Jur-WordNet

Maria Teresa Sagri[1], Daniela Tiscornia[1], and Francesca Bertagna[2]

[1] ITTIG (Institute for Theory and Techniques for Legal Information)-
National Research Council, Via Pianciatichi 56/16, 50127 Firenze, Italy
Email: sagri@Ittig.cnr.it, tiscornia@Ittig.cnr.it
[2] ILC (Istituto di Linguistica Computazionale)-National Research Council,
Via Moruzzi 1, 56100 Pisa, Italy
Email: francesca.bertagna@ilc.cnr.it

Abstract. The paper describes Jur-Wordnet, an extension for legal domain of the Italian ItalWordNet database, aimed at providing a knowledge base for the multilingual access to sources of legal information. Motivations and aims are discussed, together with details concerning the linguistic architecture and construction methodology.

1 Introduction

The subject of this paper is a description of *Jur*-WordNet (*Jur*-WN), an extension for legal domain of the Italian *ItalWordNet* (IWN) database, aimed at providing a knowledge base for the multilingual access to sources of legal information. In the first section of the paper, we will introduce the application needs that are at the basis of the demand of such a lexical resource. A brief description of IWN will be introduced, focussing on the points of contacts between the Italian general wordnet and *jur*-WordNet. Then, the strategies followed during the *jur*-WordNet construction will be describe, with special attention to the handling of lexical polisemy and to the creation of an ontological layer of description.

2 Application Needs for the Legal Sector

The starting point was the *Norme in rete* (Law on the Net) project, launched in 1999 as part of the Italian *E-government Plan. Norme in rete* involves the most important Italian institutions with the goal to "create a portal which, through a single and simple user interface, allows research on all the documentation of normative interest published free on Internet, particularly by institutional sites." [12]. The portal allows free access to normative information through standard methods of editing, processing, and distributing data; the project provides codification standards for source types, identifiers (*urn*[3]), structure, links, and *metainformation.* System design, by now consolidated, consists of classes of XML DTDs[4] for structuring normative texts and of metadata, the most relevant part of which deals

[3] Uniform References Notation, which allows the identification of the partitions of legislative texts independently of the location

[4] See: http://www.normeinrete.it/standard/standard_xml.htm;
http://www.lexml.de, http://www.legalxml.org/,
http://lri.jur.uva.nl/METALex/.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 305–310.

with the formal/structural features of each type of source, and with *urns* for the identification of the partitions of texts. The aim of Jur-WN is providing the system with a knowledge-base able to supply:

- a source of metadata for the semantic tagging of legislative texts, both at the level of articles and of dispositions. It may also be used in the legislative drafting phase as an enrichment of the specialised XMLeditor now in the development phase [18], and of others legal sources.
- A support resource for information retrieval systems, for facilitating access to heterogeneous and multilingual data.
- An interface between the *common language* approach of citizen and the specific terminology of legal standard[5]. The greatest part of legal thesauri are primarily designed for the "professional" user and not for members of the public.
- A conceptual knowledge base, which can be used for a wide variety of applications and task, such as information extraction, question answering, automatic tagging, knowledge sharing, norm comparison, etc.

3 Overall Architecture of the IWN Database

The EuroWordNet (EWN) [16] project retains the basic underlying design of WordNet [11], trying to improve it in order to answer the needs of research in the computational field, in particular extending the set of lexical relations. In the last years, an extension of the Italian component of EWN was realized with the name of IWN [13]. IWN follows exactly the same linguistic design of EWN (with which shares the Interlingual Index -ILI- and the Top Ontology -TO- as well as the large set of semantic relation[6]) and consists now of about 70,000 word senses organized in about 50,000 *synsets*. Terminological wordnets dedicated to specific domains and linked to the generic module were envisaged, but at the moment only the eco-WordNet module[7] is publicly available, while we are still building the jur-WordNet plug-in. By means of the ILI, all the concepts in the generic and specific wordnets are directly or indirectly linked to the TO. In the EWN model a Domain Ontology was foreseen and in IWN a Domain Ontology was developed for the economic domain. An ontology dedicated to the legal domain is also in construction in *jur*-WN.

3.1 The Plug-in Mechanism

During the IWN project, an innovative methodology (the so-called Plug-in model) for linking domain-independent and domain-specific wordnets was defined. The plug-in relations in jur-

[5] The Proposal for a Directive of the European Parliament and of the Council on the re-use and commercial exploitation of public sector documents(14047/02) is aimed at encouraging the re-use of *Public Sector Information* by private operators *for commercial purposes.* Legal and regulatory information, as well as information on rights and duties are a relevant part of PSI. In the regulation of public/private relationship in the market place, the "added value" is a crucial point, dealing with the assessment of Intellectual Property Right and of pricing policies, where added value is mainly conceived as capacity to improve the accessibility for citizen of relevant information, both from a technical and a subjective (content-driven) perspective.

[6] For a complete list of the available semantic relations cf. [13]

[7] developed by Istituto per la Ricerca Scientifica e Tecnologica of Trento (IRST)

WN concern only nouns, which represent the vast majority of the db lexical entries. The plug-in model is realized by means of three plug-in relations defined in order to allow the integrated consultation of the two databases: i) PLUG_SYNONYMY (connecting IWN and domain-specific wordnet whenever it is possible to find an IWN synset having the same meaning of an domain-specific synset), ii) PLUG_NEAR_SYNONYMY (connecting synsets which have 'similar' meanings but are not interchangeable in contexts or whose lists of hyponyms are not compatible) and iii) PLUG_HYPONYMY (connecting an IWN synset and a domain-specific synset with a more specific meaning). The linking via plug-in relations has two effects: (i) the creation of one or more plug-in synsets, where the pairs of synsets involved in the connections are substituted by plug-in synsets and are therefore no longer accessible in the integrated consultation; (ii) the eclipsing of certain synsets, i.e those reachable from IWN through downward links (i.e. its hyponyms) and those reachable from the domain-specific wordnet through upward links (i.e. its hyperonyms). Eclipsed synsets are no longer accessible in the integrated consultation. For a more detailed description of the plug-in model and relations, cf. [13].

4 Jur-WN As a Lexical Resource and a Content Description Model

Jur-WN is a multi-layered lexical resource [14]. First of all, a large set of semantic relations (inherited from the linguistic design of the general IWN database) can be used to link synsets within the same domain-specific module. Then, the plug-in model provides the lexicographer with the possibility to exploit the information already available in the general wordnet, without the necessity to encode general lexical-semantic information from scratch. The latter, more conceptual and abstract layer is the "ontological" one, made up of the higher level of jur-WN, which becomes a core ontology for the legal domain. The first two layers are designed to improve legal information retrieval from heterogeneous (legislation, legal cases, policies) and multilingual sources. Providing a legal lexicon, allowing the handling of linguistic phenomena as polisemy and synonymy, means also to establish a bridge between the common language – often used from the non-jurist ones in order to place legal questions – and the technical language of the law. Under this viewpoint the *plug*-relations linking Jur-WN and Italwordnet allow a more precise definition of technical meanings of terms used in the common Italian, such as *autorizzazione* (*authorisation*), *alienazione* (*alienation*), and the specification of terms acquiring specific law meaning such as *alimenti* (*alimony*) and *mora* (*delay*). Moreover, plug-relations allow the insertion of domain-specific syntagms which ihnerit the "semantics" of their domain-independent head: for instance, the *accettazione delle prove* (*evidence acceptance*), *accettazione della testimonianza* (*witness acceptance),* of the legal domain are linked, trough a plug-hyponymy relation, to the synset *accettazione* (*acceptance*) of the IWN lexicon, by means of which is also linked to the Top-Ontology shared by all the Euro-WordNet databases.

As a source of *metadata for content description,* we need a standard of metadata based on the ontological nature of the entities of the legal domain: within *jur*-WN, an ongoing effort is dedicated to the creation of an ontological level [5]: from the 1500 synsets structured so far, the higher terms/concepts (about 40) have been organised selecting concepts that, acquiring a specific meaning in the legal domain and roughly matching the classical partitions of legal theory[8], are organised in a *legal core ontology* [8], that takes into account both the new upper levels (DOLCE) [4], and the proposal in the field of legal ontologies [8,15]. For a detailed description of the results for the ontological level, cf. [5].

5 Method of Development of the Semantic Network

In the construction of Jur-WN the "citizens' perspective" was taken into account and a "bottom-up" approach from existing linguistic/terminological resources was followed, selecting as starting points the most frequent terms in user queries of the major legal information retrieval systems.[9] We have used:

- For identification of the relevant terms: the query strings of the Progetto N.I.R. and those of ITALGIURE; the lists of terms linked by AND in the queries provide about 13.000 syntagms; the lists of terms linked by OR in the queries provide the analogical chain and the identification of synonyms.
- For definition of the principal technical concepts: handbooks, dictionaries, legal encyclopedias, etc., [3,2,6,1,10,11] and the L.L.I. containing historical archive of Italian legislative language [18].
- For determination of the syntagms relative to the principal lemmas: the syntagms extrapolated by the ITALGIURE Information Service.

Each sense of the basic terms is then considered as a possible "root" of a sub-hierarchy of terms and syntagms. The general method, in part conducted using automated procedures, considers the syntagms as hyponyms every time their "head" is identical to that of the "basic terms." For example, we identify two different senses of provvedimento (ruling); that is, as public authority act and as disciplinary measure. Nine relative hyponyms are attached to sense 1 (e.g., provvedimento amministrativo -administrative ruling-, and provvedimento legislativo -legislative ruling-) while to sense 2 are linked five terms (e.g., ingiunzione -injunction-, sanzione -sanction-, arresto -arrest- and detenzione -detection-), which are semantically more specific even if lexically different. Often, the syntagms are considered more interesting if they are linked to basic terms by different semantic relations; for example, verbale d'udienza (trial transcript) is linked to udienza (trial) as 'role-instrument' and to verbale (transcript) as hyponym. Where possible, synonym variants were also included. By the end of this phase, the terms collected are about 1500. The still ongoing phase consists of connecting Jur-WN with IWN and with the ILI (Inter-Lingual Index) in order to integrate the synsets with the networks of the Italian and the other European wordnets.

5.1 Polisemy Handling

Polysemy arises in legal terms both in relation to common language and within the specific context. For example, at legal level, the Italian term *canone* can refer to a payment (in money or goods) or to a legal norm of universal character. *Alimento* considered in the singular is "nutriment" while in the plural is a compulsory payment in the field of divorce

[8] Concepts as *licenza* (*license*), *autorizzazione* (*authorisation*), and *delega* (*delegation*).

[9] We will also evaluate the coverage of the synsets labelled with "law" in MultiWordNet

(*alimony*). The WordNet model permits handling multiple senses in an explicit manner and this allows us to establish conceptual correspondences among terms in different languages. It is especially efficacious in the legal domain: in law we do not speak of the translation of a legislative text but rather of its multilingual versions. The issue concerning multilingual versions of legal texts is crucial in European Community, where a dual approach is taken: the semantic relations established *a priori* on a conceptual nucleus are integrated with the context comparison on which the Eurodicautom translator is based; for example, the term *prescrizione* corresponds to at least six English terms: *statute of limitations*, *requirement*, *inscription* etc..

Prescrizione1	**Prescrizione 2**	**Prescrizione 3**
synonym: norma, regola (norm, rule, prescription)		
has-hyperonym: diritto (law) *has-hyponym*: prescrizione medica	*has-hyperonym*: fatto giuridico (legal fact) *has-hyponym*: prescrizione speciale, prescrizione ordinaria *cause*: acquisition	*has-hyperonym*: Fatto giuridico (legal fact) *has-hyponym*: prescrizione della pena, prescrizione del reato *cause*: expiration *involved:* termini di prescrizione
equal to: **requirement**	*equal to*: **prescription**	*equal to*: **prescription of claims, limitation of action**

In the above example, we see that word sense discrimination takes into account the distinctions among common and technical meanings (between sense 1, 2 and 3), and among legal institutions (between senses 2 and 3), as well as the confusion between cause (*passage of time*) and effect (*extinction/acquisition*) and between *lapse of time* and *final term*. In other words, we need to manage "semantic overlapping" with more sophisticated linguistic and representational devices, devices that permit us to make distinctions concerning the ontological nature of the concepts. Terminological domains seem to offer a profitable test of the relations between ontology and lexicon: "it is possible that a lexicon with a semantic hierarchy might serve as the basis for a useful ontology, and an ontology may serve as a grounding for a lexicon. This may be so in particular in technical domains, in which vocabulary and ontology are more closely tied than in more-general domains." [7]

6 Future Work

The *jur*-IWN database is still under development: we expect to reach a satisfying coverage of the basic legal contents trough the definition of about 3000 synsets. The enrichment of the lexical database will probably act as a test of the ontological level, and allow refinement and completion of the work done. The European Commission has recently approved, under the E-Content Program, the Project Lois (*Lexical Ontologies for Legal Information Sharing*), aimed at the localization of WordNets for legal domain to Italian, English, German, Czech, Portuguese and Dutch, in order to allow cross-lingual retrieval across different national collection of laws. Furthermore, it will enable cross-lingual access to legislative corpora by inexperienced users and better retrieval by experienced users.

References

1. De Mauro T., *Il Grande Dizionario italiano dell'uso*, UTET, Torino, Italy (2000).
2. *Enciclopedia del diritto*, Giuffrè, Varese, Italy (1989).
3. *Enciclopedia giuridica*,Treccani, Roma, Italy (1995).
4. Gangemi A., Guarino N., Masolo C., Oltramari, A., Schneider L., *Sweetening Ontologies with DOLCE.* in *Proceedings of EKAW 2002*, Siguenza, Spain (2002) 166–178.
5. Gangemi A., Sagri M. T., Tiscornia D., Metadata for Content Description in Legal Information, Workshop Legal Ontologies, ICAIL2003, Edinburgh. In press for *Artificial Intelligence and Law Journal*, Kluwer.
6. *Grande Dizionario enciclopedico del diritto*, Fratelli Fabbri Editore, Milano, Italy.
7. Hirst G., *Ontology and the Lexicon*. In Staab, Steffen and Studer, Rudi (eds) Handbook on Ontologies in Information Systems, Berlin: Springer, 2003.
8. `http://wonderweb.semanticweb.org/deliverables/D17.shtml`.
9. *Il Dizionario della lingua Italiana*, Garzanti, Milano, Italy (2002).
10. *Il Nuovo Zingarelli, Vocabolario della lingua italiana*, Zanichelli Ed. Milano, Italy (2002).
11. Miller, G., Beckwith R., Fellbaum C., Gross D., Miller K. J., *Introduction to WordNet: An On-line Lexical Database*. In International Journal of Lexicography, Vol.3, No.4, (1990) 235–244.
12. Report on *"Il progetto Norme in rete",* Italy (2000) (`http://www.normeinrete.it/documenti`).
13. Roventini A., Alonge A., Bertagna F., Calzolari N., Girardi C., Magnini B., Marinelli R., Speranza M., Zampolli A., *ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian*, in "Linguistica Computazionale", Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma, ISSN (2003).
14. Sagri M. T., *Progetto per lo sviluppo di una rete lessicale giuridica on line attraverso la specializzazione di ItalWornet*. In *Informatica e Diritto*, ESI, Napoli, (2003).
15. Visser P., Bench Capon T., *Ontologies in the Design of Legal Knowledge Systems, towards a Library of Legal Domain Ontologies*, in Proceedings of *Jurix 99,* Leuven, Belgique (1999).
16. Vossen P. (ed.), *EuroWordNet General Document*, 1999. `http://www.hum.uva.nl/~ewn`.
17. `http://www.ittig.cnr.it/banche/LLI/`.
18. `http://www.ittig.cnr.it/organizzazione/personale/biagioli/normeinrete`.