The LOIS Project

Wim Peters*, Maria Teresa Sagri, Daniela Tiscornia**, Sara Castagnoli*****

*NLP Group, Department of Computer Science, University of Sheffield, U.K.
W.Peters@dcs.shef.ac.uk
**Istituto di Teoria e Tecniche per l'Informazione Giuridica del Consiglio Nazionale delle Ricerche, Italy
{Sagri | Tiscornia}@ittig.cnr.it
*** Scuola Superiore di Lingue Moderne per Interpreti e Traduttori, Università di Bologna, Italy
scastagnoli@sslmit.unibo.it

Abstract

The legal knowledge base resulting from the LOIS (Lexical Ontologies for legal Information Sharing) project consists of legal WordNets in six languages (Italian, Dutch, Portuguese, German, Czech, English). Its architecture is based on the EuroWordNet (EWN) framework (Vossen et al, 1997). Using the EWN framework assures compatibility of the LOIS WordNets with EWN, allowing them to function as an extension of EWN for the legal domain. For each legal system, the document-derived legal concepts are integrated into a taxonomy, which links into existing formal ontologies. These give the legal wordnets a first formal backbone, which can, in future, be further extended.

The database consists of 33,000 synsets, and is aimed to be used in information retrieval, where it provides mono- and multi-lingual access to European legal databases for legal experts as well as for laymen. The LOIS knowledge base also 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.

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)[1] 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.

2. Law and Language

Law and language are connected in many ways. First of all, they have a similar structure: each has, at his essence, rules which are constitutive of a system and which ensure its consistency. A second aspect is the dependency of law on language, since 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 .

The legal language, like law, has a multi-layered structure: according to Kalinowsky (1965), it consists of the *language of law* and *language of Jurists.* The former 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 latter is a meta-language. It is composed of the “judge's language”, which they use to speak about legal rules and about persons and behaviours bounded by legal rules; and the “language of jurisprudence”, which puts legal language and judicial interpretation into concepts, to make the structure of the system consistent and systematic[2].

Similarly to legal language, law has a hierarchical and multilevel architecture, where *primary* or *regulative norm*s (Hart, 1961) are enacted according to *secondary meta-norms* which create all the apparatus (entities, powers, procedures) necessary to produce the law. A specific kind of a two-level regulative situation is the interrelation between European and National law-making, where Member States are the addressees of European rules, committed to implement them by creating new rules in national systems.

Given the structural domain specificity of legal language and the involved concepts, 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,

[1] See http://www.loisproject.org/

[2] A good example is “negozio giuridico” (juridical act): the term never appears in Italian Legislation, but is crucial in contract law to distinguish contracts from other classes of legal acts.

nor the requirements of the different legal systems (see section 3.3). 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.

A legal "language", consisting of a complex structure of concepts, forms an abstraction from legal textual material. 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; 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). In building the Lois database, we have therefore taken in account the unavoidable intertwining between the linguistic and the content dimensions, and have distinguished *legal* and l*exical* concepts by adopting a strict notion of "legal" concepts as concepts explicitly defined in legislation.

The Lois domain ontology is populated by concepts, relations and instances extracted in a bottom-up fashion from the legal documents. This domain ontology can be classified as a lexicon, also called *lightweight ontology*. 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 order to connect linguistic expressions of concepts to the underlying conceptual domain entities, an intermediate structure (a *core ontology)* is needed, made up of units of understanding, to distinguish language-independent concepts and relations from concepts and relations which are not. 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. A *core legal ontology* such as *CLO (*Gangemi et al., 2003) and LRI-Core (Breuker et al., 2005) intends to bridge the gap between domain-specific concepts and the abstract categories of formal upper level or foundational ontologies such as DOLCE (Gangemi et al., 2002), transforming lexical relations into formal properties consistent with the top-down formal semantics imposed by the upper ontology. Foundational ontologies contain domain-independent concepts, relations and meta-properties, which provide ontology builders with a formal semantic framework, i.e. high-level formal ontological distinctions to categorize entities in a domain. The elements from the domain ontology are consistent with the top-down formal semantics imposed by the upper ontology.

Given a superset of entities and relations in six wordnets, each pertaining to a particular legal system, this logical backbone will help to distinguish language-independent concepts and relations from concepts and relations which are not. Figure 1 illustrates the interconnection of the different types of ontology.

At this stage of the project, conceptual relations derived from legal text have only partly been formalized; the ontological level has been introduced mainly to support conceptual consistency of the knowledge base. Subsumption of concepts into ontological classes makes sense distinctions explicit, e.g. "contract" as a document and "contract" as a legal transaction,, where the latter two are concepts from the core legal ontology; (Gangemi et al., 2003). They also separate classes from instances. For example, "competent authority" in the EU Directive on data protection is a class; the "*garante per la protezione dei dati personali*" in Italian legislation and the "*Agencia de Proteccion de Dato*" in Spanish legislation are instances.

It is envisaged that the formalization of LOIS will be ex-tended by transforming additional lexical relations into formal relations. This will take 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.

From a multilingual point of view, ontological classes might be even used at later stages to enhance comparison between different legal systems, by grouping similar national instantiations of a given class (e.g. the Italian "Camera dei Deputati" and the English "House of Commons" as instances of the ontological class "legal institutions") which might not – due to the *country-dependence* of the legal domain – be perfect equivalents (see section 3.3).

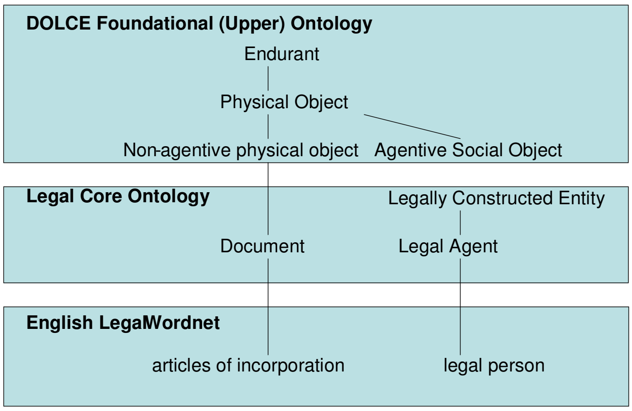

Figure 1: The interconnection of different ontologies

3. The LOIS Database Architecture

3.1. 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. EuroWordNet (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, causa, 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.

The LOIS database is compatible with the *EuroWordNet* 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 2 for a simplified view on the database structure).

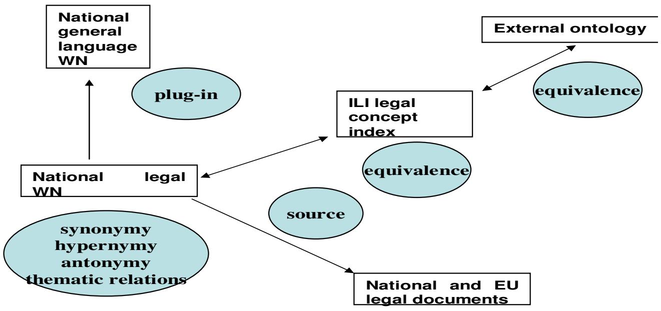

Figure 2: Modular structure of the LOIS database

Currently, the LOIS database covers six legislative systems coinciding with six languages. In line with the discussion of legal language above, each LOIS national legal wordnet is composed of two types of database modules:

- an indigenous *lexical database*, which conceptualizes general language entities pertaining to legal theory and legal dogmatics, 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 legal 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 inter-lingual index (ILI). Moreover, synsets in the English legal wordnet are linked by *plug-in* relations (Magnini and Speranza 2001), such as synonymy and hypernymy, to Princeton WordNet concepts. These WordNet concepts can be linked up to EuroWordNet ILI concepts, which will, in their turn, enable access to the other available EuroWordNet language modules. Overall, LOIS will eventually consist of a number of modules that directly or indirectly link into EWN modules through each individual language component. For the languages participating in LOIS, for which there is no EuroWordNet general language module available, this objective is beyond the scope of LOIS.

3.2. Language Internal Relations

Within each national legal wordnet, synsets are related to each other by means of 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); even if less used, LOIS also includes all EWN relations. Taxonomic relations can span across the different modules (esp. lexical and national legal) which form a LOIS wordnet: legal concepts (i.e. concepts defined within legal texts) can have hypernyms - as well as near-synonyms - in the lexical database, as legal terms bear specialised meanings which might be different from the meaning of the same words within general language. Moreover, synsets in the National Legal WN are (or shall be) linked by *plug-in* relations (Magnini and Speranza, 2001) to the general language modules, developed within the *EuroWordNet Project*.

The interrelation between EU and national concepts represents a special case of intra-lingual inter-modular relation. Even though the comparison between EU and national concepts is carried out at a mono-lingual level, it nonetheless involves a confrontation of two distinct legal systems. This is confirmed by the fact that, while being obliged to implement EU provisions in their respective national legislations, member states can still choose how such integration should be done, i.e. by “importing” concepts altogether or by adapting EU provisions to existing national (conceptual and therefore linguistic) situations. In fact, the implementation of a Directive may not correspond to its straight transposition in national law, since the same concept can be defined either in a different (more specific or more generic) way, or by a different term. The terminological transposition reflects the legal process, where several national legal orders have been converging into a new European order, which does not substitute the national traditions, but with which national orders have to interplay.

These considerations led us to adopt two kinds of relations between EU and national legal concepts: the former, expressed by “Implemented_as”, is a purely legal relation that indicates the link between European legal concepts and the nation-specific concepts which are (even if partially) based on them; conversely, the relation “Implemented_from” defines the link between national legal concepts and the EU concepts they implement. In the following section we will see how such relations - besides contributing to structure monolingual wordnets - can also enhance cross-lingual comparison.

The other relation type is expressed in terms of equivalence relations in order to measure the semantic distance between the original EU concept and the nationally implemented ones. Even if equivalence is traditionally used to link different linguistic systems, we consider it justifiable to extend the same relation to link concepts pertaining to distinct legal systems such as European and national legislations.

3.3. Cross-lingual Equivalence Relations

Cross-lingual equivalence relations – or, more precisely, equivalence relations across the legal systems under examination – are made explicit in the so-called Inter-Lingual-Index (ILI), which is, in its most explicit form, the superset of all concepts from all wordnets. Each synset in the indigenous wordnets has therefore at least one equivalence relation with a record in the ILI. 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/legal system imposes its own 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
- links with existing external ontologies (see section 5).

Synsets from different national legal wordnets are linked to the same ILI-record when they are conceptually equivalent or present some degree of similarity. Providing for relations other than complete equivalence is indispensable when dealing with concepts from the legal domain: legal concepts are the result of social and cultural traditions varying across countries, and therefore they are deeply rooted in the national legal systems which originated them (Mayr & Sandrini, 1999). Since the same word form might have different meanings within systems sharing the same language, e.g. the German and Austrian systems, a strict concept-oriented approach – rather than a linguistic one – should be adopted.

It can easily be inferred from the considerations above that full conceptual equivalence is very rare in the legal domain, especially when – as in the case of LOIS – some of the systems under examination belong to different legal traditions (civil-law vs common-law) and present therefore profoundly different cultural and knowledge backgrounds. Within LOIS, instances of absolute equivalent concepts (linked to the same ILI by means of a synonym relation) can nonetheless be found across lexical databases (which are not strictly law-dependent) and in relation to concepts which were originated by a common law-source, such as international or EU law.

As mentioned in 3.2, European directives provide measures that should be implemented in each national legislation, thus introducing equivalent or similar concepts in different legal systems. EU legislation constitutes therefore a source of legal (conceptual and functional) equivalence, thus enhancing cross-lingual information retrieval (Mommers & Voermans, 2005): even where European concepts are not implemented as such in different national legislations, similar implementations of a same EU concept can be retrieved through the "Implemented_as" relation in the different wordnets, thus favouring comparison between legal systems. Most often, however, concepts belonging to different legal systems differ from each other, even for some minor facets (legal effects, competences, duties, election procedures, and so on): near-equivalence relations indeed belong to the most frequent within LOIS, (see Table 1), since they give the compiler the possibility to associate concepts without pretending they are full equivalents, which would be misleading for the database user.

Other kinds of cross-lingual relations include equivalence as a hyponym or hypernym. The network of equivalence relations determines the interconnectivity of the indigenous wordnets. The hybrid approach of introducing both lexical and legal terms helps retrieving cross-lingual equivalents, since specific legal concepts lacking perfect equivalents in the target language might have near-equivalents or hypernyms in the other language's lexical database.

As was mentioned in section 2, the integration of ontological classes into LOIS might favour legal comparison by providing a common denominator for "comparable" concepts, i.e. concepts such as institutions or legal acts which share the same functions or other characteristics. Examples of such situational or "functional equivalents" (Pigeon, 1982) might include pairs such as "Camera dei Deputati (IT) – House of Commons (UK)" (see 2), or "Presidente della Repubblica (IT) – Sovereign (EN)" (as instantiations of the ontological class "Head of state"). Such relations being founded on comparative law, they require more manual analysis, and will be possible added at later stages.

Relation	Number of Relations
eq_near_synonym	26504
eq_synonym	4479
eq_has_hyperonym/ has_hyponym	27

Table 1: The most frequent equivalence relations in LOIS

4. Database Population

Currently, the LOIS database contains approximately 33,000 synsets, belonging to either the legal EU and national legal database modules, or the lexical module. Various methodologies have been applied to populate and structure the wordnets, of which the following were the most important:

- manual expert translation of a selected bootstrapping set of existing synsets in the Italian legal wordnet (JurWN);
- manual creation of legal synsets on the basis of authoritative resources;
- automatic extraction of explicitly defined concepts from legislative text (national and EU);
- automatic extraction of significant lexical elements from legal text;
- mapping lexical concepts onto WordNet and adopting its hierarchies;
- mapping ILI concepts to external ontologies in order to ensure a language independent ontological backbone.

The ILI forms the platform for the integration of external knowledge resources. These resources function as meta-ordering principles of the ILI concepts. At present, concepts from the Legal Core Ontology have been linked to a number of ILI concepts. This will provide a sharable, general core where concepts are formed and structured according to formal requirements.

It is to be expected that the inclusion of an increasing number of these ordering principles will allow greater complexity and refinement in knowledge representation and ontology comparison.

5. Conclusions and Prospects

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.

Further research will focus on improved techniques for information retrieval, such as further formalization of legal content through legal definition analysis and extending the links to existing formal ontologies.

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 legal 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.

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.

6. References

Breuker, J.A., Valente, A. & Winkels, R. (2005). Use and reuse of legal ontologies in knowledge engineering and information management. In Benjamins e.a., editors, *Law and the Semantic Web*, pp. 36-64. Springer Verlag, Berlin. Volume 3396,

Dini, L., Liebwald, D., Mommers, L., Peters, W., Schweighofer, E. And Voermans, W. (2005). Cross-lingual Legal Information Retrieval using a WordNet Arhitecture. In: *Proceedings of ICAIL2005*, pp.163-167. ACM, Bologna.

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

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

Gangemi, A, M.-T. Sagri, D. Tiscornia (2003). Jur-Wordnet, a Source of Metadata for Content Description in Legal Information. In: *Proceedings of the Workshop on 'Legal Ontologies & Web based legal information management', part of The International Conference of Artificial Intelligence and Law (ICAIL 2003)*, Edinburgh, June 24, 2003.

Hart.L.A. (1961). *The Concept of Law*, Oxford (UK): Clarendon Press.

Kalinowsky, G. (1965). *Introduction à la logique juridique*. Bibliotheque de Philosophie du Droit, Paris.

Macdonald, D. (1997). Legal bilingualism. *McGill Law Journal* 1997,42, pp. 50-99, McGill L.J. 119

Magnini, B & Speranza, M. (2001). Integrating Generic and Specialized WordNets. In: *Proceedings of the Euroconference RANLP 2001*, Tzigov Chark, Bulgaria.

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

Mayr, E. & Sandrini, P. (1999). Coming to Terms with Legal Information, in *TKE'99 Terminology and Knowledge Engineering. Proceedings of the 5th Int. Congress*, Innsbruck 23 - 27 August 1999. Vienna: TermNet. 455-463.

Mommers, L. & Voermans, W.J.M. (2005). Using Legal Definitions to Increase the Accessibility of Legal Documents. In M.-F. Moens and P. Spyns (eds), *Legal Knowledge and Information Systems. Jurix 2005: The Eighteenth Annual Conference*, 147-156.

Pigeon, L.P. (1982). "La traduction juridique – L'équivalence fonctionnelle". In J.C. Gémar, *Langage du droit et traduction – Essais de jurilinguistique.*

Québec: Linguatech, Conseil de la langue française. pp. 271-282.

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