personality type. In *In Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

[15] U. Undeutsch. Beurteilung der Glaubhaftigkeit von Aussagen [Veracity assessment of statements]. In U. Undeutsch, editor, *Handbuch der Psychologie: Vol. 11. Forensische Psychologie*, pages 26–181. Hogrefe, Gottingen, Germany, 1967.

Legal Thesauri Reuse

An Experiment with the U.S. Code of Federal Regulations

Núria Casellas*
Legal Information Institute
Cornell University
Ithaca, NY, 14850, USA
Institut de Dret i Tecnologia
Uni. Autònoma de Barcelona
Bellaterra, 08193, Spain
nuria.casellas@cornell.edu
nuria.casellas@uab.cat

Joan-Josep Vallbé†
Dept. of Political Science
Universitat de Barcelona
Barcelona, Spain
vallbe@ub.edu

Thomas R. Bruce‡
Legal Information Institute
Cornell University
Ithaca, NY, 14850, USA
tom.bruce@cornell.edu

ABSTRACT

This paper relates preliminary results towards assessing the possible problems related to the reuse of existing thesauri for the search and retrieval of legal information. In particular, it focuses on the study of the Code of Federal Regulations and its related thesaurus, the Thesaurus of Indexing Terms. The relationship between the content of a thesaurus and the content of the objects which represents may be of different nature depending on the thesaurus' production process. Here we would like to study the related nature of the content of the Code of Federal Regulations and the Thesaurus of Indexing Terms. In order to assess the extent to which the indexing terms and the content of the CFR headings are equally representative of the content of a CFR title, we use a scaling method that will deliver the position of each CFR title with respect to a one-dimensional space, in particular, the document scaling algorithm WORDFISH. Results show that , although there is in general a strong correlation between the scores received by the CFR titles in both experiments, the existence of mismatches suggests that part of corpus of indexing terms is not representative of the actual content of the CFR titles based on the analysis of their headings. This, in turn, may suggest that the direct reuse of an existing thesaurus for textual search and retrieval could, in some cases, lead to inefficient results.

1. INTRODUCTION

The Code of Federal Regulations (CFR) is the codification of the general and permanent rules published in the Federal Register by the executive departments and agencies of the Federal Government of the United States. Promulgated rules are is compiled by the Federal Register in the Code of Federal Regulations (CFR) under a specific title. Federal regulations are at the core of governmental action and they affect people's lives in various aspects: "[t]he air we breathe, the water we drink, the jobs we hold, and the general welfare of our families and friends are increasingly protected and defined by rules issued by federal agencies of various sorts" [14]. Thus, the content of the CFR is immensely varied.

The information contained in the CFR is varied, detailed and complex, and searching for specific regulatory material is a difficult and time consuming task. Search and retrieval of information could be enhanced if assertions could be made about:

- regulatory bodies (e.g., agencies, etc.);
- regulated objects (e.g., products, manufacturing processes, behaviors, activities etc.);
- regulated subjects (e.g., professions, etc.);
- regulated processes (e.g., administrative procedures, processes for appeal, etc.);
- regulatory scope (e.g., time, place, etc.).

The reuse of existing thesauri, controlled vocabularies or taxonomies in a machine-readable form could allow semantic search and retrieval enhancement through the indexing and annotation of the content of the text of the CFR.[1] Moreover it would allow the combination of ontology supported search, free text search, or facet search, together with the exploitation of the CFR structural information currently modeled and published in XML.

The study and use of thesauri and controlled vocabularies to support legal information search and retrieval is extensive, and the use of Semantic Web ontology languages such as RDF/RDFS (Resource Description Framework/Schema) [18] and OWL (Ontology Web Language) [12], which offer machine-readable semantic metadata, could enhance the

*Assistant professor at the School of Law of the Universitat Autònoma de Barcelona, and a researcher of the Institute of Law and Technology. She is currently a post-doctoral visiting researcher at the Cornell Legal Information Institute under a Spanish grant MEC/Fulbright 2010-2012.

†Assistant professor of Political Science at the University of Barcelona, collaborator of the UAB Institute of Law and Technology, and member of the UB Research Group of Local Studies (GREL). This research was carried out during a post-doctoral research visit at the Cornell Legal Information Institute, during September 2010-January 2011.

‡Director of the Legal Information Institute (LII), Cornell University.

[1]Previous research in this direction is discussed in: [16, 8, 9].

storage, search and retrieval of information and knowledge, together with human-computer interaction (see, for example, [20, 19, 11, 21, 2, 13, 7]). "The aim of the Semantic Web is to allow much more advanced knowledge management systems", and to overcome current limitations regarding searching information, extracting information, maintaining information, uncovering information and viewing information [3, 4]. Thus, the purpose of the Semantic Web is to expose the meaning of data on the Web in a standard machine-readable form that allows users (applications) to connect and integrate these data and to discover new information (knowledge) through its relationships. Berners-Lee and colleagues [6] described the Semantic Web as an extension of the Web "in which information is given well-defined meaning, better enabling computers and people to work in cooperation".

Therefore, the Semantic Web may be understood as an extension of the current Web, sometimes is also referred to as Web 3.0, enriched with semantic metadata, with *meaning*. These efforts also include the Web of Data (or Linked Data[2] efforts), which relies on the existence of standard formats that allow the access and query of interrelated datasets.

A specific RDFS/OWL development is SKOS (Simple Knowledge Organization System), which allows the representation of controlled vocabularies, thesauri, taxonomies and folksonomies used in knowledge organization systems. The SKOS specification acts as a thesauri development standard for Web reuse,[3] where "[t]he elements of the SKOS data model are classes and properties, and the structure and integrity of the data model is defined by the logical characteristics of, and interdependencies between, those classes and properties. This is perhaps one of the most powerful and yet potentially confusing aspects of SKOS, because SKOS can, in more advanced applications, also be used side-by-side with OWL to express and exchange knowledge about a domain. However, SKOS is not a formal knowledge representation language".

The conversion of existing thesauri into the SKOS specification is an increasingly used technique for the publication of thesauri towards reuse, and Linked Data enabling. For example, the EuroVoc Thesaurus, a multilingual thesaurus that includes terms about all the activities of the European Union, and that is used by the Eur-Lex application to enable keyword search for all legal documents produced in the EU, has been recently published in its SKOS version.[4]

This paper investigates the possibilities offered and the possible problems related to the reuse of existing thesauri for the search and retrieval of legal information. In particular, it focuses on the study of the Code of Federal Regulations and its related thesaurus, the Thesaurus of Indexing Terms. The relationship between the content of a thesaurus, defined by the ANSI/NISO Z39.19-2005 standard as a "a controlled vocabulary arranged in a known order and structured so that the various relationships among terms are displayed clearly and identified by standardized relationship indicators", and the content of the objects which represents may be of different nature depending on the thesaurus' production process. Here we would like to study the related nature of the content of the Code of Federal Regulations and the Thesaurus of Indexing Terms. This is part of a preliminary research towards developing a methodological approach to the evaluation of thesauri reuse for information management, retrieval and search.

1.1 The Code of Federal Regulations

In the United States, each federal regulation is compiled in the Code of Federal Regulations under a specific title: "The Code of Federal Regulations (CFR) is an annual codification of the general and permanent rules published in the Federal Register by the executive departments and agencies of the Federal Government".[5] The CFR is divided in 50 titles that represent broad areas subject to Federal regulation: agriculture, food and drugs, judicial administration, energy, etc.

These titles are updated once per year and on a quarterly basis:

- Titles 1-16 are updated as of January 1st
- Titles 17-27 are updated as of April 1st
- Titles 28-41 are updated as of July 1st
- Titles 42-50 are updated as of October 1st

At the same time, the Office of the Federal Register publishes daily (Monday to Friday) rules, proposed rules and notices of Federal agencies and organizations, together with executive orders and presidential documents in the Federal Register. This official publication, created in 1935, "remains not only the daily compendium of almost all activities of the executive branch agencies, but also a principal mechanism for permitting citizens to know about and participate in agency decision making in a timely, uniform manner" [10].

Regulations are compiled in the CFR under a specific title according to their subject matter, in a similar manner as the non-positive law titles of the U.S. Code (USC). "Each title is divided into chapters, which usually bear the name of the issuing agency. Each chapter is further subdivided into parts that cover specific regulatory areas. Large parts may be subdivided into subparts. All parts are organized in sections, and most citations in the CFR are provided at the section level".[6] Each of these divisions is identified by a heading that specifies generally the content of the text immediately below it.

1.2 The Thesaurus of Indexing Terms

The Thesaurus Of Indexing Terms "includes indexing terms that describe the specific program regulations of individual agencies as well as general administrative regulations common to all agencies. The indexing terms included are

[2]For more details on Linked Data efforts and community visit `http://linkeddata.org` and `http://www.w3.org/standards/semanticweb/data`.
[3]SKOS: `http://www.w3.org/2004/02/skos/`.
[4]`http://eurovoc.europa.eu`.
[5]`http://www.archives.gov/federal-register/cfr/about.html`.
[6]As described in: `http://www.gpoaccess.gov/cfr/about.html`. See also 1 CFR §8.1-8.9, [1, 10].

intended to express and organize the often technical regulatory concepts in research terms familiar to laypersons".[7]

This list of indexing terms is used by the Office of the Federal Register "as the basis for the subject entries in the Code of Federal Regulations Index which is published annually as of January 1. Federal agencies also use the Thesaurus to prepare the 'List of Subjects' which is included in rule and proposed rule documents submitted for publication in the Federal Register".[8]

These terms, subjects, are identified by agencies according to section 1 C.F.R. §18.20 "Identification of subjects in agency regulations":

(a) Federal Register documents. Each agency that submits a document that is published in the Rules and Regulations section or the Proposed Rules section of the Federal Register shall–

  (1) Include a list of index terms for each Code of Federal Regulations part affected by the document; and

  (2) Place the list of index terms as the last item in the Supplementary Information portion of the preamble for the document.

(b) Federal Register Thesaurus. To prepare its list of index terms, each agency shall use terms contained in the Federal Register Thesaurus of Indexing Terms. Agencies may include additional terms not contained in the Thesaurus as long as the appropriate Thesaurus terms are also used. Copies of the Federal Register Thesaurus of Indexing Terms are available from the Office of the Federal Register, National Archives and Records Administration, Washington, D.C. 20408.

At the moment, the National Archives/Federal Register offers online access to:[9]

- The CFR Subject List arranged by Title and Part (January 1, 2004)[10]
- An alphabetic list of all indexing terms with a series of notations under each term to refer users to preferred or related terms (November, 16 1995)[11]
- A grouping of terms under 19 subject categories[12]

[7] http://www.archives.gov/federal-register/cfr/thesaurus.html.

[8] http://www.archives.gov/federal-register/cfr/thesaurus.html. It is also mentioned that the Federal Register Index is issued monthly in cumulative form, based on a consolidation of the "Contents" entries in the daily Federal Register.

[9] The paper print publication of the Indexing Terms, revised January 1, 2010 is 785 pages long and contains references to CFR parts.

[10] http://www.archives.gov/federal-register/cfr/subjects.html

[11] http://www.archives.gov/federal-register/cfr/thesaurus-alpha.txt

[12] http://www.archives.gov/federal-register/cfr/thesaurus-categories.txt

ACCEPTANCE OF CONTRIBUTIONS	
	See: CONTRIBUTIONS
ACCOUNT	
	Allocation between federal and Levin, See: ALLOCATION OF EXPENSES
	Allocation between federal and nonfederal, See: ALLOCATION OF EXPENSES
	Credit union, disbursements from, Sec. 102.9(b)(2)(iii)
	Established by collecting agent, Sec. 102.6(c)(4)
	Federal, separate from nonfederal, Sec. 102.5(a)(1)(i) and (b)(1)(i)
	Levin, See: "LEVIN" FUNDS
	Office, See: OFFICE ACCOUNT
	Transmittal, for joint fundraising, Sec. 102.17(c)(4)
	See also: CAMPAIGN DEPOSITORY
ACCOUNTANTS' SERVICES	
	See: LEGAL AND ACCOUNTING SERVICES
ACT	
	Definition, Sec. 100.18

Table 1: Indexing information contained at the section level in CFR title 11.

Thus, we may assume that when a set of regulations is compiled under one title, all the index terms attached to each regulation may be thought of as being also attached to that title. The objective of this preliminary research is to present and discuss the possibilities offered by the reuse of these index terms—mainly, to discover the representativeness of these indexing terms with respect to the content of the Code of Federal Regulations, in particular, to the content of title, (sub)chapter, (sub)part, and (sub)section headings. The following sections will describe the method used and the results obtained.

2. METHOD

In order to assess the extent to which the indexing terms and the content of the CFR headings are equally representative of the content of a CFR title, we use a scaling method that will deliver the position of each CFR title with respect to a one-dimensional space, in particular, the document scaling algorithm WORDFISH [23, 22]. Although WORDFISH was created with the aim of estimating policy positions from political textual data, it can in principle be applied to other kinds of text assuming the existence of an underlying spatial dimension on which documents may be scaled. For instance, it has been used to measure the influence of lobbies in European legislation [15].

The main advantage of WORDFISH compared to other methods for document scaling such as WORDSCORES [17] is that while the latter relies on a reference set of documents previously coded as representing both extreme points of the policy dimension, WORDFISH simply draws on term frequency in existing texts and does not need a training set. The underlying assumption of this algorithm is that words in documents occur following a Poisson distribution. This distribution is usually employed for events that have a rare likelihood of

occurring repeatedly in a short period of time or in a reduced physical area [5]. As it is usually the case, few words occur most of the time in the text, while the vast majority of words in a text occur very few times. For example, in the case of the corpus of indexing terms—containing 1,050 different terms and a total of 64,378 occurrences—the median occurrence is 9, and the third quartile is a frequency of 41, while one term occurs 7,165 times. The basic assumption here refers to the way these terms occur in the text—i.e., that "the number of times an actor i mentions word j is drawn from a Poisson distribution" [22].

The Poisson distribution has only one parameter, λ, which represents its mean and its variance at once. The model may be represented as follows:

$$y_{ij} \sim Poisson(\lambda_{ij})$$

$$\lambda_{ij} = exp(\alpha_i + \psi_j + \beta_j * \omega_i)$$

where y_{ij} is the frequency of word j in CFR title i, α is a set of CFR title fixed effects, ψ is a set of word fixed effects, β is an estimate of a word specific weight capturing the importance ow word j in discriminating between positions, and ω is the estimate of CFR title i's position in the one-dimentional space of substantive-procedural activity.

Word fixed effects (ψ) control for the fact that some words are used much more often than other words in all CFR titles, and similarly document fixed effects (α) are used here as control for the fact that some documents contain more words than others—i.e., some agencies attach more indexing terms to their regulations than others.

Summing up, the most interesting parameters are, in principle, ω (the estimate of the position of a CFR title) and β (the estimate of the discriminant weight of a word).

3. PREPARING THE DATA

3.1 CFR subject index

The first set of data for analysis is the corpus of indexing terms from the information provided on the National Archives website,[13], a list of subjects by title revised as of January 1, 2004. As mentioned before, these indexing terms have been attached to each title by the Federal Register and the regulatory agencies.

From this data, we produced a corpus of 50 different documents—i.e., one for each CFR title—each of them containing the indexing terms attached to every CFR part under that title. For further analysis, the document related to CFR Title 2 was removed because it contained no terms. As indexing terms are, in this data, attached to the parts of the titles, once they are aggregated per title, they may be duplicated. We maintained this duplication in the analysis, since the WORDFISH algorithm relies on the relative term frequency of terms in documents. Using `jFreq`[14] we first convert all terms to lower case and remove irrelevant information such as numbers.

[13] `http://www.archives.gov/federal-register/cfr/subjects.html`

[14] `http://www.williamlowe.net/software/`

3.2 CFR headings

The second set is the one provided by the text contained in the headings of the title, (sub)chapters, (sub)parts, and (sub)sections, which was extracted querying the XML database (per title) for the 2009 revision of the CFR.[15]

While the set of indexing terms was rather small as regards the count of different terms (1,050), the number of different terms appearing in the corpus of headings is more than 25 times bigger: it has 28,137 different terms. For the sake of algorithmic efficiency, those terms that occur only in one document were removed from the analysis, as suggested by [22]. Once this operation was complete, the corpus of CFR headings contained 10,859 different terms which occurred at least in two different documents.

3.3 Two term-document matrices

The two corpus are organized in a matrix-like form for analysis, specifically a term-document matrix—i.e., a matrix in which rows are terms and columns are documents (in our case CFR titles). Therefore, we obtain two different matrices, one that contains the indexing terms and the list of CFR titles, and another that contains the terms in headings and subheadings and the list of CFR titles. jFreq is designed for this task. Once both matrices have been prepared, the WORDFISH algorithm was applied separately on them, using the statistical programming language `R`.[16]

4. RESULTS

4.1 Document positions

The distribution of CFR titles according to their indexing terms is depicted in Figure 1, while Figure 2 shows the position of CFR titles with respect to the text of their headings. The x-axis of both figures represents the ω score received by each title—i.e., its position on the one-dimensional scale—while the y-axis represents fixed effects (α)—i.e., control for the fact that some titles have more indexing terms attached than others (Figure 1) or their headings contain more text than others (Figure 2). Apart from the neat distributional pattern of titles among the x-axis, note that a decreasing (negative) pattern also exists between fixed effects (α) and ω score in both cases, represented through a dashed regression line. The results of the estimates for the CFR titles according to both corpora are presented in Table 2 (in the appendix), which also presents the differences in scores.

In order to ease the evaluation of the differences on the scores received by each CFR title in both corpora, we represent them in Figure 3. Figure 3 shows that most of these differences are close to zero, though there are some extreme values—i.e., cases in which CFR titles have received very different scores from WORDFISH in both corpora.

The median difference in the scores received by each title in each corpus is -0.27. The CFR titles that present higher differences in scores are those that contain Federal regulations on General Provisions (CFR 1), Agriculture (CFR 7), Housing and Urban Development (CFR 24), Indians (CFR

[15] Titles 3 and 35 are not analyzed because data on their headings was not available).

[16] `http://cran.r-project.org/`. The `R` code for the analysis is available from the authors upon request.

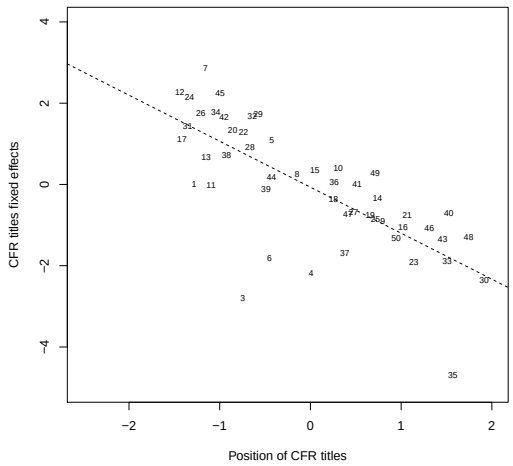


Figure 1: Plot of the positions of CFR titles according to their indexing terms.

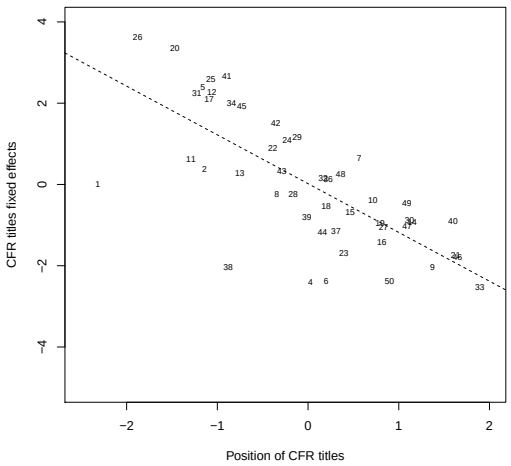


Figure 2: Plot of the positions of CFR titles according to their headings.

25), Public Contracts and Property Management (CFR 41), Public Lands, interior (CFR 43), and Federal Acquisition Regulations System (CFR 48). However, as Figure 4 shows, only CFR titles 7, 25, 41, and 43 are complete mismatches—i.e., they take completely opposed positions (positive *vs.* negative) in both classification tests.

4.2 Word positions

The two plots in Figures 5 and 6 show the distribution of terms along the estimate scale (β) for the indexing terms and the headings, respectively. One common feature is observable in both figures: a great number of terms are located around the zero position, i.e., the center of the x-axis.[17]

Scores significantly different from zero represent the extent to which a term is discriminant of a position of a particular CFR title in its position in the one-dimensional scale. Therefore the terms situated on the extremes of the x-axis are the ones that discriminate better in document scaling. The y-axis in both figures accounts for the words fixed effects (ψ), i.e., the fact that some words are used more usually than others (high values denote words that occur more times in all documents).

In the corpus of indexing terms (Figure 5) terms are more or less symetrically distributed along the x-axis, for which we expect that a similar number of terms are discriminant for negative and positive positions. In contrast, the terms from the headings (Figure 6) are heavily tailed in a positive direction (right side of the figure), for which we expect to find that a greater number of terms discriminating titles in the positive side of the scale than doing so in the negative side. The words on the top of the "Eiffel tower" are the most used terms in the corpus.

[17]The complete lists of the actual discriminant terms for each corpus—too long to be included in this paper—are available upon request. An example containing the set of the 50 most discriminant terms on the **right** side of the position axis, from the corpus of indexing terms, can be found in Figures 3 in the Appendix.

4.3 Check for correspondence

Apart from the computation of the differences in estimates CFR titles receive in both corpora, a second check directly relates the scores received by each title in each experiment. Figure 7 plots the position of the actual titles related to the scores they received in each experiment. The scores are standardized to have mean zero and standard deviation equal to 1. In the figure a positive linear relationship is clearly visible, represented by the regression line that summarizes that relationship ($r = .73$), indicating a quite strong correspondence between both scores.

5. DISCUSSIONS AND FUTURE RESEARCH

Before any discussion about the results, a comment on the assumptions of the test is in order. In the method section we briefly presented the assumptions on which WORDFISH is based—i.e., that words in a text are random variables that occur following a Poisson distribution. This has implications about occurrence independence—the fact that one particular word occurs in a particular moment is independent from the fact that another word has occurred immediately before. While the independence assumption is openly debatable in political speeches or manifestos (the ones for which WORDFISH was designed), the assumption is hardly tenable in our context, when indexing terms or even the words of the headings have been purposely chosen by agencies to mark or "represent" specific regulations (although they may still involve one or multiple stochastic processes of text generation [5]). This fact, though, just highlights the need to extend the analysis to the complete text of the CFR titles and compare its results with the ones we have obtained, this task will be conduced in the future.

Regarding the results towards establishing the extent of the correspondence of the two corpora under analysis, there is in general a strong correlation between the scores received by the CFR titles in both experiments. Nevertheless, the exis-

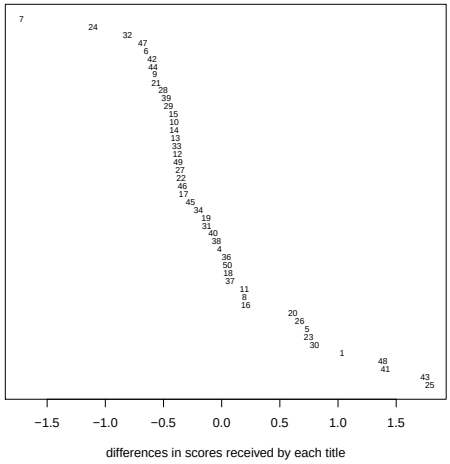


Figure 3: Representation of the difference in score received by each CFR title in both corpora. The graph shows that the differences for most titles are around zero, though some high positive differences exist (> 1, < -1) for a set of 7 titles.

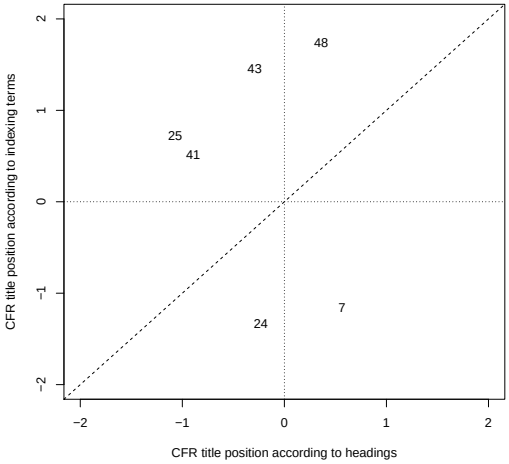


Figure 4: Representation of the position of CFR titles that present the higher differences. Titles in the first (top left) and fourth (bottom right) quadrants are considered complete mismatches.

tence of four clear mismatches suggests that part of corpus of indexing terms is not representative of the actual content of the CFR titles based on the analysis of their headings. This also suggests that the direct reuse of an existing thesaurus for textual search and retrieval could, in some cases, lead to inefficient results. In fact, this points at the existence of a gap between the representation of terms in a thesaurus and the terms contained in the text the thesaurus is linked to. The measure of this gap will be the basis of our future research.

Moreover several open issues for discussion arise from these results. First, CFR titles show similar patterns regarding their position in one dimensional space in both analysis, a) they are distributed rather uniformly on the x-axis, b) there seems to be a linear decreasing relationship between the position of a CFR title and its fixed effects. In effect, the relationship between both variables showing data points is plotted in both figures through a regression line. The Pearson's correlation coefficient between fixed effects and position is almost identical in both cases ($r = -0.74$ and -0.73). See Figures 1 and 2). This could have many interpretations, for instance, it might mean that some type of CFR titles contain more regulations, that the regulations they contain are textually longer, that more agencies are involved in the development of their contents, or even that those titles containing procedure-oriented regulations have more indexing terms attached *and* also their headings have more words.

If we consider the hypothesis that the CFR titles on the left side of the scale contain more procedure-oriented information and the ones on the right side substantive content, would the terms that discriminate both extreme CFR titles be reasonably related to the procedure/substantive distinction? Are the terms that classify CFR titles as more procedural mostly verbs or nouns related to procedural actions, while terms on the other side are more nouns relating to objects, indicating more substantial content?

Also, is there a relationship between the agencies that produce regulations in certain titles and the classification of those titles in the axis? Further research on this may also have interest for analyzing complexity contained in the information and knowledge produced by organizations. We will conduct further research on the complete corpus of the CFR and an extended version of the Thesaurus of Indexing terms to explore these questions.

6. ACKNOWLEDGEMENTS

Part of this research has been funded by grant MEC/Fulbright 2010-2012. We would like to acknowledge the feedback and assistance of Sara Frug, Daniel Nagy, David Shetland, and Wayne Weibel at the Legal Information Institute.

7. REFERENCES

[1] The federal register and the code of federal regulations. a reappraisal. *Harvard Law Review*, 80(2):pp. 439–451, 1966.

[2] G. Ajani, G. Boella, L. Lesmo, M. Martin, A. Mazzei, D. P. Radicioni, and P. Rossi. Legal taxonomy syllabus version 2.0. In N. Casellas, E. Francesconi, R. Hoekstra, and S. Montemagni, editors, *3rd Workshop on Legal Ontologies and Arificial Intelligence Techniques joint with 2nd Workshop on Semantic Processing of Legal Text (LOAIT 2009), Co-located with the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009).*, volume 2 of *IDT Series*, pages 9–17. IDT/Huygens Editorial, 2009.

[3] G. Antoniou and F. van Harmelen. Web ontology language: Owl. *Handbook on Ontologies*, pages 67–92, 2004.

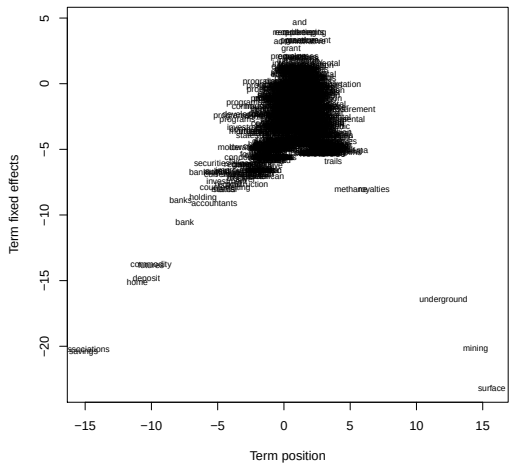


Figure 5: Plot of the positions of the terms of CFR titles indexing terms.

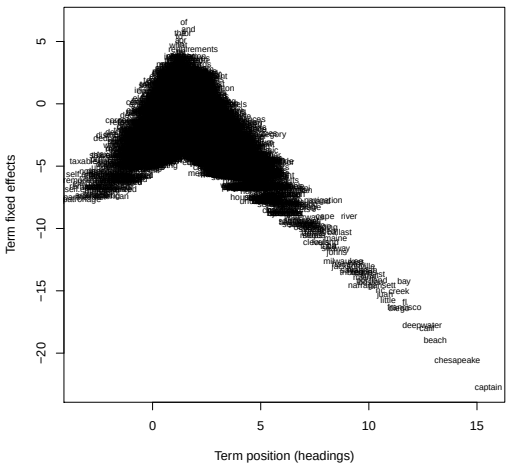


Figure 6: Plot of the positions of the terms of CFR titles headings.

[4] G. Antoniou and F. van Harmelen. *A Semantic Web Primer, 2nd Edition*. Cooperative Information Systems. The MIT Press, March 2008.

[5] K. Benoit, M. Laver, and S. Mikhaylov. Treating words as data with error: uncertainty in text statements of policy positions. *American Journal of Political Science*, 53(2):495–513, 2009.

[6] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, May 2001.

[7] N. Casellas. *Legal Ontology Engineering Methodologies, Modelling Trends, and theOntology of Professional Judicial Knowledge*. Law, Governance and Technology. Springer Verlag, [to appear] 2011.

[8] C. P. Cheng, G. T. Lau, K. H. Law, J. Pan, and A. Jones. Regulation retrieval using industry specific taxonomies. volume 16, pages 277–303, Hingham, MA, USA, September 2008. Kluwer Academic Publishers.

[9] C. P. Cheng, J. Pan, G. T. Lau, K. H. Law, and A. Jones. Relating taxonomies with regulations. In *Proceedings of the 2008 international conference on Digital government research*, dg.o '08, pages 34–43. Digital Government Society of North America, 2008.

[10] L. E. Feinberg. Mr. justice brandeis and the creation of the federal register. *Public Administration Review*, 61(3):pp. 359–370, 2001.

[11] E. Francesconi, S. Faro, and E. Marinai. A framework for semantic mapping between thesauri. In *Proceedings of the 2nd international conference on Theory and practice of electronic governance*, ICEGOV '08, pages 251–257, New York, NY, USA, 2008. ACM.

[12] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph. Owl 2 web ontology language: Primer. W3C Recommendation 27 October 2009, W3C, 2009.

[13] R. Hoekstra, R. Winkels, and E. Hupkes. Reasoning with spatial plans on the semantic web. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 185–193, New York, NY, USA, 2009. ACM.

[14] C. M. Kerwin. *The management of regulation development: out of the shadows*. Presidential Transition Series. IBM Center for The Business of Government, Washington, DC, 2008.

[15] H. Klüver. Measuring interest group influence using quantitative text analysis. *European Union Politics*, 10(4):535–549, 2009.

[16] G. T. Lau, K. H. Law, and G. Wiederhold. Legal information retrieval and application to e-rulemaking. In *Proceedings of the 10th international conference on Artificial intelligence and law*, ICAIL '05, pages 146–154, New York, NY, USA, 2005. ACM.

[17] M. Laver, K. Benoit, and J. Garry. Extracting policy positions from political texts using words as data. *The American Political Science Review*, 97(2):311–331, May 2003.

[18] F. Manola and E. M. (ed.). Rdf primer. W3c recommendation 10 february 2004, World Wide Web Consortium (W3C), February 2004.

[19] J. McClure. The legal-rdf ontology. a generic model for legal documents. In P. Casanovas, M. A. Biasiotti, E. Francesconi, and M. T. Sagri, editors, *Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007) at the International Conference on AI and Law (ICAIL'07) Stanford, USA, June 4*, pages 25–42, 2007.

[20] M. Muller. Legal rdf dictionary. In R. Winkels, editor, *Proceedings of the Second International Workshop on Legal Ontologies (LEGONT) in JURIX 2001, Amsterdam (Netherlands)*, pages 20–21, Amsterdam, Netherlands, 2001.

[21] L. Polo, J. M. Alvarez, and E. R. Azcona. Promoting government controlled vocabularies to the semantic web: EUROVOC thesaurus and cpv product classification scheme. In *Proceedings of the Semantic Interoperability in the European Digital Library workshop (SIEDL2008), co-located with 5th European*

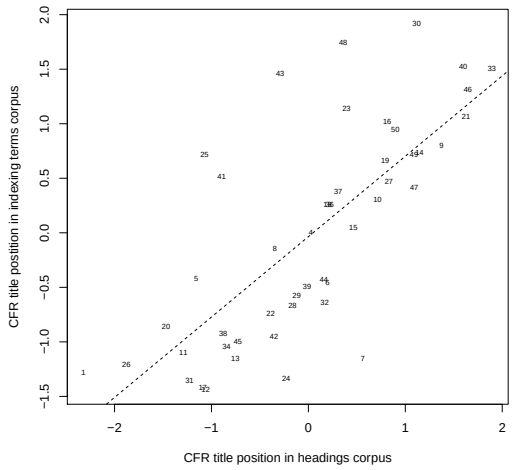


Figure 7: Plot showing the correspondence between the scores received by each CFR title in each corpus. The general trend is positive ($r = .73$).

Semantic Web Conference (ESWC2008), Tenerife, Spain, June 2, 2008., pages 111–122, 2008.

[22] S.-O. Proksch and J. B. Slapin. *WORDFISH: Scaling Software for Estimating Political Positions from Text (version 1.3)*. http://www.wordfish.org, 2009.

[23] J. B. Slapin and S.-O. Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722, 2008.

APPENDIX

A. DIFFERENCES IN ESTIMATES FOR CFR TITLE POSITION IN BOTH CORPORA, AND SETS OF MOST DISCRIMINANT TERMS

CFR titles	ω indexing terms	ω headings	difference
CFRT1	-1.28	-2.32	1.04
CFRT4	0.01	0.03	-0.02
CFRT5	-0.42	-1.16	0.74
CFRT6	-0.45	0.20	-0.65
CFRT7	-1.16	0.56	-1.72
CFRT8	-0.15	-0.34	0.20
CFRT9	0.80	1.37	-0.57
CFRT10	0.31	0.71	-0.41
CFRT11	-1.09	-1.29	0.20
CFRT12	-1.44	-1.06	-0.38
CFRT13	-1.15	-0.75	-0.40
CFRT14	0.74	1.15	-0.41
CFRT15	0.05	0.46	-0.41
CFRT16	1.02	0.81	0.21
CFRT17	-1.42	-1.09	-0.33
CFRT18	0.26	0.20	0.06
CFRT19	0.66	0.79	-0.13
CFRT20	-0.86	-1.47	0.61
CFRT21	1.07	1.63	-0.56
CFRT22	-0.74	-0.39	-0.35
CFRT23	1.14	0.39	0.75
CFRT24	-1.33	-0.23	-1.10
CFRT25	0.72	-1.07	1.79
CFRT26	-1.21	-1.88	0.67
CFRT27	0.47	0.83	-0.36
CFRT28	-0.67	-0.16	-0.50
CFRT29	-0.58	-0.12	-0.46
CFRT30	1.92	1.12	0.80
CFRT31	-1.35	-1.23	-0.13
CFRT32	-0.64	0.17	-0.81
CFRT33	1.51	1.89	-0.39
CFRT34	-1.04	-0.84	-0.20
CFRT36	0.26	0.22	0.04
CFRT37	0.38	0.31	0.07
CFRT38	-0.92	-0.88	-0.05
CFRT39	-0.49	-0.01	-0.48
CFRT40	1.53	1.60	-0.07
CFRT41	0.51	-0.89	1.41
CFRT42	-0.95	-0.35	-0.60
CFRT43	1.46	-0.29	1.75
CFRT44	-0.43	0.16	-0.59
CFRT45	-1.00	-0.73	-0.27
CFRT46	1.31	1.65	-0.34
CFRT47	0.41	1.09	-0.68
CFRT48	1.74	0.36	1.38
CFRT49	0.72	1.09	-0.37
CFRT50	0.95	0.90	0.05

Table 2: CFR titles estimates for position (ω) in both corpora. The data are ordered by title.

word	β	ψ
savings	-15.12	-20.44
associations	-14.91	-20.18
home	-11.06	-15.14
deposit	-10.37	-14.85
commodity	-10.03	-13.79
futures	-10.03	-13.79
banks	-7.79	-8.88
bank	-7.48	-10.56
holding	-6.09	-8.71
banking	-6.03	-6.81
securities	-5.43	-6.05
accountants	-5.25	-9.14
currency	-4.78	-7.03
mortgage	-4.69	-6.78
rent	-4.62	-6.68
subsidies	-4.62	-6.68
marital	-4.51	-8.03
status	-4.51	-8.03
counterfeiting	-4.43	-7.93
investment	-4.30	-7.41
gold	-3.82	-7.71
fair	-3.70	-5.85
low	-3.64	-4.86
moderate	-3.64	-4.86
block	-3.57	-7.18
register	-3.43	-6.21
improvement	-3.29	-6.65
reconstruction	-3.19	-7.62
silver	-2.88	-7.24
development	-2.81	-2.34
states	-2.75	-3.91
programs-housing	-2.65	-2.42
mortgages	-2.62	-3.68
manual	-2.61	-6.21
investments	-2.60	-3.31
diamonds	-2.59	-6.88
forgery	-2.59	-6.88
humanitarian	-2.59	-6.88
united	-2.53	-3.63
companies	-2.52	-3.54
condominiums	-2.41	-5.56
community	-2.38	-1.73
compilation	-2.35	-6.18
papers	-2.35	-6.18
presidents	-2.35	-6.18
weekly	-2.35	-6.18
code	-2.29	-5.42
african	-2.11	-7.00
asian	-2.11	-7.00
european	-2.11	-7.00

Table 3: Set of the 50 most discriminant terms on the **right** side of the position axis, from the corpus of indexing terms.

Towards the intelligent processing of non-expert generated content: Mapping web 2.0 data with ontologies in the domain of consumer mediation

Meritxell Fernández-Barrera
Cersa, CNRS-Université Paris2
10, Rue Thénard, 75005-Paris
meritxell.fernandez@cersa.cnrs.fr

Pompeu Casanovas
Institute of Law and Technology, Universitat Autònoma de Barcelona, 08193 Bellaterra
pompeu.casanovas@uab.cat

ABSTRACT

This paper presents a case study in the processing of user-generated content and its mapping with expert ontologies in the domain of consumer justice. The analysis is made in the framework of the ONTOMEDIA project, which aims at the design of a semantic platform enabling users and professional mediators to meet in a community-driven Web portal. The paper first presents the characteristics of the platform and its requirements; then describes the methodology for term-extraction from a corpus of consumer queries and finally the model proposed for the mapping with available domain ontologies. It concludes with the discussion of open issues for future work.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]; H.3.5. [**Online Information Services**]

General Terms

Management, Documentation, Performance, Design, Theory, Legal Aspects.

Keywords

Legal Electronic Institutions (LEI), mediation, user-generated content, term-extraction, domain ontologies, semantic mapping.

1. THE CATALAN WHITE BOOK ON MEDIATION (CWBM), LEGAL ELECTRONIC INSTITUTIONS (LEI), AND THE ONTOMEDIA PROJECT

The CWBM is a large research project (2008-2010)[1] aiming at the implementation of mediation as defined by the EU Directive 52/2008.[2] One of its most surprising findings is that near 18% of the population in Catalonia (7,5 million people) has pending cases in the Spanish Courtrooms. Heavy caseloads and chronic shortage of judges and magistrates, on the one side, and increasing social problems on the other (especially large immigration rates and the emergence of all kind of violence in families, schools, hospitals and institutions) have fostered the need to draw a map of dispute resolution techniques in the country, before drafting a general statute. It is worthwhile taking into account that from 2000 to 2010, more than one million people have landed in Catalonia (15.9 % of the population are newcomers). Therefore, we conceived mediation not only as an Alternative Dispute Resolution (ADR) device, but as a set of tools operating near the communities, Courts and Administrations. In this way, mediation as institution may be adapted to the nature of conflicts arising within the different environments, contexts and settings (neighborhoods, colleges, hospitals, administrations etc…).

To apply technology to mediation, we followed a twofold strategy leading to two separate models: (i) building mediation as a Legal Electronic Institution (LEI)[3]; and (ii) setting up a general platform for citizens, administrations, institutions and professionals. The first strategy (LEI) models the performative structure of mediation as a set of procedural rules. The second one (ONTOMEDIA) allows users and professional mediators to meet in a community-

[1] All the results of the Catalan White Book (Department of Justice, 2010-2011) are available at http://www.llibreblancmediacio.com in both languages, Catalan (1186 pp.) and Spanish (1206 pp.).

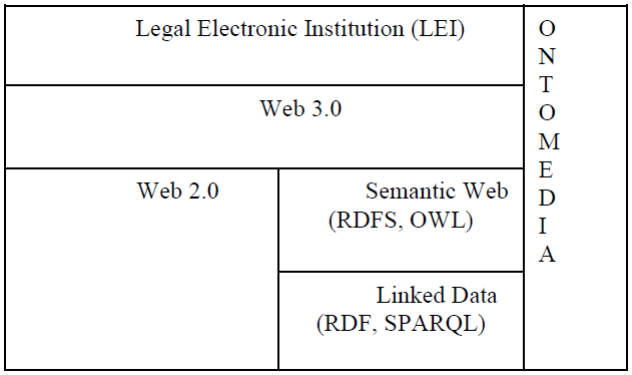
[2] Art. 3.a. "*Mediation* means a structured process, however named or referred to, whereby two or more parties to a dispute attempt by themselves, on a voluntary basis, to reach an agreement on the settlement of their dispute with the assistance of a mediator" ; art. 3.b. "*Mediator* means any third person who is asked to conduct a mediation in an effective, impartial and competent way, regardless of the denomination or profession of that third person in the Member State concerned and of the way in which the third person has been appointed or requested to conduct the mediation." It is worth to mention R. (9): 'This Directive should not in any way prevent the use of modern communication technologies in the mediation process".

[3] Electronic Institutions (EIs) organize interactions by establishing a restricted environment where all interactions take place (*e.g.* e.commerce, e-learning, or ODR). They create a virtual environment where interactions among agents in the real world correspond with illocutions exchanged by agents within this restricted environment. When an EI is entitled to perform legal acts, or at the end of successive steps may produce a result with legal value, or an agreement that can be alleged in Court or before other appropriate ruling institutions, we face a Legal Electronic Institution (LEI) See [18]. See also http://e-institutions.iiia.csic.es. See for a more detailed analysis [19]; for a comparison of the grounds of LEI and Ontomedia [7]; for the state of the art of the ODR existing platforms, [25]. The LEI software code for mediation [20] is available at http://www.llibreblancmediacio.com (Spanish version).

driven Web portal (in which contents are provided by users and annotated by the ODR web platform).

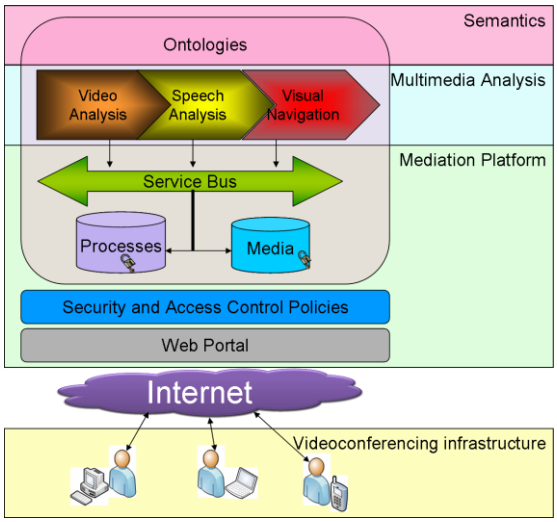
ONTOMEDIA is, then, a semantic platform for relational justice. It has been conceived as a bus of services to offer to both citizens and mediators a kit of tools and services to facilitate a better access to justice. Attention is focused on the development and synergy between different technologies stemming from Web Services (WS), the Semantic Web (SW), Social Networks (SN), Multi-agents Systems (MAS), Computer Vision (CV) and legal applications. LEI and ONTOMEDIA are orthogonally related, according to the original James Hendler's diagram on the link between Web 2.0 and the emergent Web 3.0 (*Fig.* 1):

Fig. 1. Hendler's diagram. Source: [7]



The sections of ONTOMEDIA are tailored on the domains previously identified within the CWBM: commercial and business disputes, consumer complaints, labor conflicts, family, restorative justice (adult and juvenile mediation in criminal issues), community problems, local administration, health care, environmental management, and education (*Fig.* 2).

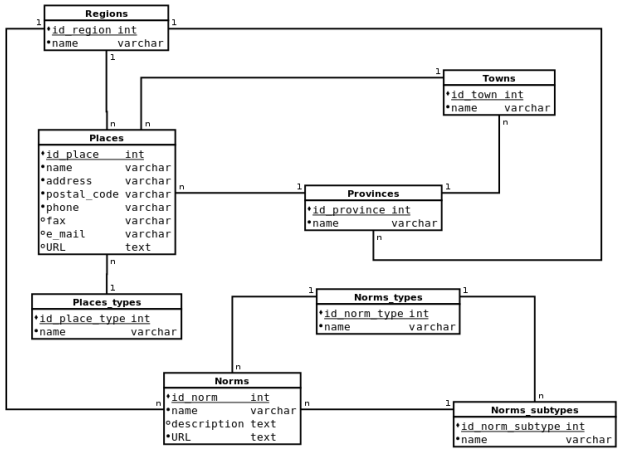
Fig. 2. ONTOMEDIA layered architecture



We have planned a lifecycle of five years to the full development of all the functionalities. We chose the consumer domain, first, to implement some of them specifically addressed to citizens. We made this decision because we had a good description of all the procedures and the precise workflow of pre-mediation, mediation and post-mediation stages [1]. Moreover, as it will be shown later, the Catalan Consumer Agency would give us access to more than 30,000 complaints and information requests to work with.

As a result of gathering consumer mediation related resources, a relational schema for a database was proposed as well. This database is a critical component of the platform's data tier. However, the proposed relational schema is only a little portion of it, storing entities and relations involving national regulations, regional regulations, soft-law, consumer offices, and so on. The database contains so far information on 19 Spanish regions, 892 towns and 52 provinces, holding 1,264 consumer mediation resources (which include consumer offices and other public and private institutions) and 75 different regulations (*Fig*. 3).

Fig. 3. Database diagram. Source: [13]



The idea behind this schema is to provide basic legal and judicial resources to citizens involved in consumer mediation processes, including users in conflict starting scenarios where the mediation process (and thus mediation resources) may be suggested by the platform. Taking into account some basic geographical data is required at this point, because the platform will be able to locate the user and an efficient and accurate response requires norms and institutions to be geographically queried.

The basic entities here are *places* and *norms*. The places relation stores registers about consumer mediation resources, like consumer offices, private consumer organizations, and public institutions supporting mediation. These types of places are references in the *places types* relation. Furthermore, a given place is located in a *town*, *province* and *region*. A given Spanish region, where a place belongs, has a set of *norms*. These can be binding or non-binding regulations, such as best practices codes. These distinctions are made in norms *types* and *subtypes* relations.

2. THE ONTOMEDIA SEMANTIC PLATFORM: A GATE TO INFORMATION AND SERVICES

One of the expected functionalities of the semantic platform is to allow citizens to present their problem in natural language and to redirect them either to relevant information already available

online or to the suitable state agency. The assumption at the basis of this process is being able to map two different conceptual systems: the user representation of a problem in the form of concrete actions, actors and contexts (non-expert model); and the regulative representation of the problem usually in the form of general classes of actions, actors and normative provisions (expert model). We propose here the existence of a middle-level which corresponds to the practices and know-how of professionals. Professionals are indeed frequently in charge of reformulating regulative information into more comprehensible texts that are subsequently published in the form of electronic leaflets in institutional websites, and they are usually as well the ones interacting directly with citizens. Thus it can be assumed that theirs is an intermediary conceptual system, bridging abstract legal provisions with concrete conflictive situations presented by non-expert citizens. Our conception of domain knowledge can thus be seen as a multidimensional figure, which, vis-à-vis flat knowledge models, takes into consideration elements such as different domain actors (citizen, professional, legislator) and communicative contexts (information request, complaint).

The technical aspects underlying this functionality are related to the automatic classification of consumer queries according to a conceptual scheme which models citizen's problems or conflictive situations on the basis of available institutional structures and procedures. This model has been described thoroughly in the Catalan White Book of Mediation [8] and an ontological representation of mediation expert knowledge has been proposed in the Mediation Core Ontology, available in OWL-DL [24]. A further representation of the domain of consumer mediation is provided by the mediation ontology [24]. The goal of this paper is to reuse these two ontological structures to map the representative terms of a corpus of consumer queries to an institutional-expert representation of the domain and thus to enable the channeling of consumer needs into the institutional infrastructure which is meant to satisfy them.

We have a diachronic corpus of around 10,000 questions and 20,000 complaints which have been addressed by consumers to the Catalan Consumer Agency[4] from 2007 to 2010. The difference between queries and complaints relies on the fact that while queries are mere requests of information, complaints are meant to initiate an administrative process of mediation between the consumer and the seller or service provider[5]. A further distinction relevant to characterize our corpus is the input language. Indeed, since both Catalan and Spanish are official languages in Catalonia and thus citizens are entitled to address state agencies in both languages, a previous step for treating automatically our corpus has been to classify documents according to their language.

[4] The mission of the Catalan Consumer Agency is to defend citizen's rights as consumers, and thus on the one hand it provides information regarding consumer affairs and on the other it has a role in the resolution of conflicts between consumers and companies through mediation and arbitration. http://www.consum.cat/qui_som/index_en.html

[5] One of the requirements for being able to initiate a mediation process is to have previously contacted the seller or service provider.

In this paper we present a case study on the mapping between consumer terminology and the available formal domain model that is aimed at testing our methodology. At this initial stage we have decided to concentrate exclusively on queries expressed in Spanish, corresponding to the year 2010. The subset of queries of 2010 has been used to extract representative terminology from subsets of consumer questions classified by topic (Internet service providers, travel agencies, vehicles, …), and the extracted terminology has been linked to the available ontological domain models.

Section 3 discusses the technical challenges of an intelligent platform able to process citizens' queries and presents the model that will be used in our case study; Section 4 details the process of terminology extraction from a set of consumer queries; Section 5 describes the extension of the available formal ontologies with consumer terminology through a has_lexicalisation property; Section 6 discusses the main contributions of the paper and identifies the issues that require being dealt with in the follow-up of the ONTOMEDIA project.

3. BRIDGING THE GAP BETWEEN KNOWLEDGE *IN ACTION* AND THEORETICAL LEGAL KNOWLEDGE: WEB 2.0 vs. WEB 3.0

Enabling the intelligent processing of non-expert generated content is strongly connected with the problem of interfacing Web 2.0 with Web 3.0. Indeed, with the advent of Web 2.0, semantic technologies face a new challenge: the processing of heterogeneous non-standardized knowledge, with unknown producers and with the absence of explicit terminological and conceptual harmonization. This problem was already highlighted in connection to the need of conceiving artificial intelligence as the ability to cope with heterogeneous and disperse data, based on different ontologies, instead of focusing on highly axiomatised and unified ontological models ([17], [11], [10]).

Coping with this challenge implies finding a way to bridge Semantic Web data structures, such as formal ontologies expressed in RDF or OWL, with unstructured implicit ontologies emerging from user-generated content. Sometimes these emergent lightweight ontologies take the form of unstructured lists of terms used for tagging online content by users. Accordingly, some works have dealt with this issue especially in the field of social tagging of web resources in online communities. More concretely, different works have proposed models for making compatible the so-called top-down metadata structures (ontologies) with bottom-up tagging mechanisms (folksonomies)[6]. Some authors, such as [31], point out that the emergent problem of linking Web 2.0 and Web 3.0 lies on the way in which emergent *collective* rationality of the Web 2.0 relates to the proposed *connective* rationality of Web 3.0.

[6] It should be highlighted that the terms *top-down* and *bottom-up* are here used as referring to the participants in the construction of the resource: while in the first case the resource is the result of an agreement on a world model reached by the members of a particular community, in the second case the resource emerges from the distributed tagging activity of a big number of anonymous users.

The possibilities range from transforming folksonomies into lightly formalized semantic resources ([16], [15]) to mapping folksonomy tags to the concepts and the instances of available formal ontologies ([30], [21]). At the basis of these works we find the notion of emergent semantics [16], which questions the autonomy of engineered ontologies and emphasizes the value of meaning emerging from distributed communities working collaboratively through the web. An important element in the model proposed by many of those works is the *actor*, who tags a specific resource with a particular tag.

This is not the case in our corpus, since users in our case study simply provide input texts describing their problems and asking for institutional assistance. Thus we do not have a ready-made folksonomy created collaboratively by users. In this context the *implicit ontology* is understood as the set of linguistic structures on which users rely to represent the concepts of the domain. Thus in this framework a further challenge consists in being able to extract recurrent linguistic structures from non-normalized texts. Indeed, while texts following the standards defined by a particular community of experts lend naturally to the extraction of patterns, this task becomes much less obvious with regard to texts which do not necessarily conform to pre-established guidelines.

This way, the terminological, argumentative and semantic patterns of texts following certain standards in the legal community (i.e. bills, acts, judgments, legal expert files) has been deeply studied (see state-of-the-art on legal ontologies, legal argumentation models and XML models for legal documents in [26], while the analysis of recurrent structures in the way citizens express their legal problems has been paid less attention. In the domain of semantic technologies for the legal domain efforts have indeed mostly concentrated on making explicit formal ontologies deriving from normative sources and from legal expert texts [27]. This work has lead to the creation of several domain ontologies but so far explicit mappings between these formal ontologies and implicit ontologies emerging from the citizens' representation of particular legal problems are not available. This implies an important drawback for legal web-based services, since the linguistic and conceptual schemas used by citizens in the expression of their needs are not taken into account. The improvement of such services requires taking into account the particularities of non-expert common discourse and above all, to connect it to specialised legal discourse.

As a first effort in this direction, this paper presents a case study in the consumer law domain. We propose to reuse the available (a) Mediation-Core Ontology (MCO) and (b) Consumer Mediation Ontology (COM) as anchors to legal, institutional and expert knowledge, and therefore as entry points for the queries posed by consumers in common language. We will follow the approach proposed by [21] and enrich the available ontologies with the terminology appearing in the consumer corpus. For so doing, Owl classes and instances will be complemented with a has_lexicalization property linking them to consumer terms.

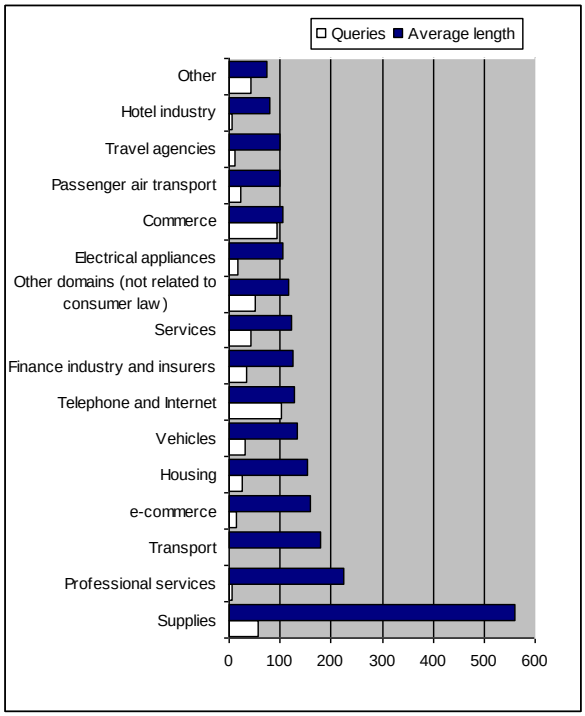
Our methodology is thus based on the following steps: *i.* extraction of relevant terminology from the consumer queries on the basis of morphological tagging; *ii.* enrichment of the ontological resources with consumer terminology through a has_lexicalization property.

4. NLP EXTRACTION OF CONSUMER TERMINOLOGY

Since the corpus of consumer queries has not been previously annotated or semantically tagged there is no available semantic representation of consumer knowledge. This is why it has been decided to semi-automatically extract a list of representative terms through NLP techniques. The goal was to see whether despite the fact that producers are unknown and do not follow explicit guidelines in the construction of their message common lexical patterns emerge.

Firstly, the set of queries of 2010 was manually classified into subsets according to a list of topics used by the Catalan Consumer Agency [8] in order to enable the extraction of contextually-related groups of terms. The topics defined by the Agency are: *commerce, e-commerce, electrical appliances, housing, hotel industry, finance industry and insurers, services, professional services*[7]*, supplies, telephone, passenger air transport, transport, vehicles* and *travel agencies*. *Fig.* 4 reports the number of queries corresponding to each topic as well as their average length in number of tokens.

Figure 4. Number of queries and query average length per domain



It can be observed that the topics concentrating the highest number of queries are, in this order, telephone supply, commerce, other supplies (gas, electricity, water, …), and services. The average question length varies according to the domain. For

[7] *Professional services* refer mostly to the services provided by liberal professionals such as lawyers, doctors, dentists. On the other hand, *services* refers in general to services such as sport facilities, hairdresser's, cultural shows (theatres, cinemas…) or educational services (for instance e-learning).

instance, the average length in the domain of supplies (560 tokens) is five times that of queries in the domain of commerce (104 tokens). This fact can be explained due to the characteristics of the problems arising in the domain of supplies, which are usually connected to technical failures which require a certain degree of precision to determine where responsibility lies (nature of electric installation, power of the electric circuit, …). On the other hand, queries in the domain of commerce require usually only a few details to present the situation (such as place of purchase or guarantee length) and therefore tend to be shorter. These are important aspects to take into account in the process of terminology extraction.

Another relevant feature to be highlighted is that a high number of queries belong to other domains which go beyond the competences of the Consumer Agency and which mainly belong to other areas of law such as private law (*i.e.* disputes between tenant and landlord; private deals) or administrative law (*i.e.* tax paying; appeals to speeding tickets or to penalties for drunk driving). This reinforces the need of a semantic platform able to classify and distribute citizens' queries to the state agencies which are able to provide useful information and assistance to solve the conflict.

Next, the questions were tagged and lemmatized using Tree-tagger. Tree-tagger [28] is a probabilistic morphosyntactic tagger and lemmatiser which estimates transition probabilities on the basis of a binary decision tree in order to avoid the limitations of probabilistic taggers based on Markov Models.

By using the *make_separate.pl* module we created an XML version of the tagged documents and imported them into the NooJ platform. NooJ is a platform that enables the linguistic processing of texts at different levels (*i.e.* morphological, syntactic, semantic) with the aid of different types of grammars (among which, inflexion grammars, morphological grammars and syntactic grammars) [29]. In order to enable NooJ to recognise tree-tagger tags as morphosyntactic annotations we adapted the original XML element and attribute to NooJ standards[8]. NooJ offers the possibility of analyzing morphologically any input text but it does not offer in-built disambiguation grammars, so whenever there is ambiguity all the possible syntactic categories will be maintained. This would have created a lot of noise in the subsequent search of morphosyntactic patterns, so we decided to rely on the probabilistic tagging of Tree-tagger, which provides disambiguated morphosyntactic tags with a low level of error.

Once the queries had been imported in the NooJ platform we extracted first simple terms and then multiword terms. Firstly, regarding simple terms, we follow the traditional trend in terminological studies that states that the most common linguistic unit carrying conceptual meaning is the noun. In this line, we extracted through the NooJ function Locate pattern all the nouns of our corpora. Nevertheless we do not rule out the possibility of extending term extraction to other linguistic units such as predicates in the future, since recent works have highlighted that units of specialized knowledge can take different syntactic forms[9].

Table 1 reports some of the simple terms (nouns) extracted in each subset of questions. Some of the extracted terms are recurrent in different topic subsets and therefore we can consider that they belong to the general domain of consumer queries. They denote: the seller, such as *firm* (*empresa*); the contractual binding between consumer and seller, such as *contract* (*contrato*), *invoice* (*factura*), guarantee (*garantía*); or the amount paid by the consumer, such as *money* (*dinero*), *euros*, *amount* (*importe*); the cause of the conflict, such as *problem* (*problema*), *abuse* (*abuso*), *failure* (*fallo*); and the expectations of the consumer, such as *return-refund* (*devolución*).

Other terms seem to be topic-specific. They denote either the actors of specific domains such as *real state agency* (*inmobiliaria*), in the domain of housing; *campsite* (*camping*), *camper* (*campista*), in hotel industry; *bank* (*banco*, *entidad*), *insurer* (*correduría*), in finance industry and insurers; *lawyer* (*abogado*, *letrado*), *dentist* (*dentista*), *psyschiatrist* (*psiquiatra*), *hospital* (*hospital*), in professional services; *customer* (*abonado*), *telephone operator* (*operador*), in the domain of telephone and the Internet services; *taxi driver* (*taxista*) in transport; *garage* (*taller*), *car dealer* (*concesionario*), in vehicles.

Actors are often denoted in the corpus through named entities, specially in the domain of telephone and Internet, such as *Jazztel*, *Teléfonica*, *Vodafone*; and the domain of passenger air transport, such as *Iberia*, *Easyjet*, *Aerlingus*. They will be integrated into the available ontologies as concept instances.

Other nouns denote actions and states which are typical of particular domains, such as *sales* (*rebajas*), in commerce; *activation* (*activación*), *prepaid card* (*prepago*), *permanence clause* (*permanencia*), *contract cancelation* (*baja*), *portability* (*portabilidad*), in the domain of phone and Internet services; *technical maintenance* (*mantenimiento*), *supply* (*suministro*), in the domain of supplies.

Finally, some nouns denote typical objects of certain domains: *appliance* (*aparato*), *washing machine* (*lavadora*), *computer* (*ordenador*), *microwave* (*microondas*), in the domain of electrical appliances; *meter* (*contador*), *boiler* (*caldera*), heating (*calefacción*), in the domain of supplies.

[8] In other words, the original element and attribute "<TOKEN tag= >" have been transformed into "<LU cat= >".

[9] More concretely, [3] highlight that units of specialised knowledge can be: morphological units (morphemes); one word units; syntagmatic units, that is to say, multiword units and phraseological units; and phrasal unit.

Table 1. Sample of extracted terms (N) by topic

Commerce	tienda, euros, dinero, producto, empresa, devolución, problema, servicio, cámara, garantía, vale, rebajas, reclamación, tarjeta, fabricante, importe
e-commerce	tarjeta, cargo, teléfono, cuenta, garantía, devolución, producto, precio, reclamación, estafa, factura, paquete, transporte, calidad
Electrical appliances	servicio, reparación, cambio, garantía, lavadora, tienda, marca, ordenador, reclamación, tele, ACER, aparato, avería, denuncia, establecimiento, fallos, microondas
Housing	piso, arras, casa, problema, puerta, vecino, contrato, inmobiliaria, vivienda, cliente, empresa, propietario, alquiler, ascensor, reparación, comunidad, parking, edificio, fianza
Hotel industry	hotel, tarjeta, reserva, nieve, noche, importe, viaje, camping, campista, agosto, autopista, cancelación, caravana, Llagostera, PortAventura, restaurante, recargo
Finance industry and insurers	abogado, abuso, banco, entidad, cargo, cobertura, coche, complemento, compraventa, conductor, correduría, deuda, dinero, escritura
Services	contrato, curso, dinero, empresa, autoescuela, bono, cambio, casa, honorario, factura, gimnasio, guardería, enseñanza, estudio, fotografía, formación, gestora
Professional services	banco, gestoría, gestión, abogado, dentista, psiquiatra, medicación, dentadura, hospital, letrado
Supplies	gas, factura, luz, suministro, consumo, contador, contrato, agua, euros, servicio, Endesa, vivienda, mantenimiento, domicilio, reclamación, canon, electricidad, abuso, aparato, caldera, inspección, subida, teléfono, recibo, suministradora, apagón, calefacción, cuota, facturación, lampista
Telephone and Internet	contrato, factura, teléfono, abonado, abuso, acceso, activación, servicio, llamada, permanencia, baja, Internet, portabilidad, operador, línea, Vodafone, Jazztel, Adel, sms, penalización, cuota, telefonía, móvil, conexión, contratación, contraoferta, prepago, blackberry, Movistar
Passenger air transport	vuelo, billete, compañía, reclamación, aeropuerto, avión, destino, retraso, salida, billete, maleta, reserva, pasajero, easyjet, Aerlingus, compensación, Iberia, espera, indignación, viaje-circuito, volcán
Transport	cinturón, taxi, taxista, trayecto, minusválidos, peaje
Vehicles	moto, taller, coche, concesionario, vehículo, garantía, fallo, problema, dinero, marca, vendedor, reparación, freno, motor, pieza, taller, airbag, fabricante, motocicleta, avería, centralita, distribuidor, Honda, caravana, carburador, ciclomotor, coche, embrague, homologación, Suzuki, volante, válvula
Travel agencies	viaje, dinero, euros, reserva, crucero, devolución, hotel, importe, reembolso, adelanto, anulación, compañía, reclamación, tour, agencia, alquiler, reembolso, Tailandia

The analysis of extracted terms according to semantic categories paves the way for their insertion into the available domain ontologies. This task will be described in Section 5.

Secondly, in order to extract complex terms, we applied a series of morphosyntactic grammars to the annotated corpus. The grammars correspond to patterns which are recurrently carriers of conceptual meaning in specialized discourse and, more concretely, in legal discourse. In the line of the approach followed for the extraction of simple terms, the grammars we chose are all syntagmatic units with a noun header[10]. By applying the grammars to our corpus we observed that not all the patterns were suitable for non-specialized discourse, specially the most complex patterns with embedded noun phrases (such as N+PREP+ART+N+ADJ) and those containing a syntactic inversion (ADJ+N, or ADJ+N+PREP+N).

The patterns that were finally applied are summed up in *Table* 2 with their corresponding examples. Similarly to simple terms, multiword terms denote either domain actors (*air company -compañía aérea-*, *motorcycle insurer -empresa aseguradora de motos-*, *phone company -compañía de telefonía, compañía de teléfono-*, *voice over IP operator -operador de voz por ip-*, *debt collector -empresa de gestión de cobros-*); events giving place to the conflict between seller and consumer (*unexpected flight connection -escala imprevista-*, *undue charging -cobro indebido-*, *erroneous fee -error de tarificación-*, *damages on a wall -desperfectos en una pared-*, *uninhabitable house -inhabitabilidad de la vivienda-*); or events creating a contractual relation (*deed signature -firma de la escritura-*, *purchase deposit -firma de las arras-*).

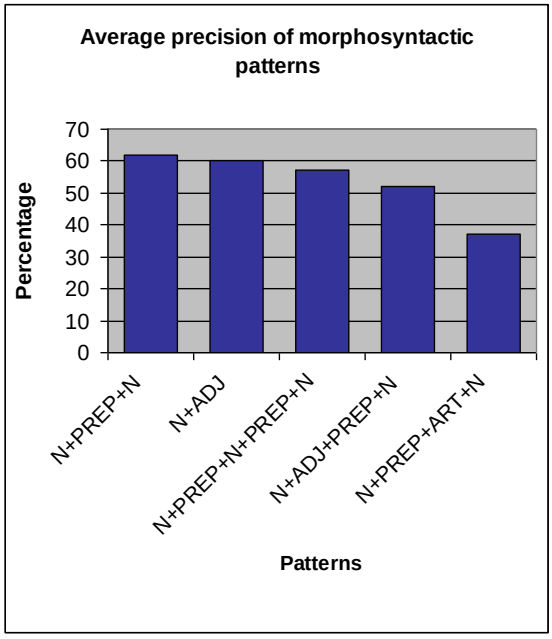
Some extracted terms deserve a particular attention. This is the case of *volcano cloud* (*nube volcánica)* and *Icelandic volcano (volcán islandés)*. A priori and out of context these terms do not belong to the domain of consumer law, but to geologic phenomena. However they appear repeatedly in consumer queries as a source of conflict in the domain of air passenger transport. This makes evident that once general normative provisions materialize in real facts, the control over concepts and vocabulary becomes more and more sophisticated, because there is no predefined domain restriction.

Table 2. Sample of multiword terms

Pattern	**Examples**
N+ADJ	compañía aérea, vuelo regional, escala imprevista, nube volcánica, volcán islandés, cobro indebido
N+ADJ+PREP+NC	acción redhibitoria por vicios, placa identificativa de voltaje, empresa aseguradora de motos
N+PREP+N	compañía de telefonía, compañía de teléfono, contrato de Adsl, error de tarificación, fecha de activación
N+PREP+N+PREP+N	fecha de fin de permanencia, operador de voz por ip, empresa de gestión de cobros
N+PREP+ART+N	desperfectos en una pared, firma de la escritura, inhabitabilidad de la vivienda, firma de las arras

[10] The set of grammars was built on the basis of a legal corpus in [12].

Figure 5. Precision of morphosyntactic patterns



The levels of precision of patterns vary, but they are mostly situated between 50% and 60% of precision. The percentage of precision shown in *Fig*. 5 has been calculated as an average of the precision of each pattern per domain, so all patterns did not have the same performance in all domains. For instance, the pattern with a higher average precision, N+Prep+N (62%), presented a considerably lower level of precision in the domain of Transport (50%), while in the domain of electrical appliances the precision reached 75%. Similarly, the pattern N+Adj, with an average precision of 60%, has a precision of 56% in transport and of 70% in Supplies. This fact might be related to the length of the corpus Transport (which is the shortest with around 400 tokens). The levels of performance of grammars will be studied in a detailed way per domain in further research.

Furthermore, in the follow-up of the project we plan to add statistical measures to reduce the levels of noise, as proposed by the most efficient current terminology extractors ([2], [22]). It is further to be noted that at this initial stage we did not set up a threshold of occurrence in the corpus, so we included all candidate terms even if they had a frequency of just 1. As it will be detailed in Section 6 one of the core issues of the ONTOMEDIA project is to redefine the notion of "term" in the light of user-generated content. Both morphosyntactic patterns and statistical measures currently applied to the detection of domain terms in a specialized text will have to be tuned to the characteristics of user-generated corpora. We plan to apply the results of the analysis of our corpus to the design of a new set of NLP tools tailored to the nature of user queries.

As an initial step in this direction, however, we consider that the results obtained are rich enough to support lexically the available ontologies. This is shown in the next section.

5. MAPPING OF CONSUMER TERMINOLOGY WITH THE MEDIATION CORE-ONTOLOGY (MCO) AND THE CONSUMER MEDIATION ONTOLOGY (CMO)

5.1. The available domain ontologies

The Mediation-Core Ontology [23] contains the basic concepts of the domain of mediation, since it is aimed at providing the conceptual anchors for the set of domain mediation ontologies that will be developed in the ONTOMEDIA platform. This way, its top classes denote the agents involved in the mediation process (MediationAgent), any information source used in the process (MediationInformationSource), the mediation process according to the domain (MediationProcess) (*Fig*. 6), the different phases of the process (MediationProcessStage), the sessions of the mediation process (MediationSession), the roles that actors might play in the mediation process (MediationRole) and the domains in which mediation can intervene (MediationTopic). It may be noticed that MCO is a structured general ontology that focuses on the mediation *system*, while the second one (COM) is a domain ontology especially focused on legal *institutional* features [5]. The underlying conceptual structure of MCO points to the social, political and economic features of the ADR, ODR and relational justice processes —including negotiation, Victim-Offender Mediation (VOM-) and transitional justice.

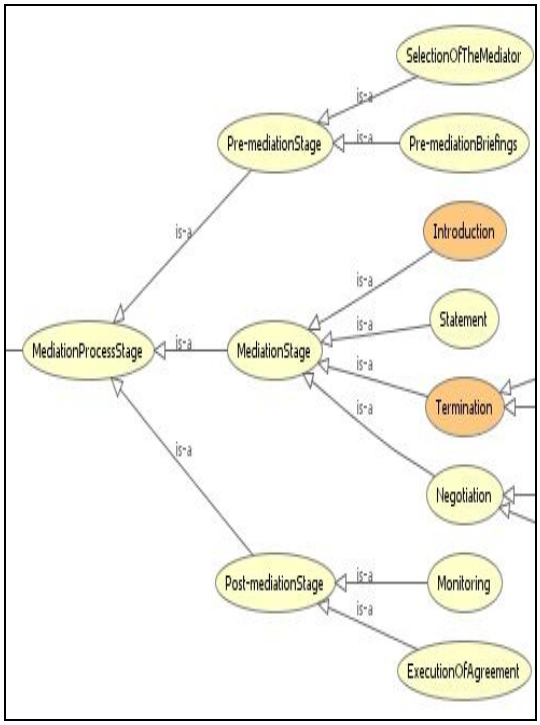
On the other hand, the (CMO) ontology [24] focuses on the particularities of mediation in the consumer domain. Its main classes denote the parties involved in the conflict (PartiesinConflict), the regulation applicable to the conflict (Regulation), the geographic area (Territory), and the type of conflict.

5.2. Integrating consumer terminology into the ontologies

Among the lists of topics provided by the Core-Mediation Ontology as subclasses of the MediationTopic top class we find ConsumerTopic. One possibility would thus be to link all the extracted terms to this class through a has_lexicalization property. However, this would imply the loss of the fine-grained classification of terms by topic presented in the previous section. This is why we decided to create 14 OWL subclasses of the class ConsumerMediation corresponding to each of the domains and to link to each of the subclasses the terms belonging to each domain.

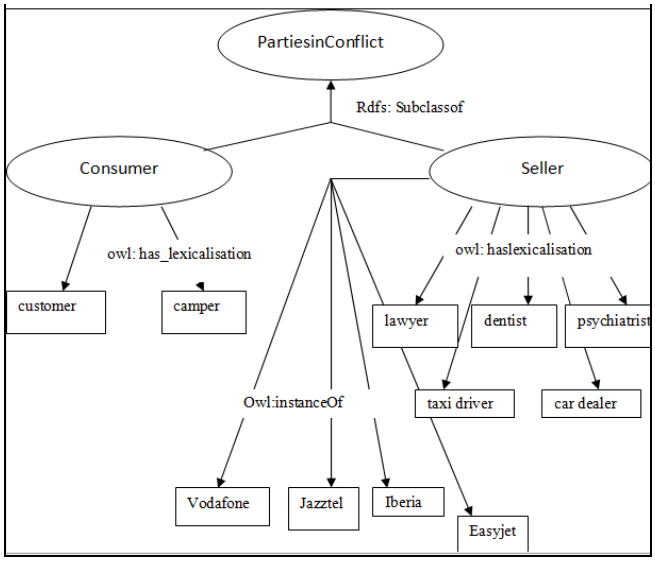
Once the extracted terms were mapped to the newly created OWL subclasses we linked the terms to the COM, this time according to their semantic nature and not to the topic they belong go. We do not reproduce here all mapped terms. The main semantic typologies of extracted terms were presented in Section 4.

Figure 6. Fragment of the Mediation-Core Ontology. Source: [24]



As an example we show in *Fig.* 7 how we linked to the class `PartiesinConflict`, and more concretely, to its subclasses `Consumer` and `Seller`, respectively, some of the terms we identified as being actors in different consumer domains. The figure shows as well the introduction of some named entities as instances of the class `Seller`.

Figure 7. Integration of consumer terminology into consumer mediation domain ontology



6. CONCLUSIONS AND FURTHER WORK

This paper has presented a case study of the trend towards the convergence between Web 2.0 and Web 3.0. More concretely, it has described the initial steps taken in the ONTOMEDIA project in order to align the knowledge and linguistic structures used by citizens to represent their conflicts in the domain of consumer law with the expert and institutional knowledge of the domain. The main result consists in a terminological extraction from a user-generated corpus and in the enrichment of two available domain ontologies with the extracted terms.

Moreover, the paper has provided some hints on the theoretical issues that underlie the Ontomedia project. More precisely, this research opens a Pandora's Box in terms of automatic processing of user-generated content. Indeed, several issues in the domain of Natural Language Processing will have to be tackled in the follow-up of the Ontomedia project in order to ensure the efficiency of the semantic platform.

First of all, as mentioned above, an in-depth analysis of the notion of *term* is required. *Term* has been traditionally defined as a linguistic unit carrying conceptual meaning in a particular domain. The morphosyntactic characteristics of terms have been widely studied with regard to technical texts, but research on the linguistic form taken by terms in common-language discourse are much less common. In this paper we provided an analysis of user-generated corpora on the basis of morphosyntactic grammars previously designed for the processing of legal texts. We saw that not all of them were reusable and this indicates that a more detailed analysis of the linguistic characteristics of user-generated content in the domain of consumer law is required. Aspects such as *unithood*, that is to say, the level of stability of syntagmatic combinations [2] and *termhood*, the extent to which terms are representative of a domain will have to be re-explored in user-generated texts.

On the basis of these observations, one of the hypothesis on which our future work will rely is that *term is any linguistic unit which is carrier of concepts relevant for the description of any type of conflictive situation in the domain*.

We expect this provisional definition to enable us to render more objective the task of annotating the relevant domain terms in a user-generated corpus. This will furthermore enable us to measure recall and thus to overcome one of the limitations of our current approach (since we were only able to measure precision and could not estimate the number of potential terms left out by our grammars).

Secondly, in our future work we will have to deal with orthographic errors and common abbreviations in short online messages (i.e. *cía* instead of *compañía –company-*). We will have to deal as well with language mixture in some queries, since both Catalan and Spanish being official some citizens mix both languages in their message (this occurs specially when they are

using reported speech and literally quoting what was said by the seller or by another state agency in Spanish).

Thirdly, we observed two potentialities in our corpus that will have to be exploited in the future. The first one refers to the presence of terms in more than one topic subset. Exploiting this multiple occurrence as links between terms expressed in the form of graphs might give us an idea of the semantic relations between different "consumer topics". The second one refers to the presence of a large number of expressions denoting psychological states (*powerlessness -situación de impotencia-*, *leg-pull -tomadura de pelo-*). The construction of a database of those expressions might be useful in other domains of the mediation platform.

In terms of ontological models, it should be noted that we were able to find anchors in the available domain ontologies for linking the terms extracted from the user-generated corpus. This indicated that even if domain ontologies are difficult to use in an open environment, in a relatively restrained legal-institutional environment we build on them, because citizens, in a way, are already adapting their discourse to what they believe are the available institutional mechanisms.

Finally, we might consider adding to the available domain ontology an ontological representation of the workflow of treatment of queries and complaints by the Agency, and of the specific services dealing with them in order to enhance the semi-automatic redirection of questions.

ACKNOWLEDGEMENTS

ONTOMEDIA: Platform of Web Services for Online Mediation, Spanish Ministry of Industry, Tourism and Commerce (Plan AVANZA I+D, TSI-020501-2008, 2008-2010); ONTOMEDIA: Semantic Web, Ontologies and ODR: Platform of Web Services for Online Mediation (2009-2011), Spanish Ministry of Science and Innovation (CSO-2008-05536-SOCI).

7. REFERENCES

[1] Barral, I. and Suquet, J. (2010-2011). "La mediación en el ámbito de consumo", in P. Casanovas, J. Magre, M.E. Lauroba (Dir.), *Libro Blanco de la Mediación en Cataluña*, Generalitat de Catalunya, Dpt. de Justícia, Cap. 3, Barcelona, pp.7 ff.

[2] Cabré, M.T.; Estopà, Rosa; Vivaldi, Jordi (2001). "Automatic term detection: a review of current systems". In *Recent Advances in Computational Terminology*. John Benjamins, Amsterdam; Philadelphia, pages 53-87.

[3] Cabré, R.; Estopà, R. (2005) "Unidades de conocimiento especializado, caracterización y topología". Cabré, M. T.; Bach, C. (2005) *Coneixement, llenguatge i discurs especialitzat.* p. 69-94.

[4] Casanovas, P. and Poblet, M. (2008). "Concepts and Fields of Relational Justice", in P.Casanovas, G. Sartor, N. Casellas, R. Rubino (Eds.) *Computable Models of the Law: Languages, Dialogue, Games, Ontologies*, LNAI 4884, Springer Verlag, Berlin, Heidelberg, pp. 323-342.

[5] Casanovas, P.; Poblet, M. (2009). "Esquema general de los conceptos y ámbitos de la justicia relacional", in P. Casanovas *et al.*, (Eds.) *Materiales del Llibro Blanco de la Mediación en Cataluña. Vol. I. La mediación: conceptos, ámbitos, perfiles, indicadores.* Generalitat de Catalunya, Departament de Justícia, Centre d'Estudis Jurídics i Formació Especializada, Justícia i Societat, n. 32, pp. 21-33.

[6] Casanovas, P. (2009). "The Future of Law: Relational Law and Next Generation of Web Services", in M. Fernández-Barrera, P. de Filippi, N. Nuno Andrade; M. Viola de Azevedo Cunha; G. Sartor; P. Casanovas (Eds.). *The Future of Law and Technology: Looking into the Future.Selected Essays.*, European Press Academic Publishing, Legal Information and Communication Technologies Series, vol 7, Florence, 2009, pp. 137-156.

[7] Casanovas, P. (2010) 'Legal Electronic Institutions and ONTOMEDIA: Dialogue, Inventio, and Relational Justice Scenarios', in P. Casanovas, U. Pagallo; G. Sartor; G. Ajani (Eds.) , *AI Approaches to the Complexity of Legal Systems (AICOL I-II) The Semantic Web, multilingual ontologies, multiagent systems, distributed networks*, LNAI 6237, Springer Verlag, 2010 (forthcoming).

[8] Casanovas, P.; Magre, J.; Lauroba, M.E. (Dir.) (2010) *Llibre Blanc de la Mediació a Catalunya*. Generalitat de Catalunya, http://www.llibreblancmediacio.com, 1184 pp.

[9] Casanovas, P.; Magre, J.; Lauroba, M.E. (Dir.) (2011) *Libro Blanco de la Mediación en Cataluña.* Generalitat de Catalunya, http://www.llibreblancmediacio.com, 1206 pp. (with the code of LEI for mediation).

[10] D'Aquin,M.; Motta, E.; Sabou, M.; Angeletou, S.; Gridinoc, L; Lopez V.; Guidi, D.; "Toward a new Generation of Semantic Web Applications", IEEE *Intelligent Systems* (2008), May/June, pp. 20-28.

[11] Fensel, D. STI Technical Report 2008-01-10, STI Innsbruck, http://www.sti-innsbruck.at/fileadmin/documents/SemanticTechnology.pdf

[12] Fernández-Barrera, M. (*forthcoming*) *From specialised legal knowledge to user-generated knowledge through legal ontologies: paving the way towards a Semantic Web 2.0.* PhD dissertation, European University Institute, Florence, Italy

[13] González-Conejero, J. and Meroño, A. (2011). "Mediation Tools for eGovernment: the MediWeb and MediApp applications" (unpublished paper).

[14] Hendler, J. (2009). "Web 3.0 emerging" (january 2009), *IEEE Intelligent Systems*, pp. 88-90.

[15] Lux, M. and Dsinger, G. (2007) From folksonomies to ontologies: Employing wisdom of the crowds to serve learning purposes. *International Journal of Knowledge and Learning (IJKL)*, 3(4/5): 515-528.

[16] Mika, P. (2005) Ontologies are Us: a Unified Model of Social Networks and Semantics. In *Proc. of Int. Semantic Web Conf.*, volume 3729 of *LNCS*, pp. 522-536. Springer.

[17] Motta, E. and M. Sabou (2006). Next Generation Semantic Web Applications. In R. Mizoguchi et al. (Eds.) [ASWC

2006], *The Semantic Web*, LNCS 4185, Springer, Heidelberg, Berlin, pp. 24-29.

[18] Noriega, P. (2007). "Regulating Virtual Interactions", in P. Casanovas, P. Noriega, D. Bourcier, and F. Galindo, *Trends in Legal Knowledge, the Semantic Web and the Regulation of Electronic Social Systems.* Papers from the B-4 Workshop on Artificial Intelligence and Law. May 25th-27th 2005. XXII WorldCongress of Philosophy of Law and Social Philosophy. *IVR 05',* Granada, May 24th-29th 2005. European Press Academic Publishing, Florence, 2007, pp. 55-77.

[19] Noriega, P and López, C. (2009). "Toward a platform for Online Mediation", in M. Poblet, U. Shield, J. Zeleznikow (Eds.) Proceedings of the Workshop on Legal and Negotiation Support Systems 2009, in conjunction with the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009), Barcelona, June 12th (2009), IDT Series 5, 67-75 http://www.huygens.es/site/service4.html

[20] Noriega, P.; Lopez de Toro, C. (2011). *Software de Desarrollo*, en P. Casanovas, J. Magre, E. Lauroba (Dirs.), *Libro Blanco de la Mediación en Cataluña*, http://www.llibreblancmediacio.com

[21] Passant, A. (2007) Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs. In Int. Conf. on Weblogs and Social Media, 2007.

[22] Pazienza, M.T., Pennacchiotti, M., Zanzotto, F.M. (2005) "Terminology Extraction: An Analysis of Linguistic and Statistical Approaches". In Studies in Fuzziness and Soft Computing. Vol 185, pages 255-280.

[23] Poblet, M., Casellas, N., Torralba, S., Casanovas, P. (2009) "Modeling Expert Knowledge in the Mediation Domain: A Mediation Core Ontology", in N. Casellas et al. (Eds.) *LOAIT- 2009. 3rd Workshop on Legal Ontologies and Artificial Intelligence Techniques joint with 2nd Workshop on Semantic Processing of Legal Texts*. Barcelona, IDT Series n. 2.

[24] Poblet, M.; Casanovas, P.; López-Cobo, J.M. (2010). "Online Dispute Resolution for the Next Web Decade: The Ontomedia Approach", *Journal of Universal Computer Science*, Proceedings of the 10th International Conference on Knowledge Management and Knowledge Technologies, Graz, Austria, pp. 117-125.

[25] Poblet, M.; Noriega, P.; Suquet, J.; Gabarró, S.; Redorta, J. (2011). "Capítulo 16: Tecnologías para la mediación en línea: estado del arte, usos y propuestas", in P. Casanovas, J. Magre, E. Lauroba (Dirs.) *Libro Blanco de la Mediación en Cataluña*, pp.

[26] Sartor, G. et al. (2008). Computable Models of the Law and ICT: State of the Art and Trends in European Research. In P. Casanovas et al. (Eds.) *Computable Models of the Law: Languages, Dialogues, Games, Ontologies*, LNAI 4884,Springer, Heidelberg, Berlin, 2008, 1-20.

[27] Sartor, G.; Casanovas, P.; Biasiotti, M.A.; Fernández-Barrera, M. (Eds.) (2011). *Approaches to Legal Ontologies. Theories, Domains, Methodologies*. LGT Series n. 1, Springer VerlagDordrecht, Heidelberg, London, New York, 2011. ISBN: 978-94-007-0119-9, pp. 1-15.

[28] Schmid, H. (1994) "Probabilistic Part-of-Speech Tagging Using Decision Trees". In Proceedings of the International Conference on New Methods in Language Processing (1994), pp. 44-49.

[29] Silberztein, M. (2003) NooJ Manual. Available for download at: www.**nooj**4nlp.net/

[30] Specia, L. and Motta, E. (2007) Integrating folksonomies with the semantic web. *Proc. Euro. Semantic Web Conf.*, 2007.

[31] Spivack, N. (2007). "The Semantic Web, Collective Intelligence and Hyperdata", at http://novaspivack.typepad.com/nova_spivacks_weblog/2007/09/hyperdata.html

Formal Models of Sentences in Dutch Law

Emile de Maat[1] and Radboud Winkels[1]

[1] Leibniz Center for Law, University of Amsterdam
{e.demaat, r.g.f.winkels}@uva.nl

Abstract. A main issue in the field of artificial intelligence and law is the translation of sources of law that are written in natural language into formal models of law. This article describes a step in that transformation: the creation of models for individual sentences in a source of law. The approach uses a natural language parse to analyse the sentence, and then translates the resulting parse tree to a formal model, using both generic and law-specific attributes. We show how the formal models can be expressed as OWL statements for legal reasoning using HARNESS.

Keywords: Automated Modelling, Natural Language Processing.

1 Introduction

A main issue in the field of artificial intelligence and law is the transformation of sources of law that are written in natural language (and therefore rather informal) into formal models of law that computers can reason with. This is a time and effort consuming process, and error prone. Also, different knowledge engineers will arrive at different models for the same sources of law. In addition, these models should be closely linked to the original sources (and at the right level of detail, i.e. isomorphic) since these sources tend to change over time and maintenance of the models is a serious problem. This calls for tools and a method for supporting this modelling process and increasing inter-coder reliability.

We have been researching a method to create isomorphic models semi-automatically, focusing on (Dutch) laws. This article presents a next step in this creation process.

1.1 General Approach

In order to achieve (semi-)automatic modelling of legal sources, we follow a number of steps, as shown in figure 1.

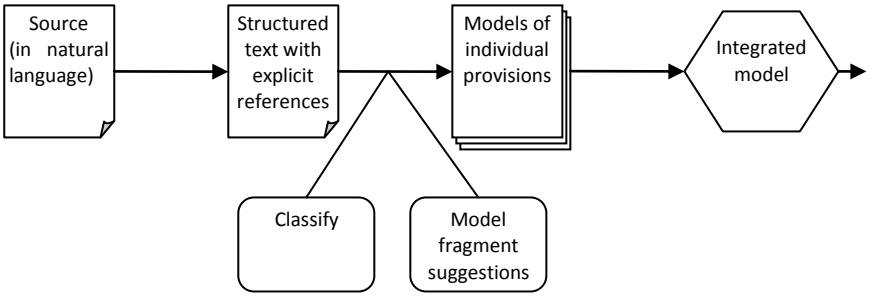


Fig. 1. Steps in automatic modelling of legal texts

The process starts with the source document, written in natural language (Dutch). Currently, we focus on laws, though we hope to expand to other types of sources of law later on. We first make the structure of the document explicit, by marking up the different parts, such as chapters, paragraphs and sentences, and assigning identifiers to each part. We then proceed to mark all references to other legal sources that are contained in the text, using a parser based on patterns for references [6]. This structure and reference information is stored in CEN/MetaLex XML[1].

The next step is to create models for each individual statement in the text. In most cases, each sentence in Dutch law forms a complete statement (though possibly part of a bigger construct), so we are, in fact, creating a model for each sentence in the text. In the last step, these individual models are integrated with each other to come to a complete model.

In order to create the models, we start by classifying each sentence in the text as a specific provision, such as a definition, a duty, or a modification of an earlier law. In total, we recognise ten different main categories. As with the references, this is done by recognising certain patterns in the text [7].

For several types of statements, such as modifications and setting the enactment date or citation title, recognising the pattern and classifying the sentence is also nearly sufficient for creating a model of the sentence. For example:

> **Aliens Act 2000**
> This law is referred to as: Aliens Act 2000.

This sentence is classified by the pattern "is referred to as", which splits the sentence in two parts: a reference (recognised by the reference parser) to "this law" and a citation title. This is all the information that is needed to represent the meaning of this sentence[2]:

Citation Title	
Target	This law
Citation Title	Aliens Act 2000

[1] See http://www.metalex.eu/

[2] As said, this also holds true for sentences containing modifications to other legal sources. However, for such sentences, analysis of the modified text is needed to determine the full impact (not meaning) of such a sentence.

Another example is an insertion of text somewhere, e.g.:

> **Law of May 13th, 2004 (Stb. 2004/220), article I, sub A**
> After article 7:1 a new article is inserted in section 7.1, reading: ...

We need to model two pieces of information. Firstly, the text to be inserted, which can be found after the colon. Then, we need to know where to insert it. The (part of the) document is denoted by a reference that is preceded by *in* (in this case: *in section 7.1*). The location is given by another reference, preceded by either *before* or *after*[3].

Insertion	
Target	Section 7.1
Location	Article 7:1
Position	After
Text	...

More elaborate sentences, that contain terms relating to the subject matter that the law is about, require more detailed analysis[4]. A natural language parser can provide such a more detailed analysis. This paper describes our initial experiences while using a natural language parser to enhance the input for our modeller.

For this research, we have used the Alpino parser for Dutch [3] to parse the sentences. The Alpino parser assigns a dependency structure to the sentence. These structures are described in [3]:

> Dependency structures make explicit the dependency relations between constituents in a sentence. Each non-terminal node in a dependency structure consists of a head-daughter and a list of non-head daughters, whose dependency relation to the head is marked.

The dependency structure can be stored as an XML file, which is the format we use as input for our modeller.

2 Creating Model Fragments

The idea to extract meaning from (parsed) sentences is not new. For example, Bos et al. [2] translate parse trees to first order logic, such as this sentence:

> *The school-board hearing at which she was dismissed was crowded with students and teachers*

This results in the following first-order logic statement:

[3] More complex positioning sometimes occurs, but will not be discussed here.

[4] This applies to norms, definitions and many application provisions. Earlier research (de Maat and Winkels, 2008) suggests that these comprise about 64% of the sentences encountered.

```
∃a((school-board(a) & hearing(a)) & ∃b(female(b) &
∃c(dismiss(c) & (patient(c,b) & (at(a,c) & ∃d(crowd(d) &
(patient(d,a) & ((∃e(student(e) & with(d,e)) & ∃f(teacher(f)
& with(d,f))) & event(d)))))))))
```

A similar approach has been applied to legal texts by McCarthy [9], who transforms parse trees to quasi-logical form.

As a basis for computer models, these logical statements seem too fine-grained. For our goals, we need a model that represents the situation described by the sentence, and not necessarily the sentence itself. So, our models will more resemble those of Sarwar Bajwar et al. [11] who generate UML models from parse trees, or those by Biagioli [1] et al., who has modelled Italian laws. Biagioli et al. used fixed fields for their frames; for example, for an obligation, their approach attempts to fill the slots addressee, action and third-party. This ignores the remainder of the text. We aim to include all elements of each sentence, though this means that their role will sometimes be not as clearly defined.

For normative sentences, this means that we see each normative sentence as describing a situation that is allowed or disallowed. We consider the main verb of a sentence as the action that is allowed or disallowed, with the other elements being modifiers or properties of that action. A number of these other elements are labelled according to their semantic role (or thematic relation) in the sentence. The other elements are considered as generic modifiers. At the moment, we distinguish only the agent, patient and recipient of the action ([10]). Other researchers have already been working at classifiers to assign semantic roles [8], and we hope to adopt one of those in the future, but for the moment, we use two simple schemes for labelling them, one for active sentences, and one for passive sentences.

In an active sentence, we assume that:

- the subject is the agent of the action;
- the direct object is the patient of the action;
- the indirect object is the recipient of the action.

For example:

Our Minister issues a warrant to the negligent person.

The main verb of this sentence is *to issue*, so that is considered the action. Properties of this action are the agent (*Our Minister*), the patient (*a warrant*) and the recipient (*the negligent person*). All these elements are distinguished by the Alpino parser (as subject, direct object and indirect object), allowing us to extract them for our model.

Within Dutch law, this sentence format expresses an obligation, so the action as a whole is classified as an obligation.

Obligation	
Action	Issue
Agent	Our Minister
Patient	Warrant

Recipient	negligent person

The articles (*the*, *a*) are left out of the model, though they are stored internally, as they are of importance during later integration of the model; *the negligent person* often is a reference to an earlier sentence, whereas *a negligent person* is not.

The example above is an active sentence, but many sentences in Dutch law are phrased in the passive voice, such as this instruction:

An English translation is added to this report.[5]

A sentence in the passive voice cannot be modelled in the same way as a regular sentence, as the subject of the sentence is not the agent, but the patient, and should be modelled as such. Again, the parse of the sentence gives us an easy way to do this:

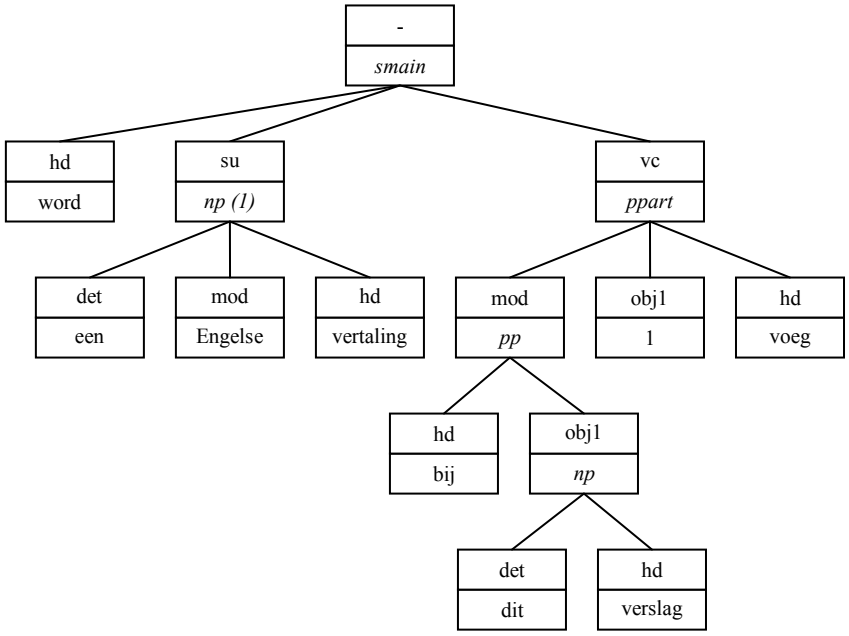


Fig. 2. Alpino parse tree (with reduced information) for "An English translation is added to this report" (in Dutch).

The verb clause (vc) of the sentence holds the sentence in active voice, with the subject re-cast in the role of object. By modelling the verb clause instead of the sentence as a whole, we get the correct model, with the correct object, and without the auxiliary verb.

If the agent is present in the sentence (for example, if the sentence would read *An English translation is added to this report by the organiser)*, then this prepositional

[5] Law for the protection of Antarctica, article 33, sub 3

object is not re-cast in the role of object in the tree. We will have to detect its presence by scanning for signal words like *by*. As this does not always indicate the agent, this will be one of the cases were human validation is necessary. Further detail can be added by splitting of adjectives and relative clauses from the noun they modify. For example, *negligent person* has two properties: being a person and being negligent. Splitting adjectives from nouns is not always desirable; it is preferable to leave multiword expressions intact. *European Union* is not any union that is also European; *Our Minister of Finance* is not any minister that is also ours, and of finance[6]. Instead, these are references to concepts that have been defined elsewhere: the common sense domain, the juridical domain or elsewhere in this law. Common multiword expressions are recognised by the Alpino parser; juridical domain or law-dependent expressions need be filtered out separately.

Relative clauses are more complex then adjectives, as they contain a complete new sentence. In this case, we repeat the procedure for the main sentence, identifying the main action and all properties of that action. For example:

Our Minister issues a warrant to the person that neglected his duties.

This sentence yields a frame like:[7]

<table>
<tr><th colspan="2">Obligation</th></tr>
<tr><td>Action</td><td>issue</td></tr>
<tr><td>Agent</td><td>Our Minister</td></tr>
<tr><td>Patient</td><td>warrant</td></tr>
<tr><td>Recipient</td><td>person

<table><tr><th colspan="2">SubjectOf</th></tr><tr><td>Action</td><td>neglect</td></tr><tr><td>Direct Object</td><td>his duties</td></tr></table></td></tr>
</table>

2.1 Filtering Out Signal Words

The sentences we showed above are examples of normative sentences that do not use signal words; only the desired situation is described, and it is left implicit that this is an obligation. Other sentences in the law use signal words to make the kind of norm explicit, such as:

The buyer is obliged to pay the price.[8]

[6] In Dutch laws, *Our Minister of Finance* is a reference to the (Dutch) Minister of Finance. No more detailed model is needed, as no derivations need to be made.

[7] For the moment we use a frame-like representation. These look somewhat like the frames presented by Van Kralingen (1995), but these were more legally oriented and had a fixed number of slots, while our structures are more dynamic and language oriented.

This sentence uses *is obliged* to make it clear that this is an obligation. Other examples of signal words are *must*, *may* and *is allowed*. These sentences require a different approach than the sentences without signal words. If we were to use the same approach, the result would be something like:

Obligation	
Action	is obliged to pay
Agent	Buyer
Patient	the price

This is not a desirable outcome, as the action that this norm deals with is *pay* rather than *is obliged to pay*. When modelling these sentences, these signal words should not be included in the model of the situation (their meaning is translated into whether the situation is allowed or disallowed). Ideally, after we've categorised the sentence (based on the signal words), we would like to transform the sentence to a sentence without signal words, like:

The buyer pays the price.

We could then model that sentence to come to a correct frame. Simply leaving out the signal words may lead to errors, since the role of the other words might need to shift as well. However, the parse of the sentence actually contains this "transformed sentence" that we want to model. This is shown in figure 2.

[8] Dutch Civil Code, BW7, article 6 sub 1.

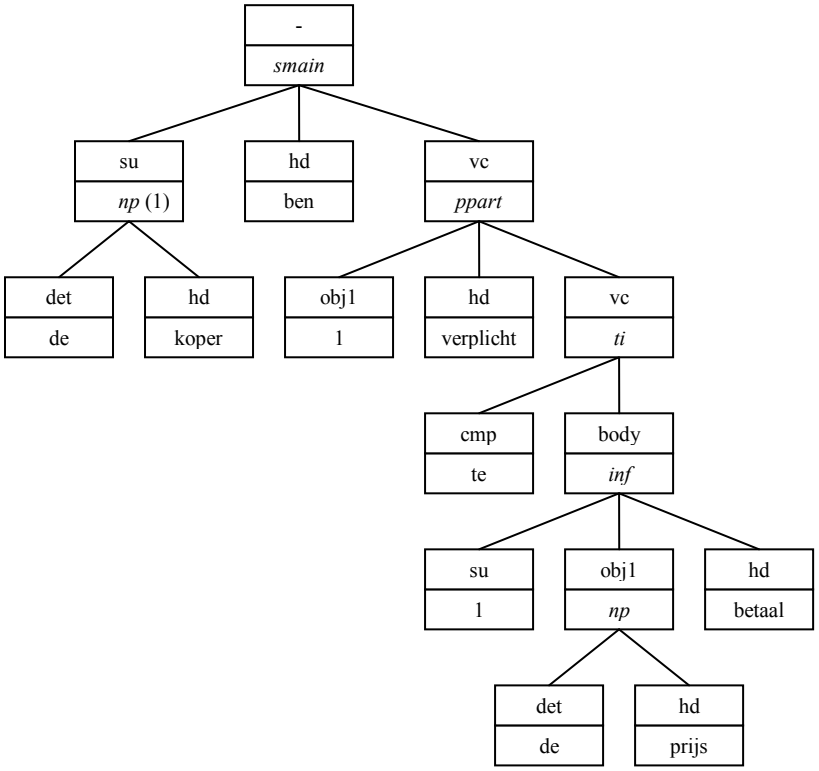


Fig. 3. Alpino parse tree (with reduced information) for "The buyer is obliged to pay the price" (in Dutch).

Beneath the body node, we find exactly the sentence that we are looking for. Alpino assigns this dependency structure to any sentence that follows this pattern. This makes it easy to filter out the signal words by simply focusing on the part of the parse tree that contains the transformed sentence.

For each pattern we use for classification, it seems possible to define a part of the parse tree that should be ignored in order to come up with a correct model.

2.2 Lists

Lists are also recognised by the Alpino parser, and can therefore easily be added to our models as the union or intersection of the different list items, depending on the conjunction used. However, though the conjunction *and* suggests an intersection, it often expresses a union instead. For example:

Advances and duties are paid in cash.

In this sentence, it is the union of *advances* and *duties* that is meant. Our current approach is to translate *and* with a union if it appears in a relative clause, and with an intersection otherwise.

2.3 Negation

Negative sentences should also be recognised, and modelled as the "positive" sentence, with the additional notion that it is inverted. This can usually be done by not including certain signal words as element in the model, but by inverting the model if it is encountered.

The most common signal word is *not*. If it is encountered, it is not added to the frame, but instead, the containing element is marked as inverted.

The determiner *no* is another example of a signal word for negation. However, it can affect more than its containing element. For example:

No bodies are interred on a closed cemetery.

This is an obligation, and the direct object of this sentence is *no bodies*. However, if we apply the negation simply to the object, i.e. the object is "not a body", it would imply the obligation to bury something that is not a body on the cemetery. Instead, we need to apply the negation to the entire sentence: One is obliged not to bury bodies at a closed cemetery.

2.4 Explicit exceptions

Sometimes, a normative sentence in a Dutch law includes a prefix to denote that it is an exception to some other rule, like:

In exception to article 12, ...

Alternatively, some sentence start with a prefix to denote that it is not an exception, like:

Without prejudice to article 12, ...

These prefixes differ from other elements in the sentence in the sense that they do not describe the situation that is allowed or obliged, but instead tell us something about how this rule interacts with some other rule. Hence, this element should not be added to the frame describing the rule.

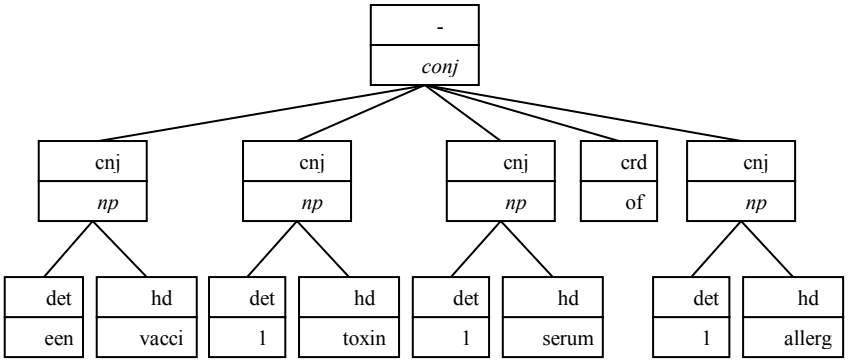
2.5 Definitions and Deming Provisions

Definitions and deeming provisions attach a meaning to a specific concept. At top level, a definition contains three elements: the *definiendum* and the *definiens*, and, optionally, a scope declaration stating for which sources of law this definition applies. For example:

> **Medication Act, article 1, introduction and item c**
> In this law and stipulations based upon it, it is understood by immunological drug: a vaccine, toxin, serum or allergen.

This definition has the scope "this law and stipulations based upon it". The *definiendum* is "immunological drug" and the *definiens* is "a vaccine, toxin, serum or allergen". Like the sentences presented in section 1.1, these top elements can easily be extracted by means of the pattern used to classify the sentence (in this case "it is understood by") and some additional features. The scope, if present, will follow the word "in*"* (and end at the text "it is understood by"). The *definiens* will follow the word "by*"* and end at the colon, and the *defiendum* will follow the colon. Thus, we can easily extract a top level frame:

Definition	
Scope	This law and stipulations based upon it
Definiendum	immunological drug
Definiens	a vaccine, toxin, serum or allergen



However, modeling the *definiens* in this way is unsatisfactory, as it gives insufficient detail to use this model for practical purposes. To create a more useful model, we need to split up the *definiens*. To do so, we use the same methods used for concepts in normative sentences. This requires a parse tree; we can either parse the entire sentence or just the *definiens*.

3 Example: Converting Frames to OWL

In HARNESS, a norm is represented as a deontic qualification of a generic case [13]. Such a generic case is a conjunction of conditions that together form a description of

the situation expressed by the norm. Such a generic case is defined as a set of conditions in conjunctive normal form. An individual case is a set of grounded propositions that describe a certain state of affairs (cf. [12]). The norm itself is qualified using the deontic notions Permission, Obligation and Prohibition, which have been defined in the LKIF Core ontology [4], as follows:

Norm ⊑ Qualification ⊓ qualifies **some** Normatively_Qualified
Permission ⊑ Norm
≡ allows **some** Allowed ⊓ allows **only** Allowed
Obligation ⊑ Permission
≡ allows **some** Obliged ⊓ disallows **some** Disallowed
⊓ allows **only** Allowed ⊓ disallows **only** Disallowed
Prohibition ≡ Obligation

So, in order to add a normative sentence to the system, we need to specify the generic case as a set of conditions. We continue with an earlier example:

Our Minister issues a warrant to the negligent person.

We can describe the generic case using the elements from the resulting frame after parsing (see above). The generic case is an action *issue* with agent *Our Minister*, patient *warrant* and recipient *negligent person*:

GC
≡ issue ⊓ ∃agent Our_Minister ⊓ ∃patient warrant ⊓ ∃recipient negligent_person

This generic case is allowed by the article, and its negation is disallowed by the article, which leads to the following complete statement in HARNESS:

GC_P ⊑ Generic_Case ⊓ ∃allowed_by{*article*}
≡ issue ⊓ ∃agent Our_Minister ⊓ ∃patient warrant ⊓ ∃recipient negligent_person
GC_F ⊑ Generic_Case ⊓ ∃disallowed_by{*article*} ≢ GC_P
article_obligation ⊑ Obligation ⊓ ∀allows(GC_P)⊓ ∀disallows(GC_F)
≡ {*article*}

4 Experiences

At this moment, we do not have a fully automated process to create the models, and have not yet tested this method on a large body of sentences. Instead, random sentences have been selected, parsed using Alpino and then fed into our modeller.

There is a clear difference between the computer generated models and those created by a human expert with regard to the granularity of the model. Our method will create models with model elements that represent one word from the original sentence, whereas a human expert is more likely to include some sentence fragments as a whole. For example, one Dutch law defines an alcoholic drink as *the drink that, at a temperature of twenty degrees Celsius, consists of alcohol for fifteen or more volume percents, with the exception of wine*. Our algorithm will dissect this sentence, whereas most human modellers will leave the first subordinate sentence intact and add it to the model as a single attribute (most likely abbreviated to *alcohol by volume*). A more detailed model seems not necessarily wrong, but quite possibly over-the-top and inconvenient for many applications.

The method assigns rather broad categorisations to each object (it is either a direct, indirect or prepositional object), but does not yet assign a legal meaning to such an

object. It may be a third party involved or the instrument. Perhaps this is not an obstacle; users dealing with a system based on such models are likely to recognise the roles from the context and language used, whereas a computer does not need this information for the derivations we currently want to make. For future projects, though, the information may be required, and some way to automatically recognise it is desired.

For the modelling of norms, we have been focussing on the sentences that represent an obligation, duty or right. For those sentences, the method seems adequate. However, for other types of sentences, such as delegation, we have not come to an acceptable approach yet. Dealing with these sentences will require first of all that we recognise them. Currently, our classifier distinguishes only between obligation/prohibition and right/permission. Several of the patterns used clearly indicate delegations, but we have not yet established whether these patterns cover all delegations in Dutch laws.

A minor problem with regard to the parses made by Alpino is that most often, the correct parse is not the one preferred by Alpino, but second, third or fourth. If we make several suggestions (each suggestion based on a parse by Alpino), this means that it will often not be the first suggestion that is correct, which means more effort is needed by a human expert who is verifying the models.

We expect that by expanding the lexicon used by Alpino, and perhaps by recalibrating the disambiguation on a written legal corpus, these problems will disappear.

4 Conclusion

We have presented a next step towards a method and tools for supporting the semi-automatic modelling of sources of law, necessary for an efficient, effective, and more reliable and pragmatic use of knowledge technology in the legal domain. We were already able to reliably detect structure in sources of law, find and resolve references in and between them, and classify individual sentences. Now we are able to suggest formal model fragments for certain types of the classifications. Though we are convinced that these model fragments will be a useful in supporting human experts creating models, we do feel that the approach is still too general. A more elaborate method is needed to create appropriate model fragments for different subtypes of sentences. Some method to avoid too detailed models is desirable as well.

References

1. Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S., Soria, C.: Automatic semantics extraction in law documents. In: Proceedings of the Tenth International Conference on Artificial Intelligence and Law (ICAIL '05), ACM Press, New York (2005) 133-140.
2. Bos, J., Clark, C., Steedman, M., Curran, J.R., Hockenmaier, J.: Wide-Coverage Semantic Representations from a CCG Parser. In: Proceedings of the 20th international conference on Computational Linguistics (2004) 1240–1246.

3. Bouma, G., van Noord, G., Malouf, R.: Alpino: Wide Coverage Computational Analysis of Dutch. In: Daelemans, W., Sima'an, K., Veenstra, J., Zavrel, J. (eds.): Computational Linguistics in the Netherlands CLIN 2000. Selected Papers from the Eleventh CLIN Meeting. Rodopi, Amsterdam (2001) 45-59.
4. Hoekstra, R., Breuker, J., Di Bello, M., Boer, A. (2007) The LKIF Core Ontology of Basic Legal Concepts, Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007),
5. van Kralingen, R. W.: Frame-based Conceptual Models of Statute Law. PhD thesis. Kluwer Law International, The Hague (1995).
6. de Maat, E., Winkels, R., van Engers, T.: Automated Detection of Reference Structures in Law. In van Engers, T.M. (ed.): Legal Knowledge and Information Systems. Jurix 2006: The Nineteenth Annual Conference. IOS Press, Amsterdam (2006) 41-50.
7. de Maat, E., Winkels, R.: Automatic Classification of Sentences in Dutch Laws. In: Francesconi, E., Sartor, G., Tiscornia, D. (eds.): Legal Knowledge and Information Systems. Jurix 2008: The Twenty-First Annual Conference. IOS Press, Amsterdam (2008) 207-216.
8. Gildea, D. and Jurafsky, D. Automatic Labeling of Semantic Roles,. Computational Linguistics, 28(3):245–288, 2002.
9. McCarty, L.T.: Deep semantic interpretations of legal texts. In: Proceedings of the 11th International Conference on Artificial Intelligence and Law. ACM Press, New York (2007) 217-224.
10. Payne, Th. E. Describing morphosyntax: A guide for field linguists. Cambridge; New York: Cambridge University Press (1997).
11. Sarwar Bajwa, I., Samad, A., Mumtaz, S.: Object Oriented Software Modeling Using NLP Based Knowledge Extraction. European Journal of Scientific Research, Vol. 35, No. 1 (2009) 22-33.
12. Valente, A. (1995). Legal knowledge engineering: A modelling approach. IOS Press, Amsterdam.
13. Ven, S. van de, Hoekstra, R., Breuker, J., Wortel, L., El-Ali, A. (2008) Judging Amy: Automated Legal Assessment using OWL 2, Proceedings of OWL: Experiences and Directions (OWLED 2008 EU).

Eunomos, a legal document management system based on legislative XML and ontologies (position paper)

Guido Boella
University of Torino
and Nomotika s.r.l.
Italy
guido@di.unito.it

Llio Humphreys
University of Torino
Italy
humphreys@di.unito.it

Marco Martin
University of Torino
Italy
notmart@gmail.com

Piercarlo Rossi
Università del Piemonte
Orientale
Italy
piercarlo.rossi@unipmn.it

Leon van der Torre
University of Luxembourg
Luxembourg
leon.vandertorre@uni.lu

ABSTRACT

In this position paper, we illustrate the ongoing work and future developments on the Eunomos software, an advanced legal document management system to classify norms, based on legislative XML representation of laws which are retrieved automatically from institutional legislative portals, which complements the tool for building legal ontologies called Legal Taxonomy Syllabus.

1. INTRODUCTION

To operate efficiently, a law firm needs to regularly create and update legal documents, classify them according to the different domains it operates into, access reliable information on the state of the law and keep track of changes in legislation, regulations and contracts.

Currently much of this work is done by hand. Law firms employ personnel who assiduously trawl through various sources to find relevant legislation and influential cases. Law management in law firms, in-house legal offices and law departments is today more complicated than ever due to the number of laws that have to be considered from various sources at many different levels - international, European, national, regional, or even internal regulations and standards. The complexity of researching the state of the law on a particular topic from various sources is difficult not only for businesses in countries such as Italy, which is well known for legislative over-production, but for any business that needs to operate in an international context and deal with multiple legislations.

Another problem is managing regulations, contracts and other documents. Most law firms today do not use dedicated document management systems even for their own legal documents. Law firms typically use folder trees as repositories and folder names as classification tags. They use master contracts to help formulate actual contracts for clients, but no links are made between elements in master contracts and derived contract instances. Different versions of contracts are often maintained using basic versioning features of standard editors for word processing.

Legal document management is more complicated than general document management, particularly regarding the requirement to continually review documents in the light of regulatory changes. This requirement means that documents need to keep track of the laws referred to by the various parts. Such links may be implicit or made explicit in the text using a legal reference. In the first case, the document must be annotated with references by a knowledge engineer. Once all legal references are made explicit, it is possible to automatically identify documents that need to be revised because the legal text referred to has been changed. Most commercial legal document management systems fail to address these issues. As far as we know, no commercial legal document management systems offer an integrated workspace for classifying relevant laws, drafting legal documents and monitoring changes in laws, norms and concepts.

Lately, articles have begun to appear in specialist[1] and even mainstream[2] press about an increased interest in bespoke ITC solutions, and in particular, human language technologies, for legal domains.

There are several factors that enter into play at this point in time:

[1]http://legalinformatics.wordpress.com/2009/08/07/susskind-on-the-end-of-lawyers/
[2]http://business.timesonline.co.uk/tol/business/law/ article7003373.ece

- subsidiarity and a significant increase in legislative activity at different levels,
- the availability of legislation online, albeit from different sources and in different formats,
- the growth of legislative XML in the public sector, to enable expert tools to access legal information,
- the cost of clerical, research and professional legal work without the support of an integrated legal knowledge management system.

The research question of this paper is thus: How to extend knowledge management systems to deal with specific needs of lawyers and law scholars?

The methodology we use is to take inspiration from the technologies developed in the related fields of legislative drafting for parliaments, so called legislative XML, and legal ontologies, and export them in the context of applications for lawyers and law scholars.

In this paper we illustrate for the first time how this methodology resulted in the Eunomos software being developed in the context of the project ICT4LAW[3] to address these needs, and compare the product with other systems in the field. Eunomos is an advanced legal document management system based on legislative XML representation of laws which are retrieved automatically from institutional legislative portals, and extending a tool for building legal ontologies called Legal Taxonomy Syllabus[2, 1].

In the next section we describe the background technologies from which Eunomos emerged: legislative XML and legal ontologies. In Section 3.1 we describe the main functionalities of the software and in Section 3.2 we describe the semi-automated classification mechanism of Eunomos. Related and future work and conclusions end the paper.

2. BACKGROUND

2.1 Legislative XML

In many regions in Europe and beyond, there are now online portals making laws and decrees available to citizens. These portals are updated on a regular, often daily basis. Some initiatives, such as legislation.gov.uk by The National Archives in the UK aim to go beyond being a legislative portal, providing a co-operative editorial tool, thereby giving others a stake and an incentive to work with them to create and maintain, open, free to use, up to date revised legislation. Every document published on their website will be available in machine readable XML format, as well as PDF.

Over the last few years, several XML specifications for legal documents have arisen in Europe with a view to make laws accessible to citizens and suitable for processing by specialist applications. Examples of legislative XML in use are FORMEX for the EU Publications Office and the NormeInRete (NIR) project defining several DTDs for Italian legislation and identifiers through URNs (Uniform Resource Names). In Denmark the government is working on LexDania and the Swiss and Austrians are also busy trying to provide better access to their legal sources with the use of XML. Boer and Winkels [5] argue that there is a need for an interchange XML standard for describing legal documents. Such a standard should be language and jurisdiction independent, but law specific. It should enable external knowledge models about (the content of) legal documents to link to text from the original sources at the right level of granularity - i.e. legally relevant subparts. The XML standard they developed, CEN Metalex, is an interchange format which defines standards for naming conventions and cross referencing for information exchange and interoperability. It is not intended to replace jurisdiction-specific standards and vendor-specific formats.

The NormeInRete (NIR) standard is a well-established legislative XML used in Italy. It specifies the structure of legal documents in terms of XML tags for metadata, articles, paragraphs, etc. and that such components of legislations should be identified through URNs (Uniform Resource Names). URNs are designed specifically for the Internet community to provide unique identifiers, unambiguous and persistent network resources, regardless of their physical location. Assigning a uniform name for each legal document aims to assign a unique identifier, in a standardized format, which depends only on the characteristics of the document itself and is independent of availability in the network, physical location and means of access. This identifier is used as a tool to represent references - and more generally any kind of relationship - between acts. It facilitates the construction of a global hypertext among legal documents in a network environment with computer resources distributed among several publishers. It also allows the construction of knowledge bases containing the relationships between these documents.

An URN for a document constructed according to the NormaInRete standard will have the following components:

1. An ID for the original document, comprising the authority responsible for publishing the law (e.g., Ministry, Region, City, Court), the type of measure (e.g., law, decree, order, decision, etc.), the date and number and IDs for any annexes.
2. A version identifier, including the date of issue.
3. The ID of the press publishing the law.
4. An identifier of the fragment of the resource itself the URN refers to (e.g., article, paragraph, etc.). The URN can be used in a HTML (<META name="nir.urn" content="urn:nir:stato:legge:1996-12-31;675̈>) or XML (<urn valore="urn:nir:stato:legge:1996-12-31;675"/>) file.

[3]ICT4LAW: "ICT Converging on Law: Next Generation Services for Citizens, Enterprises, Public Administration and Policymakers" funded by Regione Piemonte 2008-2013, call Converging Technologies 2007, website: http://www.ict4law.org

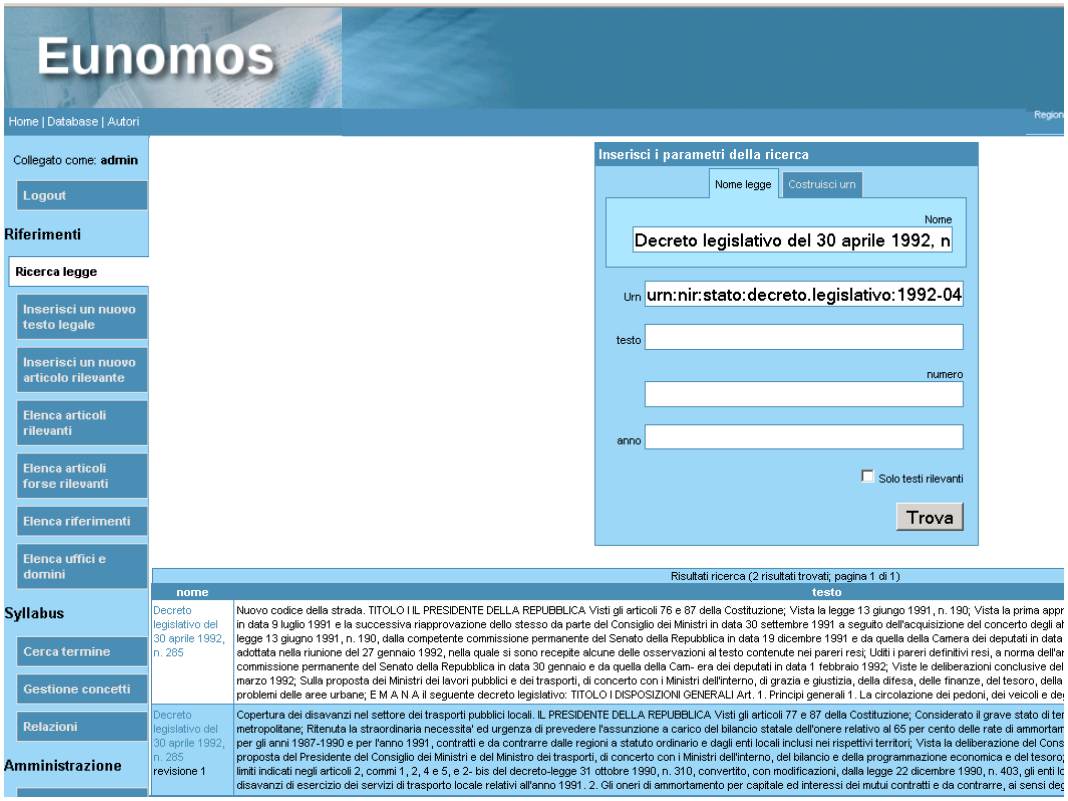


Figure 1: The search interface of Eunomos.

2.2 Ontology

The main assumptions of the Legal Taxonomy Syllabus ontology on top of which Eunomos is built come from studies in comparative law [17] and ontologies engineering [12].

- Terms –*lexical entries* for legal information–, and concepts must be distinguished; for this purpose we use lightweight ontologies [8], i.e. simple taxonomic structures of primitive or composite terms together with associated definitions.
- We distinguish the ontology implicitly defined by EU Directives (EUD), the *EU level*, from various national ontologies. Each national legislation refers to a distinct national legal ontology. We do not assume that the transposition of an EUD automatically introduces in a national ontology the same concepts that are present at the EU level.
- Corresponding concepts at the EU level and at the national level can be denoted by different terms in the same national language.

A standard way to properly manage large multilingual ontology is to make a clear distinction between terms and their interlingual acceptions (or *axes*) [19, 13]. The basic idea in our system is that the conceptual backbone consists in a taxonomy of concepts (ontology) to which the terms can refer to express their meaning. We do not assume the existence of a single taxonomy covering all languages. In fact, the different national systems may organize the concepts in different ways. For instance, the term *contract* corresponds to different concepts in common law and civil law, where it has the meaning of *bargain* and *agreement*, respectively [18]. In most complex instances, there is no correspondance between terms-concepts such as *frutto civile* (legal fruit) and *income*, but respectively civil law and common law systems can achieve functionally the same operational rules thanks to the functioning of the entire taxonomy of national legal concepts [9]. Consequently, the Legal Taxonomy Syllabus includes different ontologies, one for each involved national language plus one for the language of EU documents. Each language-specific ontology is related via a set of *association* links to the EU concepts.

3. DESCRIPTION OF EUNOMOS

3.1 Features of Eunomos

We have developed a sophisticated legal document management system based on ontology and legislation monitoring system called Eunomos with the following features:

- A large database of laws (about 70,000 Italian national laws in the current demo) maintained in XML format in accordance with the NormeInRete (NIR) standard for Italian laws.[4]

[4]The software does not depend on the specific NIR DTD, and can be used for other XML standards for other languages.

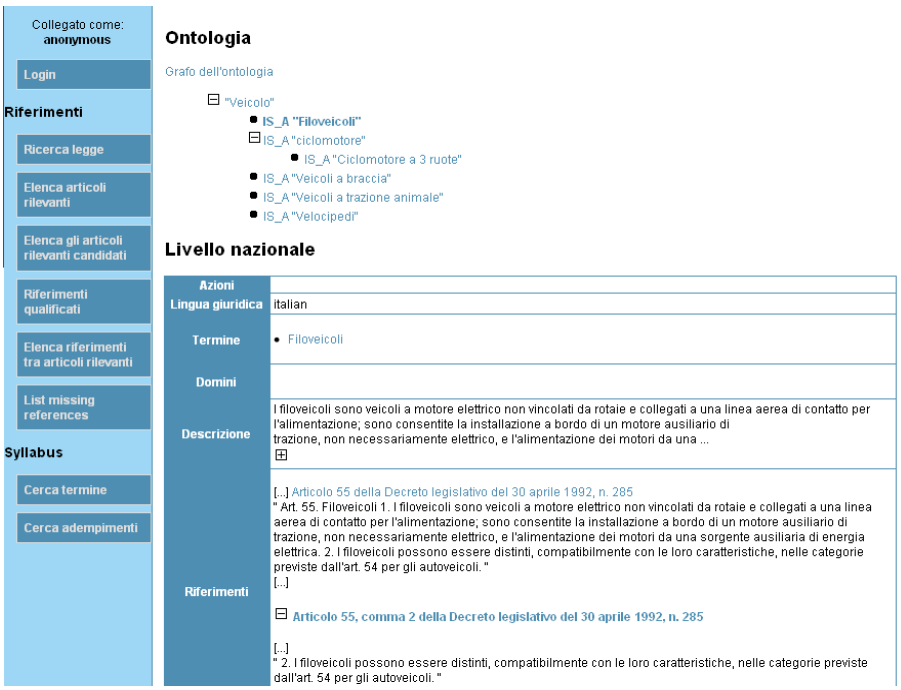


Figure 2: The Eunomos integrated ontology

- Automatic downloads of laws from institutional legal portals via dedicated spiders. Currently the software harvests the Italian national portal http://www.normattiva.it including over 50,000 laws, the portal Arianna of Regione Piemonte http://arianna.consiglioregionale.piemonte.it/ and a portal of regulations from the Italian Ministry of Economy.

- The conversion of laws into NIR XML if they are in pure textual format.[5]

- Automated parsing of legal references using the URN format of NIR.[6] This enables legal references to be transformed into hypertext links to the relevant legislation, thus facilitating automated linking and reasoning and user navigation.

- Semi-automated classification of laws at the level of paragraphs or articles according to domains specified by the expert user.

- An alert messaging system, using URN references and semantic similarity tools, that informs users of new laws downloaded into the database and suggests which existing laws could be affected by the new legislation.

- Enabling concepts from the Legal Taxonomy Syllabus ontology to be linked via URN to legal definitions within relevant legislation.

Figure 1 shows the legislation search page. The user can search legislation via name, year, or URN. The research results are displayed below in the table below the search box. The first column contains the name of the law, and a link to the full text of the relevant legislation. The second column contains a summary of the law. If coordinated versions of the norms are available, they are shown besides the original ones. The navigation on the right hand side enables the expert user to view paragraphs and articles relevant to a particular domain, view similar pieces of legislation, analyse usage of terms within the legislation, and make links between terms within legislation and concepts in the ontology.

Figure 2 shows a concept with its place in the taxonomy and a link to relevant legislation (with a link expressed as a URN to the shown article). The ontology can be created contextually to a piece of legislation, thus facilitating the creation of the link and of the description. Here we find the concept of vehicle, and sub-categories such as trolleybus, motorcycle etc. By clicking on the plus/minus signs, the user can view definitions and references for each concept displayed in a table below.

Ontology and legislation document management is designed to be an online service provided by Eunomos to several clients, information and costs are shared. Another advantage of having several clients using the model is that with more people using the system, the higher higher the like-

[5]The Arianna portal already exports documents to NIR XML format. The conversion in the current version of the software is done using the XMLeges Marker tool developed by Istituto di Teoria e Tecniche dell'Informazione Giuridica (ITTIG) of Florence (http://www.xmleges.org).

[6]This is done using the XML Leges Linker tool developed by ITTIG.

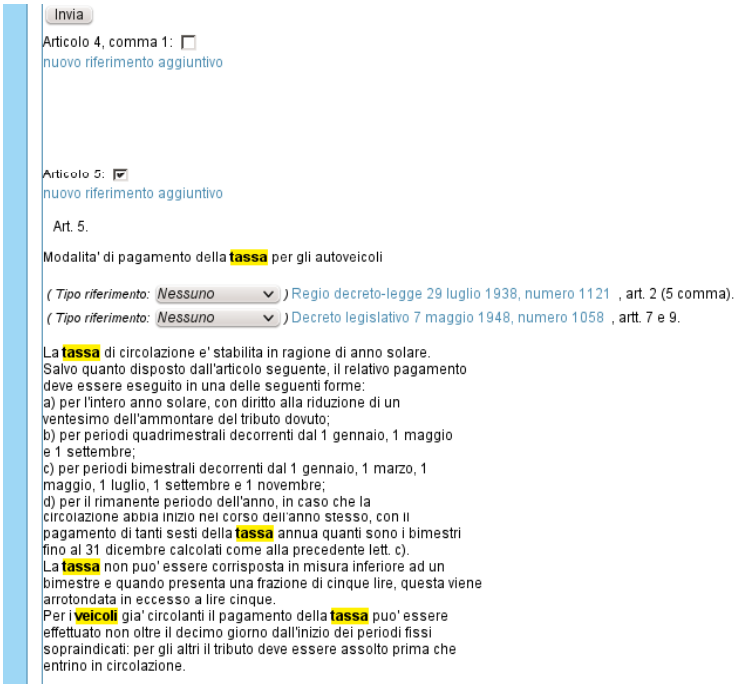


Figure 3: The law classification form.

lihood that errors are quickly detected and corrected. Putative links are verified by domain experts as a matter of course. This means that when users need to find related legislation or concepts, or various definitions for the same concept in different contexts or time-frame, they can do so with confidence that the information that the system will provide will be thorough and accurate. Users can find the information they need quickly while the task of maintaining and updating information is left to the professionals.

3.2 Eunomos and law classification

By connecting ontologies and legislation structured in XML within the database framework, Eunomos provides a powerful knowledge base for keeping up to date with legal changes. But this is a system that requires expert users to manage the information. In Italy, there are two major challenges for expert legal knowledge management systems:

1. some laws include various norms for a variety of different and unrelated topic areas;
2. some laws contain norms that implicitly override norms in other laws, but fail to include references to the norms they override.

Eunomos uses natural language technologies to help the expert user with the labour-intensive work of categorisation of norms and retrieval of implicit references. The support is based on two techniques: analysis of outgoing references, and semantic similarity. The Eunomos product provides suitable interfaces for the expert user to create a set of category labels representing domains like taxation, immigration, etc. and to associate each component of a law (identified by a URN) to a particular category. In Figure 3, we can see annotated articles from a piece of legislation. The expert user uses this interface to assign domains to each article and a type (modification, overriding, etc.) for each reference to other legislation. Terms which name concepts in the ontology are highlighted.

Where articles and paragraphs contain references to the articles and paragraphs they talk about or override, this information is used not only to link the relevant legislation via URN, but also to suggest to which category a new piece of legislation belongs. The rationale is that where paragraphs or articles contain references to classified paragraphs or articles in previous legislation, it is more than likely that the new paragraph or article belongs to the same domain. The user can check and deselect the suggested classifications.

For articles and paragraphs that do not contain explicit references, it can be useful to find relevant domains and implicit references by referring to a list of the ten most similar pieces of legislation in the whole database. Eunomos generates this list using Cosine Similarity text classification.

4. RELATED WORK

Our solution has some similarities with Pazienza et al. [4] but has a different aim, since it is not a precompetitive project, and is more wide-ranging in scope. While Pazienza et al. [4] takes XML files as input, Eunomos downloads text-based laws automatically from portals and converts them into XML, generates automatic alerts concerning possible legislative updates, and identifies norms and concepts within

new laws which can be integrated with a sophisticated, multilevel and multilingual ontology tool. The use of ontology in the two systems are also quite different. Pazienza et al. [4] use the Semantic Turkey [10] ontology, where definitions can be taken from any source and arranged in any order. The Eunomos product is more careful, encouraging the expert user to create links to definitions in legislation, judgement and official journals, and to track the evolution of terms in a systematic manner. On the other hand, Eunomos requires considerable maintenance work, as web spiders need to keep up to date with any modifications made to online legal portals, and expert users are required to verify classification and find implicit references. Pazienza et al. [4]'s text similarity tool working at a paragraph level is very interesting, and we intend to add a similar feature in the next development phase of our product.

It is instructive also to refer to de Maat et al. [6]'s comparison of machine learning versus knowledge engineering in classification of legal sentences, since Eunomos uses machine learning and knowledge engineering techniques. de Maat et al. [6] uses knowledge engineering to find standardised patterns suggestive of a particular class, while we use knowledge engineering to find standardised patterns for references to classified norms in previous legislation, which provides a clue as to the classification of new norms. On the machine learning side, de Maat et al. [6] uses Support Vector Machines for text classification, while we use Cosine text similarity to find the most similar pieces of legislation, which provides clues on relevant domains and norms that may be overridden implicitly. The conclusion of de Maat et al. [6]'s research (ibid, page 16) was that 'a pattern based classifier is considered to be more robust in the categorization of legal documents at a sentence level.' However, the classification task is quite different since that research was concerned with classifying sentences as norms, delegations, penalizations, value assignments, application provisions etc, while our classification task is to categorise norms as belonging to domains such as taxation. The author (ibid. page 14) noted that Support Vector Machines were better than knowledge engineering at categorisation where word order was less restricted, and as such may be more suitable for our work.

Concerning text classification techniques, there are a number of different solution to evaluate [14] They work on the principle of labelling a collection of documents in various categories, training classifiers on the various categories, and using these classifiers to select the most appropriate topic for a new document. Most classifiers (Naive Bayes Classifier, Bernoulli, Vector Space Model) use as features keywords that have high frequency within the topic but not in general. Some implementations remove stop-words. Some give different weights to different keywords in terms of how representative they are of the topic. Mutual information and Chi2 are popular measures for ranking keywords. Compression-based classifiers are usually character-based. Adaptive Prediction by partial matching (PPM) is a lossless compression technique that assigns different codelengths to different letters based on their frequency within a document. The optimum coding will vary for each language, sub-language and topic. A new document can thus be classified by selecting classifiers trained on a collection of related documents that can compress the new document most efficiently [7]. Of the keyword-based classifiers, the Vector Space Model is widely regarded as the most accurate, but is also the most computationally expensive. Biagioli et al. [3] achieved an accuracy of 92% in the task of classifying 582 paragraphs from Italian laws into ten different categories.

5. FUTURE WORK

Eunomos is an ongoing piece of work, and we are always interested in finding promising technologies that we can include in our research and products. Eunomos could be improved by applying text categorisation besides text similarity techniques. During the construction phase of the Eunomos database of norms, we did not have much labelled data, and the Cosine text similarity technique was useful for suggesting domains for unclassified norms as well as for finding norms that implicity override other legislation. In developing and testing the Eunomos system, we are building more and more labelled data, and we will soon be in a position to use this data to bootstrap a new topic-based classifier for paragraphs and articles. Text similarity for finding norms that implicitly override other legislation could also be more useful at paragraph and article level. But the task is more challenging with shorter text, and we need to compare the Cosine Similarity with other algorithms such as Latent Semantic Analysis. The WEKA toolkit [11] contains various machine learning algorithms for text categorisation and text similarity which we can use for our tests. These new requirements place a high performance demand in terms of precision, recall and speed, and careful analysis is required to select the most appropriate technique for each task. Given the size of the database we must firstly take into account efficiency considerations. To cope with the problem of the size of the dataset since laws are considered paragraph by paragraph, we propose to build topic-based classifiers on a small subset of representative norms that have already been classified. We may use results from text categorisation to aid text similarity and vice versa. For example, references to classified norms can be included as a factor in the text categorisation algorithm. Even in cases where a norm refers to a general law containing several topics rather than to an article assigned to a specific category, the reference can be included as a factor and the classifier should be able to take the information into account and assign it the appropriate weight. On the other hand, more efficient retrieval of similar norms could be achieved by limiting the text similarity searches to classified norms within the same domain as assigned by the topic-based classifier.

We will also be evaluating the accuracy of the automated translation of legal text into NIR XML, even if at first sight it seems sufficient for the requirements of the clients. Another development on the NIR XML side is to analyse explicit references. Currently Eunomos can find most explicit references but an expert user needs to specify whether the reference is a simple reference or whether it modifies or overrides other legislation. By incorporating the natural language technologies developed by Mazzei et al. [15], the type of modification can be discerned automatically.

Another area for future development is to exploit Eunomos's potential to cater for multilingual and multilevel legal research, since some clients may be interested in specialist

databases for foreign legal systems. While Eunomos uses the NormaInRete standard internally, as standards are developed for interchange between different legislative XML formats [5], it should be possible to use Eunomos in other jurisdictions. This would require suitable parsers to structure laws in XML in different languages. It is already possible, however, to model EU directives and their national implementations, and the Legal Taxonomy Syllabus ontology is already multilingual.

The Eunomos software could also be adapted to manage contracts and other legal documents. An integrated document management system that incorporates legislation, ontology and contracts could be very attractive to law firms. After a change in legislation, changes may also need to be made to contracts. In some cases even signed existing contracts need to be checked to ensure that new regulations do not invalidate them. The same mechanism based on reference recognition to find regulations affected by modifications can be applied to contracts.

It may be possible to integrate editors designed for drafting legislation to draft and edit legal contracts. The Norma-System legislative XML editor developed by the Università di Bologna [16] works as a plug-in for Microsoft Word. Additional menu items make it possible to: create an XML version of documents valid for NIR DTDs,mark up the structures of documents with automatics tools, view and compare structures with a text mapping, consolidate documents in fully automated mode or manually, manage integrative, or informative acts as attachments, automatically recognise and mark up normative references. The open-source Bungeni editor for drafting legal text[7] has been designed to work with the Akoma Ntoso standards. The Word-like editor can be integrated with Open Office. The workspace has an attractive and user-friendly interface to enable legislative drafting staff to import and mark up debates and legislation, review metadata and create links between referenced legislation.

6. CONCLUSIONS

Legal informatics is experiencing growth in activity. There is a place for experimentation and cross-fertilisation of ideas from other domains. There is good research within legal informatics, knowledge management, natural language processing and artificial intelligence which can help make the legal process more effective and efficient. Now is the right time to apply this research to products for law firms and not just legislative bodies.

In this position paper we illustrate ongoing work on the Eunomos software. The software is being developed to support the work of law firms, in-house legal offices and law scholars by offering them an environment which makes laws easier to navigate, annotate and understand, using automatically generated hyperlinks to referenced legislation, an extensible and updatable ontology which provides current and previous definitions for norms and concepts within any specific context, and an alert system that specifies existing legislation affected by new legislation.

[7]http://code.google.com/p/bungeni-editor/

Eunomos is being developed as a commercial software part of a wider suite distributed by Nomotika s.r.l., a spinoff of University of Torino. Eunomos has a clear business model: a combined software and services package that effectively means that legislation monitoring is outsourced. The roles, permissions and technologies have been carefully selected to address real business needs. The software and related services will be provided by experts with sound technological and business expertise.

7. REFERENCES

[1] G. Ajani, G. Boella, L. Lesmo, A. Mazzei, and P. Rossi. Multilingual Ontological Analysis of European Directives. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 21–24. ACL, 2007.

[2] G. Ajani, L. Lesmo, G. Boella, A. Mazzei, and P. Rossi. Terminological and ontological analysis of european directives: multilinguism in law. In *The 11th International Conference on Artificial Intelligence and Law, Proceedings of the Conference (ICAIL)*, pages 43–48. ACM, 2007.

[3] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria. Automatic semantics extraction in law documents. In *The Tenth International Conference on Artificial Intelligence and Law, Proceedings of the Conference*, pages 133–140. ACM, 2005.

[4] M. Bianchi, M. Draoli, G. Gambosi, M. Pazienza, N. Scarpato, and A. Stellato. ICT tools for the discovery of semantic relations in legal documents. In *Proceedings of the 2nd International Conference on ICT Solutions for Justice (ICT4Justice)*, 2009.

[5] A. Boer and R. Winkels. What's in an interchange standard for legislative XML? *I Quaderni*, 18:32–41, 2005.

[6] E. de Maat, K. Krabben, and R. Winkels. Machine learning versus knowledge based classification of legal texts. In *Proceeding of the 2010 conference on Legal Knowledge and Information Systems: JURIX 2010: The Twenty-Third Annual Conference*, pages 87–96, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press.

[7] D.V.Khmelev and W.J.Teahan. Verification of text collections for text categorization and natural language processing. Technical Report AIIA 03.1, University of Wales, Bangor, 2003.

[8] F. Giunchiglia and I. Zaihrayeu. Lightweight Ontologies. Technical Report DIT-07-071, University of Trento, Department of Information and Communication Technology, October 2007.

[9] M. Graziadei. Tuttifrutti. In P. Birks and A. Pretto, editors, *Themes in Comparative Law*. Oxford University Press, 2004.

[10] D. Griesi, M. T. Pazienza, and A. Stellato. Semantic turkey - a semantic bookmarking tool (system description). In *4th European Semantic Web Conference (ESWC 2007)*, volume 4519 of *Lecture Notes in Computer Science*, pages 779–788. Springer, 2007.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data

mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November 2009.

[12] M. Klein. Combining and relating ontologies: an analysis of problems and solutions. In *Workshop on Ontologies and Information Sharing at IJCAI'01*, 2001.

[13] V. Lyding, E. Chiocchetti, G. Sérasset, and F. Brunet-Manquat. The LexALP information system: Term bank and corpus for multilingual legal terminology consolidated. In *Proc. of the Wokshop on Multilingual Language Resources and Interoperability at ACL'06*, pages 25–31, 2006.

[14] C. D. Manning, P. Raghavan, and H. Schuetze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[15] A. Mazzei, D. P. Radicioni, and R. Brighi. NLP-based extraction of modificatory provisions semantics. In *The 12th International Conference on Artificial Intelligence and Law, Proceedings of the Conference (ICAIL)*, pages 50–57. ACM, 2009.

[16] M. Palmirani and R. Brighi. Norma-system: A legal document system for managing consolidated acts. In *Database and Expert Systems Applications, 13th International Conference, DEXA 2002, Proceedings*, volume 2453 of *Lecture Notes in Computer Science*, pages 310–320. Springer, 2002.

[17] P. Rossi and C. Vogel. Terms and concepts; towards a syllabus for european private law. *European Review of Private Law (ERPL)*, 12(2):293–300, 2004.

[18] R. Sacco. Contract. *European Review of Private Law*, 2:237–240, 1999.

[19] G. Sérasset. Interlingual lexical organization for multilingual lexical databases in NADIA. In *Proceedings of the 15th conference on Computational linguistics (COLING)*, pages 278–282, 1994.

From Spelling Checkers to Robot Judges?

Some Implications of Normativity in Language Technology and AI & Law

Anna Ronkainen
University of Helsinki, Department of Modern Languages
PO Box 24, 00014 Helsingin yliopisto, Finland
Anna.Ronkainen@helsinki.fi

ABSTRACT

In language technology, the process of turning language analysis tools into language checkers offers a viable model for the development of decision support or judicial automation tools based on current software for electronic discovery. This paper presents arguments for why the cases are more analoguous than one might think and based on that presents some lessons learned from normative language technology and the implications they might have for AI software projects in a judicial setting.

Keywords

e-discovery, judicial decision support, judicial automation, language technology, language checking, normativity, project management

1. INTRODUCTION

The idea of 'robot judges' is older than the field of AI & law itself. In more recent years the idea has been largely abandoned, and certainly these days nobody would suggest that 'The Supreme Court might be chosen for this purpose.'[1, p. 398]. The expectations of the early enthusiasm proved overambitious and as a result, today AI & law remains largely unutilized within the judiciary, with the exception of a fairly small number of experimental decision support systems, and, of course, mainstream legal information systems. Nonetheless, new solutions, such as those used for e-discovery, offer considerable potential for streamlining judicial decision-making, especially if, rather than a Supreme Court, a first-instance court dealing largely or exclusively with a document-based procedure and with a docket consisting mainly of routine cases, such as the Administrative Courts in Finland, would be chosen for this purpose. The need for greater efficiency is particularly crucial in jurisdictions currently struggling to meet the requirements for the duration of trials as required by Article 6 of the European Human Rights Convention. This is of course not to say that the complex cases leading to excessively long trials are the ones that should be solved automatically, but rather that by processing easy cases more efficiently, more resources can be made available for deciding the hard cases within an acceptable timeframe.

Language technology plays a central role in most if not all e-discovery solutions [2, 3]. More to the point for this paper, however, language technology also offers a well-established and commercially succesful precedent in technology transformation that can be seen as an analogy for the move from e-discovery to judicial automation which is already currently available on our desktops, namely spelling and grammar checkers. A number of such tools have been made based on preexisting software for morphological and syntactic analysis of a given language, though of course it is also possible to construct a language checker directly. The relationship between language analysis (descriptive) to language checking (prescriptive) is basically the same as finding the facts of a case (e-discovery, descriptive) and deciding the case based on the facts (judicial automation, prescriptive). The analogy may seem far-fetched and the distinction between descriptive and prescriptive applications trivial but in the following I will do my best to argue that this is indeed the case and future constructors of judicial automation should be able to avoid repeating a number of mistakes thanks to a better understanding of the task at hand based on previous experiences from language technology software engineering.

The distinction between descriptive and prescriptive is of course also one of the central issues in legal theory. For a Kelsenian positivist this distinction between *Is* and *Ought* is fundamental a priori, whereas for a (Scandinavian neo)realist such as myself, most of what is said about this distinction seems to be just highfalutin metaphysical nonsense. By examining the practical impact of the transition between *Is* and *Ought* in both existing software implementations of language technology as well as potential future implementations in AI & law we can see its real effects in a new perspective. At a first glance this connection may seem tenuous, but upon closer inspection natural language and law are quite similar in this respect after all. To begin with, of course law and language are closely related, with law as a system built on top of language: any particular legal system is practically inseparable from its language(s). Furthermore, both law and language can be seen as normative systems. As far as law is concerned, the role of normativity generally goes without saying, whereas for the case of language, the matter is still somewhat controversial and open for debate [4]. Perhaps the ongoing debate in linguistics could provide some fresh ideas for legal theory as well.

It should be borne in mind that the real long-term goal of judicial automation or decision support cannot really be the idea of replacing human judges with software, but rather a reconfiguration of the tasks within the judiciary in order to let the computers and the humans excel in what they each are good at, and not the other way around. Just like a spelling checker is in some respects no match

for a human proofreader, some types of errors are nonetheless captured more reliably by software. Similarly, a computer should at least be able to make decisions consistently both before and after lunch, which may not always be the case with human judges, as shown by a recent study [5].

2. OBSERVATIONS ON NORMATIVITY

In this paper, I present a number of central observations concerning the development of prescriptive language technology applications, in particular spelling and grammar checkers, combined with conjectures regarding the possible impact of comparable circumstances on judicial decision support or automation software. These observations are pronounced by a legal theorist but first and foremost they are based on an average Finnish life sentence's worth of industry experience of linguistic development and project management in language checker projects, many of which have also led to a number of academic publications by my former colleagues often dealing with many of the same topics [6, 7, 8, 9]. Unfortunately the issues raised in many of these papers are of limited academic interest and typically only become visible in commercial projects on a larger scale. In general, points such as those raised in this paper are also not typically discussed in most AI & law work, either.

For the purposes of this paper it is particularly important to note that all the language checker projects with which I have been involved have been originally based on software tools for descriptive analysis: spelling checkers based on morphological analyzers using two-level morphology [10] and grammar checkers based on syntactic parsers using the Constraint Grammar framework [11]. In most cases, the analysis components existed in their own right beforehand whereas some of them were expressly made as the first stage of a language checker project. In either situation, the vast majority of the work put into the analysis component, whether by itself or as a part of the language checker, was in the end done based on the demands of language checking.

1. Know what you are building

The importance of advance planning in software projects can hardly be overestimated. In a typical language checker project, up to one fourth of the entire project should be spent at the specification and planning stage, before the actual implementation work even starts (and, since the specification is typically a part of the contract for the project, before the final contract is even signed). As a rule, approximately another fourth consists of implementing an alpha version according to the specification, and the other half of the project consists of turning that alpha version into something acceptable as a final release.

The role of the specification is particularly important when dealing with software in a complex and open-textured domain such as language (or law). The limitations of what is technically feasible with current technology and under whatever other constraints in terms of capacity, speed and so on may exist. For instance, a swarm of parking attendant drones would operate under environmental constraints vastly different from those for an administrative court decision support machine stuck somewhere in a basement. For the sake of all the parties involved in the project, it is crucial that all the parties agree as to what the software is supposed to do and how, and just as importantly, what it is not and cannot be supposed or expected to do. Merely trying to define 'the law in effect' as it is supposed to be implemented as a contractual provision limiting the scope of the project will undoubtedly turn out to be quite a challenge, and possibly even interesting from a theoretical point of view.

In considering future judicial applications based on e-discovery, the planning process must be carried out in several stages. In the initial stages, the scope of the project will be made more specific gradually based on factors such as the number of cases within potential target domains and complexity of the target domain in terms of both implementation cost and potential savings within the judiciary, all while considering what is technically feasible in the near future. Once a target domain has been established and financing for the rest of the project secured, the planning process for the actual software engineering implementation process can start. First at this point it can properly be established to a reasonable certainty, what the system is supposed to do and when it can be expected to be deployed.

2. Reality is messy, especially when rules are broken

Norms play a crucial role in creating structures that help us to make sense of reality. Linguistic (and legal) structures are complex enough when all the rules are being followed, but when this cannot be relied on, the situation only becomes even more complex. For instance, in a grammar checker it is impossible to rely on the correctness of the punctuation for correct syntactic analysis if the checker is expected to check said correctness and punctuation, it must therefore by definition be assumed to be at least occasionally incorrect. Or if the input is likely to contain spelling errors, one cannot rely entirely on lexicon-based morphological analyses for the individual words but rather must also use some sort of heuristics to make as educated guesses for the misspellings as possible. (Ideally spelling check is always run before the grammar check, but in practice this cannot be relied upon.) As a consequence, the parser used in a grammar checker is very different from one built to only understand completely correct sentences. The best strategy is to start by handling individual words together with their most proximate contexts and then try to parse the sentence as a whole if possible, rather than starting with the presumption of a well-formed sentence.

Translating this into the judicial context, this means that the system in its analysis of the materials of the case should not make too many assumptions about the facts of the case, especially not in terms of normative structures which may or may be present in the individual case in the expected form. It will probably be better to take a bottom-up approach to the analysis task by first identifying the details of the case as reliably as possible before moving on to the structure of the case as a whole. Of course even the analysis of details will to some degree be done in terms of the totality of the case, but it is better to keep one's options as open as possible during the early stages of the process.

Another consequence of this issue is that language technology approaches succesful in extracting information out of legislative texts should not automatically be expected to work in the judicial context. In terms of both correctness and consistency, legislation is in a class of its own, yet even there inconsistency and occasional errors (most clearly visible in legislation issued in several languages) do occur. Natural-language input to a judicial system, on the other hand, will not be subject to checking and standardization at the same level.

This may seem to be a fairly trivial concern, but in terms of system architecture it is one of the major design decisions. More impor-

tantly, it is a decision whose consequences will not become apparent until later on in the project, either through the cumbersome workarounds it makes necessary or, in the worst case, through a complete mid-project redesign of the system. In some recent work using context-free grammars (eg. [12, 13]), the question of the order in which rules are applied is not explicitly discussed (or a top-down order is implied). In toy grammars the order may not matter, but in real-world applications even small changes can make a big difference.

3. Even though something is right in theory, it may be wrong in practice

A morphological analyzer may take a liberal approach to compositional phenomena such as compounding and derivation: anything that can be considered a well-formed word according to the rules of the language can and possibly even should be given an analysis, no matter how nonsensical it is semantically. In a spelling checker for almost any Germanic (or Finno-Ugric) language but English, however, this approach does not work in practice. For instance, the Swedish word *komission* is a theoretically correct compound noun formed by the nouns *ko* and *mission*, thus meaning 'cow mission' or 'mission of cows'. When this word appears in a text, however, it is almost certainly (except for scientific texts dealing with this very topic, naturally) a misspelling for *kommission* 'commission'. Worst of all, it is a very common misspelling and something a spelling checker most reasonably can be expected to detect. Therefore, productive mechanisms must take this into account and strike a balance between not accepting misspellings such as *komission* but also not flagging too many probably correct ad hoc compounds of the same type which are not to be found in dictionaries.

Something similar must also be taken into account in the grammar checker, especially when a common misspelling coincides with an actual word which must remain in the lexicon. This is particularly problematic when the misspelling has a different part of speech, which in turn may throw the syntactical analysis of the sentence all upside down. And, as shown in the previous item, since correct syntactic analysis is not something that can be taken for granted when dealing with language in need of checking, the grammar checking rules had better be prepared for incorrectly disambiguated analyses of important words in their contextual conditions.

For judicial decision support, the message is clear. Aiming at total coverage of a given domain may result in the inclusion of esoteric combinations of states of affairs (such as case factors which are mutually exclusive in practice but not by definition) that, while somehow theoretically possible, are likely to be something different altogether yet still quite proximate, and the availability of such a theoretical situation may prevent a correct analysis of the more probable scenario. It is better to aim at a comprehensive coverage of the most typical cases that cover the bulk of the domain, and make sure that more marginal types of cases are identified and dealt with separately.

4. Know the limitations of the system, and make sure all the users know them too

A typical spilling chucker operates without context, that is, it receives only each word one by one as its input. Because of this, a verdict of no errors from a spelling checker does not mean that the text is free of spelling errors, but rather that it only contains words which are acceptable at least in some context. Similarly a grammar checker cannot guarantee that a checked text will be free of grammatical errors, the most it can do is to check the correctness of the text in terms of defined categories of errors and try to make sure that those particular error types are detected as well as practicable.

The corollary to this is that the limitations of the system should not become visible for the user, either. False flags, that is, correctly spelled words or correct grammatical constructions marked as errors, considerably undermine the user's ability to trust the system. The same can be said for actual errors in the text which are detected by a rule designed to catch a completely different type of errors, thus producing an incomprehensible error message or nonsensical suggested corrections, both of which will quite simply appear to the user as though the system is malfunctioning. On the other hand, many users seem to take even the nonsensical suggestions at face value, which in turn lead to expressions that must essentially be reverse-engineered in this respect to become understandable again.

In a judicial setting this means that considerable care must be taken to ensure that everyone is aware of the limitations of a decision support system, including both its users as well as the system itself. That is, to begin with, users should know better than to feed the system cases from outside its domain, but as an additional safeguard, the system itself must also be able to recognize when it is given such a case. And should such a system be put into production use to produce suggested verdicts for actual new cases, the verdicts must of course be examined very carefully in general, but with extraordinary care in this respect. It is quite easy to produce a convincing argument for an absurd outcome when something central is being ignored because it falls outside the domain of the system's expertise, and if it at the same time is also omitted from the proposed verdict from the some reason, it may be quite difficult to detect at that point without reading all the documents of the case.

5. The rules are alive

The law changes constantly. Language changes over time as well, but is the development of a language anything at all like the constant shifts of the law? For a native speaker of English, the idea that the orthography of language might be subject to legislative (or administrative) fiat and therefore change, at times even quite radically, at the stroke of a pen may seem strange. Such a situation is nonetheless current reality for some of the closest relatives of the English language, as the following examples clearly illustrate.

The orthography of Danish is regulated by the Danish Language Council *(Dansk Sprognævn)*. The Danish spelling is reviewed approximately once per decade through the publication of the Spelling Dictionary *(Retskrivningsordbogen)*, with the latest edition from 2001 [14] and the next edition due within the next year or so. The dictionary consists of the actual lexicon followed by a rather statute-like (but of course much more readable) description of the general rules of the orthography regarding issues undescribable in lexicon form, such as punctuation, morphology and the productive mechanisms of compounding and derivation. From a legal point of view, the Spelling Dictionary is an administrative decree issued by the Minister of Culture of Denmark. The ministerial approval is not a total formality, and as a curiosity it can be pointed out that for the 1986 edition, the minister himself ordered some changes to a paradigm, and the addition of very common forms such as *akvarie* to stand beside forms such as *akvarium* was thus cancelled. Or possibly merely postponed, since the same addition is once again expected to be proposed for the 2011/12 edition.

The orthography of German was officially regulated first towards

the end of the 20th century through the adoption of the 1996 German spelling reform by the Ministers of Culture of the German-speaking countries and states. Some aspects of the reform were quite controversial and the question of reform even faced the German Federal Supreme Court as well as a referendum in the state of Schleswig-Holstein [15]. A transitional period ran from 1998 to 2005, during which both the new and the old forms were acceptable at schools. Of course, as a practical consequence of this, language checkers had to be maintained at least during this period in both 'pre-reform' and 'post-reform' versions, in addition to the preexisting national variants to cope with some peculiarities of Swiss and Austrian German orthography. Widespread criticism led the Ministers of Culture to reconsider the reform, first by establishing the Council for German Orthography *(Rat für deutsche Rechtschreibung)* as a permanent organ for maintaining the German orthography. The Council issued an updated proposal for the reform in 2004, which was yet once more revised before it entered into force in 2006 [16]. Structurally the reform document is quite similar to the Danish one, though the lexicon part is more limited, and thus commercially published spelling dictionaries are normally used in practice. Also its legal status is comparable, though the details vary from country to country and state to state.

The orthography of Norwegian is unique in the world in the complexity of its situation [17]. There are two official orthographies for Norwegian: Bokmål and Nynorsk. Bokmål has evolved as a continuation of the Danish orthography which has been gradually norvagized over the past two centuries, whereas Nynorsk (originally Landsmaal) was developed by Ivar Aasen in the 1860s based on Norwegian dialects. To complicate matters even further, in addition to the two official orthographies, there is also the privately maintained orthography Riksmål, a more conservative version of Bokmål (closer to Danish) used eg. by many academic lawyers. And an extremely conservative orthography is still being used in amendments to the Norwegian constitution to maintain consistency with the original sections from 1814. The Norwegian orthography is now maintained by the Language Council *(Språkrådet)* and it is revised continuously, until 2009 through Annual Reports *(Årsmeldingar)*, and from 2010 through the annual Language Status *(Språkstatus)* report, in which changes to the orthography are announced incrementally word by word or rule by rule. Spelling dictionaries are published only commercially based on these decisions, separately for each orthography. The extent of the changes varies considerably from year to year. Also the underlying general tendencies have varied both regarding the distance between the two orthographies and the variations permitted within them. The last major revision took place in 2005, though some major simplifications were only implemented for Bokmål at that time, and a corresponding proposal should be presented for Nynorsk later this year. Changes to the orthography are authorized by the Minister of Culture, who also has to submit a report on the state of the language for the approval of the Parliament once every fourth year. As an aside note, any commercial e-discovery provider contemplating entry into the Norwegian market should take the orthographical situation into consideration as the different orthographies are used interchangeably (though subject to strict legal regulation in some contexts), and many texts, though mainly written in Bokmål, may contain passages in Nynorsk, such as quotes from statutes enacted only in Nynorsk.

These examples show already a considerable diversity in the way different language authorities handle change. What they all however show is that changes in the norms are something that should be taken into account in the planning process from the start. This is equally important from an architectural point of view (distribution of updates, allowing different versions to run concurrently when dealing with different periods) as well as from a project management point of view (the changes do not implement themselves, staffing has to be planned for the whole product lifecycle). This may be ignored only in exceptional cases, such as systems with a very limited scope of application. In a legal setting this kind of long-term perspective should be easier to justify, whereas in the linguistic setting the timeliness of the updates may be less crucial and subordinate to a more general product release cycle. Unpredictability of the future changes makes staffing difficult to plan and/or costly, especially if changes must be implemented quickly.

In addition to changes required by external factors, post-deployment changes to real-world software are also required due to bugs. The performance of the system must be monitored constantly and questionable recommendations or decisions reviewed more closely. Strictly from the software project management point of view, legislative changes can be viewed simply as a bug with a particular kind of source: the decision produced by the system no longer matches what is expected by the law in effect. Modifying a rule-based part of the system to deal with this is comparatively trivial: the rules must be modified correspondingly, and possibly some rules must be added or deleted. Modifying a part of the system based entirely on machine learning may be more complicated, it may for instance be necessary to review all the cases affected by the change, change the outcomes of the cases in the training set as though how they would most likely be decided under the new legislation, and then redo the learning process with the updated cases.

3. DISCUSSION

The proposed analogy between language checkers and automated judicial decision-making is on a very abstract level of function and as such it cannot be used to support the transformation of specific individual techniques from the linguistic to the legal context, nor is this the purpose of the present paper. The possibility of borrowing specific new solutions from language technology to AI & law must still be examined the old-fashioned way, through careful experimentation on a one-by-one basis. No such experiments were made in preparation for this position paper. The choice of the best tool for the job at hand depends of course on the nature of the task to a great deal. Furthermore, it also depends on the specific language of the application, both in terms of what alternatives are readily available as well as the specific demands of that particular language.

In the light of these observations taken as a whole, however, I propose that the best way of turning e-discovery-like solutions into judicial decision support or even supervised automation requires that we adopt and implement a dual-process model of legal reasoning [18]. That is, I suggest that we start with the hypothesis that human cognition consists of two very different yet interdependent systems: the evolutionarily much older 'heuristic' *System 1*, which is fast, effortless, automatic, non-conscious, etc., and the more recent and 'logical' *System 2*, which is slow, effortful, controlled, conscious, etc.[19] Applied to legal reasoning, the dual-process theory suggests that System 1, as formed by a lifetime of socialization into a particular legal culture, together with years of specialization over the course of one's legal training and subsequent professional experience, is in charge of coming up with the right outcome for the case, and System 2, which is able to look up the relevant statutes and precedents to construct a syllogistic argument seemingly inevitably leading to said outcome, is in charge of com-

ing up with the justifications for the decision, thereby opening up the reasoning process (or rather parts thereof) for external scrutiny. Of course, in some cases the process must be carried out in several iterations, in the situation when the outcome initially produced by System 1 proves out to be unjustifiable (incorrect). After the failed attempt at justification, System 1 may try again, but this time within a restricted search space no longer containing the initial decision, which of course is trivial if there are only two possible outcomes of which one has already been eliminated.

Somewhat paradoxically, more iterations are more likely to be required in easy rather than hard cases, since in easy cases it is much more likely that a proposed outcome can be shown to be demonstratably wrong, whether through explicit statutory language or a veritable forest of nearly identical precedents with a different result, whereas in hard cases it is much more difficult to demonstrate conclusively that some particular outcome is undoubtedly incorrect. This difficulty is indeed the very nature of a hard case, in that a judge is required to produce an authoritative decision based on conclusive arguments, yet he or she may only have some general principles and distant precedents on which to base one's arguments, together with one's imagination. In legal theory, the justification process has been made practically synonymous with legal reasoning, with the consequence that the process of finding the outcome of the case, which is after all what everyone is really interested in, has been largely neglected, with some recent exceptions [20, 21, 22]. As a further consequence, current legal theory has very little to say that could be considered useful in deciding easy cases.

From a judicial automation point of view, on the other hand, for the time being, hard cases are uninteresting and can be all but ignored. Their scarcity together with their one-of-a-kind nature mean that they cannot be expected to be decided reliably with computational methods, but it also means that in economical terms there is very little to be accomplished by automation, as the cost of system implementation per case decided is very high and likely to exceed the costs for a conventional procedure. Easy cases are where the volume is, together with the greatest potential for savings through more efficient methods. What a judicial automation system must however know about hard cases is how to identify them reliably or even overcautiously and pass them on to a conventional bench.

The dual-process model also has significant implications for the software architecture. When System 1 heuresis and System 2 justification are seen as distinct processes, the System 1 part responsible for aligning the particulars of an individual case with the legal system can be implemented using any method best suited for the job. At this point we do not have to care about the details on how it is done, as long as the results it produces are reasonably correct and usable. For instance, in one pioneering experiment, a collection of over 6 million patent texts received a meaningful arrangement through self-organization without the involvement of any particular type of legal knowledge in the process [23].

Once the case has found its place in the system, the System 2 part using GOFAI(&law) techniques can take over [24]. Once the case has found its tentative solution and thereby its factors of legal importance have been identified, those factors can be inserted into a model of statutory and case law of the domain in order to validate the result. If the result stays the same, it is quite likely to be correct, and the statutory rules and/or precedents used in order to compute the result the second time provide the justification. If the result however changes, that is, the rules or the nearest precedents produce a different outcome, the system may quite reliably identify that it is dealing with a hard case (or possibly a case from outside its domain). Alternatively, the same procedure can be carried out for a number of the best outcomes to see whether any single one of them stands out from the crowd or whether they all produce the same material result but on different grounds. In this variant, if there are considerable differences between the alternatives, we are probably again dealing with a hard case.

The dual-process architecture is likely to be able to handle changes in rules more easily. If statutory rules are encoded more or less explicitly in the System 2 part of the decision support system, the places in need of an update can also be identified and subsequently modified without difficulty. And verdict generation can be done based on the System 2 analysis with a limited number of fill-in-the-blanks templates, rather than trying to assemble something out of the texts of previous cases.

4. CONCLUSIONS

The example of turning morphological and syntactic analysis software into spelling and grammar checkers presents a viable roadmap for the transition from e-discovery to highly automated judicial decision support. In simple cases, where the use of e-discovery does not make economical sense at present, similar technologies can be integrated into the judicial decision support system instead. In a more long-term perspective, in preparation for a time when decision support systems reach the level of sophistication necessary for dealing with cases of the kind of complexity typical for the situations in which e-discovery is already being used, the idea of a standardized interchange format for communication between the parties to the case (or rather their e-discovery systems) and the judge (or rather the decision support system) should be considered seriously. Translating the results from an internal format to natural language and back again is bound to result in the loss of information or even veritable errors. Of course the standardized format must also be readable in or translatable into natural language.

The mere thought of 'robot judges' is bound to raise many red flags, and as far as the terminology goes, the use of the expression should for strategic reasons certainly be avoided. From a technical standpoint there does not have to be any substantial difference between an advanced decision support system and a completely automatized bench, but from a legal standpoint the difference is huge. By calling the same system a decision support system and requiring that an actual judge always review the verdict of each case it is possible to carry out (initially of course very limited) experiments with judicial automation, whereas leaving the system to do its work by itself would require legislative changes amounting to a considerable upheaval of some fundamental principles of law regarding personhood, agency etc. Firm political support for the project must be established in any case, already in order to arrange financing for it, but starting small and keeping one's goals realistic is the only way forward. Any system put into production use must perform at least as well as a human judge in terms of the percentage of decisions overturned on appeal (by design, the current court system does not expect perfection from the first instance), which of course varies considerably. A system with poor performance will only ruin the reputation of the idea in general for many years to come.

In addition to AI & law, also legal theory must for its own part take up the challenge presented by this endeavour. Apart from being a challenge, it is also a great opportunity to examine the nature of the legal reasoning process. By concentrating only on the cases

that present a real challenge for the human reasoner, legal theory has managed to ignore the vast majority of easy cases almost completely. Whether the present state is a hollow but structurally robust sphere or a mere façade is up for each individual to decide for themselves, but certainly the idea of looking into legal reasoning beyond the hard-case surface must be appealing for all.

Acknowledgments

This research has been funded by grants from the Finnish Cultural Foundation and the Hilkka and Otto Brusiin Foundation. Powered by *Spontaani Vire*.

5. REFERENCES

[1] Harold D. Lasswell. Current Studies of the Decision Process: Automation versus Creativity. *The Western Political Quarterly*, 8:381–399, 1955.

[2] Douglas W. Oard et al. Evaluation of Information Retrieval for E-Discovery. *Artificial Intelligence and Law*, 18:347–386, 2010.

[3] Christopher Hogan et al. Automation of Legal Sensemaking in E-Discovery. *Artificial Intelligence and Law*, 18:431–457, 2010.

[4] Esa Itkonen. The Central Role of Normativity in Language and Linguistics. In Jordan Zlatev, editor, *The Shared Mind: Perspectives on Intersubjectivity*, pages 279–305. John Benjamins, Amsterdam, 2008.

[5] Shai Danziger et al. Extraneous Factors in Judicial Decisions. *Proceedings of the National Academy of Sciences of the United States of America*, 108 (17):6889–6892, 2011.

[6] Antti Arppe. Developing a Grammar Checker for Swedish. In T. Nordgård, editor, *Proceedings from the 12th Nordiske datalingvistikkdager*, Trondheim, 2000.

[7] Antti Arppe. The Very Long Way from Basic Linguistic Research to Commercially Succesful Language Business: the Case of Two-Level Morphology. In Antti Arppe et al., editors, *Inquiries into Words, Constraints and Contexts*, pages 2–17. Helsinki, 2005.

[8] Jussi Birn. Detecting Grammar Errors with Lingsoft's Swedish Grammar Checker. In T. Nordgård, editor, *Proceedings from the 12th Nordiske datalingvistikkdager*, Trondheim, 2000.

[9] Kimmo Koskenniemi and Mariikka Haapalainen. GERTWOL: Ein System zur automatischen Wortformerkennung deutscher Wörter. In R. Hausser, editor, *Linguistische Verifikation. Dokumentation sur Ersten Morpholympics 1994*, pages 121–140. Max Niemayer, Tübingen, 1994.

[10] Kimmo Koskenniemi. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. University of Helsinki, Department of General Linguistics, Helsinki, 1983.

[11] Fred Karlsson et al., editors. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin, 1995.

[12] Raquel Mochales and Marie-Francine Moens. Argumentation Mining. *Artificial Intelligence and Law*, 19:1–22, 2011.

[13] Emile de Maat, Radboud Winkels, and Tom van Engers. Making Sense of Legal Texts. In Günther Grewendorf and Monika Rathert, editors, *Formal Linguistics and Law*, pages 225–255. Mouton de Gruyter, Berlin, 2009.

[14] Dansk Sprognævn. *Retskrivningsordbogen*. Alinea, København, 3. udgave edition, 2001.

[15] Sally Johnson. *Spelling Trouble? Language, Ideology and the Reform of German Orthography*. Multilingual Matters, Clevedon, 2005.

[16] Rat für deutsche Rechtschreibung. *Deutsche Rechtschreibung: Regeln und Wörterverzeichnis*. Tübingen, 2006.

[17] Gregg Bucken-Knapp. *Elites, Language, and the Politics of Identity: The Norwegian Case in a Comparative Perspective*. State University of New York Press, Albany, 2003.

[18] Anna Ronkainen. Dual-Process Cognition and Legal Reasoning. Paper to be presented at the *Between Interpretation and Intuition* workshop at IVR-2011, forthcoming.

[19] Jonathan St.B. T. Evans and Keith Frankish, editors. *In Two Minds: Dual Processes and Beyond*. Oxford University Press, Oxford, 2009.

[20] Richard A. Posner. *How Judges Think*. Harvard University Press, Cambridge, MA, 2008.

[21] Frederick Schauer. *Thinking like a Lawyer: A New Introduction to Legal Reasoning*. Harvard University Press, Cambridge, MA, 2009.

[22] Lawrence M. Solan. *The Language of Statutes: Laws and Their Interpretation*. The University of Chicago Press, Chicago, 2010.

[23] Teuvo Kohonen et al. Self Organization on a Massive Document Collection. *IEEE Transactions on Neural Networks*, 11:574–585, 2000.

[24] David D. Lewis. Afterword: Data, Knowledge, and E-Discovery. *Artificial Intelligence and Law*, 18:481–486, 2010.