Integrating a Bottom–Up and Top–Down Methodology for Building Semantic Resources for the Multilingual Legal Domain

Enrico Francesconi[1], Simonetta Montemagni[2], Wim Peters[3], and Daniela Tiscornia[1]

[1] Institute of Legal Information Theory and Techniques, CNR, Italy
[2] Istituto di Linguistica Computazionale, CNR, Italy
[3] Natural Language Processing Research Group, University of Sheffield, UK

Abstract. This article presents a methodology for multilingual legal knowledge acquisition and modelling. It encompasses two comlementary strategies. On the one hand, there is the top–down definition of the conceptual structure of the legal domain under consideration on the basis of expert jugdment. This structure is language–independent, modeled as an ontology, and can be aligned with other ontologies that capture similar or complementary knowledge, in order to provide a wider conceptual embedding. Another top–down approach is the exploitation of the explicit structure of legal texts, which enables the targeted identification of text spans that play an ontological role and their subsequent inclusion in the knowledge model.

On the other hand, the linguistically motivated, text-based bottom–up population and incremental refinement of this conceptual structure using (semi-)automatic NLP techniques, maximizes the completeness and domain-specificity of the resulting knowledge.

The proposed methodology is concerned with the relation between these two differently derived types of knowledge, and defines a framework for interfacing lexical and ontological knowledge, the result of which offers various perspectives on multilingual legal knowledge.

Two case-studies combining bottom-up and top-down methodologies for knowledge modelling and learning are presented as illustrations of the methodology.

Keywords: Knowledge Modelling, Knowledge Acquisition, Natural Language Processing, Ontology Learning.

1 Introduction

Since the legal domain is strictly dependent on its own textual nature, a methodology for knowledge extraction should take into account a combination of theoretical modelling and text analysis. Such a methodology expresses, in a coherent way, the links between the conceptual characterization, the lexical manifestations of its components and the universes of discourse that are their proper referents.

E. Francesconi et al. (Eds.): Semantic Progressing of Legal Texts, LNAI 6036, pp. 95–121, 2010.

The aim of this article is, therefore, to set out, through a description of some of the projects that have been implemented, the methodological routes for constructing legal ontologies in applications that, due to the tasks they intend to achieve, should maintain a clear reference to texts. The article is structured in the following way: in Section 2 we analyse the interconnections between language and law and the semantic relations among levels of the legal discourse; in Section 3 we outline the methodological issues inspiring the implementation of the DALOS knowledge modelling and the approach of ontology learning from legal texts; in Section 4 a complementary method for ontology learning is presented dealing with a legal rule learning approach; finally, in Section 5 we comment on the lessons we have learnt.

2 Language and Law

There is a strict connection between law and language, characterised by the coexistence of two autonomous but structurally similar systems: both are endowed with rules that underlie the construction of the system itself, that guide its evolution and guarantee its consistency. Both are conditioned by the social dimension in which are placed, whereby they dynamically define and fix their object in relation to a continually evolving social context.

Law is strictly dependent on its linguistic expression: it has to be communicated, and social and legal rules are mainly transmitted through their written (and oral) expression. Even in customary law there is almost always a phase of verbalisation that enables it to be identified or recognised; even if the law cannot be reduced to language that expresses it, nonetheless, it cannot escape its textual nature.

Another characteristic of law is that it is expresses through many levels of discourse:

- the legislative language is the "object" language because it is the principal source of positive law that, in its broad sense, also includes contracts and so-called soft law; the constitutive force of written sources originates from the stipulative nature of legislative definitions, that assign a conventional meaning of legal concepts in relation to the domain covered by the law that contains them.
- Judges interpret legal language in an 'operative' sense to apply norms to concrete cases: the main function of judicial discourse lies, therefore, in populating the extensional dimension of the object language, instantiating cases throughout judicial subsumption. This involves the linking of general and abstract legislative statements to their linguistic manifestation, or, in others words, the mapping of legal case elements to the kinds of descriptions that may classify them.
- the language of dogmatics is a reformulation of legislative and jurisprudential language aimed at the conceptualisation of the normative contents. Although it is a metalanguage with respect to legislative and judicial language, it is

still based on the analysis of the universe of discourse and it is dependent on specific normative systems.
- At a more abstract level, legal theory expresses the basic concepts, the systemic categories common to (almost) all legal systems (for example, duty, permission, right, liability, sanction, legal act, cause, entitlement etc.). Legal theory may, therefore, be constructed as a formal and axiomatic system, made up of concepts and assertions in the theory, whose truth is based not on a semantic model of reality, but on syntactic rules, derived from inferential, deductive resoning whose scope is explaining positive legal systems [19].
- At the highst level of abstraction, the role of philosophy of law is to express both general principles and value judgements, as well as their ordering criteria.

At the (meta)theoretical level, the border between legal theory and dogmatics may be seen as a genus/species relationship, or as a semantic relation between a logical theory and its models; legal theory has an explanatory and prescriptive function (in the broad sense) because it constructs concepts independently of the normative enunciations and interpretative operations, while the conceptual models of dogmatics arise from the analysis of legal texts, which produces interpreted knowledge.

One of the most obvious areas that demonstrate this distinction is the creativity of legal translation, perched halfway between term equivalence and concept comparison. Legal terminology used in the various legal systems, both European and non-European, expresses not only the legal concepts which operate there, but further reflects the deep differences that exist between the various systems and the different legal perspectives of the lawyers in each system. Given the structural domain specificity of legal language, we cannot speak about "translating the law" to ascertain correspondences between legal terminology in various languages, since the translational correspondence of two terms satisfies neither the semantic correspondence of the concepts they denote nor the requirements of the different legal systems.

Transferred into the computational context, the boundary between the conceptualisations of legal theory and legal concepts built by dogmatics becomes purely methodological. The former entities, the kernel legal concepts, are modelled in the so-called core ontologies, while the latter provide content to domain ontologies, conceived as a possible, non-exclusive interpretation of linguistic objects.

These peculiarities should be taken into account while designing a methodology for ontology construction in the legal domain. They are also relevant for the combination of linguistic and ontological knowledge, in order to best reformulate the process of pure legal scholar conceptualization in a computational context, strongly based on legal language analysis. A general methodology for meaning extraction must be set up within a modular architecture, where different aspects refer to specific analytic models and appropriate Natural Language Processing (NLP) tools.

3 Legal Ontology Construction

Legal ontologies are increasingly becoming a popular field of research, as testified by the list of existing ontologies built for the legal domain which is growing rapidly over the years (for an extensive survey of existing legal ontologies, see [52], [13]). They differ in their purpose or subject-matter, they exhibit varying degrees of generality, formality or richness of internal structure; other relevant differences reflect the methodologies followed for their development, as well as the tools and knowledge representation language used.

Among these different parameters for classifying ontologies, a particularly interesting but often neglected one, deserving, in our opinion, specific attention, concerns the construction process: how was the ontology built? Unfortunately, as pointed out by Paslaru and Tempich [39] in a survey regarding several aspects of ontology development (i.e., methodology and tools used), it appears that "only a small percentage of ontology–related projects follow a systematic approach to ontology building, and even less commit to a specific methodology. Most of the projects are executed in an ad–hoc manner". Most developers do not offer an account of the followed methodological steps, and even when this is the case, it turns out that an ad hoc rather than an established methodology has been followed. This cannot be considered a marginal issue, because it has consequences at different levels on the final result of the construction process.

In this section, we would like to address the methodological issue of how a legal ontology ought to be built. In particular, in Section 3.1 we discuss general methodological issues associated with the construction of an ontology and introduce our approach to legal ontology building. The proposed approach will be illustrated in detail through its implementation within the European DALOS project (in Section 3.2).

3.1 Approaches to Ontology Design and Development

In principle, two different approaches can be recognized as far as the construction of ontologies is concerned: top–down and bottom–up.

In a top–down approach, ontology construction starts by modelling top level concepts, which are then subsequently refined. This approach is typically carried out manually by domain experts and leads to a high–quality engineered ontology. On the other hand, a bottom–up approach to ontology construction starts from the assumption that most concepts and conceptual structures of the domain, as well as the terminology used to express them, are contained in documents. In this approach, the terminological and conceptual knowledge contained in document collections is semi–automatically extracted from texts, thus creating the basis for ontology construction.

There are pros and cons connected with both approaches. Among the advantages usually associated with the top–down construction approach there is the fact that top–down ontologies may be reused across different application scenarios, and can serve as a starting point for developing new ontologies. Among the drawbacks typically associated with top–down ontologies it is worth to mention

here that they necessarily require an expert–based approach. Their development is costly in terms of both time and effort. Due to this fact, their coverage is typically rather restricted, and this is a disadvantage when they are used in the framework of real knowledge management applications. Other central problems connected with a top–down approach are the linking of textual information to the ontology, which requires linguistic knowledge about the terminology used to convey domain–specific concepts. Furthermore, there is the highly dynamic and constantly evolving nature of ontologies in different domains, including the legal one, which continuously need to be updated and refined.

When compared with top–down ontology construction, bottom–up approaches have the main advantage of making it possible to discover ontological knowledge at a larger scale and a faster pace; they can also be of some help for detecting and revising human–introduced biases and inconsistencies. Moreover, bottom–up approaches can support the refining and expanding of existing ontologies by incorporating new knowledge emerging from texts. Another crucial aspect is concerned with the fact that they create the prerequisites for the alignment between the ontology and texts: with ontologies boostrapped from texts the linking with textual information is made easier. Among the cons usually ascribed to this class of approaches, there is the fact that a bottom–up approach results in a very high level of detail which makes it difficult to spot commonality between related concepts and increases the risk of inconsistencies [51].

This short characterization of the top–down and bottom–up approaches to ontology construction shows their comlementarity. Preferring one approach over the other means ignoring complementary information that can help creating a better product. This fact is more and more acknowledged in the literature, where it is claimed that any comprehensive domain ontology needs work from top–down and bottom–up. Only by proceeding in this way, the resulting ontology reflects domain knowledge and is at the same time anchored to texts. From a general perspective, this is explicitly claimed by Uschold and Grüninger [51], who include among their guidelines for ontology construction and merging the so–called "middle–out approach", based on the combination of top–down and bottom–up ontology modelling. More recently, scholars advocating a middle–out approach to ontology construction started explicitly mentioning the "support of automatic document analysis" through which relevant lexical entries are extracted semi–automatically from available documents (see, for instance, [49]). The (semi–)automatic support in ontology development is nowadays referred to as "ontology learning". Ontology learning represents a promising line of research which is concerned with knowledge acquisition from texts as a basis for the construction and/or extension of ontologies, and in which the learning process is typically carried out by combining NLP technologies with machine learning techniques. Ontology learning is attracting increasing attention as a way to support the task of developing and maintaining ontologies [11] [12].

In the legal domain, the number of ontologies being constructed is rapidly increasing. Most of them still focus on an upper level of concepts and were mostly hand–crafted in a top–down manner by domain experts on the basis of

insights from legal theory. More recently, there have been few ontology learning experiments focused on concept extraction as a primary step of the ontology development process. Among them it is worth mentioning here: the work on definitions in a large collection of German court decisions by [54] [55]; the extraction of domain relevant terminology from normative texts on the basis of which domain relevant concepts are derived together with relations linking them (see, [32], [33], [34]). To our knowledge, relatively few attempts have been made so far to build legal ontologies following a middle–out approach: this is the case, for instance, for the LKIF Core ontology [29], the lexical ontology LOIS [50] [40], the Ontology of Professional Judicial Knowledge [13], and the DALOS ontology [22], where only the latter two appear to resort to ontology learning techniques as far as the bottom–up acquisition process is concerned [41]. Last but not least, a kind of middle–out approach to legal ontology construction is proposed by Saias and Quaresma [46], who exploit NLP tools in order to identify and extract legal concepts and properties: the new domain ontology bootstrapped from texts is then integrated and merged with an externally defined upper foundational legal ontology, with the result of creating a new domain ontology combining low–level concepts with top–level ones.

On the basis of what has been said so far, we believe that the most promising way to build legal ontologies is through the integration of top–down and bottom–up approaches. Such an integrated approach leads to accurate ontology construction, which cannot be achieved by either bottom–up or top–down approach alone. This is particularly true in the legal domain, where ontology construction should follow insights provided by legal theory but at the same time should guarantee textual grounding. Although it is a widely acknowledged fact that ontology building is primarily concerned with the definition of concepts and relations holding between them, it should also include the extraction of linguistic knowledge about the terms used in texts to convey a specific concept, and their relations such as synonymy. In the following section, we will detail how bottom–up and top–down approaches to ontology construction have been combined together into a single construction process in the framework of the DALOS project.

3.2 Knowledge Modelling in the DALOS Project

DALOS[1] was a project launched within the "eParticipation" framework, the EU Commission initiative aimed at promoting the development and use of Information and Communication Technologies in the legislative decision–making processes. The aim of this initiative was to foster the quality of the legislative production, to enhance accessibility and alignment of legislation at European level, and to promote awareness and democratic participation of citizens in the legislative process.

In particular, DALOS aimed to ensure that legal drafters and decision–makers have control over the legal language at national and European level, by providing

[1] DrAfting Legislation with Ontology–based Support.

law–makers with linguistic and knowledge management tools to be used in the legislative processes, in particular within the phase of legislative drafting. To this specific end, a knowledge resource was designed and implemented within the project, the DALOS Knowledge Organization System (KOS).

The DALOS KOS is organized in two layers:

- the *Ontological layer*, containing the conceptual modelling at a language-independent level;
- the *Lexical layer*, containing multi–lingual terminology conveying the concepts represented at the Ontological layer.

Concepts at the Ontological layer are linked by taxonomical as well as object property relationships (e.g.`has_object_role`, `has_agent_role`, `has_value`, etc.). On the other hand, the Lexical layer aims at describing the language–dependent lexical expression of the concepts contained in the Ontological layer. At this level, lexical units can be linked through linguistic relationships such as `synonymy`, `hypernymy`, `hyponymy`, `meronymy`, etc.

In the DALOS KOS, the two layers are connected by relationships mapping concepts to their linguistic counterpart, i.e. terms: this mapping is implemented through the `hasLexicalization` relationship, which from a monolingual perspective maps a given concept to the term(s) expressing it, whereas from a cross–lingual perspective it maps a given concept to the multilingual terminological variants conveying it.

In this two–layer architecture, the Ontological layer acts as a layer that aligns concepts at the European level, independently from the language and the legal order, where possible. Moreover, the Ontological layer allows to reduce the computational complexity of the problem of multilingual term mapping (N–to–N mapping). Concepts at the Ontological layer act as a "pivot" meta–language in an N–language environment, allowing the reduction of the number of bilingual mapping relationships from a factor N^2 to a factor $2N$. Entries and relationships at both levels are described by exploiting the expressiveness of RDF/OWL semantic Web standards.

The two–level knowledge architecture is illustrated in Figure 1, where it can be noticed that the Ontological layer provides a detailed semantic description of the defined concepts and their relationships and properties, and the Lexical layer describes its linguistic counterpart through the domain terms and the linguistic relationships linking them.

The terms at the Lexical layer are linked by different types of linguistic relationships: for instance, the English term *supplier* is linked to its hyponyms *supplier of goods* and *supplier of services* as well as to its Italian translation equivalent *fornitore*. Another type of lexical relationship, so–called `fuzzynym`, appears to hold between the terms *consumer* and *supplier*: such a relationship refers to a wider associative relation linking words which may share a number of salient features (in the case at hand, of being involved in a commercial transaction) without being necessarily semantically similar. At the Ontological layer the defined concepts are linked through different types of relationships, namely

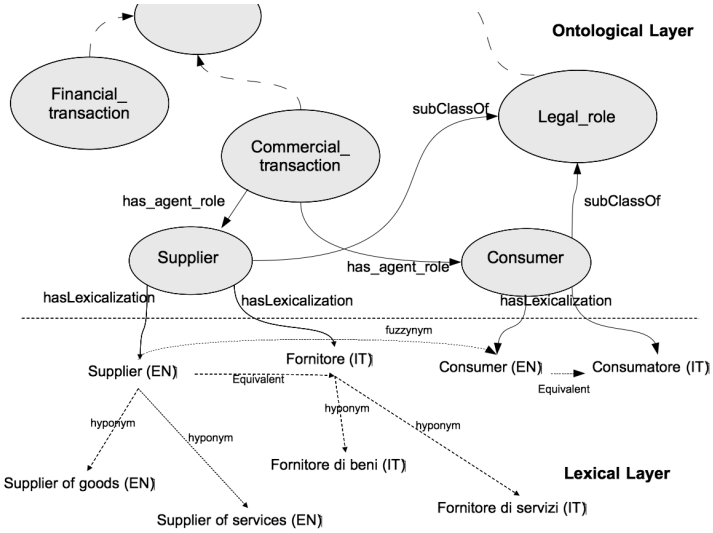

Fig. 1. Knowledge Organization System (KOS) of the DALOS resource

`subClassOf` (such as the one holding between the SUPPLIER and LEGAL_ROLE concepts) and `has_agent_role` (linking the COMMERCIAL_TRANSACTION concept to the SUPPLIER one). It is interesting to note that the semantic relatedness between the terms *supplier* and *consumer* captured by the `fuzzynym` relationship at the lexical level is assigned an explicit semantic interpretation at the ontological level, where it can be noticed that the corresponding concepts a) relate as agents to the COMMERCIAL_ TRANSACTION concept, and b) are subclasses of the LEGAL_ROLE concept.

The DALOS KOS was built following the middle–out approach sketched in Section 3.1. In particular, the DALOS KOS construction was articulated into three main lines of activity:

1. the top–down construction of a (core) domain ontology;
2. the semi–automatic extraction of terminology from domain corpora in different languages by using Natural Language Processing technologies combined with Machine Learning techniques;
3. the refinement of the Ontological and Lexical layers and well as the linking between the two, driven by the terminological and ontological knowledge extracted from the domain corpora.

Whereas the first activity line refers to a top–down process carried out manually by domain experts, the second one corresponds to a bottom–up process aimed at boostrapping the domain terminology from legal document collections. The third activity line refers to the linking of the Ontological layer and the Lexical layer as well as to the refinement of both of them on the basis of the lexical

and ontological knowledge bootstrapped from texts. It is at this level that the results of the top–down and bottom–up processes are combined together through an incremental process. For instance, the results of the term extraction process can play an important role by suggesting ontology concepts which were not originally included in the top–down ontology. In principle, the reverse could also hold, in the case where no terms have been acquired that denote some of the concepts included in the Ontological layer.

In what follows, the three activity lines will be illustrated in detail, with particular emphasis on their interaction. Note that for the DALOS case study the "consumer protection" domain has been selected.

Construction of the DALOS Domain Ontology. The Ontological layer of the DALOS resource is aimed at providing an alignment of concepts at language-independent level. It acts not only as a pivot structure for language-dependent lexical manifestations, but it provides an ontologically characterized description of the chosen domain in terms of concepts and their relations, exploiting the expressiveness and reusability of the RDF/OWL semantic Web standards for knowledge representation. This allows also to validate the developed knowledge resource with respect to existing foundational or core ontologies.

As discussed above, the Ontological layer is the result of an intellectual activity aimed at describing the consumer protection domain, chosen for the pilot case. Within the project constraints, an intellectual approach has been chosen to manually capture ontological relations between concepts, relying on expert judgment.

Classes and properties have been implemented on the basis of the terminological knowledge extracted from the chosen Directives on consumer protection law, in particular from the "definitions" contained, maintaining coherence to the design patterns of the Core Legal Ontology (CLO)[2] [25] developed on top of DOLCE foundational ontology [36] and the "Descriptions and Situations" (DnS) ontology [24] [35] within the DOLCE+ library[3]. The DALOS ontology covers the entities pertinent to the chosen domain and their legal specificities. In this knowledge architecture the role of a core legal ontology is to provide entities/concepts, which belong to the general theory of law, bridging the gap between domain-specific concepts and the abstract categories of formal upper level or foundational ontologies such as, in our case, DOLCE.

As regards domain-specific concepts, the DALOS Ontological layer is designed to stress the distinction identified by the "Descriptions and Situations" ontology, extended by CLO within the legal domain, between *intensional specifications* like norms, contracts, roles, and their *extensional realizations* in the same domain, such as cases, contract executions, and agents. This distinction is formally captured by the so called *Norm* ↔ *Case* design pattern [26] (CODeP[4]). According to the *Norm* ↔ *Case* CODeP, *intensional specifications* like norms use tasks, roles, and parameters, while *extensional realizations* like legal cases conform to

[2] `http://www.loa-cnr.it/ontologies/CLO/CoreLegal.owl`
[3] DOLCE+ library, `http://dolce.semanticweb.org`
[4] Conceptual Ontology Design Pattern.

norms when actions, objects and values are classified by tasks, roles, and parameters respectively. The matching is typically performed when checking if each entity in a legal fact is compliant to a concept in a legal description [26].

The distinction stressed by DALOS is strictly linked to the activity of legislative drafting addressed by the project. Apart from more technical provisions like 'amendments' on existing norms, legislative drafting can in fact be considered as an activity that creates norms on generic situation descriptions, qualifying them by, for example, deontic terms [29]. According to CLO, this activity deals with descriptions (*intensional specifications*) of generic situations (also called "situational frameworks" in [29]), giving them a normative perspective. For example the Directive 97/7/EC of 20 May 1997, at Art. 7 paragraph 1 states that "Unless the parties have agreed otherwise, the supplier must execute the order within a maximum of 30 days from the day following that on which the consumer forwarded his order to the supplier". This states that, unless differently agreed, the generic situation in which the supplier is obliged to execute an order to the consumer, following a consumer request, and this obligation has to be satisfied within a maximum of 30 days from the consumer request.

A normative perspective on generic situations is the result of the legislative drafting activity; it results in legislative text paragraphs grouped into articles, which can be semantically qualified as *provisions* [7], i.e. fragments of a regulation (for example an *obligation* for a role with respect to a task).

A support for legislative drafting can therefore include: 1) a taxonomy of provision types able to give a normative perspective to generic situations; 2) a knowledge resource supporting the description of generic situations in a specific domain, as well as giving an ontological perspective on entities involved in such situations [9]. The DALOS Ontological layer aims at representing this second kind of knowledge resource, tailored to the consumer protection domain pilot case.

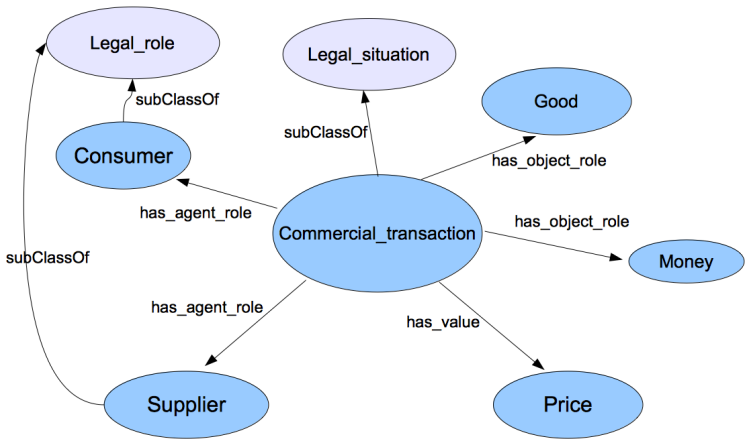

Fig. 2. Excerpt of the DALOS Ontological Layer

The Ontological layer is therefore populated by the conceptual entities, which characterize the consumer protection domain. The first assumption is that all concepts *defined* within consumer law are representative of the domain, and, as a consequence, that several concepts *used* in the definitional contexts pertain to the ontology as well, representing the basic properties or, in other words, the 'intensional meaning' of the relevant concepts. Similarly, the Ontological layer contains generic situations having a legal relevance in the chosen domain.

Such domain-specific concepts are classified according to more general notions, imported from CLO, such as LEGAL_ROLE and LEGAL_SITUATION. Examples of some concepts obtained by the definitions from the consumer law domain are COMMERCIAL_TRANSACTION, CONSUMER, SUPPLIER, GOOD, PRICE. The specific roles they play ([35]) are illustrated in Fig. 2.

On the other hand, the main entities derived from CLO are axiomatized, disjoint classes, characterized by meta properties, such as Identity, Unity and Rigidity. The most relevant distinction is between Roles (anti-rigid) and Types (which are rigid). Roles, according to [35], are anti-rigid since they are "properties that are contingent (non-essential) for all their instances". Types on the other hand can play more roles at the same time. For instance, a legal subject (either a natural or artificial person) can be a seller and a buyer. Domain-specific requirements are expressed by restrictions over ontological classes, for instance by defining CONSUMER as a role that can be *played by* NATURAL_PERSON only.

The first version of the DALOS Ontological layer contains 121 named classes with necessary & sufficient definitions, implemented in the OWL-DL language.

Terminology Extraction in the DALOS Project. Term extraction is the first and most–established step in ontology learning from texts. Terms are surface realisations of domain–specific concepts and represent, for this reason, a basic prerequisite for ontology construction as well as more advanced ontology learning tasks. In principle, they need to be recognized whatever the surface form they show in context, irrespectively of morpho–syntactic and syntactic variants. A term can be a common noun as well as a complex nominal structure with modifiers (typically, adjectival and prepositional modifiers). Term extraction thus requires some level of linguistic pre–processing of texts.

In the DALOS project, term extraction was performed on the English and Italian parts of the DALOS consumer law multi–lingual corpus, including Directives, Regulations and case law on protection of consumers' economic and legal interests. The corpus was built by legal experts and includes 16 Directives and 42 Case Law texts, a total of 292,609 Italian and 273,667 English word tokens.

Term extraction was performed with two different acquisition systems, which were used for dealing with English and Italian texts respectively. For English, GATE[5] [15] was used, a framework for language engineering applications, which supports efficient and robust text processing. GATE uses NLP based techniques to assist the knowledge acquisition process for ontological domain modelling, applying automated linguistic analysis to create ontological knowledge from

[5] http://www.gate.ac.uk

textual resources, or to assist ontology engineers and domain experts by means of semi-automatic techniques. For Italian, T2K (Text–to–Knowledge)[16] [34] was used, a hybrid ontology learning system combining linguistic technologies and statistical techniques.

In both cases, term extraction was carried out on the results of a linguistic pre–processing stage, in charge of enriching the original corpus with valuable linguistic information, which is added to the text by means of annotations, in turn used in the subsequent analysis stages. The linguistic pre–processing modules are in charge of:

1. tokenisation of the input text;
2. sentence splitting, segmenting the text into sentential units;
3. morphological analysis (including lemmatisation);
4. part of speech tagging;
5. shallow syntactic parsing (so–called "chunking").

The starting point of the term extraction process is different for the two systems: whereas term extraction in GATE is perfomed against the pos–tagged text (i.e. the output of step 4 above), T2K starts from the shallow parsed text (step 5). To be more concrete, for what concerns English, term candidates are extracted from the text by first selecting either individual pos–tags or sequences of part of speech tags constituting noun phrases, as exemplified below:

- noun (e.g. *creditor*, *product*);
- adjective–noun (e.g. *current account*, *local government*);
- noun–noun (e.g. *credit agreement*, *product safety*);
- noun–preposition–adjective–noun (e.g. *purchase of immovable property*, *principle of legal certainty*);
- noun–preposition–noun–noun (e.g. *cancellation of credit agreement*, *settlement of consumer dispute*).

For Italian texts, candidate terms are identified in the shallow parsed texts on the basis of a set of chunk patterns encoding syntactic templates of candidate either simple or complex terms. For what concerns the latter, chunk patterns were defined to cover the main modification types observed in complex nominal terms: i.e. adjectival modification (e.g. *organizzazione internazionale* 'international organisation'), prepositional modification (e.g. *commercializzazione di autovetture* 'marketing of cars'), including more complex cases where different modification types are compounded (e.g. *commercio di prodotti fitosanitari* 'trade of fitosanitary products'). The set of chunk patterns used to identify candidate complex terms was tailored to meet the specific needs of the legal domain, characterised by the frequent use of deep PP-attachment chains including a high number of embedded prepositional chunks [53].

Having identified both single and multi–word term candidates from texts, the following step consists in filtering through the candidates to separate terms from non–terms. This step involves the use of statistically–based measures to compute whether and to what extent a term candidate qualifies as a terminological unit.

In the literature, measures for identifying terms range from raw frequency to Information Retrieval measures such as Term Frequency/Inverse Document Frequency (TF/IDF) [45], the C/NC–value method [23], and lexical association measures such as log likelihood [17], mutual information, or entropy.

In GATE term filtering was perfomed on the basis of the TF/IDF measure, a technique widely used in information retrieval and text mining taking into account term frequency and the number of documents in the collection, and yielding a score that indicates the salience of term candidates for each document in the corpus. All term candidates with a TF/IDF score higher than an empirically determined threshold have been selected: in the DALOS case, a TF/IDF threshold value of 5 yielded 3000 selected terms.

T2K adopts a different term filtering strategy. If on the one hand single terms are identified on the basis of raw frequency in the source document collection (after discounting stop–words), on the other hand multi–word terms are selected on the basis of the log–likelihood measure, an association measure that quantifies how likely the constituents of a complex term are to occur together in a corpus if they were (in)dipendently distributed, where the (in)dependence hypothesis is estimated with the binomial distribution of their joint and disjoint frequencies. The lists of acquired potential single and complex terms are then ranked according to raw frequency and the associated log–likelihood ratio respectively. The selection of the final set of terms (both single and complex ones) requires some threshold tuning, depending on the size of the document collection and the typology and reliability of expected results. In T2K, thresholds define *a)* the minimum frequency for a candidate term to enter the lexicon, and *b)* the overall percentage of terms that are promoted from the ranked lists. For the DALOS corpus, we adopted the following thresholds: minimum frequency threshold equal to 5 for both single and complex terms; selected single terms cover the topmost 20% in the ranked list, whereas selected multi–word terms correspond to the topmost 70% of the ranked list of candidate complex terms. With this configuration, we obtained a term list of 1,443 terms (both single and multi–word terms), of which 1,168 are multi-word terms of different complexity corresponding to the 80% of the acquired term list.[6]

Evaluation of acquired English and Italian term lists was carried out with respect to a subset of 56 of the European Union Legal Concepts (EULG concepts) from LOIS (see [40] and [37] for the complete list) which were selected as a gold standard. The selection of these EULG concepts was based on the fact that they are explicitly listed and defined in the directives included in the DALOS corpus, and are therefore considered to play an important role in their conceptual characterization. Achieved results are promising in both cases: for English, the percentage of correctly acquired terms with respect to all terms appearing in the gold standard terminology is 73.2%, for Italian 80.69%.

[6] This peculiar distribution of single vs complex terms follows from the fact that multi–word terms appear to cover the vast majority of domain terminology (85% according to [38]).

For Italian, another evaluation type was carried out, to assess the precision of acquired results, calculated as the percentage of correctly acquired terms with respect to all acquired terms. Automatically acquired terms were evaluated against two reference resources, namely the *Archivio DoGi (Dottrina Giuridica)*[7] and *JurWordNet* [24], containing respectively 9,127 keywords and 5,353 lemmata; note that these resources could not be used for an evaluation in terms of recall (calculated as the percentage of correctly acquired terms with respect to all terms in the reference lexicon) due to their wider coverage, which is not limited to the selected domain. By considering both full and partial[8] matches, the observed precision corresponds to 85.38%, with only 14.62% cases of non–matching terms. Manual inspection of non–matching cases showed that only 6.1% of the cases were to be considered as real errors.

Semi–automatic Refinement and Linking of the Ontological and Lexical Layers. The result of the first two activity lines consists of a hand–crafted core domain ontology and of multilingual term lists. It goes without saying that, when considered separately, the two results cannot effectively be used to support legal knowledge management applications. Only the linking of the domain–specific terms extracted from texts to their description in the ontology provides a usable platform for semantic interpretation of textual information. In this section, we will briefly illustrate the strategy followed within the DALOS project for term to concept mapping, where the results of the bottom–up acquisition process are used both to define the mapping between the Lexical and Ontological layers and to refine the already defined ontology.

First, acquired terms were carefully evaluated by domain experts and linked to the concepts they express in the top–down ontology. It may be the case that newly acquired terms do not find a counterpart at the ontological level; if judged as relevant by domain experts, the ontology is revised accordingly.

However, term extraction is not the only contribution of bottom–up approaches to ontology construction. Extracted terms need to be organized into proto–conceptual relational structures, for them to be exploited in the ontology refinement by domain experts. At this level, different types of relations linking acquired terms can be discovered, based on their distribution in texts.

Starting from the lists of acquired English and Italian terms, different types of lexical relations holding between them were extracted. Acquired relations were in turn used to model and refine the Ontological layer, both at the level of defined concepts and of the relationships linking them.

First, for both English and Italian, the acquired terms were organized into fragments of taxonomical chains, whereby terms such as *time–share contract*,

[7] `http://nir.ittig.cnr.it/dogiswish/dogiConsultazioneClassificazioneKWOC.php`

[8] Partial matches refer to the following cases: a) the same term appears both in the extracted termbank and in the gold standard resource under different prototypical forms; b) the gold reference resource contains a more general term whereas the extracted list includes one of its hyponyms; c) the gold reference resource contains a more specific term with respect to the extracted list which includes its hypernym.

credit contract and *consumer contract* were classified as co–hyponyms of the general term *contract*. In both cases, taxonomical relationships between terms (typically, single and multi–word terms) were reconstructed by exploiting the internal structure of noun phrases [10]: under this approach, a taxonomic relation is acquired as holding between a single term and all complex terms with this term as the headword.

For English, a second acquisition technique has been experimented with, based on Hearst patterns [28], i.e. a set of lexico–syntactic patterns typically conveying information about hyponymic relations in unrestricted texts. Consider the following example pattern, i.e. "NP such as (NP,)* (or—and) NP" where NP stands for a Noun Phrase and the regular expression symbols have their usual meanings, matching the following context: *advertising and marketing practises, such as product placement, brand differentiation or the offering of incentives ...*. From contexts like this one it is possible to acquire hyponymic relations such as the one holding between the term *product placement* and the more general term *advertising and marketing practices*. Taxonomical relations acquired with this technique are not limited to head–sharing terms only. Typically, with this technique a high level of precision can be achived, but quite low recall [14]. Unfortunately, this turned out not to be the case with the DALOS corpus; as reported in [41], Hearst patterns appear very rarely in legal corpora.

The identification of taxonomic relations between terms allows the ontology engineer to create concept hierarchies that represent the backbone of the ontology under construction. These linguistic relations can then be reformulated in terms of ontological relations, by means of the OWL SubClassOf relation. Examples from the DALOS ontology are:

DistanceContract `SubClassOf` Contract
CommercialActivity `SubClassOf` Activity

whose linguistic counterpart (namely, *distance contract* is hyponym of *contract* and *commercial activity* is hyponym of *activity*) has been extracted from both the English and Italian corpora.

Yet, taxonomic relations do not exaust the typology of linguistic relations holding between terms which can be automatically extracted from running texts.

For Italian, T2K also acquires clusters of semantically related terms on the basis of distributionally–based similarity measures [1]: following this approach, two terms are semantically related if they can be used interchangeably in a statistically significant number of syntactic contexts. For all terms (both single and complex ones) in the acquired list, we extracted a set of 1,071 semantically related terms referring to 238 terminological headwords. Clusters of automatically acquired semantically related terms are exemplified below:

```
disposizioni 'provision'
  norme, disposizioni legislative, decisione, atto, prescrizioni
legge 'law'
  regolamento, protocollo, accordo, statuto, amministrazioni comunali
```

```
pubblicità ingannevole 'misleading advertisement'
   pratiche commerciali, procedimento, pubblicità comparativa, clausole abusive,
   pubblicità
```

It should be appreciated that in these clusters of semantically related words different classificatory dimensions are inevitably collapsed; they include not only quasi–synonyms (as in the case of *disposizioni* 'provision' and *norme* 'regulations'), hypernyms and hyponyms (e.g. *pubblicità* 'advertisement' and *pubblicità ingannevole* 'misleading advertisement'), but also looser word associations. As an example of the latter we mention the relation holding between *legge* 'law' and *amministrazione comunale* 'municipal administration', or between *comitato* 'committee' and *membri* 'members'.

Acquired clusters of semantically related words can be usefully exploited for the linking between the Lexical and Ontological layers as well as for refining the Lexical and Ontological layers of the DALOS KOS. At the lexical level, whenever possible semantic relatedness between words detected through distributionally–based measures is encoded in terms of lexical aradigmatic relationships such as synonymy, hyponymy/hypernymy, meronymy, antonymy, etc. Remaining relations, which should rather be ascribed to a generic syntagmatic relatedness between words (due to any kind of functional relationship or frequent association), have been encoded in the Lexical layer of DALOS KOS in terms of a rather vague relationship, so–called `fuzzynym` (see Figure 1). For what concerns the Ontological layer, acquired relations have been carefully evaluated by domain experts and encoded in terms of new classes and/or properties (see Figure 1 and its discussion above).

For what concerns English, experiments have been carried out with respect to two syntagmatic relation types, namely a) verbal complementation patterns and b) syntagmatic relations detected through association measures.

Verbal patterns typically reflect lexicalized semantic relations between its arguments. In order to investigate the nature of the semantic contribution from verbal patterns the user needs to be enabled to browse a text according to predefined patterns. Patterns defined in the GATE interface can consist of any type of text annotation that has been added in GATE, e.g. part of speech, string value, lemma etc. The corpus indexing and querying tool in GATE, called ANNIC (ANNotations In Context) [3], allows the evaluator to enter search patterns over text annotations, and detect semantic relations between ontology elements at the fine-grained text level. As proof of concept, the following simple pattern was defined, which identifies pairs of elements from the DALOS ontology that are mentioned in the texts as verb arguments. The surface representation restricts the verb context to a two-token window on either side.

DalosConcept(Token)*2Token.category=="VERB"(Token)*2 DalosConcept

A graphical user interface allows the user to query a corpus and inspect the results from the query. The screenshot in Figure 3 below illustrates how the

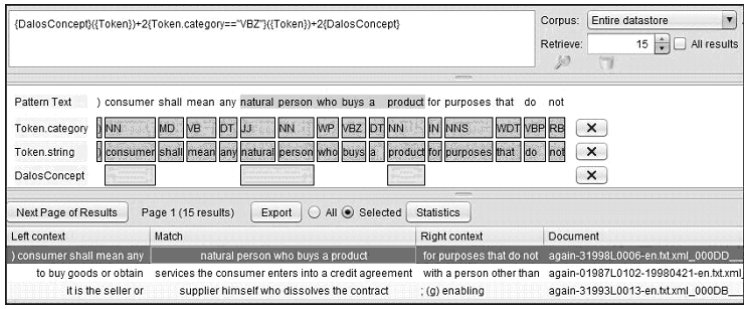

Fig. 3. Snapshot of ANNIC functionality

results are displayed in the GATE interface. Annotations over spans of text are displayed as rows with coloured blocks indicating part of speech, string and DalosConcept. Contexts to the left and right of the text matching the search pattern are displayed at the bottom.

Using this query, 56 patterns were extracted, of which 37 (66%) were evaluated as deserving expert attention. For example:

NaturalPerson	conclude	Contract	with Seller or Supplier
NaturalPerson	buy	Product	
Seller/Supplier	dissolve	Contract	
Consumer	enter into	CreditAgreement	

For what concerns the second experiment, pointwise mutual information (PMI) is a well-known technique that measures the mutual dependence of the two variables as an expression of a syntagmatic relation. It is commonly used as a significance function for the computation of collocations in corpus linguistics [48], measuring the statistically-based strength of relatedness through collocation within the same document. Overall, forty PMI relations were found between existing concepts from the DALOS ontology after matching DALOS ontology labels onto textual elements. Nine (22.5%) of the forty are not connected by any relation or concatenation of relations in the ontology. Consider, for instance, the following pairs with their associated PMI value:

ConsumerGoods	ConsumerProtection	4.10099
ConsumerProtection	Consumer	3.37321
FinancialService	Supplier	2.56943

It turned out that 77.5% of the extracted MI relations are already attested in the ontology. The 22.5% of the MI pairs without ontological confirmation make ontological sense in that they express fine–grained relations that should be expertly evaluated for inclusion into the ontology, and linked to existing ontology elements by means of existing or new object properties. In general, the significant overlap between pointwise mutual information results and existing ontological relations indicate the relevance of such a measure for ontology acquisition.

From the work discussed so far, it should be clear that automatic knowledge acquisition cannot be seen as a stand–alone method for ontology creation, refinement, expansion and population, but rather as a support to the engineering activity of domain experts. In this section, we have shown how the results of text–driven knowledge extraction, which is just a phase in the ontology development cycle, can be used for the manual development, refinement or extension of domain ontologies.

4 Legal Rules Learning

In this section an approach to ontology learning in the legal domain is presented, which is comlementary to the DALOS methodology.

Domain ontologies assume a specific importance in the legal domain since they provide knowledge, in terms of concepts and their relationships, on scenarios addressed by legal rules, expressed in legal texts. Domain concepts addressed by legal rules are particularly relevant for the legal domain. In fact, in this domain users are mainly interested in accessing concepts regulated by norms. They look for legal reasoning and consultancy support, as, for example, instruments to check compliance with procedures with respect to specific statutes and regulations.

The approach presented in this section addresses the identification of domain concepts addressed by legal rules, as derived from knowledge extraction techniques, aimed at legal rules learning from legislative texts. The extracted domain concepts as well as the established relationships can represent a starting point for the implementation of domain ontologies.

An approach to support the acquisition of legal rules contained in legislative documents has been recently proposed [8] [21]. It is based on a semantic model for legislation and implemented by using knowledge extraction techniques over legislative texts. This methodology is targeted at providing a contribution to bridge the gap between consensus and authoritativeness in legal rule representation, because it contributes to reaching consensus by limiting human intervention in the descrituion of legal rules, which are extracted from authoritative texts as the legislative ones.

The proposed approach to legal knowledge acquisition is based on learning techniques targeted at extracting legal rules from text corpora. Legal rules are essentially "speech acts" [47] expressed in legislative texts regulating entities of a domain: their nature therefore justifies an approach aimed at the analysis of such texts. Therefore, the proposed knowledge acquisition framework is based on a twofold approach:

1. Knowledge modelling: definition of a semantic model for legislative texts able to describe legal rules;
2. Knowledge acquisition: instantiation of legal rules through the analysis of legislative texts, being driven by the defined semantic model.

This approach traces a framework which combines top-down and bottom-up strategies: a top–down strategy provides a model for legal rules, while a bottom-up

strategy identifies rules instances from legal texts. The bottom–up knowledge acquisition strategy in particular can be carried out manually or automatically. The manual bottom-up strategy consists, basically, of an analytic effort in which all the possible semantic distinctions among the textual components of a legislative text are identified. On the other hand, the automatic (or semi-automatic) bottom-up strategy performs the previous activities with support from automatic tools that are able to classify rules, according to the defined model, and identify the involved entities.

The knowledge model proposed in this work reflects this orientation and is organized into the following two components:

1. Domain Independent Legal Knowledge (DILK)
2. Domain Knowledge (DK)

DILK is a semantic model of Rules expressed in legislative texts, while DK is any terminological or conceptual knowledge base (thesaurus, ontology, semantic network) able to provide information and relationships among the Entities of a regulated domain. The combination of DILK with one or more DKs is able to provide a formal characterization of Rules instances. For this reason the proposed methodology to legal knowledge modelling has been called *DILK-DK* approach [21].

DILK. DILK is conceived as a model for legal Rules, independently from the domain they apply to. In literature several models (classification) of legal rules have been proposed, from the traditional Hohfeldian theory of legal concepts [30] until more recent legal philosophy theories due to Rawls [43], Hart [27], Ross [44], Bentham [6], Kelsen [31].

In this respect, the work of Biagioli [7] deserves particular attention. Combining the work of legal philosophers on rules classification with the Searlian theory of rules preceived as "speech acts", as well as the Raz's lesson [42] to perceive laws and regulations as a set of *provisions* carried by speech acts, Biagioli underlined two views or *profiles* according to which a legislative text can be perceived: a) a structural or *formal profile*, representing the traditional legislator habit of organizing legal texts into chapters, articles, paragraphs, etc.; b) a semantic or *functional profile*, considering legislative texts as composed by *provisions*, namely fragments of regulations [7] expressed by speech acts. Therefore, a specific classification of legislative provisions was carried out by analysing legislative texts from a semantic point of view, and grouping provisions into two main families: *Rules* (introducing and defining entities or expressing deontic concepts) and *Rules on Rules* (different kinds of amendments). Rules are provisions which aim at regulating the reality considered by the including act. Adopting a typical law theory distinction, well expressed by Rawls, rules consist of:

- *constitutive rules*: they introduce or assign a juridical profiles to entities of a regulated reality;
- *regulative rules*: they discipline actions ("rules on actions") or the substantial and procedural defaults ("remedies").

On the other hand, Rules on Rules can be distinguished into:

- *content amendments*: they modify the literal content of a norm, or their meaning without literal changes;
- *temporal amendments*: they modify the times of a norm (come-into-force and efficacy time);
- *extension amendments*: they extend or reduce the cases on which the norm operates.

In Biagioli's model each provision type has specific arguments describing the roles of the entities which a provision type applies to (for example the *Bearer* is argument of a *Duty* provision). *Provision types* and related *Arguments* represent a semantic model for legislative texts [7]. They can be considered as a sort of metadata scheme able to analytically describe fragments of legislative texts. For example, the following fragment of the Italian privacy law:

> *"A controller intending to process personal data falling within the scope of application of this act shall have to notify the "Garante" thereof, . . . "*

besides being considered as a part of the physical structure of a legislative text (a *paragraph*), can also be viewed as a component of the logical structure of it (a *provision*) and qualified as a *provision* of type *Duty*, whose arguments are:

Bearer: "Controller"; *Object*: "Process personal data"
Action: "Notification" *Counterpart*: "Garante"

The specific textual anchoring of Biagioli's model represents, in our opinion, its main strength. Since the DILK-DK approach aims at representing Rules instances as expressed in legislative texts, we consider Biagioli's model, limited to the group of rules, as a possible implementation of DILK. "Rules on rules" affect indirectly the way how the reality is regulated, since they amend Rules in different respects (literally, temporarily, extensionally): therefore such provision types are not part of DILK model. On the other hand, their effects on Rules has to be taken into account for knowledge acquisition purposes.

DK. In legislative texts *Entities* regulated by provisions are expressed by lexical units. These can be provided by a *Domain Knowledge* (DK) repository providing conceptualization of entities consisting of the language-dependent lexical units[9]. Information on such entities at language-independent level, as well as their lexical manifestations in different languages needs to be described by DK. A possible architecture for describing DK has been proposed within the DALOS project[10]

[9] "Typically regulations are not given in an empty environment; instead they make use of terminology and concepts which are relevant to the organisation and/or the aspect they seek to regulate. Thus, to be able to capture the meaning of regulations, one needs to encode not only the regulations themselves, but also the underlying ontological knowledge. This knowledge usually includes the terminology used, its basic structure, and integrity constraints that need to be satisfied." [2].

[10] http://www.dalosproject.eu

(see Section 3.2). More details on the DALOS DK architecture, as well as a possible implementation of it for the domain of consumer protection, can also be found in [22] (see also previous section).

Knowledge Acquisition. Knowledge acquisition within the DILK-DK framework consists of two main steps: 1) DILK instantiation, 2) DK construction.

DILK instantiation. The DILK instantiation phase is a bottom-up strategy for legislative text paragraphs classification into *provision types*, as well as specific lexical units identification, assigning them roles in terms of *provision arguments*. The automatic bottom-up strategy, here proposed, consists in using tools able to support the human activity of classifying provisions, as well as to extract their arguments. Three main steps can be foreseen:

- Collection of legislative texts and conversion into an XML format [5]
- Automatic classification of legislative text paragraphs into provisions [20]
- Automatic argument extraction [8]

Legislative documents are firstly collected and transformed into a jurisdiction-dependent XML standard (NormeInRete in Italy, Metalex in the Netherlands, etc.). For the Italian legislation a module called xmLegesMarker, of the xmLeges[11] software family, has been developed [5]. It is able to transform legacy content into XML in order to identify the formal structure of a legislative document.

For the automatic classification of legislative text paragraphs as provison types, a tool called xmLegesClassifier of the xmLeges family has been developed. xmLegesClassifier has been implemented using a Multiclass Support Vector Machine (MSVM) approach, which provided the best results in preliminary experiments compared to other machine learning approaches [20]. With respect to [20], in this work MSVM is tested on the Rules provision family, as the first step of DILK instantiation[21].

A tool called xmLegesExtractor[12] [8] of the xmLeges family has been implemented for the automatic detection of provision arguments. xmLegesExtractor is realized as a suite of NLP tools for the automatic analysis of Italian texts (see [4]), specialized to cope with the specific stylistic conventions of the legal parlance. A first prototype takes as input legislative raw text paragraphs, coupled with the categorization provided by the xmLegesClassifier, and identifies text fragments (lexical units) corresponding to specific semantic roles, relevant for the different types of provisions (Fig. 4). The approach follows a two–stage strategy. The first stage consists in a syntactic pre–processing which takes in input a text paragraph, which is tokenized and normalized for dates, abbreviations and multi–word expressions; the normalized text is then morphologically analyzed and lemmatized, using an Italian lexicon specialized for the analysis of legal

[11] http://www.xmleges.org

[12] xmLegesExtractor has been developed in collaboration with the Institute of Computational Linguistics (ILC-CNR) in Pisa (Italy).

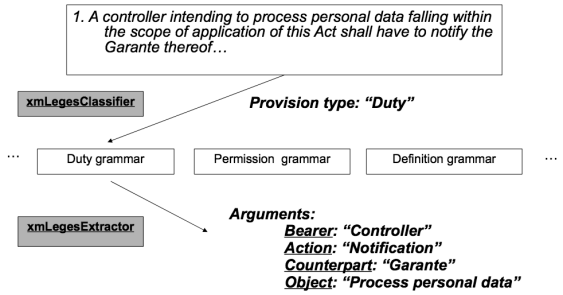

Fig. 4. xmLegesClassifier combined with the grammar approach used by xmLegesExtractor

language; finally, the text is POS-tagged and shallow parsed into non–recursive constituents called "chunks" [18]. The second stage consists in the identification of all the lexical units acting as arguments relevant to a specific provision type. It takes in input a chunked representation of legal text paragraphs, locating relevant patterns of chunks which represent entities with specific semantic roles within a provision type instance, by using a specific provision type oriented grammar (Fig. 4).

DK construction. Lexical units identified by xmLegesExtractor represent language-dependent lexicalizations of provision arguments. More information on related entities, as well as their relations within a specific domain, can be obtained by mapping lexical units to concepts in any existing DK repositories. On the other hand, the extracted information can be considered as a bisis on which to construct DK repositories (in terms of thesauri or domain ontologies). Actually, their construction is not a specific task of *legal* ontologists, but of ontologists *tout court*, since a DK repository has to contain information on entities of a domain independently from a legal perspective. This aspect is important in order to conceive a legal knowledge architecture whose components can be reused. A DILK-DK learning approach only suggests that only language-dependent lexical units are contained in DK repositories, which can be implemented by projecting lexical units onto a large text corpora of a specific domain, inferring conceptualizations by term clustering, as well as using statistics on recurrent patterns for discovering term relationships. This issue is out of the paper scope; a vast literature exists on this topic, therefore the interested reader can refer to [12].

Benefits of the DILK-DK bottom-up learning approach. The proposed learning approach for legal knowledge acquisition can provide the following benefits: a) it assists the implementation of taxonomies, or suggests concepts for hand-crafted ontologies [55]; b) it contributes to bridging the gap between authoritativeness and consensus for legal rule representation, since it is able to extract rules directly form legislative texts, which are authoritative sources (by definition), while promoting consensus, since rules are automatically extracted from legal sources, limiting human interaction.

5 Discussion

In analysing legal documents, the first aspect that must be elicited is the relation between meaning (norm) and form (text). Norms are conceived as the interpreted meaning of written regulations that correspond to a partition in legal text, such as articles and paragraphs. Additionally, a norm can be built by interpretative activities on a set of linguistic expressions logically entailed, for instance, the decision in a judgement, or set of legislative statements in a judgement, or set of legislative statements. Only in few cases (definitions, deeming provisions) legal concepts are elicited from the core meaning of a single norm, but more frequently they are built on sets of norms, through a process of abstraction and generalization, by collecting sets of normative conditions, to be linked to sets of legal effects.

To respect the peculiarities of the legal domain, different approaches should be adopted in the process of legal concept extraction. On one side, a conceptual model of the domain needs to be created, either by means of manual ontology engineering or the extraction of the intensional definition of legal concepts from linguistic contexts. Legislative definitions are generally expressed by fixed linguistic structures within a legislative text, and therefore they can be easily identified and isolated.

On the other side, the analysis of text containing legislative provision instances can identify relevant concepts as well as relationships pertaining to a regulated domain, thus providing effective hints for the construction of a domain ontology as well as linking the related concepts to core and fundamental ontologies.

Techniques such as term extraction, lexical analysis, parsing and statistical collocations (as discussed and illustrated in previous sections) yield textually derived information, which can then be re-engineered into ontological concepts, concept properties and relations. The work performed in both the DALOS project and the approach followed within the DILK-DK framework is illustrative of this type of activity, by means of bottom–up knowledge acquisition in the former case, or as a result of provisions categorization in the latter case.

The application of both top–down and bottom–up knowledge acquisition techniques to the legal domain enables the adoption of several perspectives on legal knowledge and the formulation of various aims within the legal field. For example, in the field of legal comparison, different conceptualizations (resulting from the bottom-up analysis of texts from different legal systems) can be compared throughout a shared reference ontology, conceived as a level of abstraction, which legal experts can agree upon. In addition, the same model can be exploited in European law-making, where conceptual equivalence of multiligual entities is assumed. The role of a reference ontology is, in this case, to assess (multilingual) terminological consistency, i.e. whether different lexicalization reflects the same normative conceptual meaning.

References

1. Allegrini, P., Montemagni, S., Pirrelli, V.: Example-Based Automatic Induction of Semantic Classes Through Entropic Scores. In: Linguistica Computazionale, XVI-XVII, Tomo I, pp. 1–45 (2003)
2. Antoniou, G., Billington, D., Governatori, G., Maher, M.: On the modeling and analysis of regulations. In: Proceedings of ACIS, pp. 20–29 (1999)
3. Aswani, N., Tablan, V., Bontcheva, K., Cunningham, H.: Indexing and Querying Linguistic Metadata and Document Content. In: Proceedings of 5th International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria (2005)
4. Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V., Soria, C.: Automatic classification and analysis of provisions in italian legal texts: a case study. In: Meersman, R., Tari, Z., Corsaro, A. (eds.) OTM-WS 2004. LNCS, vol. 3292, pp. 593–604. Springer, Heidelberg (2004)
5. Bacci, L., Spinosa, P., Marchetti, C., Battistoni, R.: Automatic mark–up of legislative documents and its application to parallel text generation. In: Proceedings of LOAIT Workshop, Barcelona, Spain, pp. 45–54 (2009)
6. Bentham, J., Hart, H.L.A.: Of Laws in General (1st edn., 1872). Athlone, London (1970)
7. Biagioli, C.: Towards a legal rules functional micro–ontology. In: Proceedings of LEGONT (1997)
8. Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S., Soria, C.: Automatic semantics extraction in law documents. In: Proceedings of ICAIL, pp. 133–139 (2005)
9. Biagioli, C., Cappelli, A., Francesconi, E., Turchi, F.: Law making environment: perspectives. In: Proceedings of the V Legislative XML Workshop, pp. 267–281. European Press Academic Publishing, Firence (2007)
10. Buitelaar, P., Olejnik, D., Sintek, M.: A protégé plug–in for ontology extraction from text based on linguistic analysis. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS, vol. 3053, pp. 31–44. Springer, Heidelberg (2004)
11. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: an Overview. In: Buitelaar, et al. (eds.) Ontology Learning from Text: Methods, Evaluation and Applications. Frontiers in Artificial Intelligence and Applications, vol. 123, pp. 3–12 (2005)
12. Buitelaar, P., Cimiano, P. (eds.): Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press, Amsterdam (2008)
13. Casellas, N.: Modelling Legal Knowledge through Ontologies. OPJK: the Ontology of Professional Judicial Knowledge. PhD thesis, Faculty of Law, Universitat Autònoma de Barcelona (2008)
14. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Learning taxonomic relations from heterogeneous sources. In: Proceedings of the ECAI 2004 Ontology Learning and Population Workshop (2004)
15. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: Gate: A framework and graphical development environment for robust nlp tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, ACL 2002 (2002)

16. Dell'Orletta, F., Lenci, A., Marchi, S., Montemagni, S., Pirrelli, V., Venturi, G.: Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio. In: Atti del Convegno Nazionale Ass.I.Term I-TerAnDo, Università della Calabria, 5-7 giugno 2008, Roma, AIDA Informazioni, n. 1-2/2008, pp. 185–206 (2008), http://www.aidainformazioni.itpubdellorletta-atal122008.pdf
17. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1), 61–74 (1993)
18. Federici, S., Montemagni, S., Pirrelli, V.: Shallow Parsing and Text Chunking: a View on Underspecification in Syntax. In: Proceedings of the Workshop on Robust Parsing, European Summer School on Language, Logic and Information (ESSLLI 1996), Prague (1996)
19. Ferrajoli, L.: Principia iuris Laterza Bari (2007)
20. Francesconi, E., Passerini, A.: Automatic classification of provisions in legislative texts. International Journal on Artificial Intelligence and Law 15, 1–17 (2007)
21. Francesconi, E.: An Approach to Legal Rules Modelling and Automatic Learning. In: Proceedings of the JURIX Conference (2009)
22. Agnoloni, T., Bacci, L., Francesconi, E., Peters, W., Montemagni, S., Venturi, G.: A two-level knowledge approach to support multilingual legislative drafting. In: Breuker, J., Casanovas, P., Klein, M.C.A., Francesconi, E. (eds.) Law, Ontologies and the Semantic Web. Channelling the Legal Information Flood. Frontiers in Artificial Intelligence and Applications, vol. 188, pp. 177–198. IOS Press, Amsterdam (2009)
23. Frantzi, K., Ananiadou, S.: The C-value/NC-value domain independent method for multiword term extraction. Journal of Natural Language Processing 6(3), 145–179 (1999)
24. Gangemi, A., Sagri, M.T., Tiscornia, D.: Jur-wordnet, a source of metadata for content description in legal information. In: Proceedings of the ICAIL Workshop on Legal Ontologies & Web based legal information management (2003)
25. Gangemi, A., Sagri, M.T., Tiscornia, D.: A constructive framework for legal ontologies. In: Benjamins, V.R., Casanovas, P., Breuker, J., Gangemi, A. (eds.) Law and the Semantic Web. LNCS (LNAI), vol. 3369, pp. 97–124. Springer, Heidelberg (2005)
26. Gangemi, A.: Design patterns for legal ontology construction. In: Casanovas, P., Noriega, P., Bourcier, D. (eds.) Trends in Legal Knowledge. The Semantic Web and the Regulation of Electronic Social Systems, pp. 171–191. European Press Academic Publishing, Firence (2007)
27. Hart, H.: The Concept of Law. Clarendon Law Series. Oxford University Press, Oxford (1961)
28. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th International Conference on Computational Linguistics, pp. 539–545 (1992)
29. Hoekstra, R., Breuker, J., Bello, M.D., Boer, A.: The lkif core ontology of basic legal concepts. In: Casanovas, P., Biasiotti, M., Francesconi, E., Sagri, M.T. (eds.) Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques. CEUR Workshop Proceedings, vol. 321, pp. 43–63 (2007) ISSN 1613-0073, http://CEUR-WS.org/
30. Hohfeld, W.N.: Some fundamental legal conceptions. Greenwood Press, Westport (1978)
31. Kelsen, H.: General Theory of Norms. Clarendon Press, Oxford (1991)

32. Lame, G.: Knowledge acquisition from texts towards an ontology of French law. In: Proceedings of the International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks (EKAW 2000), Juan-les-Pins (2000)
33. Lame, G.: Using NLP techniques to identify legal ontology components: concepts and relations. In: Benjamins, V.R., Casanovas, P., Breuker, J., Gangemi, A. (eds.) Law and the Semantic Web. LNCS (LNAI), vol. 3369, pp. 169–184. Springer, Heidelberg (2005)
34. Lenci, A., Montemagni, S., Pirrelli, V., Venturi, G.: Ontology learning from Italian legal texts. In: Breuker, J., Casanovas, P., Klein, M.C.A., Francesconi, E. (eds.) Law, Ontologies and the Semantic Web, Frontiers in Artificial Intelligence and Applications, vol. 188, pp. 75–94. IOS Press, Amsterdam (2009)
35. Masolo, C., Vieu, L., Bottazzi, E., Catenacci, C., Ferrario, R., Gangemi, A., Guarino, N.: Social roles and their descriptions. In: Welty, C. (ed.) Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning, Whistler (2004)
36. Masolo, C., Gangemi, A., Guarino, N., Oltramari, A., Schneider, L.: Wonderweb deliverable d18: The wonderweb library of foundational ontologies. Technical Report (2004)
37. Montemagni, S., Marchi, S., Venturi, G., Bartolini, R., Bertagna, F., Ruffolo, P., Peters, W., Tiscornia, D.: Report on Ontology learning tool and testing. In: Progetto Europeo DALOS (Drafting Legislation with Ontology–Based Support). Deliverable 3.3. DALOS Project (2007)
38. Nakagawa, H., Mori, T.: Automatic Term Recognition based on Statistics of Compound Nouns and their Components. Terminology 9(2), 201–219 (2003)
39. Paslaru, E., Tempich, C.: Ontology Engineering: A Reality Check. In: Meersman, R., Tari, Z. (eds.) OTM 2006. LNCS, vol. 4275, pp. 836–854. Springer, Heidelberg (2006)
40. Peters, W., Sagri, M.T., Tiscornia, D.: The Structuring of Legal Knowledge in LOIS. In: Proceedings of 10 th International Conference of Artificial Intelligence and Law (ICAIL 2005), Bologna, June 6-11 (2005)
41. Peters, W.: Text–based Legal Ontology Enrichment. In: Proceedings of LOAIT 2009, 3rd Workshop on Legal Ontologies and Artificial Intelligence Techniques joint with 2nd Workshop on Semantic Processing of Legal Texts, Barcelona, Spain, June 8, pp. 55–66 (2009)
42. Raz, J.: The Concept of a Legal System, 2nd edn. Clarendon Press, Oxford (1980)
43. Rawls, J.: Two concepts of rule. Philosophical Review 64, 3–31 (1955)
44. Ross, A.: Directives and Norms. Routledge, London (1968)
45. Salton, G., Buckley, C.: Term–Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24(5), 513–523 (1988)
46. Saias, J., Quaresma, P.: A Methodology to Create Legal Ontologies in a Logic Programming Based Web Information Retrieval System. In: Benjamins, V.R., Casanovas, P., Breuker, J., Gangemi, A. (eds.) Law and the Semantic Web. LNCS (LNAI), vol. 3369, pp. 185–200. Springer, Heidelberg (2005)
47. Searle, J.: Speech Acts: An Essay in the Philosophy of Language. CUP, Cambridge (1969)
48. Smadja, F.A., McKeown, K.R.: Automatically extracting and representing collocations for language generation. In: Proceedings of ACL 1990, Pittsburgh, Pennsylvania, pp. 252–259 (1990)

49. Sure, Y., Studer, R.: A methodology for ontology-based knowledge management. In: Davies, J., Fensel, D., van Harmelen, F. (eds.) Towards the Semantic Web. Ontology-driven Knowledge Management, pp. 33–46. John Wiley & Sons, LTD, Chichester (2003)
50. Tiscornia, D.: The LOIS project: Lexical ontologies for legal information sharing. In: Biagioli, C., Francesconi, E., Sartor, G. (eds.) Proceedings of the V Legislative XML Workshop, pp. 189–204. European Press Academic Publishing, Firence (2007)
51. Uschold, M., Grüninger, M.: Ontologies: Principles, methods, and applications. Knowledge Engineering Review 11(2), 93–155 (1996)
52. Valente, A.: Types and Roles of Legal Ontologies. In: Benjamins, V.R., Casanovas, P., Breuker, J., Gangemi, A. (eds.) Law and the Semantic Web. LNCS (LNAI), vol. 3369, pp. 65–76. Springer, Heidelberg (2005)
53. Venturi, G.: Legal Language and Legal Knowledge Management Applications. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.) Semantic Processing of Legal Texts. LNCS (LNAI), vol. 6036, pp. 3–26. Springer, Heidelberg (2010)
54. Walter, S., Pinkal, M.: Automatic extraction of definitions from german court decisions. In: Proceedings of the International Conference on Computational Linguistics (COLING 2006), Workshop on Information Extraction Beyond The Document, Sidney, pp. 20–28 (2006)
55. Walter, S., Pinkal, M.: Definitions in court decisions – automatic extraction and ontology acquisition. In: Breuker, J., Casanovas, P., Klein, M.C.A., Francesconi, E. (eds.) Law, Ontologies and the Semantic Web. Channelling the Legal Information Flood. Frontiers in Artificial Intelligence and Applications, vol. 188, pp. 95–113. IOS Press, Amsterdam (2009)