Artificial Intelligence and Law (2007) 15:117–135
DOI 10.1007/s10506-007-9034-4
© Springer 2007

The structuring of legal knowledge in LOIS

WIM PETERS[1], MARIA-TERESA SAGRI[2], and DANIELA TISCORNIA[2]

[1]*NLP Group, Department of Computer Science, University of Sheffield, Sheffield, UK*
E-mail: peters@dcs.shef.ac.uk
[2]*Istituto di Teoria e Tecniche per l'Informazione Giuridica del Consiglio, Nazionale delle Ricerche, Rome, 00185, Italy*
E-mail: tiscornia@ittig.cnr.it

Abstract. Legal information retrieval is in need of the provision of legal knowledge for the improvement of search strategies. For this purpose, the LOIS project is concerned with the construction of a multilingual WordNet for cross-lingual information retrieval in the legal domain. In this article, we set out how a hybrid approach, featuring lexically and legally grounded conceptual representations, can fit the cross-lingual information retrieval needs of both legal professionals and laymen.

Key words: multilingual lexicon, legal ontologies, legal information retrieval

1. Introduction

Today, search engines for legal information retrieval do not include legal knowledge into their search strategies. These strategies include keyword and metadata search, but do not address the semantics of the keywords, which would allow, for instance, conceptual query expansion. In other words, there is no semantic relationship between information needs of the user and the information content of documents apart from text pattern matching. Often, query formulation by either legal practitioners or laymen users is only an imperfect description of an information need (Matthijssen 1999).

The LOIS project (EDC 22161) aims to remedy this semantic lacuna by means of the development of a multi-language legal thesaurus, whose structure is based on existing de facto standards for semantic thesaurus construction. From the start, the project integrated a number of methodologies, in order to cope with the acquisition and combination of multilingual domain-specific terminology and existing general language repositories. Our architecture ensures the coverage of the semantic peculiarities of the legal dominion, and facilitates the capture of essential semantic differences between the legal systems involved.

This paper is structured as follows. Section 2 discusses the issue of law and translation. Sections 3 and 4 examine the relationships between general language and terminological lexicons and between lexicons and ontologies. It addresses the various roles that a common core ontology may fulfil according to the envisaged tasks of the LOIS database.

In Section 5 the methodological steps for the development of the LOIS database are clarified, and the main problems to be addressed in the integration of the multilingual legal database are outlined.

Among the various solutions, the possible role of a common ontology is discussed, and the alignment between one of the LOIS ontological modules and an existing candidate formal ontology is evaluated.

2. Law and translation

It can generally be stated that law depends on language: regulatory knowledge must be communicated, and the written and oral transmission of social or legal rules passes through verbal expression. Therefore legal conceptual knowledge is closely related to language use within the legal domain. Legal discourse can never escape its own textuality (MacDonald 1997). This means that linguistic information plays an important role in its definition, which may lead to the postulation that there is, as in other terminological domains, a relatively high level of dependence between legal concepts and their linguistic realization in the various forms of legal language (see below).

Law uses the common language, usually narrowing the connotation to technical meaning, or creating new terms. In both cases, regulatory language defines the semantic area covered by the terms. In common language, the concept of *home* may be expressed by several terms in the same language (dwelling, house, accommodation, etc.). In law (as in all terminology languages), the term *rent*, as a single term without variants, expresses a univocal concept, defined by regulations pertaining to a regulatory system in which term, concept and legal institution converge.

The law selects within social phenomena and behaviours only those that it means to regulate, for the sake of its consociates. Thus, often, law terms are open-textured, with the purpose of extensionally covering the constant evolution of social reality, such as, for instance, law and order and common morality.

Juridical language includes several levels. According to Kalinowsky (1965), the *language of law* is the language in which legal rules are written: not any linguistic expression in a legal text is a legal term, but every legal term is a linguistic expression. The *language of Jurists* is a meta-language, which they use to speak about legal rules and about persons and behaviours bounded by legal rules.

- The law-maker's language defines objects and agents of the selected reality and regulates, or better, describes the 'ideal' vision of such reality according to the law.
- The judge's language interprets the law-maker's: the regulation is the interpreted content of law expressions; to apply a regulation, the judge qualifies facts as an instance of the semantic extension of the regulatory concepts.
- The language of jurisprudence is a meta-language; it puts legal language and judicial interpretation into concepts, to make the structure of the system consistent and systematic. The judge defines the extension, the law-maker describes the intension of the law terms, in terms of necessary, if not sufficient, conditions. Usually, we differentiate, within doctrinal language, a more specific level depending on the single regulatory systems where dogmatics operate, from the level of theory of law, where universal concepts, common to all the systems – law, penalty, damage, obligation, etc. – are processed.

Given the structural domain specificity of legal language and the involved concepts, we cannot speak about 'translating the law' to ascertain correspondences between law 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. Sacco (2005) points out: ''Defining law may in addition mean to reconstruct that which comes from written law [.]. Whereas the law is the basis of the system, a first dislocation of the relationship between law and language takes place on the basis of law interpretation. The effort of judges and theoreticians multiply the number of linguistic formulas employed to speak about the legal element at issue. Therefore, the legal researcher faces different levels of legal language use: there is the written formula by the legislator (first level), and those proposed by experts and judges (second level). And eventually there is the 'legal rule' that he deduces from these attested formulas, which represents the legal truth according to him, and which he formalizes as a third level of legal speech.''

Overall, there is a lack of a clear language level where the equivalence has been set up. In ''translating law'' we have to negotiate the distance between the statute and the law or, more generally, between the law and its verbalization.

In planning the LOIS database, we formalized the distinction between translational equivalence on the one hand and legal equivalence on the other as a distinction between linguistic and terminological knowledge, thus adopting two separate notions of juridical concept, to deal with concepts pertaining to the doctrine (we call them 'lexical', since their descriptions still pertain to general language) and 'legal' concepts, defined in legislative text.

We do not take into account the judge's language, as concepts in the lexicon are intended as classes of each possible instance of world objects, expressed by a word in a given language. Therefore, the level of legal decisions, the area of interpretation, where instances of the reality are assigned to a class of meaning, was left out.

3. Ontology and lexicon

Contrary to other academic disciplines (such as biology and genetics), taxonomies are rarely inherent in law. Legal vocabularies contain open-textured terms, they are inherently dynamic, and the norms in which legal terms are used, are syntactically ambiguous. Such legal structural knowledge does not only contain interpretations of the meaning of legal terms, but also shows the (supposed) logical and conceptual structure (Dini et al. 2005). Therefore, when examining the legal vocabulary, we encounter two different types of semantic information associated with elements from legal text. On the one hand, there is ontological structuring in the form of a conceptual model of the legal domain. A legal 'language', consisting of a complex structure of concepts, forms an abstraction from legal textual material. On the other hand, there is a vocabulary of lexical items that lexicalize concepts (a lexicon), which are not necessarily restricted to the legal domain, and are associated with specific linguistic information (e.g., nouns versus verbs and syntactic preference).

Before discussing relations between lexicon and ontology, some terminological clarifications are needed, as within the Semantic Web Community, the term 'ontology' has acquired several meanings and specifications, such as *lightweight, core, domain*, and *foundational (or upper) ontology*. In the Wonderweb project[1] differences are explained as follows: "In most practical applications, ontologies appear as simple taxonomic structures of primitive or composite terms together with associated definitions. These are the so-called *lightweight* ontologies, used to represent semantic relationships among terms in order to facilitate *content-based access* to the (Web) data produced by a given community. In this case, the *intended meaning* of primitive terms is more or less known in advance by the members of such community.

On the other hand, however, the need to establishing precise agreements as to the meaning of terms becomes crucial as soon as a community of users evolves, or multicultural and multilingual communities need to exchange data and services.

To capture (or at least approximate) such subtle distinctions we need an explicit representation of the so-called *ontological commitments* about the meaning of terms, in order to remove terminological and conceptual ambiguities. A rigorous logical axiomatization seems to be unavoidable in this

case, as it accounts not only for the relationships between terms, but – most importantly – for the formal structure of the domain to be represented.

Axiomatic ontologies come in different forms and can have different levels of generality, but a special relevance is enjoyed by the so-called *foundational ontologies*, which address very general domains. ''We see the role and nature of foundational ontologies (and axiomatic ontologies in general) as complementary to that of lightweight ontologies: the latter can be built semi-automatically, e.g., by exploiting machine learning techniques; the former require more painful human labour, which can gain immense benefit from the results and methodologies of disciplines such as philosophy, linguistics, an cognitive science.'' (Masolo et al. 2003).

A formal ontology can be considered a theory about several views (i.e., models) of reality. Formal ontologies have a multi-layered structure: *foundational ontologies* contain domain-independent concepts, relations and meta-properties, which provide ontology builders with a formal semantics, that is, formal ontological distinctions to categorize entities in a domain. A *domain ontology* is populated by concepts, relations and instances extracted in a bottom-up fashion from the domain and consistent with the top-down formal semantics imposed by the upper ontology. In complex domains such as the legal one, a *core ontology* is part of a layered architecture; a core ontology intends to bridge the gaps between domain-specific concepts and the abstract categories of upper ontologies; it expresses the basic concepts that are common across a variety of domains, providing a global and extensible model into which data originating from distinct sources or different vocabularies can be mapped and integrated (Doerr et al. 2003).

Lexicons are therefore considered *lightweight ontologies*, linguistic expansions of the description of a way of perceiving reality, with limited formal modelling. Lightweight ontologies are generic and based on a weak abstraction model, since the elements (classes, properties, and individuals) of the ontology depend primarily on the acceptance of existing lexical entries. In a lexical ontology, such as WordNet (Fellbaum 1998), many of the hyper/hyponymy links are not logically consistent, as it was designed as a lexical resource, not as a formal ontology. In lexical ontologies constraints over relations and consistency are ruled by the grammatical distinctions of language.

It is possible that a lexicon with a semantic hierarchy might serve as the basis for a useful ontology, and that an ontology may serve as a grounding for a lexicon. This is particularly the case in technical domains, in which vocabulary and ontology are more closely tied than in more general domains (Hirst 2004, p. 14). In terminological lexicons, terms and concepts usually coincide, which creates an intersection between linguistic meaning and formal ontological conceptual structure.

On this basis, ontology experts consider the ontological level as a more abstract layer than the lexicon. Both types of knowledge structures can combine into a hybrid structure, where lexical and ontological knowledge are integrated. The resulting integration still pertains to the discourse level, where the passage from lexicon to formal ontology takes place in the following ways (Gangemi et al. 2003):

- Transforming lexical definition into formal description;
- Interpreting lexical relations from a thesaural structure as ontological relations;
- Checking the consistency of a hybrid knowledge base on the base of the meta-properties of ontological classes;
- Modularizing the resulting hybrid ontology into a structure that is consistent with the relations between entities defined in a core ontology.

These possible migrations and interdependencies between ontological and linguistic knowledge indicate that there is a trade-off between linguistic ontology design and abstract ontology design.

In general, the more detailed the adopted linguistic constraints on ontology design are, the more detailed and explicitly justifiable that ontology design becomes (Bateman 1992), but also there will be an increasingly strong connection between abstract concepts and their linguistic realization.

Variable dependency of ontological structure on language constraints indicates a continuum between formal structure and linguistic description. On the formal end of the continuum are maximally language independent ontologies reflecting the structure of the domain in question. On the linguistic end there is a complete ontological dependency on domain-specific lexis and grammar. This means that, in the latter case, the structure of the ontology is fully determined by the structure of the language.

A knowledge base that models legal knowledge needs to take both types of information into account, and establish a modularly organized integration of the two. This is the objective of LOIS.

4. Task-driven ontologies

Another very important determinant in ontological modelling is the task for which the ontology is created.

It has been argued that task neutral ontologies are unrealistic (Bench-Capon 2001): this is effectively true for *domain ontologies*, i.e., ontologies that cover a specialized area of knowledge, where the process of knowledge modeling is affected by the task the system is expected to perform.

Individual tasks vary heavily in their requirements for information and typical sources of information.

Legal advisory systems use a variety of legal domain ontologies as a knowledge repository for the legal qualification of cases and for detecting anomalies in legal systems. The domain ontology engineer defines the ontology scope, acquires knowledge about the domain, identifies and describes key concepts and assigns a term to them. The linguistic realization of the concepts may be ignored, while the characterization of the knowledge model is mainly influenced by reasoning patterns, such as legal assessment or planning (Valente 1995; Breuker and Hoekstra 2004; Breuker et al. 2005). Applications of such type are, for instance the Clime Project (Boer et al. 2001), the IPRONTO system (Delgado et al. 2003) and case study analysis in Dolce applications (Gangemi et al. 2003; Gangemi et al. 2004).

Core Ontologies lay at a more general ontological level where it can be claimed that its task-dependency is of less importance. A *core legal ontology is* a complete and extensible ontology that expresses the basic concepts of Law, and that can provide the basis for specialization into domain-specific concepts and vocabularies. These key basic concepts form as it were the basic conceptual vocabulary of the legal domain, and cover the ontological requirements of the majority of tasks at a general level. The task-specific extension of the core vocabulary is expected to use these core concepts as general legal classes for e.g., hypernymic relations and preference encoding for legal actions. Another way of extending a core legal ontology (CLO) is to associate lexical information with it. A project where lexical resources are integrated into core ontological categories is shown in (Despres and Szulman 2005), where terms, automatically extracted from legislative texts, are transformed into concepts and syntactic relations are interpreted as functional properties or *roles*. Alignment between micro-ontologies created from similar, partially overlapping sources is driven by means of categories grounded in the core ontology.

Similarly, in the FFPOIROT Project a 'terminographic' analysis supports the ontology building process, (Kerremans and Temmerman 2004), to add a multilingual layer to a domain ontology for the financial forensic domain. The method used by terminology lexicon builders shares many aspects with ontology engineers, if we don't consider the fact that ontologists often acquire content by means of the expert's knowledge, instead of by terms taken from multilingual texts. To connect ontological concepts, 'termino-ontographers' need an intermediate structure of the dominion, made up of units of understanding, to distinguish language-independent concepts and relations from concepts and relations which are language-dependent.

If the ontology is to be used in order to process text, then obviously a more linguistically oriented ontology will be needed that enables natural language processing. The main application area of the LOIS database is information retrieval, and therefore textual units are its building blocks. The database will be used in order to find relevant documents relevant to a query,

in the same language or in different languages. For this particular task, linguistic descriptions, in particular vocabulary and lexical information in the form of lexical semantic relations such as synonymy and hypernymy, are very important. This is a reason to opt for the (Euro)WordNet architecture (Vossen et al. 1997). This structure is also suitable for other purposes, such as comparative law and didactic tasks, and it allows a certain level of legal knowledge modelling through the concepts and the defined semantic relations.

If the ontology is to be used for reasoning and maintaining a formal level of conceptual consistency, then a more abstract and formal ontology is needed. Formal legal ontologies offer a solid support for legal information systems, because they make explicit the underlying assumptions, as well as the formal definition of the components of legal knowledge (Masolo et al. 2004).

There are a number of formal legal ontologies available that approach formal modelling from different perspectives, such as the Functional Ontology (Valente, 1995) and the Frame Based Ontology (van Kralingen 1995). Furthermore, CLO previously integrated in the Italian Legal Wordnet (Gangemi et al. 2005), organizes juridical concepts and relations on the basis of formal properties defined in DOLCE.

Although the aim of the LOIS project is primarily oriented towards information retrieval, more specifically the retrieval of relevant documents on the basis of multilingual and ontological expansion of query terms, it is envisaged that the multilingual database will form the basis of further development within the legal domain in terms of other tasks for information retrieval and extraction purposes. These will involve more refined knowledge modelling and automated reasoning. Therefore the architecture should be extensible and able to accommodate knowledge objects imported from other resources. This will enable the LOIS database to adapt to more than one possible usage scenario. For this reason, the LOIS architecture enables the modular integration of ontologies at different positions on the scale between linguistic and conceptual, and offers the possibility to organize them into one single model. The envisaged end result will be a superset of ontological and lexical structures, which will enable an incremental integration into the knowledge base of the ontological requirements of targeted application tasks. The incremental growth of the knowledge base makes it possible to observe general patterns across tasks and contexts, which will, in its turn, allow a flexible adaptation to new tasks, where increasing amounts of existing concepts are reused and the conceptual coverage of the database is extended with the necessary task- and domain-specific vocabulary.

5. The LOIS database

5.1. CHOICE OF DATABASE STRUCTURE

As its methodological starting point, LOIS adopts the structure of two widely known and used thesauri. WordNet (Fellbaum 1998) is a lexical database which has been under constant development at Princeton University. Euro-WordNet (EWN) (Vossen et al. 1997) is a multilingual lexical database with WordNets for eight European languages, which are structured along the same lines as the Princeton WordNet. Both thesauri are organized around the notion of a *synset*. A synset is a set of one or more uninflected word forms (lemmas) with the same part-of-speech that can be interchanged in a certain context. For example, {*case, cause, causal, law suit*} form a noun synset because they can be used to refer to the same concept. A synset is often further described by a gloss. Synsets can be related to each other by semantic relations, of which the most important are hypernymy/hyponymy (between specific and more general concepts), meronymy (between parts and wholes), and antonymy (between semantically opposite concepts). Cross-lingual equivalence relations are made explicit in the so-called Inter-Lingual-Index (ILI). Each synset in the monolingual WordNets has at least one equivalence relation with a record in this ILI. Language-specific synsets from different languages that are linked to the same ILI-record by means of a synonym relation are considered conceptually equivalent.

The ILI is in principle the superset of all concepts from all WordNets, and the concepts from indigenous WordNets are linked into one or more ILI records by means of equivalence relations. These relations indicate several levels of equivalence: complete equivalence, near equivalence, or equivalence as a hyponym or hypernym. The network of equivalence relations determines the interconnectivity of the indigenous WordNets.

In principle, the ILI is an unordered list of concepts, i.e., it does not have any internal structuring. The reason behind this is that we assume that each language imposes its own language-specific structural constraints on the concepts. Therefore, any ordering of ILI concepts needs to be retrieved from knowledge bases that link into the ILI. ILI concepts enter into relations with each other by means of:

- the equivalence relations between indigenous concepts and ILI concepts
- traversal through the relations within the indigenous WordNets

The LOIS database is compatible with the EWN architecture, and forms an extension of the EWN semantic coverage into the legal domain. Overall, LOIS consists of a number of modules that directly or indirectly link into EWN modules through each individual language component (see Figure 1 for a simplified view on the database structure).

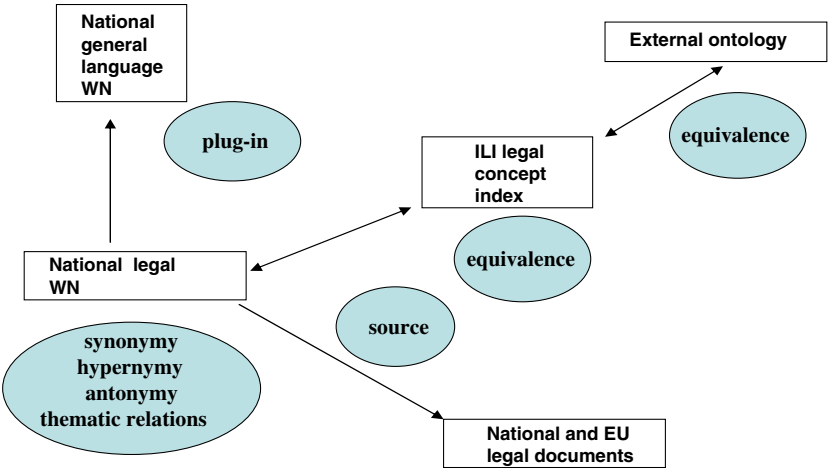

Figure 1. Modular Structure of the LOIS database.

5.2. BUILDING THE WORDNETS

From a methodological perspective, there are two basic methods for building multi-language thesauri. The first one assumes that the internal semantic relations between terms may be transferred to another language on the basis of the translational correspondence of the concepts expressed. This methodology, applied in the MultiWordNet project (Pianta et al. 2002), allows the automatic development of multi-language lexicons, through the mapping of dictionaries. This approach is less effective in the treatment of legal terminology, and can only be applied to law to a limited extent, on account of the reflections in the previous section.

In LOIS's initial phase, we needed to distinguish a nucleus of pilot concepts offering a reference structure for the building of the WordNets in the other languages. To allow greater sharing, only the general level of doctrine definitions could prove effective. This entailed the inclusion of common sense concepts that are used in doctrine, but are not confined solely to the domain of legal terminology. Thus we ''translated'' only a nucleus of common sense knowledge from the Italian legal wordnet (JWN), in order to bootstrap the localization into other languages. The Italian Concepts were manually selected from the frequency list of the Italian Legislation corpus.[2] Their selection was based on the assessment of experts. Descriptions (glosses) were extracted from legal handbooks.

The second method entails creating individual national legal WordNets in parallel, and setting up correspondences between them through an integrated bottom-up and top–down approach. This method is based upon the

automatic selection of terms from parallel corpora and other resources, on the clustering of terms in synsets, and on their intersection, in an attempt to define a series of basic concepts on which a correspondence relation may be defined (Palmer et al. 2000).

The identified core allows the integration and homogenization of local lexicons. This approach was adopted in the second step of LOIS development. In this stage the emphasis was on the detection of legal conceptual terminology, i.e., terminology that is specific to the legal domain, as opposed to the common sense concepts described above. In order to identify this legislative knowledge, a parallel corpus was created from the European Directives in the EU languages. Semi-automatic alignment techniques enabled the selection of a multilingual set of legal terms. Legal terms were only selected if they had an explicit definition in the text. This criterion sets these terms apart from the lexical terms described above.

Once alignment had been established, conceptual equivalence was assumed and each set of corresponding terms in different languages were automatically linked to one unique ILI concept. As for legal concepts from European legislation, the unique *Identifier* acts as the Interlingual Index item.

Automatic extraction of legal concepts from national legislation (limited to the consumer law domain) took place on the basis of explicit definition patterns in the legal text (e.g., "'*citizens' means all members of the public in the United Kingdom.*"). Furthermore, salient multiword units were extracted after linguistic pre-processing, and subsequently manually evaluated.

5.3. CONTENT OF LOIS

In correspondence with the two building approaches described in the previous section, the main module of each LOIS national wordnet is composed of:

- An indigenous *lexical database*, which conceptualizes general language entities pertaining to legal theory and legal dogmatic, a set of patterns (models) in line with which law is formed and operates, and which is structured according to the EWN methodology;
- A *legislative database*, populated by concepts defined in European and national legislation and structured according to purely legal (supra)national models.

The entries of the two types of legal knowledge link into the interlingual database component: the ILI. The relation *implemented_as* defines the link between a European legal concept and a national legal concept based on it. Moreover, synsets in the National Legal WN are (or shall be) linked by *plug-in* relations (denoting for instance equivalence and hypernymy, see

Magnini and Speranza 2002) to the general language modules, developed within the *EuroWordNet Project*.

The ILI forms the platform for the integration of external knowledge resources. These resources will function as meta-ordering principles of the ILI concepts. Inclusion of an increasing number of these ordering principles will allow greater complexity and refinement in knowledge representation and ontology comparison.

6. Project results and open methodological questions

The database holds 33,000 synsets (around 5000 per legal system), which originate from European Community definitions, national legislation and lexical databases. The expansion of the lexicon requires the integration of the bottom-up strategy described above with a top–down validation from the side of formal ontologies as described in Section 3, in order to expand the coverage and consolidate the structure of the overall model. The two aspects are linked, since, to meet some requirements that satisfy the completeness of the dominion, we need to refer to a shareable general *core*, where conceptual entities are related according to ontology-based properties. On the other side, we need a back-bone structure, consistent with the synsets (and semantic relations) in local lexicons to drive the classification and the equivalence setting.

For this purpose, two ontologies have been integrated to a greater or lesser extent. These are the foundational ontology DOLCE2.1-Lite-Plus and the CLO (Gangemi et al. 2003), a shareable core ontology for the legal domain.

DOLCE (*a Descriptive Ontology for Linguistic and Cognitive Engineering*) is a *foundational ontology* (FO) developed originally in the EU WonderWeb project. Foundational ontologies are domain-independent axiomatic theories, contain a rich axiomatization of their vocabulary, and are used in order to make the rationales and alternatives underlying different ontological choices as explicit as possible. In DOLCE+, basic DOLCE is extended by means of the Description and Situation(D&S) ontology, suited to conceptualize domains (such as Law) that are mainly constituted by non Physical (Mental, Social) objects. In D&S, DOLCE is taken as a ground ontology, i.e., an ontology that is used to represent the entities in a domain; the main classes in Dolce are *Endurant* (including Objects or Substances), and *Perdurant* (including Events, States, or Processes), linked by the relation of *participation, Qualities inhere in* either Endurants (as Physical or Abstract Qualities) or in Events (as Temporal Qualities).

The current version of CLO is based on the D&S distinction between (Legal) Descriptions (the main subclass being that of *norms*), and Situations

(objects *participant-in* events or processes), which encompass legal and non-legal states of affairs (*legal facts* or *cases*).

A *Description* in Dolce D&S is a social object, which represents a conceptualization. Like physical objects, social objects have a lifecycle, can have parts, etc. Different from physical objects, social objects are dependent on some agentive physical object that is able to *conceive* them. Descriptions have typical components, called *concepts*. *Concept* is also a social object, which is *defined by* a description. Once defined, a concept can be *used in* other descriptions. *Figures*, or *social individuals* (either agentive or not) are other social objects, defined by descriptions, but differently from concepts, they do not classify entities. Typical agentive figures are societies, organizations, and in general all *socially constructed persons*.

The *classified* relation relates concepts in Description and entities in Situation. There are several kinds of concepts reified in D&S, the primary ones (*role*, *course*, and *parameter*) being distinguished by the categories of entities they classify in DOLCE: *Role c*lassifies *Endurant*, *Course c*lassifies *Perdurant, Parameter* Classifies Region (of Qualities*)*.

Requisites are constraints over the attributes of entities. When a situation satisfies a description that uses parameters, endurants and perdurants that constitute the situation must have attributes that range within the boundaries stated by parameters (in DOLCE terms, entities must have qualities that are mapped to certain value ranges of regions).

Therefore, a norm is a description (the legislator's perspective), *which* defines (in constitutive norms) or uses (in prescriptive norms) *legal persons*, as well as *legal roles*, *legal courses of events*, and *legal parameters:* all these *concepts* classify entities (objects *participant-in* events or processes), which constitute a legal Situation.

The existing alignment of WordNet 1.6 Noun Synsets (to which LOIS synsets are linked by means of plugin relations) with the DOLCE2.0-Lite-Plus ontology (Gangemi et al. 2002) allows greater conceptual clarification, cognitive transparency and effective re-use. For LOIS, the integration of this alignment ensures a tight interconnection between its modules. The flexible modular architecture of LOIS acts as a bridge between different kinds of knowledge sources, and its modular integration allows a regulated and incremental growth in available legal knowledge.

One of the open questions is the management of the semantic/equivalence relations via the ILI in the integration of legislative and lexical/common sense knowledge. With respect to legal concepts from national legislation, the ILI can be automatically generated, i.e., for each legal concept a corresponding ILI equivalent is created. If a legal concept from a European directive is implemented in indigenous legislation, and the local legal concepts are deemed (legally) equivalent to their European counterparts, then an equivalence relation between the two local concepts may also be established. In all

other cases, the creation of semantic links between local synsets does not necessarily imply the creation of equivalence relations with the ILI, except in cases where concepts from more than one indigenous wordnet coincide, in which case these will all be related to one ILI record. Since the ILI will be the superset of all concepts in all legal wordnets, the semantic structures peculiar to each wordnet will be preserved within the LOIS architecture, and will overlap through the ILI. External ontologies such as the CLO will structure the ILI concepts, classifying concepts according to explicit and logically consistent subsumption relations.

Another issue that needed to be addressed in LOIS is that of polysemy. For instance, a lexical term may have more than one meaning, or a legal term may have more than one definition in legislation. Polysemy detection and resolution is in particular one of the aspects that an ontology may solve, as pointed out by Gangemi et al. (2002) and Vossen et al. (1997).

Polysemy is expressed in LOIS by the association of one synset to each sense of a polysemic word. To assign for each sense of a word in the source language the right equivalent in the target language (or to create a new synset when a sense in source or target language is missing), ontological distinctions can be necessary to make meaning commitments explicit. For instance, one of the typical ambiguities of legislation is the distinction between the regulatory and physical existence of the legal phenomena. The Italian term *contratto* is, in terms of CLO concepts, a *legal description*, an *information conten*t and a *physical object* (the material support of the information content. A legal institution, for instance the *Prime Minister*, is a figure, created by norms, but it is also a social role: in complex figures, like organizations or institutions, an endurant (a physical person) plays a *delegate*, or *representative* role of the figure.

In addition to the association of word senses of polysemic terms with ontologically well-formed concepts, a key step in the process of the methodological refinement of ontological categorization will be the consistent distinction of degrees of equivalence between *contexts*, in which the word occurs. The aim of conceptual description (either formalized or not) is to identify a core conceptual category by expressing prototypic features, relations and condition of use, to which contexts add meaning specifications. The importance of contexts is stressed in the field of computational terminology (Temmerman and Tummers 2003), who agree on the necessary anchoring of term extraction, term definition and inter-term relation identification on the contexts of use. The traditional 'standardization oriented' and 'concept centred' approach, where (ideally) only one term is assigned to a concept, has proved to fail in cross-lingual conceptualizations.

In law, legislative definitions are contexts which have a prescriptive force. This fact influences the determination of the number of senses of terms, and the equivalence setting between legal concepts and lexical concepts.

The most common situation is the case of 'apparent polysemy' generated by the integration of the legal and lexical databases, because meanings of the legal concepts are usually more specific than the meanings of corresponding lexical items, and legal senses can display degrees of ontological overlap or even taxonomic ordering. For instance, from EU Legislation texts, obtained from Celex,[3] four senses of *'worker'* are defined:

1. Any worker as defined in Article 3(a) of Directive 89/391/EEC who habitually uses display screen equipment as a significant part of his normal work.
2. Any person employed by an employer, including trainees and apprentices but excluding domestic servants;
3. Any person carrying out an occupation on board a vessel, including trainees and apprentices, but excluding port pilots and shore personnel carrying out work on board a vessel at the quayside;
4. Any person who, in the Member State concerned, is protected as an employee under national employment law and in accordance with national practice; The corresponding lexical entry is defined in the lexical part as follows:
5. A person who works at a specific occupation.

The lexical sense, taken from WN (sense no. 5), can be considered to be the most general, and therefore it is classified as a hypernym of all legal senses. For the legal concepts, a taxonomic ordering can be perceived where senses 2 and 4 are more general than the other two, and sense 4 is more general than sense 2.

A *legal equivalence* holds between legal concepts extracted from European legislation in different languages, because parallel texts (and terms) are assumed equivalent by law. By the same criterion, if a legal concept from a European directive is implemented in indigenous legislation, and the local legal concept is deemed (legally) equivalent to its European counterparts, then an equivalence relation between the two local concepts may also be established.

As for legal concepts extracted from national legislation, these reflect normative diversity between legal systems, and therefore in establishing conceptual equivalence across language systems there is no context dependence.

7. Conclusion and future work

In this paper we have described theoretical, practical and structural aspects of the LOIS multilingual legal knowledge base. This legal knowledge repository contains legal terminology from national and European legislation within the domain of consumer law. It also holds significant lexical, general language

concepts that occur in the legal documents. These concepts are interlinked within each language and between languages by means of an extended set of EWN relations.

The structure of the LOIS database allows a user to perform a concept-based search for monolingual and cross-lingual legal information retrieval, which uses keywords obtained from query expansion through the structured hierarchies of the legal wordnets and the equivalence relations with the ILI.

Furthermore, the LOIS architecture will allow users to investigate a wide range of legal research issues, such as the comparison of national legal systems through translation, equivalence and ontological structure across the different legal wordnets, the investigation of relations between EU and national legislative documents, and an empirical inventory of the differences between common language meaning and legal meaning.

The structure of the LOIS database enhances the interoperability of multilingual legal data, and allows the incremental integration of additional legal information. The role of top-level formal ontology is fundamental in this process. This ontological level not only reinforces the existing structure (polysemy detection, ILI structuring, etc.), but also assists the automatic integration of the database through ontology-building techniques. This has several advantages, among which are the following:

7.1. INCREASING LANGUAGE INDEPENDENCE

Equivalence relations in a multilingual lexicon are based on the assumption that terms in lexical resources are linguistic representations of the same concept. Equivalence links connect language-specific WordNets to the Interlingual Index, a language-independent structure, which, in principle, consists of the superset of all language-specific concepts. Each ILI concept is associated with inter-language information allowing the connection to be made. In linguistic ontologies such as EWN this role is carried out by the ILI gloss, which is a natural language concept description (i.e., an English linguistic expression in our database) expressing the minimal features, characteristics and/or conditions. This gloss allows localizers to understand the concept and to link the ILI concept to a concept in their own language. The fact that ILI concepts only have an informal characterization in English does not warrant complete language independence. Another methodological step, anchoring upper level concepts to ontological classes from formal ontologies, is a means to overcome indeterminacy or ambiguity in linguistic definitions, and assign logical properties and structure to the concepts. In formal ontologies, semantic and ontological information are truly language independent, since their properties are formalized, and formal properties and relations between concepts are inherited by sub-classes.

7.2. CONSISTENCY CHECKING

The main entities in DOLCE (and consequently in CLO) are axiomatized, disjoint classes, characterized by meta-properties, such as Identity, Unity and Rigidity. As for CLO, the most relevant distinction is between *Roles* (anti-rigid) and *Type*s, which are rigid. For example, every instance of a role (e.g., *student*, *plaintiff*, *guilty*) can possibly be a non-student, not guilty, etc. without loosing its identity. Every instance of a type (e.g., a person) must be a person. A type can play more roles at the same time. For instance, a legal subject (either a natural or artificial person) can be an owner, a tax-payer, or a murderer. In the CLO taxonomy, roles cannot subsume types, and therefore in LOIS lexical concepts that are anchored to roles should not have hyponyms pertaining to types. This constraint can detect inconsistencies in automatically created relations from LOIS ILI records to WN synsets as shown in the examples below:

Consumer is a person, *is a* living thing, *is a* physical entity in WN; *is a* social role, *is a* non physical entity in CLO; *lease is a* is a contract, *is a* communication, *is a* an abstraction in WN; *is a* contract, *is a* social description, *is a* social concept in CLO.

In conclusion, the LOIS knowledge base provides a flexible, modular architecture that allows integration of multiple classification schemes, and enables the comparison of legal systems by exploring translation, equivalence and structure across the different legal wordnets.

Notes

[1] http://wonderweb.semanticweb.org
[2] http://www.normeinrete.it/
[3] http://europa.eu.int/documents/index_en.htm;

References

Bateman, J. A. (1992). The Theoretical Status of Ontologies in Natural Language Processing. In Preuss, S. and Schmitz, B. (eds.), Text Representation and Domain Modelling ideas from linguistics and AI, Berlin.

Bench-Capon, T. J. M. (2001). Task Neutral Ontologies, Common Sense Ontologies and Legal Information Systems, Second International Workshop on Legal Ontologies; in conjunction with JURIX 2001: The 14th Annual International Conference on Legal Knowledge and Information Systems. Amsterdam: The Netherlands.

Boer, A., Hoekstra R., and Winkels R. (2001). The CLIME Ontology. Proceedings of the Second International Workshop on Legal Ontologies. Amsterdam: The Netherlands.

Breuker, J. and Hoekstra, R. (2004). Epistemology and Ontology in Core Ontologies Exemplified by Two Core Ontologies for Law FOLaw and LRI-Cor. In Coront-Wes Ekaw.

Breuker, J., Valente, A., and Winkels, R. (2005). Use and Reuse of Legal Ontologies in Knowledge Engineering and Information Management. In Benjamins, V. R., Casanovas, P., Breuker, J., and Gangemi, A. (eds.), Law and the Semantic Web. Springer Verlag: Berlin.

Delgado, J., Gallego, I., Llorente S., and García, R. (2003). IPROnto: An Ontology for Digital Rights Management. In JURIX 2003 Frontiers in Artificial Intelligence and Applications, vol. 106, IOS Press.

Despres S. and Szulman S. (2005). Merging of Legal Micro-Ontologies from European Directives. In Biasiotti et al. (eds.), *Proceedings of Loait, Workshop on Legal Ontologies and Artificial Intelligence*, Wolf Legal Publisher: The Netherlands.

Dini, L., Liebwald, D., Mommers, L., Peters, W., Schweighofer, E., and Voermans, W. (2005). Cross-lingual Legal Information Retrieval Using a WordNet architecture. In *Proceedings of ICAIL* '05, 163–167. ACM: Bologna.

Doerr, M., Hunter, J., and Lagoze, C. (2003). Towards a Core Ontology for Information Integration. Journal of Digital Information, 4(1).

Fellbaum, C. (ed.) (1998). WordNet: An Electronic Lexical Database. MIT Press: Cambridge, Mass.

Gangemi, A., Prisco, A., Sagri, M. T., Steve, G. and Tiscornia, D. (2003). Some Ontological Tools to Support Legal Regulatory Compliance, with a Case Study: In Jarrar, M. et al. (eds.), Proceedings of the Worm 2003 Workshop at OTM Conference. Springer Verlag: Berlin.

Gangemi, A., Sagri, M.-T., and Tiscornia, D. (2004). An Ontology-based Approach for Representing "Bundle of rights": In Jarrar, M. and Gangemi, A. (eds.), Proceedings of the Worm 2004 Workshop at OTM Conference. Springer Verlag: Berlin.

Gangemi, A., Sagri, M.-T., and Tiscornia, D. (2005). A Constructive Framework for Legal Ontologies. In Benjamins, V. R., Casanovas, P., Breuker, J., and Gangemi, A. (eds.), Law and the Semantic Web. Springer Verlag: Berlin.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). Sweetening Ontologies with DOLCE. In Proceedings of EKAW 2002.

Hirst, G. (2004). Ontology and the Lexicon, In Staab, S. and Studer R. (eds.), Handbook on Ontologies. Springer.

Kalinowsky, G. (1965). Introduction à la logique juridique. Paris.

Kerremans, K. and Temmerman, R. (2004). Towards Multilingual, Termonto logical Support in Ontology Engineering. In *Proceeding of Termino 2004, Workshop on Terminology*.

van Kralingen R. W. (1995). Frame-based Conceptual Models of Statute Law. The Hague et al., Kluwer Law International.

Macdonald, D. (1997). Legal Bilingualism, McGill Law Journal 1997, 42 McGill L.J. 119 pp. 50–99.

Magnini, B. and Speranza, M. (2002). Merging Global and Specialized Linguistic Ontologies. In Proceedings of the Workshop Ontolex-2002 Ontologies and Lexical Knowledge Bases, 43–48. LREC-2002.

Masolo, C., Borgo, S., Gangemi A., Guarino, N., and Oltramari, A. (2003). WonderWeb Project, Deliverable D18: Ontology Library.

Masolo, C., Vieu, L., Bottazzi, E., Catenacci, C., Ferrario, R., Gangemi, A., and Guarino, N. (2004). Social Roles and their Descriptions. In Dubois, D., Welty, D., and Williams, M. (eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference on Principle of Knowledge Representation and Reasoning (KR).

Matthijssen, L. (1999). Interfacing between Lawyers and Computers: An Architecture for Knowledge-based Interfaces to Legal Databases. The Hague et al., Kluwer Law International.

Palmer, M., Grishman, R., Calzolari, N., and Zampolli, A. (2000). Standardizing Multilingual Lexicons, Paper presented at the workshop on Web-Based Language Documentation and Description, 12–15, December 2000, Philadelphia, USA.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: Developing an Aligned Multilingual Database. In Proceedings of the First International Conference on Global WordNet, Mysore, India, January 21–25, 2002.

Sacco, R. (2005). Ordinary Language and Legal Language: In Barbara, P. (eds.), Ordinary Language and Legal Language. Giuffre: Milano.

Sowa (J.), Building, Sharing, and Merging Ontologies, http://users.bestweb.net/~sowa/ontology/ontoshar.htm.

Temmerman, R. and Tummers, J. (2003). Representing Multilingual and Culture-Specific Knowledge in a VAT RegulatoryOntology: In Jarrar, M. et al. (eds.), Proceedings oof the Worm 2003 Workshop at OTM Conference. Springer Verlag: Berlin.

Valente, A. (1995). Legal Knowledge Engineering: A Modelling Approach. IOS Press: Amsterdam.

Vossen, P., Peters W., and Díez-Orzas, P. (1997). The Multilingual design of the Euro-WordNet Database. In Mahesh K. (ed.), Ontologies and multilingual NLP, Proceedings of IJCAI-97 workshop, Nagoya, Japan, August 23–29.