The Construction of a Juridical Ontology

Isa Mara da Rosa Alves
Universidade Estadual Paulista (UNESP/Ar),
PhD Student of Department of Linguistics.
Postal Code 174 – 14.800-901 – Araraquara – SP – Brazil
+55 16 33016212
isamralves@gmail.com

Rove Luiza de Oliveira Chishman
University of Vale do Rio dos sinos (UNISINOS) – Associated Professor at Department of Linguistics
Unisinos Avenue, 950 – 93.022-000 São Leopoldo – Brasil
+55 51 35911122 (r.1337)
rove@icaro.unisinos.br

Paulo Miguel Torres Duarte Quaresma
University of Évora (UÉVORA) - Associated Professor at Department of Computer Science
Romão Ramalho Street, 59 – 7000 – Évora – Portugal
+351 21 2948536
pq@di.uevora.pt

The need for the representation of both semantics and common sense and its organization in a lexical database or knowledge base has motivated the development of large projects, such as Wordnets, CYC and Mikrokosmos. Besides the generic bases, another approach is the construction of ontologies for specific domains. Among the advantages of such approach there is the possibility of a greater and more detailed coverage of a specific domain and its terminology. Domain ontologies are important resources in several tasks related to the language processing, especially in those related to information retrieval and extraction in textual bases. Information retrieval or even question and answer systems can benefit from the domain knowledge represented in an ontology. Besides embracing the terminology of the field, the ontology makes the relationships among the terms explicit.

We argue for a corpus based methodology for an ontology construction that seeks for the rigorous linguistic analysis aiming at formalization. The proposed methodology is an integrated representation of the verbal content from the perspective of the Formal or Logic Semantics, Lexical Semantics, Grammatical Semantics and Pragmatics heading for the construction of an ontology.

In short details, the results of the corpus pre-analysis stage are the following. Six judicial judgments of the Supreme Court of Justice of Portugal, homologated within the 2002-2003 period, were randomly chosen. These judgments referred to the theme `traffic accidents` and were available electronically. After this, we studied the way the judgments were organized textually as well as the role of the communication contract in which they are included. The main objective of this stage was to obtain a contextualized comprehension of the texts at issue. We considered discursive approaches while carrying out this investigation. This global comprehension of the corpus helped to identify the semantic roles of the arguments, as well as other specific semantic issues of the analysis.

ICAIL '07 June 4-8, Palo Alto, CA USA Copyright 2007 ACM 978-1-59593-680-6/07/0006 ...$5.00.

From a group of 359 verbs extracted automatically from 6 texts – using the XTRACTOR – the following were selected: *recorrer* (to appeal from - 21 occurrences in 6 judgments), *condenar* (to condemn - 14 occurrences in 6 judgments), *julgar* (to judge - 12 occurrences in 5 judgments), provar (to prove - 10 occurrences in 4 judgments), *revogar* (to revoke - 10 occurrences in 5 texts), *concluir* (to conclude - 9 occurrences in 5 judgments), *acordar* (to accord - 8 occurrences in 6 judgments), *alegar* (to allege - 7 occurrences in 4 judgments), *absolver* (to acquit - 6 occurrences in 5 judgments), *conceder* (to concede - 4 occurrences in 4 judgments), summing up a total of 99 concordances analyzed.

The first step of the corpus observation was to analyze, with the help of the WordSmith Concorder, the sentences in which the verbs at issue occurred in order to proceed the ontological representation in four levels (described below) for each one of the verbs.

Having the information described above, next step was to analyze, with the help of the WordSmith Concorder, the sentences in which the verbs at issue occurred in order to proceed the ontological representation in four levels for each one of the verbs. This refers to the selection of (i) a definition, (ii) the logic-semantic relationships, (iii) the semantic roles, and (iv) the frame elements. The result of the analysis for the verb *to condemn* is presented below as an illustration of the study carried out.

The first level, *a definition*, is a level of analysis, which is useful for people who will work with the ontology, not for the system itself. The definition was selected from the WordNet, as well as from a PB (Brazilian Portuguese) dictionary, Borba (2002), and from a PE (European Portuguese) dictionary, the Dictionary of Contemporary Portuguese (Dicionário da Língua Portuguesa Contemporânea) (2001).

Systematizing information related to the logic-semantic relationships, the second level, enables the system to recognize the meanings of the concepts defined in the ontology through the relationships expressed among them.

Considering the relationships proposed by the wordnets as the basis, the following relationships that allow the structuring of the lexicon in the juridical domain were selected: (a) antonymy (b) entailment, (c) cause, (d) hyponymy, and (e) synonymy. Through these relationships, it was possible to increase the verbal entities from 10 to 120. As an example of the analysis carried out, we observed relationships among three verbs of the corpus: *to condemn*, *to acquit, to judge* and their synonyms. (to condemn, to pronounce a judgment against, to pronounce a sentence, to make guilty) is antonym of (to acquit, to pronounce a judgment in favor of, to pronounce innocent) and both are hyponyms of (to pronounce a sentence, to judge, to

label), which are hyponyms of (to speak, to say, to declare), which are hyponyms of (to verbalize, to express something).

In the third level, the analysis on a lexical level is left aside and the analysis is done on a sentence level, in the syntax-semantic interface. In order to classify the participants of the situations, which means, the arguments of the verbs, different authors contributed to the conclusions (e.g. Fillmore, 1968; Frawley, 1992; Borba, 1996). The task of classifying the arguments according to the semantic (or thematic) role to which they refer is not simple since it is highly subjective. The semantic roles identified in the corpus were: (a) agent, (b) instrument, (c) beneficiary, (d) patient, (e) goal, (f) source, (g) location, (h) purpose, (i) reason. The designation of proprieties such as thematic roles and frame elements enabled us to enlarge the scope of the analysis into nominal and some adjectives, originating 74 non-verbal classes. As an example of the studies conducted, let us take the case of `*to condemn*`. The argumentative structure of *to condemn* (VTDI) expects that *someone* (external argument – ARG 0) condemns *someone else* (internal argument – ARG 1) *to something* (internal argument – ARG 2). Within the 14 sentences analyzed, in only one of them the ARG 0 is explicit. This fact does not seem to jeopardize the description of the arguments, since the structure of the verb accepts it. A sentence that shows two semantic roles attributed to the verb *to condemn* are showed below.

`A new sentence was uttered (pages 241 to 252). And, essentially, with the same meaning as the previous one, being the defendant [patient] condemned to pay the authors the same global sum of 6151000 escudos [goal]` (Source: Judgment 02B2159).

For the fouth level, the frame elements, we adopted Fillmore´s conception of frame, expressed in the FrameNet (FN), which is an approach related to the syntax, although it has an interface with semantics and with pragmatics. By doing so, an approach based on frames enables us to classify the existing relationships between entities that go beyond the relationships among isolated lexical units and among units of the same sentence. The frames allow us to classify the entities related to the extralinguistic context. In order to represent the frame elements of the ten verbs of the corpus (to condemn, to acquit, to judge, to accord, to concede, to revoke, to conclude, to allege, to prove and to appeal from), 9 semantic frames of the FN were selected and, sometimes, it was necessary to combine different frames to describe the participants of the situation at issue in a more comprehensive way. The frame elements identified in the corpus are: (a) appraiser (judge, magistrate, court); (b) appraised (defendant, representative, author); (c) arguer (defendant, representative, author); (d) recognizer (defendant, representative, author // judge, magistrate, court); (e) means; (f) legal base; (g) reason; (h) purpose; (i) evidence; (j) topic; (k) content; (l) message; (m) request. In addition to this, we also consider *time, condition, location and manner*, which are not always lexicalized in the sentence or in the body of the text, but which are always in the heading of all the judgments. Referring to them is fundamental, because they provide information that locate the document in relation to when the process was judged, the condition of the judgment, where the process was judged and the kind of process it was.

Going back to the example provided in the previous section, we can see that an approach based on frames attaches distinct labels to the verbal arguments and enables us to classify the elements, which is not possible using the semantic roles approach.

`A new sentence was uttered [means] (pages 241 to 252) [legal basis]. And, essentially, with the same meaning as the previous one, being the defendant [appraised] condemned to pay the authors the same global sum of 6151000 escudos [topic]` (Source: Judgment 02B2159).

In the specific sense of the juridical domain, *to condemn* can be considered a verb of *judgment communication,* because there is a *communicator* (which is implicit in the sentence) that communicates a judgment upon an appraised (the author) to a subject (the non-explicit authors and defendants in the sentence). This semantic information about the situational role of ARG 0 and ARG 1 are valuable for the functioning of the ontology, since they enable the insertion of restrictions such as *communicator agent* and *appraised patient.* An entity means (the sentence) can also be seen, which identifies the abstract location through which a condemnation is uttered. The *topic* identifies the penalty suffered by the *appraised* (pay an annual law alimony).

We highlight it that the semantic description of the verbs presented here was prompted by the information provided by the corpus, but was not limited to the possibilities expressed there. By doing so, a semantic relationship among the ten verbs of the basis with others independently of their occurrence in the judgments was established, considering only domain restrictions. In the case of the semantic roles, we followed the same path: we represented the participants of the situations that did not occur with the verbs at issue in the sample of the texts from the corpus (e.g. agents implicit in the sentences).

Having finished the corpus analysis, the ontology reached its final stage. It consists of formalizing the linguistic information in an ontology editor, Protégé in OWL format. Although we emphasized the importance of each of these levels in the semantic representation, undoubtedly, one of the difficulties in the task of constructing the ontology is exactly the definition of the semantic information that have to be included. The construction of UNIVERBUE started from six judicial judgments of the PRG-PT. Then, 359 different verbal occurrences were extracted and 120 referring to the verbs in question were selected. At the end of the stage presented here, the ontology had 120 verbal entities and 74 non-verbal entities.

REFERENCES

[1] Borba, F. S. *Uma Gramática de Valências para o Português*. São Paulo: Editora Ática, 1996.

[2] Fillmore, C.J. The Case for Case. In.: Bach and Harms (Ed.): *Universals in Linguistic Theory*, Holt, Rinehart, and Winston, New York, 1968.

[3] Frawley, W. *Linguistic Semantic*. London: Lawrence Erlbaum Associates, Publishers, 1992