A two-level knowledge approach to support multilingual legislative drafting

T. AGNOLONI, L. BACCI, E. FRANCESCONI, W. PETERS, S. MONTEMAGNI, G. VENTURI

Abstract. The quality of legislative drafting process at European and national levels is highly influenced by the legal drafters control over the multilingual complexity of European legislation and over the linguistic and conceptual issues involved in its transposition into national laws. The DALOS project aims at ensuring coherence and alignment in the legislative language, providing law-makers with a knowledge (ontological-linguistic) resource and knowledge management tools to support the multilingual legislative drafting process. This paper outlines the activities within DALOS, aiming at the definition of the characteristics of the knowledge resource, at its implementation, at its integration in a legislative drafting environment for the project prototype.

Keywords. Multilingual legal drafting, Legal ontologies, NLP techniques

1. Introduction

Quality in European and national legislation is one of the main purposes of the current initiatives of the European Commission. In the Mandelkern report on Better Regulation [1] the need for a coordinated action by Member States was solicited to simplify the EU regulatory environment, to enhance the quality of EU legislation as well as to rationalise the transposition of Community legislation into national law. The Mandelkern Group on Better Regulation [1] in particular stressed on “quality of regulation” as an essential precondition to enhance the “credibility of the governance process” and to contribute to “welfare of citizens, business and other stakeholders”. The analysis of [1] identifies some significant problems in the “lack of simplicity, clarity and accessibility of European provisions – such as unclear, confusing terminology, incomplete or inconsistent regulations or use of vague terms”. Such problems, together with different legislative cultures within the Commission, are causes of accessibility difficulties of EU regulation. Problems of conceptual misalignment arise in particular when transposing Euro-

1Corresponding Author: E. Francesconi, via de’ Barucci 20, Florence, Italy; E-mail: francesconi@ittig.cnr.it
ITTIG-CNR group involved in DALOS includes also Daniela Tiscornia and Pierluigi Spinosa.
pean directives into national laws: they are not only caused by the content and the quality of European legislation, but they are also the result of national practices, national legal differences and different cultures.

Coherence, interoperability and harmonization in the legislative knowledge of, and control over, the legal lexicon is therefore considered 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 different linguistic versions of the same text, as well as their transposition in national laws.

Better regulation guidelines therefore have been given [1], which are targeted to guarantee more coherence in legal texts to enhance understandability and uniform interpretation of EU law in each national legal system [2]. In particular they aim to encourage:

- Coherence within EU law (internal coherence [3]), avoiding inconsistent definitions of EU legal terms as well as avoiding contradictory use of legal terms within different sectorial legislative interventions [2] (see, in EC Directives on consumer law, terms like Timeshare, Distant Contract, Unfair [3]);
- Coherence outside the EU law (external coherence [3]), avoiding that a same legal concept can be expressed in different ways in a Directive and in its transposition into a national law [2].

To face these problems the DALOS 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. The aim of such initiative is 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 to ensure 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 management 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 [4], by which it is possible to characterize the conceptual meaning of lexical units, as well as to provide a detailed description of the semantic properties of the related concepts as well as their relationships.

In this paper the development of a two level (ontological and linguistic) knowledge resource for DALOS is described. In particular, in Section 2 the complexity of the multilingual legal scenario is addressed; in Sections 3 the design principle of the DALOS knowledge resource and the characteristics of its Knowledge Organization System (KOS) are presented; in Section 4 the phases for implementing the DALOS resource are introduced; in Section 5 and 6 the implementation of the DALOS knowledge resource organized in two levels (lexical and ontological)
is illustrated; in Section 7 a description of the prototypical integration of the DALOS resource in a legislative drafting environment is given; finally in Section 8 some conclusions are reported.

2. Approaching the multilingual EU legal scenario for legal drafting

In legal language any vocabulary originated by a law system is an autonomous linguistic resource. To approach the complexity of the multilingual scenario as well as to align concepts of the EU legal domain one cannot transfer the conceptual structure of a legal system to another, because of different national legal contexts and legislative cultures within EU Member States, as discussed in [5]. One of the widely accepted approach consists in developing parallel alignment between different terms in different languages representing the same concept.

A similar problem arises as regards the obligation of EU Member States to implement European Directives into national laws. Far from being a straight transposition, this process usually includes a further step in which European Directives are subject to interpretation which can lead to questionable results (see [5] for interesting examples).

In the literature different methods exist for approaching the multilingual complexity of European law, for example controlled vocabularies, implemented in terminology database (such as IATE run by all the main EU Institutions), thesauri (as EUROVOC, maintained by OPOCE3), semantic lexicons or lightweight ontologies (as WordNet [6], EuroWordNet [7] and, in the legal domain, JurWordNet [8]). Alignment of multilingual terminologies can be effectively obtained by using a pivot language. More expressive descriptions of concepts associated with lexical units can be contained in domain ontologies (or statute specific ontologies), representing concepts used in a specific statute (as IPROonto [9]). More general organizations of domain concepts are addressed in literature as core ontologies (as LRI-Core [10], LKIF [11] and CLO [12] for the legal domain), while foundational concepts categories applicable to all domains are usually addressed in top or foundational ontologies (as SUMO [13], DOLCE [14]).

The integration of different lexical resources (heterogeneous because belonging to different law systems, or expressed in different languages, or pertaining to different domains) can be carried out in different fashions:

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

The methodological approach chosen in DALOS is the third one: it requires the definition of mapping procedures between semantic lexicons, driven by the reference to an ontological layer where the basic entities which populate a legal domain are described.

This approach has been chosen because:

3 Office for Official Publications of the European Communities
• it provides a higher degree of modularity for the knowledge architecture: lexical units of a generic lexicon can be associated to concepts pertaining to various domains, thus allowing to reuse the same lexical resource in different domains;
• while lexical units are characterized by lexical properties (hyponymy, hyperonymy, fuzzynymy, etc.), an ontological layer provides more detailed description of domain-specific concepts (i.e. **Buyer**, **Seller**, etc.) and semantic relations between concepts (i.e. **has_object_role**, **has_agent_role**, etc.).

A previous experience to approach the multilingual EU scenario for information retrieval tasks is the LOIS project whose aim was the creation of a lexical resource\textsuperscript{4} for the legal field: it is composed by about 35,000 concepts in six European languages (Dutch, English, German, Portuguese, Czech, and Italian, linked by English) [15].

The LOIS database contains terms extracted from EU Directives on the consumer protection domain; it uses the semantics of WordNet and EuroWordNet, which is centered around the notion of “synset”. A Synset is a set of one or more uninflected word forms (lemmas) with a synonymous meaning: for example **action**, **trial**, **proceedings**, **law suit** form a noun synset because they can be used to refer to the same concept. Because each synset denotes a particular meaning, polysemous words (with more than one meaning) occur in more than one synset. A synset is often further described by a gloss, explaining the meaning of the concept.

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

Cross-lingual linking is based on equivalence relations between synsets from the individual language wordnets and English synsets, which function as the pivot. These relations denote a.o. complete equivalence, near equivalence, or equivalence as a hyponym or hypernym. The network of equivalence relations in LOIS through the pivot Inter-Lingual-Index [7], determines the interconnectivity of the indigenous wordnets.

The structure of the LOIS database can be sufficient for cross-lingual retrieval tasks. When using this resource for different purposes, as for example for legislative drafting, more detailed views of concepts and their domain-specific relations might be needed at language or jurisdiction-independent levels.

Legal drafting of European law in particular can be effectively supported by a jurisdiction-independent representation of legal concepts and their relationships in order to obtain a view of the actors involved in a specific domain to be regulated, the roles they play in the same domain, the relationships with other actors, which can support the activity of regulating a typical situation, for example a **transaction on the Internet** between a **consumer** which wants to **buy a good** or a **service** from a specific type of **seller**. This information is not included in a semantic

\textsuperscript{4}created within the European project LOIS (Legal Ontologies for Knowledge Sharing, EDC 22161, 2003-2006)
multi-lingual lexicon, which, on the other hand, contains language dependent lexical manifestations of such concepts as well as lexical relations. What is needed therefore is a “distinction between conceptual modeling at a language-independent level and a language and culture specific analysis and description of discourse-related units of understanding” [16].

3. Design of the DALOS resource

As discussed, DALOS is targeted to providing a knowledge resource for legislative drafting. The identification of the scenario of use is particularly important because it contributes to identify the characteristics of the knowledge to be described and to avoid the common tendency to indiscriminately mix domain knowledge and knowledge on the process for which it is used (drafting, reasoning, searching, etc.). Such a mixing prevents knowledge representations from being automatically reusable outside the specific context for which the knowledge representation was originally developed [17].

The peculiarity of the “legislative drafting” activity can be identified in its function of norms creation on specific domains. In this context the use of an ontology is of primary importance. Laws in fact usually contain provisions [18], which deal with entities, expressed by lexical units, 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 and the relationships with other entities within the considered domain [19]. A formalized description of entities using an ontology of the domain to be regulated allows to obtain such additional information.

For the DALOS resource we want to avoid that the knowledge to be used as support for the activity of norms creation on a specific matter is mixed with the knowledge on the general process of drafting which, obviously, is matter independent (see also [19]). Therefore what is needed as a support for legislative drafting is a knowledge and linguistic resource giving a description of a domain to be regulated (domain knowledge), namely concepts of such domain, as well as their lexical manifestations in different languages.

For the aim of developing a project pilot, the “consumer protection” domain has been chosen, for its particular interest as a regulatory environment, which influences the competitiveness of EU Member States businesses and their ability to grow and create jobs.

In the first phase of the project we addressed the provision of the specification for the DALOS resource, in particular the definition of its Knowledge Organization System (KOS).

As discussed in Section 2, the DALOS resource is 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 lexical manifestations in different languages of concepts at the Ontological layer.
Basically, the Ontological layer acts as a layer that aligns concepts at the European level, independently from the language and the legal order, where possible. Moreover the Ontological layer allows to reduce the computational complexity of the problem of multilingual term mapping (N-to-N mapping). Concepts at the Ontological layer act as a “pivot” meta-language in 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 as well as object property relationships (has_object_role, has_agent_role, has_value, etc.).

On the other hand, the Lexical layer aims at describing language-dependent lexical manifestations of the concepts of the Ontological layer. At this level lexical units are linked by pure linguistic relationships (hypernymy, hyponymy, meronymy, etc.).

The connection between these two layers is represented by the relationship between concepts and their lexical manifestations:

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

In the DALOS KOS this link is represented by the hasLexicalization relationship.

The use of a two-level knowledge architecture provides a higher degree of modularity to the knowledge resource:

1. the Lexical layer is a lexical database which can be upgraded and reused in different domains with respect to the one considered in DALOS;
2. the Ontological layer provides a more detailed semantic description of the lexical units at the Lexical layer, as well as relationships between concepts.
Entries and relationships at both levels are described by exploiting the expressiveness of RDF/OWL semantic Web standards. The combination of these two levels of knowledge allows synsets at the Lexical layer to obtain two different kinds of properties:

- **linguistic properties**, namely EuroWordNet lexical relations that come directly from a linguistic and statistical analysis of texts;
- **semantic properties**, coming from the classification of synsets into classes of the Ontological layer, describing the consumer protection domain for the pilot case; such semantic properties are valid only within this particular domain.

For example, a statistical analysis of the texts on consumer protection law revealed that consumer and supplier are related lexical units, therefore, in our model, at the Lexical layer they have been linked by a *fuzzynym* WordNet property. At the Ontological layer such relation can be more semantically characterized as regards the consumer protection domain: such lexical units are considered as lexicalizations of the classes Consumer and Supplier respectively, while the relations between them pass through the concept Commercial_transaction which links Consumer and Supplier by a has_agent_role property (see Fig. 1).

As discussed in Section 2, for European law drafting purposes the use of the Ontological layer can provide law-makers with a detailed view of a specific scenario, providing support in identifying actors, roles and relationships involved in a particular situation to be regulated. On the other hand, for national law drafting implementing EU directives, a view on language-dependent lexical manifestations of concepts at the Ontological layer is necessary to guide legal drafters in choosing the most appropriate term to express a specific concept (see for example the term right of withdraw which in Italian legislation has been translated into two ways: *recesso* and *risoluzione*; in these cases the DALOS service may support the drafter in choosing the right translation, providing information on the contexts in which the different translations are used as well as on the related frequency of use).

4. Phases of the DALOS resource implementation

The DALOS ontological-linguistic resource is implemented by three main activities:

1. Semi-automatic term extraction on the domain of consumer protection law from a set of selected texts using NLP tools (Lexical layer implementation);
2. Manual ontology construction on the “consumer protection” domain (Ontological layer implementation);
3. Connection between Ontological layer and Lexical layer by the hasLexicalization property implementation.

The first activity (implementation of the Lexical layer) is carried out using different NLP tools: in particular GATE5, owned/provided and maintained by

---

5General Architecture for Text Engineering
the Department of Computer Science of the University of Sheffield, specifically used in this task to process English and Dutch texts, as well as T2K\textsuperscript{6}, jointly developed by CNR-ILC and University of Pisa, tailored to process Italian texts.

The second activity (implementation of the Ontological layer) is an intellectual one, which aims at describing a scenario to be regulated.

The third activity (connection between the two knowledge levels) is implemented by an intellectual activity targeted to provide links between concepts at the Ontological layer and synsets at the Lexical layer by the \texttt{hasLexicalization} property.

5. Lexical layer implementation

Hand-crafted lexical and ontological resources usually need to be continuously extended and refined in order to incorporate up-to-date knowledge. Term acquisition from legal texts, based on NLP techniques, can play an important role in this process of implementing lexical resource and suggesting concepts for the Ontological layer on the basis of significant terminological entries. Terms acquisition from legal texts, based on NLP techniques, can assist this process (see, for instance, \cite{20}, \cite{21}, \cite{22} \cite{23}) by suggesting lexical relations. In the DALOS project we concentrate on the semi-automatic implementation of a lexical resource by means of the acquisition of terminological knowledge from texts belonging to the “consumer protection” in Italian and English for the pilot case (extensions to other languages, in particular Dutch and Spanish, are currently under development).

DALOS Lexical layer is built up by using NLP tools on a data set composed by 16 directives (9 original versions; 7 consolidated texts) and 42 judgements (33 from the Court of Justice, 9 from the Court of First Instance)\textsuperscript{7} on the domain of interest.

5.1. Term Extraction from Italian Texts

To implement the Italian version of the DALOS Lexical layer we used T2K (Text-to-Knowledge), a hybrid ontology learning system combining linguistic technologies and statistical techniques \cite{24}.

T2K does its job in two basic steps:

1. extraction of domain terminology, both single and multi-word terms, from a document base;
2. organization and structuring of the set of acquired terms into proto-conceptual structures, namely
   - fragments of taxonomical chains, and
   - clusters of semantically related terms.

As far as term acquisition is concerned, a pipeline of computational tools for linguistic analysis (AnIta \cite{25}) is used to extract candidate terms from texts. These linguistic processing modules are in charge of:

\textsuperscript{6}Text-to-Knowledge
\textsuperscript{7}see the DALOS corpora at http://www.dalosproject.eu, “Documentation” section.
1. tokenisation of the input text;
2. morphological analysis (including lemmatisation);
3. syntactic parsing, articulated into two different steps:
   (a) chunking (including morpho-syntactic disambiguation);
   (b) dependency analysis.

Candidate terms may be one word terms (“single terms”) or multi–word terms (“complex terms”). The acquisition strategy differs in the two cases.

Single terms are identified on the basis of frequency counts in the shallow parsed texts, after discounting stop–words. The acquisition of multi–word terms, on the other hand, follows a two–stage strategy. First, the syntactically chunked text is searched for on the basis of a set of chunk patterns. Chunk patterns encode syntactic templates of candidate complex terms: for instance, adjectival modification (e.g. *organizzazione internazionale* ‘international organisation’), prepositional modification (e.g. *commercializzazione di autovetture* ‘marketing of cars’), including more complex cases where different modification types are compounded (e.g. *commercio di prodotti fitosanitari* ‘trade of fitosanitary products’). Secondly, the list of acquired potential complex terms is ranked according to their log–likelihood ratio [26].

In T2K recognition of longer terms is carried out by iterating the extraction process: acquired complex terms are projected back onto the original text and the acquisition procedure is iterated on the newly annotated text. The method proves helpful in reducing the number of false positives consisting of more than two chunks [27]. The iterative process of term acquisition yields a list of candidate single terms ranked by decreasing frequencies, and a list of candidate complex terms ranked by decreasing scores of association strength. The selection of a final set of terms to be included in the TermBank requires some threshold tuning, depending on the size of the document collection and the typology and reliability of expected results.

In what follows we exemplify what we discussed so far with preliminary results obtained from the DALOS consumer law corpus, including Directives, Regulations and case law on ‘protection of consumers’ economic and legal interests, for a total of 284,795 word tokens. With the best configuration we obtained a TermBank of 2,074 terms, namely 174 single and 1,900 multi–word terms, which is currently being evaluated by domain experts.

Table 1 contains a fragment of the automatically acquired TermBank. For each selected term, the TermBank reports its prototypical form (in the column headed “Term”), its frequency of occurrence in the whole document collection, and the lemma of the lexical head of the chunk covering the term (see column “Lemmatised headwords”). The choice of representing a domain term through its prototypical form rather than the lemma (as typically done in ordinary dictionaries) follows from the assumption that a bootstrapped glossary should reflect the actual usage of terms in texts. In fact, domain-specific meanings are often associated with a particular morphological form of a given term (e.g. the plural form). This is well exemplified in Table 1 where the acquired terms headed by *diritto* ‘right’ can be parted into two groups according to their prototypical form: either singular (e.g. *diritto di revoca* ‘termination right’) or plural (e.g. *diritti*
It should be noted, however, that reported frequencies are not limited to the prototypical form, but refer to all occurrences of the abstract term.

5.2. Italian term organization and structuring

In the second extraction step, proto-conceptual structures involving acquired terms are identified. We envisage two levels of conceptual organization. Terms in the TermBank are first organized into fragments of head-sharing taxonomical chains, whereby *commercio dei medicinali* ‘trade of medicines’ and *commercio elettronico* ‘electronic trade’ are classified as co-hyponyms of the general single term *commercio* ‘trade’. In this way, single and multi-word terms are structured in vertical relationships providing fragments of taxonomical chains such as the one reported below:

applicazione
applicazione dei paragrafi
applicazione dell’articolo
applicazione della direttiva
applicazione della legge
applicazione della tariffa
applicazione delle disposizioni
applicazione delle sanzioni
applicazione delle sanzioni amministrative
applicazione delle sanzioni previste
applicazione del presente decreto
applicazione del regolamento

where the acquired direct and indirect hyponyms of the term *applicazione* ‘enforcement’ are reported. In this example, it can be noticed that terms sharing the head only are the direct hyponyms of the root term. Further hyponymy levels
can be detected when two or more multi-word terms share not only the head but also modifiers, as in the case of the *applicazione delle sanzioni amministrative* ‘enforcement of administrative sanctions’ with respect to the more general term *applicazione delle sanzioni* ‘enforcement of sanctions’. The number of extracted hyponimic relations from the DALOS corpus is 911 referring to 172 hypernym terms.

The second structuring step performed by T2K consists in the identification of clusters of semantically related terms which is carried out on the basis of distributionally-based similarity measures [28]. For each term (both single and complex) in the TermBank, we extracted a set of 1,071 semantically related terms referring to 238 terminological headwords.

In what follows, clusters of semantically related terms are exemplified:

*disposizioni* ‘provision’
- norme, disposizioni relative, decisione, atto, prescrizioni
- legge ‘law’
- regolamento, protocollo, accordo, statuto, amministrazioni comunali

*pubblicità* ingannevole ‘misleading advertisement’
- pratiche commerciali, procedimento, pubblicità comparativa, clausole abusive, pubblicità

*cmv* (comitato per i medicinali veterinari) ‘committee for veterinary medicines’
- comitato, cpmp (Committee for Proprietary Medicinal Products), commissione, membri, consiglio

It should be appreciated that in these clusters of semantically related words different classificatory dimensions are inevitably collapsed; they include not only quasi-synonyms (as in the case of *disposizioni* ‘provision’ and *norme* ‘regulations’), hypernyms and hyponyms (e.g. *comitato* ‘committee’ and *cmv* (comitato *per i medicinali veterinari*) ‘committee for veterinary medicines’), but also looser word associations. As an example of the latter we mention the relation holding between *legge* ‘law’ and *amministrazione comunale* ‘municipal administration’, or between *comitato* ‘committee’ and *membri* ‘members’.

### 5.3. Term Extraction from English Texts

To implement the English version of the DALOS Lexical layer two term extraction applications are used. The overall focus of the extraction effort was on nominal term candidates. The set will be further expanded in the future with verbal term candidates.

The first application, TermExtractor[^8] [29] offers a comprehensive package of algorithms for the selection of relevant terms from any text corpus. It extracts a list of “syntactically plausible” term candidates (e.g. compounds, adjective-nouns, etc.), and determines the termhood on the basis of two entropy-based measures: Domain Relevance and Domain Consensus, which are used to select only the terms which are relevant to the domain of interest and consensually referred throughout the corpus documents. Domain Relevance is computed with reference to a set of

[^8]: [http://lcl2.di.uniroma1.it/termextractor/](http://lcl2.di.uniroma1.it/termextractor/)
contrastive terminologies from different domains. It filters term candidates using Lexical Cohesion, which measures the degree of association of all the words in a terminological string. Domain Consensus computes the overall significance of the term candidates across documents. Furthermore, it allows various other settings, e.g. lemmatization and exclusion of nested terms. Tab. 2 below illustrates the result from the TermExtractor tool. The weight is the overall score (between 0 and 1) computed on the basis of a weighted mean of the scores for domain relevance, domain consensus and lexical cohesion.

<table>
<thead>
<tr>
<th>term</th>
<th>weight</th>
<th>domain relevance</th>
<th>domain consensus</th>
<th>lexical cohesion</th>
</tr>
</thead>
<tbody>
<tr>
<td>member state</td>
<td>0.945</td>
<td>1.000</td>
<td>0.863</td>
<td>1.000</td>
</tr>
<tr>
<td>free movement</td>
<td>0.754</td>
<td>1.000</td>
<td>0.797</td>
<td>0.330</td>
</tr>
<tr>
<td>official journal</td>
<td>0.747</td>
<td>1.000</td>
<td>0.722</td>
<td>0.545</td>
</tr>
<tr>
<td>protection of consumers</td>
<td>0.731</td>
<td>1.000</td>
<td>0.783</td>
<td>0.139</td>
</tr>
<tr>
<td>financial service</td>
<td>0.729</td>
<td>1.000</td>
<td>0.729</td>
<td>0.288</td>
</tr>
<tr>
<td>internal market</td>
<td>0.728</td>
<td>1.000</td>
<td>0.673</td>
<td>0.477</td>
</tr>
<tr>
<td>council directive</td>
<td>0.722</td>
<td>1.000</td>
<td>0.705</td>
<td>0.346</td>
</tr>
<tr>
<td>national legislation</td>
<td>0.713</td>
<td>1.000</td>
<td>0.762</td>
<td>0.060</td>
</tr>
<tr>
<td>natural person</td>
<td>0.712</td>
<td>1.000</td>
<td>0.673</td>
<td>0.412</td>
</tr>
<tr>
<td>community level</td>
<td>0.711</td>
<td>1.000</td>
<td>0.740</td>
<td>0.116</td>
</tr>
<tr>
<td>national law</td>
<td>0.707</td>
<td>1.000</td>
<td>0.672</td>
<td>0.282</td>
</tr>
<tr>
<td>personal data</td>
<td>0.704</td>
<td>1.000</td>
<td>0.677</td>
<td>0.314</td>
</tr>
<tr>
<td>interest of consumers</td>
<td>0.698</td>
<td>1.000</td>
<td>0.712</td>
<td>0.111</td>
</tr>
</tbody>
</table>

Table 2. TermExtractor Results

The second application, TermRaider, has been developed by the University of Sheffield in GATE\(^9\)\[^{30}\]. GATE is an architecture, a framework and a development environment for Language Engineering (LE) applications, which support efficient and robust text processing.

GATE uses NLP based techniques to assist the knowledge acquisition process for ontological domain modelling, applying automated linguistic analyses to create ontological knowledge from textual resources or to assist ontology engineers and domain experts by means of semi-automatic techniques \[^{31}\] \[^{32}\].

For term extraction from the DALOS corpus, the following actions are performed within GATE:

- Tokenization and sentence splitting divide up the text into manageable units
- Part of speech tagging
- Lemmatization yields the citation form for each word form
- A multi word unit grammar defines the sequences of part of speech tags that constitute noun phrases
- The computation of term frequency/inverted document frequency (TF/IDF) \[^{31}\], a technique widely used in information retrieval and text mining taking into account term frequency and the number of documents in the collection, computes the saliency of term candidates for each document

\[^{9}\]http://www.gate.ac.uk
• All term candidates with a TF/IDF score higher than an empirically determined threshold are selected.

The part of speech sequences defined by the grammar are exemplified in Tab. 3. The threshold for the tf-idf score was experimentally set to 5:

<table>
<thead>
<tr>
<th>Phrase</th>
<th>term</th>
<th>tf/idf score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single nouns</td>
<td>consumer</td>
<td>10.3</td>
</tr>
<tr>
<td>Multiple noun combinations:</td>
<td>credit agreement</td>
<td>7.5</td>
</tr>
<tr>
<td>Noun-preposition-noun</td>
<td>distance sales</td>
<td>10.2</td>
</tr>
<tr>
<td></td>
<td>package travel contract</td>
<td>11.2</td>
</tr>
<tr>
<td></td>
<td>code of conduct</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>protection against victimization</td>
<td>11.2</td>
</tr>
<tr>
<td></td>
<td>seller of goods</td>
<td>10.1</td>
</tr>
</tbody>
</table>

Table 3. Part of speech sequences

The combination of TermExtractor and GATE enables us to obtain more reliable term candidates (where both agree, termhood is very likely), and GATE allows us to extend our techniques to other languages such as Dutch, which forms part of DALOS. For further processing such as determining relations between terms, GATE gives us much more in terms of NLP potential for later analysis of the terms in their textual environment.

The overall result of the English term extraction is a set combining the TermExtractor and tf/idf results. This set contains 2664 terms, of which 833 are multi word units, and 1831 are single nouns.

5.4. English term organization and structuring

The conceptual structuring of newly acquired terms exploits various linguistic properties of the candidate terminology. First, the analysis of the internal structure of term candidates enables the creation of links between term candidates themselves, or between term candidates and existing terminology from the LOIS database. For nominal phrases, modifiers are separated from their heads, and a hyponymic relation is automatically established if both the full phrase and the head are part of the term candidate set or the LOIS terminological vocabulary. For example, “circulation of information” has “circulation” as hypernym, and “arrangement for payment” is a subconcept of “arrangement”. Further, identity relations between modifiers of multi-word units and term candidates indicate thematic relations between the two terms, as in the case of “reference” and “reference period” on the one hand, and “mutual recognition” and “mutual recognition procedure” on the other. In addition, lexico-syntactic patterns (so-called Hearst patterns, e.g. “such as”, “including”, “and other” [32]) are indicative of semantic relations. When applied to the legal texts, they define hyponymic relations between candidates. For instance, the definition: “‘total cost of the credit to the consumer’ means all the costs, including interest and other charges, which the consumer has to pay for the credit;” supports the creation of a hyponymic link between the term candidates “interest” and “charge”. Our intention is to extend these techniques in the near future with the extraction of verbal relations between term candidates and statistical collocational information.
6. Ontological layer implementation

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

As discussed in Section 4 the Ontological layer is the result of an intellectual activity aimed at describing the domain of the consumer protection, chosen for the pilot case. An intellectual approach has been chosen to strictly reach the project objectives with a predictable degree of reliability in describing semantically qualified relations between concepts.

Classes and properties have been implemented on the basis of the terminological knowledge extracted from the chosen Directives on the consumer protection law (see Section 5), in particular from the “definitions” contained, maintaining coherence to the design patterns of the Core Legal Ontology (CLO)\(^\text{10}\) developed on top of DOLCE foundational ontology [33] and on the “Descriptions and Situations” (DnS) ontology [34] [35] within the DOLCE+ library\(^\text{11}\). The DALOS ontology copes with the entities of the chosen domain and their legal specificities. In this knowledge architecture the role of a core legal ontology is to provide entities/concepts which belong to the general theory of law, bridging the gap between domain-specific concepts and the abstract categories of formal upper level or foundational ontologies such as, in our case, DOLCE.

As regards domain-specific concepts, the DALOS Ontological layer is designed to stress the distinction identified by the “Descriptions and Situations” ontology, extended by CLO within the legal domain, between intensional specifications like norms, contracts, roles, and their extensional realizations in the same domain as cases, contract executions, agents. This distinction is correlated by the so called Norm ↔ Case design pattern \([36]\) (CODeP\(^\text{12}\)). According to the Norm ↔ Case CODeP, intensional specifications like norms use tasks, roles, and parameters, while extensional realizations like legal cases conform to norms when actions, objects and values are classified by tasks, roles, and parameters respectively. The matching is typically performed when checking if each entity in a legal fact is compliant to a concept in a legal description \([36]\).

The distinction stressed by DALOS is strictly linked to the activity of legislative drafting addressed by the project. Apart from more technical provisions like ‘amendments’ on existing norms, legislative drafting can in fact be considered as an activity that creates norms on generic situation descriptions, qualifying them by, for example, deontic terms \([11]\). Speaking according to CLO, this activity deals with description (intensional specifications) of generic situations (also called “situational frameworks” in \([11]\)), giving them a normative perspective. For

\(^{10}\)http://www.loa-cnr.it/ontologies/CLO/CoreLegal.owl

\(^{11}\)DOLCE+ library, http://dolce.semanticweb.org

\(^{12}\)Conceptual Ontology Design Pattern
example the Directive 97/7/EC of 20 May 1997, at Art. 7 paragraph 1 states that “Unless the parties have agreed otherwise, the supplier must execute the order within a maximum of 30 days from the day following that on which the consumer forwarded his order to the supplier”: it states that, unless differently agreed, the generic situation in which the supplier executes an order to the consumer, following a consumer request, is obliged, and this obligation has to be satisfied within a maximum of 30 days from the consumer request.

A normative perspective of generic situations is the result of the legislative drafting activity; it results in legislative text paragraphs grouped in articles, which can be semantically qualified as provisions [18], namely fragments of a regulation (for example an obligation for a role towards a task).

A support to legislative drafting can therefore include: 1) a taxonomy of provision types able to give a normative perspective to generic situations; 2) a knowledge resource supporting the description of generic situations in a specific domain, as well as giving an ontological perspective to entities involved in such situations [37]. The DALOS Ontological layer aims at representing this second kind of knowledge resource, tailored for the consumer protection domain pilot case.

![Figure 2. Excerpt of the DALOS Ontological Layer.](image)

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

Such domain-specific concepts are classified according to more general notions, imported from CLO, such as Legal_role and Legal_situation. Examples of some concepts obtained by the definitions from the consumer law domain are Commercial_transaction, Consumer, Supplier, Good, Price. The specific roles they play ([35]) are illustrated in Fig. 2.
On the other hand, the main entities derived from CLO are axiomatized, disjoint classes, characterized by meta properties, such as Identity, Unity and Rigidity. The most relevant distinction is between Roles (anti-rigid) and Types (which are rigid). Roles, according to [35], are anti-rigid since they are “properties that are contingent (non-essential) for all their instances”. Types on the other hand can play more roles at the same time. For instance, a legal subject (either a natural or artificial person) can be a seller and a buyer. Domain-specific requirements are expressed by restrictions over ontological classes, for instance by defining Consumer as a role that can be played by Natural_person only.

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

7. The application prototype

An application prototype has been developed within the project in order to show how the DALOS resource can be accessed and exploited to provide multilingual lexical and semantic support in legislative documents drafting concerning the “consumer protection” domain.

xmLegesEditor is an open source legislative drafting environment developed at ITTIG-CNR [38] for supporting the adoption of legal national standards (XML and URN NIR\textsuperscript{13} standards). Briefly, xmLegesEditor is a visual XML editor able to support legislative drafters in the production of standard compliant normative documents, providing advanced features for structural and semantic markup as well as user-friendly tools for constructing persistent hyperlinked normative references [39] [40]. xmLegesEditor is one of the three regulation-drafting environments being evaluated in the SEAL project (Smart Environment for Assisting the drafting and debating of Legislation)\textsuperscript{14}. See [41] [42] for a comparison with the other environments.

The DALOS extension of xmLegesEditor provides integrated access from the drafting environment to the knowledge resource produced in DALOS.

As discussed in Section 6, the DALOS resource is provided in the W3C standard format for semantic resources RDF/OWL. This allows easy and flexible integration in an application through specifically suited software libraries. Jena \textsuperscript{43} is the most popular framework for the development of semantic Web applications written in Java. It provides, among many other features, high level methods for accessing and manipulating RDF/OWL resources. Its use inside xmLegesEditor, also written in Java, allows flexible access to the DALOS resource. It is possible for example to upgrade the Lexical layer with lexical units in different languages, in a dynamic and transparent way.

7.1. Accessing the DALOS resource within the application prototype

The DALOS version of xmLegesEditor provides additional panels for accessing the DALOS knowledge resource. Different kinds of access have been considered

\textsuperscript{13}NormeInRete, http://www.normeinrete.it

\textsuperscript{14}SEAL is a project in the e-Participation initiative of the European Commission. See http://www.eu-participation.eu/seal
The Ontology Browser panel presents to the user the Ontology hierarchy in a tree view allowing the browsing through the ontological classes down to the classified terms. This provides the drafter with a smart overview of the domain ontology and the related term classification.

Lexical units grouped in synsets according to the strategy described in previous sections, is also accessible from a plain list Terms panel, from which the user can obtain all the domain terms in the chosen language, or express a textual query over the lexical forms of the extracted synsets specifying simple query parameters. Defined terms (in the domain corpus) are highlighted in the list for their particular relevance in preventing ambiguity and favoring harmonization (see [44]).

Details and Sources panels present to the user general lexical information on the selected synset. In particular term definitions (if present), as its textual fragment from the domain corpus, and a list of different variants (in a WordNet sense) are shown in the Details panel. The Sources panel shows the list of hyperlinks to external document fragments in the domain corpus where the selected term has been defined and where it has been used, as well as its variants (Fig. 4). This provides the drafter with easy access to the contexts of use and definitions of each term in the existing legislation, obtaining a valuable feedback on its pertinence in a current context.

Multilingual support is managed by offering to the user the possibility to choose the main language, i.e. the language in which the lexical units are shown within the editor. It is possible to change the main language by reloading the
whole knowledge resource inside the editor. In the Details panels it is then possible
to look at the corresponding definition, relations and sources of the selected synset
in one of the other supported languages provided by the term alignments.

In the **Lexical Properties** panel (Fig. 4), the linguistic relations established
in the Lexical layer, based on the EuroWordNet semantic model, are dynami-
cally accessed from the OWL resources and presented to the user. In this panel
WordNet relations like “hyponymy” and “fuzzynymy” (as “Related Term”) for the
selected synset, with respect to other synsets in the selected language, are shown.

Similarly, the **Semantic Relations** panel shows semantic relations of the se-
lected synset with other synsets inferred from their classification in the domain
ontology as described in Section 3. These are dynamically inferred by accessing
the OWL resources through the reasoning engine provided by the Jena library
and presented to the user in a hierarchical view.

![Figure 4. Additional DALOS panels](image)

### 7.2. Using the DALOS resource in the application prototype

The integration of xmLegesEditor with the DALOS knowledge resource is com-
pleted by providing functions for the interaction with the editing panels. Once a
synset is selected, it is possible to choose one of its lexical variants as a normalized
term to be inserted in the text being drafted at the cursor position, or to markup
the selected word with a reference to the chosen synset.

The inserted term will be highlighted in the text as being taken from the
DALOS resource as well as marked up in the underlying XML document with
a reference to the DALOS resource it belongs to. This means that the text in-
cludes information on term origin. For example, by opening a document within
xmLegesEditor and clicking on the highlighted terms, all the previously described
panels are populated with several types of information about the selected term
or concept. This provides a self explaining view of the document. It is to be
noticed that the DALOS term markup, though implemented in the prototype
application for Italian standard (NormeInRete standards [39]) compliant documents, is not constrained to any particular format and can be applied to different XML standards, using either xmLegesEditor for DTD or XMLSchema compliant XML documents, or different XML editors after having integrated the DALOS linguistic-ontological resource.

7.3. Using the DALOS resource to improve the quality of legislative texts

The integration of the DALOS resource into xmLegesEditor provides users with facilities which aims at enhancing the quality of legislative documents in different scenarios, which can be grouped into three main working situations:

1. new legislative documents drafting;
2. existing documents checking;
3. transposition of European directives into national laws.

The first situation deals with drafting new provisions; in this case the user has the possibility to browse the Ontological layer, identifying concepts which are relevant for a situation to be regulated, as well as clusters of related concepts and semantic relationships which give a view of the actors and their relationships on the situation to be regulated. Going down to the Lexical layer the user may choose, in a specific language, the more appropriate lexical manifestations of such concepts, checking also the contexts in which they are used within the European legislation (Source panel), and insert the chosen lexical unit, with a proper XML annotation, within the text. Similarly, the user may search if a term is included in the Lexical layer and which is his conceptual description at the Ontological layer.

The second situation deals with checking the terminological accuracy of an existing text. The system may check all the lexical units of the text also contained in the DALOS Lexical layer, as well as verifying their pertinence to the context, and, in case, substitute them with more appropriate terms.

The third situation deals with facilities able to guarantee terminological and conceptual coherence in the transposition of European legislation into national laws. Usually European directives use general concepts and terminology to better cope with different Member States legal cultures; the use of the same term in a transposition law and in the transposition language version of the related directive might not be appropriate for a national legal culture and legislation (an example is the case of the term ‘electronic signature’, used in the Directive 1999/93/EC, and terms used in the Italian legislation (D. Lgs. 7 March 2005, n. 82) where three variants are distinguished: a more general concept of ‘firma elettronica’ (electronic signature), and more specific ones as ‘firma elettronica qualificata’ (qualified electronic signature) and ‘firma digitale’ (digital signature)). By using the DALOS resource the legal drafter can be supported in choosing the more appropriate concept and terminology to be used within the transposition law.

8. Conclusions

The main purpose of the DALOS project is to provide law-makers with linguistic and knowledge management tools to be used in the legislative processes, in par-
ticular within the phase of legislative drafting. The aim is to keep control over the legal language, especially in the EU legislation multilingual environment, enhancing the quality of the legislative production, as well as the accessibility and alignment of legislation at European level.

In this paper we presented the DALOS resource, which is organized into two knowledge layers (the Ontological and Lexical layers). The motivations of this kind of architecture and the methodologies for its implementation have been presented. In particular, we have discussed the principle on the basis of which the Ontological layer has been developed, as well as NLP techniques used to implement the Lexical layer. Finally we have illustrated the use of the DALOS resource in the xmLegesEditor legislative drafting environment, along with facilities aiming at enhancing the quality of legislative texts.

As future developments the extension of the Lexical layer to Dutch and Spanish is being carried on, as well as methodologies to build automatically or semi-automatically the DALOS Ontological layer can be investigated. Similarly xmLegesEditor can be adapted to serve as the interface for modification of the DALOS knowledge resource, for example to locally enrich the lexicon or to classify terms, to correct or align terms in different languages and, in general, to modify the knowledge resource, which can then later be submitted to a community of experts for its dynamic upgrade according to users’ contributions.

The availability of a document archive marked up with a vocabulary of normalized terms derived by DALOS modules can also be useful in documents indexing to provide Semantic Web-oriented retrieval services. Moreover, as terms in XML texts will be linked to the Ontological layer through the Lexical layer, it will be possible to provide more advanced query features exploiting semantics for extracting norms or document fragments using more complex retrieval inferences.

The application prototype (the integrated environment composed by the drafting tool and the knowledge resource) is under test and evaluation by legislative offices of the Italian Parliament and CNIPA. Following the extension to Dutch and Spanish, the system will be also tested by other public administration users in the Netherlands and Spain.

References


Italian Minister of Reforms and Innovations in Public Administration / National Center for Information Technology in Public Administration


