

Towards an Italian Legal FrameNet

Exploring evidence for extending and specializing FrameNet to the legal domain

The work presented here is part of a wider on-going study aimed at investigating methodological issues and strategies towards the construction of an Italian FrameNet-like resource specialized for the legal domain, through extending and specializing the general FrameNet¹ [1].

This study stems from the idea that FrameNet can be seen «as a backbone of several domain-specific FrameNets» [2]. To the author's knowledge, little research has been done on the use of Fillmore's *Frame Semantics* to investigate the grammatical and semantic properties of specialized languages. A noteworthy exception is [2], who successfully developed a BioFrameNet, an extension of FrameNet to the molecular biology domain. It is aimed at producing a semantically insightful analysis of this language through examining the syntactic and semantic combinatorial possibilities exhibited within a corpus of scientific writings.

Interestingly, more than in the case of other domains expressed through specialized languages, the law is invariably conveyed through natural language. It is claimed that «il diritto non si serve della lingua, ma è fatto di lingua»² [3]. Consequently, a better understanding of legal language features results in a better understanding of law. Moreover, both linguists and legal experts state that within the legal field the domain-specific language is closely intertwined with common language. According to their studies (see among others [4], [5], [6], [7], [8]), legal language, still differing from ordinary language, is in fact not dramatically independent from every day speech; it is rather an *extension* of it. As reported in [4], an example of such phenomenon is the use of the bare infinitive in the pronominal form. Differently from the ordinary language usage, within the Italian legal texts it tends to occur in subordinative phrases not uniquely in conjunction with verbs of perception (such as in the following sentence: *il difensore chiede applicarsi all'imputato la diminuzione della pena* 'the defense asks to apply to the defendant the reduction of the punishment').

The present study would suggest that the Fillmore's *Frame Semantics* theory seems particularly appropriate to the analysis of legal language. Namely, noteworthy grammatical constructions and links with their corresponding semantic properties are pointed out thanks to the approach of *frame semantics* to lexical and semantic organization³. The value of laying out to what extent these constructions are privileged to others in law texts is made fundamental by considering that the technical language used is closely intertwined with common language.

A study which documents the *range of semantic and syntactic combinatory possibilities of each word in each of its senses* supported by corpus evidence - given by law text collections - aims at being between a number of research areas. Firstly, lexicographic studies on legal terms and linguistic (grammatical) investigations on legal language are involved. Studies on legal semantics, legal theory and legal reasoning mainly carried out by scholars of jurisprudence or legal philosophers can contribute, as well.

Moreover, within the Artificial Intelligence and Law community the need for automatically processing and extracting semantic information implicitly contained in law texts drew new attention to Natural Language Processing methods and techniques. The complex conceptualizations being

¹ <http://framenet.icsi.berkeley.edu/>

² Lit. "law doesn't use natural language, rather it is made by language".

³ It «makes it possible to separate the notion of the conceptual underpinnings of a concept from the precise way in which the words anchored in them get used.» (Fillmore, C. J. and B. Atkins (1992). **Towards a frame-based lexicon: the semantics of RISK and its neighbors.** *Frames, fields, and contrasts*. A. Lehrer and E. F. Kittay. Hillsdale, Erlbaum, pp. 75-102, p.101).

expressed in natural language, legal text modelling requires facing with and solving the natural language problem. Similarly to the field of open-domain research, by providing both a syntactic and a semantic description of relevant lexical units, a FrameNet-like resource specialized for this domain may offer advanced capabilities in higher level language processing, advancing the results of domain-specific Knowledge Management applications.

The work presented here aims at pointing out that applying *Frame Semantics* theory to the legal domain seems appropriate in order to investigate i) lexical semantic properties of legal texts, identifying noteworthy *argument structures* (i.e. syntactic and semantic valences) of frame-evoking lexical units and ii) how concepts and relations are realized linguistically within law texts, providing a *frame-based* representation of the events and situations depicted in legal texts.

The need for a FrameNet extension to the legal domain originates, on one hand, from the chance to use the organization principles of the FrameNet project as a basis for linguistic studies; on the other hand, it originates from the lack of a domain-specific lexical-semantic resource i) founded on corpus evidence and ii) accounting for the semantic combinatorial properties of lexical items contained. Today, the existing lexical resources for the legal domain are based on the WordNet model [9]. Such is the case of JurWordNet [10] and its multilingual extension LOIS [11], where words expressing legal concepts are organized in *synsets* (i.e. sets of synonyms) in turn linked by hierarchical or taxonomical relations such as hyponymy and hyperonymy. Under this view, the meaning of a word is intended as a distinct, atomic semantic object, fully identified by its position in the general semantic network.

However, the taxonomical organization of legal concepts is not the only possible one. Views selecting properties, legal experts aim at constructing Knowledge Organization systems (Legal Ontologies) which can take into account the **context** where legal entities move. From the knowledge modelling point of view, it is raised the need for «capturing and handling all possible stereotypical situations distinguished by law» [12]. Moreover, it is affirmed that, despite their utility, WordNet-like resources are not completely adequate and satisfactory in order to represent events and situations typically expressed in legal documents (e.g. under which *Circumstances*, which *State of affairs* is sanctioned by which *Principle*). This is a consequence of the WordNet model they follow: it shows that certain groups of words are semantically related but it shows nothing about a word's *combinatorial behaviour*. As opposed to a WordNet-like resource, by providing schematic script-like organization of knowledge, a FrameNet-like lexical resource for the legal domain meets the legal experts' needs to access the inner structure of events expressed by laws and norms.

Moreover, as Fellbaum noted in [9], «WordNet reflects the structure of frame semantics to a degree, but suggested that its organization by part of speech would preclude a full frame semantic approach». On the contrary, in FrameNet the lexical units that evoke a frame are not restricted to a single part of speech. This is a very important FrameNet feature when dealing with corpora of legal language. Linguistic studies (see among others [4]) affirm that is very common in law texts that events are expressed through nominal rather than verbal constructions. This is in line with the results of an analysis of the main morpho-syntactic and syntactic peculiarities of legal language carried out by the author comparing corpora of Italian law texts with a corpus of ordinary Italian⁴. For this purpose, NLP techniques have been used, mainly exploiting a shallow parsing approach, i.e. *chunking*. Even quite rudimentary, this first level of syntactic grouping has allowed detecting some main peculiarities of legal language. Among the others, the quite high occurrence of

⁴ In this study, two different legal corpora were taken into account, differing at the level of the regulated domain (i.e. the environmental and the consumer protection domains) and of the enacting authority (i.e. European Union, Italian national state and Piedmont local authority). The Environmental Corpus consists of 824 legislative, institutional and administrative acts for a total of 1,399,617 word tokens. The Consumer Law Corpus is made up of 18 European Union Directives in consumer law, for a total of 74,210 word tokens. These two legal corpora have been compared with an Italian reference corpus, the PAROLE corpus [14], made up of about 3 million words including texts of different types (newspapers, books, etc.).

prepositional chunks and the fairly low presence of *finite verbal chunks* have been considered as two of the more visible syntactic phenomena proper of legal sub-language. Being closely connected, these findings may reflect a noteworthy linguistic realisation of events within law texts. The Italian law texts seem to be broadly biased towards a nominal realization rather than a verbal one; in particular, the use of deverbal nouns typically embedded in PP-attachment chains may be mostly involved. In this sense, a *Frame Semantics* description of the grammatical structure expressing events within law texts can prove and continue such an hypothesis.

Interestingly enough, the need for a frame-based lexical-semantic resource is in line with the Van Kralingen's proposal, within the Artificial Intelligence and Law community, of a *frame language* as a plausible method for the conceptual representation of legal knowledge [13]. In spite of the fact that it dates back to the early '90s, it still represents a need commonly felt. The proposed *frame language* is based on the concept of a *norm* and of an *act* as legal conceptual primitives of the legal domain which can be conceived as *frames*, i.e. data-structures for representing a stereotyped situation in which each element is represented. Thus, the focus is on the inner structure of a *norm* and of a *legal act*, i.e. on what their building elements are. That is, a *norm frame* is defined as a template in which each element of a norm is represented as a slot of the norm frame (e.g. 'Scope', 'Conditions of application', 'Subject', etc.). Since every legal action has many different aspects, a *legal act* has also been conceived as a frame; each aspect of an action is represented as a slot of the *act frame* as well (e.g. 'Act type', 'Manner', 'Circumstances', 'Agent', etc.). Even though the Van Kralingen's approach is mostly based on domain-theoretical assumptions, it would be taken into account as the only one within that community which has envisaged a frame-based organization of legal knowledge.

Currently, the author is analyzing a corpus of Italian law texts which regulate the environmental domain as well as a corpus of legislative texts in consumer law. Supported by corpus evidence, such an initial work is aimed at selecting a set of Semantic Frames already existing in FrameNet relevant to semantically describe both the legal domain and the regulated domain at hand. The investigation of valence properties of frame-evoking Lexical Units and support expressions – such as, e.g., support verbs lexicographically relevant - worth describing for insightful analysis of legal language purposes follows. The methodology followed towards the construction of the Italian Legal FrameNet mostly consists in maintaining and reusing the Semantic Frames and Frame Elements already defined in FrameNet. Nevertheless, extensions and specializations of the general English FrameNet with respect to both the Italian language – if needed – and the legal domain have been foreseen. For this purposes, jointly with a pool of legal experts, the author is currently discussing a number of both language and domain-specific customization strategies. They mainly include i) the introduction of one or more new Frame Elements within an existing Frame, ii) the exploitation of domain-specific Semantic Types which classify Frame Elements from the general FrameNet repository and iii) the creation of a new Frame to specify domain-specific information - only when it is absolutely needed.

In designing a FrameNet extension for the legal domain, categorizing the sort of lexical fillers that is expected in a Frame Element through the use of Semantic Types is considered as a fundamental stage. For this purpose, the link of the FrameNet-like resource with an already existing Legal Ontology has been taken into account. For example, the lexical filler of the Frame Element 'Principle' – part of the Frame IMPOSING_OBLIGATION and defined in FrameNet as 'a regulating idea (which may be instantiated as a document) that the Responsible_party is subject to' - can be domain-specifically categorized with the Semantic Type 'LegalNorm', which is a class (i.e. a juridical concept) in the Core Legal Ontology (CLO)⁵ [15], fully specified by its relationships with other ontology classes (i.e. nodes of the ontology). Such an approach will result in a frame-based and hierarchic-based combined resource, complementing the type of information we have so far for the legal domain.

⁵ <http://www.loa-cnr.it/>

References

- [1] Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., and Scheffczyk, J. (2006), **FrameNet II: Extended Theory and Practice**, available online at <http://framenet.icsi.berkeley.edu/>
- [2] Dolbey A, Ellsworth M, Scheffczyk J. (2006), **BioFrameNet: A Domain--Specific FrameNet Extension with Links to Biomedical Ontologies**, O. Bodenreider (ed.), Proceedings of KR-MED, 87-94.
- [3] Cortelazzo M.A. (1997), **Lingua e diritto in Italia. Il punto di vista del linguisti**, in L. Schena (a cura di), **La lingua del diritto. Difficoltà traduttive. Applicazioni didattiche. Atti del primo Convegno internazionale, Milano, 5-6 ottobre 1995**, Cisu (Centro d'Informazione a Stampa Universitaria), Roma, pp. 35-50.
- [4] Mortara Garavelli B. (2001), **Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani**, Einaudi, Torino.
- [5] Rovere G. (2005), **Capitoli di linguistica giuridica. Ricerche su corpora elettronici**, Edizioni dell'Orso, Alessandria.
- [6] Fiorelli P. (1993), **Premessa a I.L.L.I.** in P. Mariani Biagini (a cura di), *Indice della Lingua Legislativa Italiana. Inventario lessicale dei cento maggiori testi di legge tra il 1723 e il 1973*, vol. I (A-E), I.D.G. (Istituto per la Documentazione Giuridica del Consiglio Nazionale delle Ricerche), Firenze, pp. V-XII.
- [7] Cassese S. (1992), **Introduzione allo studio della normazione**, in «Rivista trimestrale di diritto pubblico», n.2, pp. 307-330.
- [8] Scarpelli Uberto, *Semantica giuridica*, in Azara A. Eula E. (dir), *Digesto Italiano*, vol. XVII, 1969.
- [9] Fellbaum, C. (ed) (1998), **WordNet: An electronic lexical database**. MIT Press.
- [10] Sagri M-T., Tiscornia D., Bertagna F. (2004), **Jur-WordNet**, in Proceedings of the Second Global WordNet Conference, pp. 305-310, Brno, Czech Republic, January 20-23.
- [11] Peters, W. , Sagri, M. T. , Tiscornia, D. (2007), **The structuring of legal knowledge in lois**, Artificial Intelligence and Law, vol. 15, pp. 117-135.
- [12] Breuker J. (2009), **Dreams and awakenings about legal ontologies**, talk taken on the occasion of the 3rd Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 09), June 8th, 2009, Barcelona, Spain.
- [13] Kralingen R.W. van (1995), **Frame-based conceptual models of statute law**, Kluwer Law International, The Hague.
- [14] Marinelli R., Biagini L., Bindi R., Goggi S., Monachini M., Orsolini P., Picchi E., Rossi S., Calzolari N., Zampolli A. (2003), *The Italian PAROLE corpus: an overview*, in A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa - Linguistica Computazionale a Pisa. Linguistica Computazionale, Special Issue, XVI-XVII, Pisa-Roma, IEPI. Tomo I, 401-421.
- [15] Gangemi, A., Sagri, M.T and Tiscornia, D. (2005), A constructive framework for legal ontologies, in Law and the Semantic Web, Benjamins, R., Casanovas, P., Breuker, J., and Gangemi, A., (eds), Springer Verlag.