

A linguistic-ontological support for multilingual legislative drafting: the DALOS Project

Enrico Francesconi, Pierluigi Spinosa, Daniela Tiscornia

Institute of Legal Information Theory and Techniques, Italian National Research Council (ITTIG-CNR), Italy

{francesconi,spinosa,tiscornia}@ittig.cnr.it

Abstract. Coherence and alignment of the legislative language highly contribute to the quality of legislative processes, to the clarity of legislative texts and to their accessibility. DALOS aims at ensuring that legal drafters and decision-makers have control over the multilingual language of European legislation, and over the linguistic and conceptual issues involved in its transposition at national levels. The project will contribute to this goal by providing law-makers with linguistic and knowledge management tools to support the legislative drafting activity.

Keywords: Legislative drafting, multilingualism, domain ontology, lexical taxonomy

1. Introduction

Coherence, interoperability and harmonization in the legislative knowledge of, and control over, the legal lexicon is a precondition for improving the quality of legislative language and for facilitating access to legislation by legal experts and citizens. In a multilingual environment, and in particular, in EU regulations, only the awareness of the subtleties of legal lexicon, in the different languages, can enable drafters to maintain coherence among the different linguistic version of the same text. This is as much important for the EU Member State legal orders, strongly influenced by the obligation to implement EU directives.

To face this problem recently the DALOS¹ project has been 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, with the aim to foster the quality of the legislative production, to enhance accessibility and alignment of legislation at European level, as well as to promote awareness and democratic participation of citizens to the legislative process.

In particular DALOS aims at ensuring that legal drafters and decision-makers have control over the legal language at national and European level, by providing law-makers with linguistic and knowledge manage-

¹ DrAfting Legislation with Ontology-based Support

ment tools to be used in the legislative processes, in particular within the phase of legislative drafting.

Nowadays the key approach for dealing with lexical complexity is the ontological one, by which we mean a characterisation (understood both by people and processed by machines) of the conceptual meaning of the lexical units and of their connection with other terms. On the basis of an ontological characterisation of legal language DALOS wants to provide law-makers with linguistic and knowledge management tools to support legislative drafting in a multilingual environment.

In this paper an overview of the DALOS project is given. In particular in Section 2 the complexity of the multilingual legal scenario is addressed; in Section 3 the characteristic of the DALOS linguistic-ontological approach is discussed; in Section 4 the specification of the DALOS Knowledge Organization System (KOS) is presented; in Section 5 the methodologies to implement the DALOS ontological-linguistic resource are shown; finally in Section 6 some conclusions are reported.

2. Interfacing multilingual legal terminologies

In legal language every term collection belonging to a language system, and any vocabulary originated by a law system, is an autonomous vocabulary resource and should be mapped through relationships of equivalence with the others. Based on the assumption that in a legal domain one cannot transfer the conceptual structure from one legal system to another, it is obvious that the best approach consists in developing parallel alignment with the same methodology and the same conceptual model. Different methods may be applied, depending on the characteristic of the domain, the data structure and on the result to achieve.

As regards the data structure, the first consideration is that unstructured list of terms (as for instance traditional flat terminologies) cannot be mapped in a consistent way, but only connected by a one-to-one correspondence among terms, which is an invalid approach for a context dependent technical terminology, such as law vocabulary. Among structured data different degrees of formalization can be distinguished:

- controlled vocabularies (such as thesauri, classification trees, directories, key-words lists): terms are organized in taxonomic trees, linked by generic associative relations, and concepts are implicitly expressed by lists of preferred and variant terms (descriptors/non-descriptors);

- semantic lexicons, also called computational lexicons or lightweight ontologies are based on commonly accepted semantic definitions and on a limited formal modeling;
- foundational, core, and domain ontologies are formal models (logical theories) of a conceptualization of a given domain, often based on axiomatic definitions.

The integration of lexical resources (heterogeneous because belonging to different law systems, or expressed in different languages, or pertaining to different domains) leads to different final results depending on the desired results:

- generate a single resources covering both (merging);
- compare and define correspondences and differences (mapping);
- combine different levels of knowledge representation, basically interfacing lexical resources and ontologies.

Of the three strategies, the methodological approach for DALOS requires the definition of mapping procedures among semantic lexicons, driven by the reference to an ontological level where the basic entities which populate the legal domain are described. In the next section the semantic structure of the lexical component is outlined.

2.1. A LEGAL SEMANTIC LEXICON: THE LOIS DATABASE

Semantic lexicons are a means for content management which can provide a rich semantic repository. Compared to formal ontologies, semantic lexicons are lightweight ontologies as they are based on a weak abstraction model, with limited formal modeling, since constraints over relations are based on the grammatical distinctions of language (noun, verbs, adjectives, adverbs), for instance the agent-role relation holds between a noun (agent) and a verb or event denoting nouns (action) ((Castagnoli et al., 2006)) In the legal field, one of the wider semantic lexicons currently available is the LOIS database² composed by about 35.000 concepts in five European languages (English, German, Portuguese, Czech, and Italian, linked by English).

In LOIS a concept is expressed by a synset, the atomic unit of the semantic net. A synset is a set of one or more uninflected word forms (lemmas) with the same part-of-speech (noun, verb, adjective,

² created within the European project LOIS (Legal Ontologies for Knowledge Sharing, EDC 22161, 2003-2006)

and adverb) that can be interchanged in a certain context. For example *action, trial, proceedings, law suit* form a noun synset because they can be used to refer to the same concept. More precisely each synset is a set of wordsenses, since polysemous terms are distinct in different wordsenses. A synset is often further described by a gloss, explaining the meaning of the concept. English glosses drive cross-lingual linking.

In monolingual lexicons terms are linked by lexical relations: synonymy (included in the notion of synset), near-synonym, antonym, derivation. Synsets are linked by semantic relations of which the most important are hypernymy/hyponymy (between specific and more general concepts), meronymy (between parts or wholes), thematic roles, instance-of.

Cross-lingual linking is based on equivalence relations of each synsets with an English synset: these relations indicate complete equivalence, near equivalence, or equivalence as a hyponym or hyperonym. The network of equivalence relations, the Inter-Lingual-Index (ILI), determines the interconnectivity of the indigenous wordnets. Language-specific synsets from different languages linked to the same ILI-record by means of a synonym relation are considered conceptually equivalent. The LOIS approach are not completely language-independent, since the equivalence setting passes throughout the English wordnet and the English translation of glosses support the localization process.

The lesson learned from the LOIS experience is that a limited language independence could be enough for cross-lingual retrieval tasks, but that it could be a weak point when considering re-using, extending, updating the semantic connections or when integrating external lexical resources (for instance multilingual thesauri) within the framework. What is needed is “the distinction between conceptual modeling at a language-independent level and a language and culture specific analysis and description of discourse-related units of understanding” (Kerremans and Temmerman, 2004).

These considerations led us to make clear distinction, when designing the overall model of DALOS and the system architecture, among:

- types of knowledge
- layers of knowledge representation
- classes of semantic relationships between knowledge elements.

3. Which knowledge for the DALOS service?

DALOS aims at providing a knowledge resource on the basis of the LOIS experience.

The two projects however address two different scenarios: while the LOIS knowledge resource is addressed to multilingual legal information retrieval, the DALOS knowledge resource is expected to support legislative drafting.

This distinction of the addressed scenario is particularly important because it contributes to identify the type of knowledge to be described within the DALOS service, so to avoid the so called *epistemological promiscuity* addressed by Breuker and Hoekstra (Breuker and Hoekstra, 2004), namely the common attitude to “indiscriminately mixing epistemological knowledge and domain knowledge in ontologies” which prevents knowledge representations from being automatically reusable outside the specific context for which the knowledge representation was originally developed.

As underlined by (Boer et al., 2004) the “*norm* is an epistemological concept identified by its role in a type of reasoning and not something that exclusively belongs to the vocabulary of the legal domain”. As argued, “knowledge about reasoning – *epistemology* – and knowledge about the problem domain – *domain ontology* – are to be separated if the knowledge representation is to be reusable” (Boer et al., 2004).

The DALOS case addresses the legislative drafting process, namely a process that creates norms on specific domains to be regulated. What is needed therefore is a knowledge and linguistic support giving a description of concepts, as well as their lexical manifestations in different languages, in specific domains *before* they are regulated.

In particular, for the DALOS knowledge resource, avoiding *epistemological promiscuity* means to avoid that the knowledge to be used as support for legislative drafting (*domain knowledge*) is mixed with the knowledge on the general process of drafting (*epistemological knowledge*) which, obviously, pertains to different domains (see also (Biagioli and Francesconi, 2005)).

According to previous works (Biagioli, 1997) the epistemological knowledge related to the legislative drafting process can be modelled by the *Model of Provisions* which establishes a taxonomy of provision types (rules as *definition, obligation, prohibition, sanction*) and amendments (*insertion, repeal, substitution*) which describe legislative texts irrespective to the domain addressed, and pertain to the process of legislative drafting. Such kind of knowledge therefore will not be described by the DALOS resource, which, on the contrary, will contain knowledge on a

domain of interest. In particular for the aim of developing a project pilot, the “consumer protection” domain has been chosen.

4. KOS of the linguistic-ontological resource

In this phase of the project the most part of the activities are addressed to provide the specification for the DALOS resource. Chosen the domain of interest (“consumer protection”) currently the activities for domain knowledge specification are oriented to:

- the standards to be used for knowledge representation;
- the Knowledge Organization System (KOS).

As regards the standards, the RDF/OWL standard conversion of WordNet approved by the W3C standards will be used for the linguistic resource (), thus guaranteeing interoperability as well as scalability of the solution.

As regards KOS, on the basis of the arguments expressed in Section 2.1, the DALOS resource is expected to be organized in two layers of abstraction (see Fig. 1):

- the *ontological layer* containing the conceptual modeling at a language-independent level;
- the *lexical layer* containing the lexical manifestations in different languages of the concepts at the ontological layer.

Basically the ontological layer acts as a knowledge layer where to align concepts at European level independently from the language and the legal order, according to the EU Commission recommendations for Member State legislations. 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 a “pivot” meta-language in a N-language environment, allowing the reduction of the number of bilingual mapping relationships from a factor N^2 to a factor $2N$. Concepts at the ontological layer are linked by taxonomical (*is_a*) as well as object property relationships.

On the contrary the lexical layer aims at describing language-dependent lexical manifestations of the concepts of the ontological layer. At this level terms will be linked by linguistic relationships as those ones used for the LOIS database (*hyperonymy*, *hyponymy*, *meronymy*, etc.). In particular, to implement the lexical layer, the subset of the LOIS

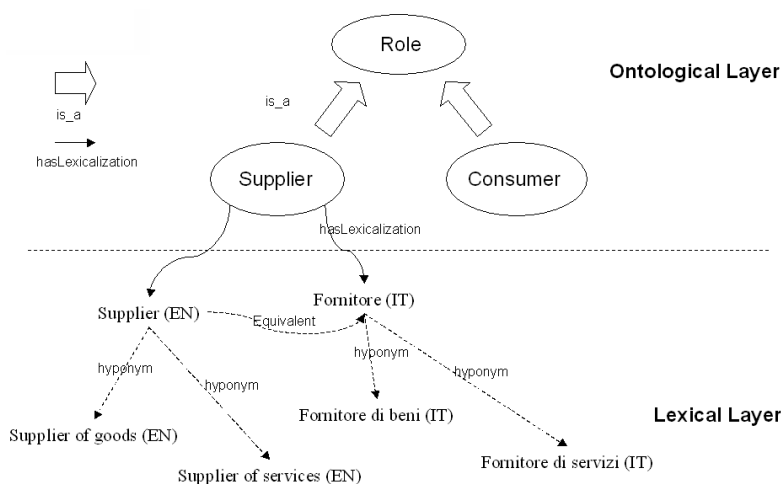


Figure 1. Knowledge Organization System (KOS) of the DALOS resource.

database pertaining to the “consumer protection” lexicon will be used. Moreover this database will be upgraded by using further texts where to extract pertaining terms from.

The connection between these two layers is aimed at representing the relationship between concepts and their lexical manifestations:

- within a single-language context (different lexical variations (lemmas) of the same meaning (concept));
- in a cross-language context (multilingual variations of the same concept).

In the DALOS KOS such link is represented by the `hasLexicalization` (and its inverse `hasConceptualization`) relationship.

5. Implementation of the DALOS resource

In order to implement the DALOS linguistic-ontological resource three main activities are foreseen:

1. Extracting terms of the domain of “consumer protection” law from a set of chosen texts by using NLP tools; this activity is aimed at upgrading the LOIS database (Lexical layer);
2. Construction of a Domain Ontology on the “consumer protection” domain (Ontological layer);

3. Semi-automatic connection between the LOIS database selection and the Domain Ontology by the `hasLexicalization` property implementation and its inverse `hasConceptualization` [Lexical layer \leftrightarrow Ontological layer]). This activity will be supported by automatic tools and validated by humans.

The first activity (implementation of the Lexical Layer) will be carried out using different NLP tools specifically addressed to process Italian texts (T2K) as well as English and other EU language texts (GATE).

T2K³ is a terminology extractor and ontology learning tool jointly developed by CNR-ILC⁴ and University of Pisa which combines linguistic and statistical techniques. It performs the following tasks: a) acquisition of domain terminology, both simple and multi-word terms, from a document collection; b) organisation and structuring of the set of acquired terms into taxonomical chains and clusters of semantically related terms. It works on Italian document collections; in principle it could be applied to document collections in languages other than Italian provided that NLP resources and tools for those languages exist (i.e. taggers, chunkers, dependency parsers).

GATE⁵ is a tool to support advanced language analysis, data visualisation, and information sharing in many languages, owned/provided and maintained by the Department of Computer Science of the University of Sheffield.

The second activity (construction of a Domain Ontology) will be an intellectual one which aims at describing the scenario to be regulated. In this context the use of an ontology is of primary importance. Laws in fact usually contain provisions (Biagioli, 1997) which deal with entities (arguments) but they do not provide any general information on them: for example the Italian privacy law regulates the behaviour of the entity “Data controller” who is the owner of a set of personal data, but such law does not give any additional information on this role in the real domain-life (Biagioli and Francesconi, 2005). Therefore a formalized description in terms of an ontology of the domain to be regulated will allow the possibility to obtain such additional general information on the entities a new act will deal with. Moreover, the use of an ontology, and particularly of the associated lexicon, allows to obtain a normalized form of the terms with which entities are expressed, enhancing the quality and the accessibility of legislative texts.

³ Text-to-Knowledge

⁴ Institute of Computational Linguistic of the Italian National Research Council

⁵ General Architecture for Text Engineering

The third activity will deal with the connection between the two level of abstractions (the *ontological layer* and the *lexical layer*). This activity is expected to be particularly time consuming, since it will implement the legal concept alignment on the basis of their lexical manifestations in a multilingual environment. A tool to support such semi-automatic mapping is expected to be implemented within the project.

6. Conclusions

In this paper an overview of the DALOS project has been presented. The main purpose of the project is to provide law-makers with linguistic and knowledge management tools to be used in the legislative processes, in particular within the phase of legislative drafting. The aim is to keep control over the legal language, especially in a multilingual environment, as the EU legislation one, enhancing the quality of the legislative production, the accessibility and alignment of legislation at European level, as well as to promote awareness and democratic participation of citizens. The ontological approach designed for the project has been presented.

References

- Breuker J. and R. Hoekstra, *Epistemology and ontology in core ontologies: FOLaw and LRICore, two core ontologies for law*. In Proceedings of EKAW Workshop on Core ontologies. CEUR, 2004.
- Boer A., T. van Engers, and R. Winkels, *Using Ontologies for Comparing and Harmonizing Legislation*, In Proceedings of the International Conference on Artificial Intelligence and Law, Edinburgh (UK), 2003. ACM Press.
- Boer A., T. van Engers, and R. Winkels, *Mixing Legal and Non-legal Norms*, In Moens, M.-F. and Spyns, P., editors, *Jurix 2005: The Eighteenth Annual Conference.*, Legal Knowledge and Information Systems, pages 25–36, Amsterdam. IOS Press.
- Biagioli C. and E. Francesconi, *A Visual Framework for Planning a New Bill*, In Quaderni CNIPA (Proceedings of the 3rd Workshop on Legislative XML), n. 18, p.83-95, 2005.
- Biagioli C., *Towards a legal rules functional micro-ontology*, Proceedings of workshop LEGONT '97.
- Castagnoli S., W. Peters, M. T. Sagri, D. Tiscornia, *The LOIS Project*, in Proceedings of the LREC 2006 Conference, Genova, May 2006.
- Kerremans K. and Temmerman R., *Towards Multilingual, Termonological Support in Ontology Engineering*, in Proceeding of Termino 2004, Workshop on Terminology, (2004).