

Formalising and bottom-up enriching the Ontology of a Generative Lexicon

Antonio Toral and Monica Monachini
Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1 - 56124 Pisa, Italy
antonio.toral, monica.monachini@ilc.cnr.it

Abstract

This paper presents on-going research to formalise the ontology of a computational lexicon in OWL (W3C standard) as well as to enrich it by applying a bottom-up approach that extracts semantic information from the lexicon. The resource used follows the Generative Lexicon (GL) theory and therefore (1) puts a challenge to ontology design as its semantic types are multidimensional and (2) enables the acquisition of further knowledge on concepts from semantic units. The formalisation allows the ontology to be processed by Description Logics reasoners as well as to be employed in Semantic Web applications. Moreover, the lexicon-driven enrichment increases the semantic information present in the ontology making it appropriate for ontology-driven Natural Language Processing. Finally, the paper studies the application of these procedures to a subsequent GL-based biological resource.

Keywords

Ontologies, OWL, Web Ontology Language, Generative Lexicon, Lexicon, Qualia Structure, Natural Language Processing, Semantic Web

1 Introduction

Ontologies are recognised as an important component in Natural Language Processing (NLP) systems that seek to deal with the semantic level of language. In fact, most, if not all of the semantic lexical resources within the area (e.g. WordNet [3], CYC [5], SIMPLE [6]), have in common the presence of an ontology as a core module.

The Web Ontology Language (OWL) is a W3C recommendation and a major technology for the Semantic Web. It is defined by [1] as “a semantic markup language for publishing and sharing ontologies on the World Wide Web”. OWL allows applications to process the content of information instead of just presenting it to the user [7].

The fact that OWL is the ontology language for the Semantic Web and that it provides a formal semantic representation as well as reasoning capabilities has encouraged the NLP community to convert existing resources to this language. Work in this area includes, for example, the conversion of WordNet [13] and MeSH

[12] and, moreover, the proposal of a general method for converting thesauri [14].

This paper deals with the conversion into OWL of the ontology of a lexico semantic resource based on the Generative Lexicon (GL) theory. The ontology design presents a challenge as the nodes of the ontology are not only defined by their formal dimension (taxonomic hierarchy), but also by additional dimensions: constitutive, telic and agentive. Besides, we take advantage of the generative possibilities of the resource in order to enrich the converted ontology with further semantic information extracted from the lexicon. The final objective of this research is to derive a formalised and semantically rich ontology which could be used for Information Extraction and Knowledge Acquisition tasks.

The rest of this paper is organised as follows. Section 2 introduces PAROLE-SIMPLE-CLIPS, the GL resource used in this research. Next, section 3 deals with the formalisation and enrichment of the ontology. Subsequently, section 4 discusses the application of these techniques to a GL-based biological resource. Finally, section 5 presents conclusions and future work lines.

2 PAROLE-SIMPLE-CLIPS: a computational Generative Lexicon

SIMPLE [6] is a large-scale project sponsored by the European Union devoted to the development of wide-coverage multipurposed and harmonised computational semantic lexica for twelve European languages (Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish). A language-independent ontology of semantic types and a set of templates were designed and developed in order to guarantee uniformity and consistency among the monolingual dictionaries. In the framework of this project, 10,000 word meanings were annotated for each language.

SIMPLE should be considered as a follow up of a previous European project, PAROLE [10], as it adds a semantic layer to a subset of the morphologic and syntactic layers that were developed by the latter. SIMPLE provides thus multi-layered lexica, as the information is encoded at different descriptive levels (morphological, syntactic and semantic). Although the information included for these levels is mutually independent, the layers are connected by one-to-one, one-

to-many or many-to-one links (e.g. a syntactic unit is linked with one or more semantic units depending on the number of meanings that the syntactic entry conveys).

CLIPS is an Italian national project which enlarged and refined the Italian PAROLE-SIMPLE lexicon [11]. The core data encoded within SIMPLE was extended in CLIPS with a new set of lexical units selected from the PAROLE corpus according to frequency-based criteria. The resulting lexical resource contains 387,267 phonetic units, 53,044 morphological units, 37,406 syntactic units and 28,346 semantic units.

From a theoretical point of view, the linguistic background of PAROLE-SIMPLE-CLIPS is based on the Generative Lexicon (GL) theory [9]. In the GL, the sense is viewed as a complex bundle of orthogonal dimensions that express the multidimensionality of word meaning. The most important component for representing the lexical semantics of a word sense is the qualia structure which consists of four qualia roles:

- Formal role. Makes it possible to identify an entity.
- Constitutive role. Expresses the constitution of an entity.
- Agentive role. Provides information about the origin of an entity.
- Telic role. Specifies the function of an entity.

Each qualia role can be considered as an independent element or dimension of the vocabulary for semantic description. The qualia structure enables to express different or orthogonal aspects of word sense whereas a one-dimensional inheritance can only capture standard hyperonymic relations. Within SIMPLE, the qualia structure was extended by assigning subtypes to each of the qualia roles (e.g. *Usedfor* is a subtype of the telic role).

The formal entities involved in the semantic description of PAROLE-SIMPLE-CLIPS which are significant for the purposes of the current research are semantic types, templates, qualia relations and features. The description of each of these elements follows.

The semantic types are the nodes that make up the ontology. A peculiar trait of the adopted ontology is the fact that it consists of both simple types, which identify only a one-dimensional aspect of meaning expressed by hyperonymic relations, and unified types, which express multidimensional aspects of meaning by combining subtyping relations and orthogonal semantic dimensions.

The ontology consists of 153 language-independent semantic types. The top types are mappable to the ontology of EuroWordNet [15]. The design of the ontology is highly influenced by the GL model. In fact, the top nodes are the semantic type *ENTITY* and three other types named after the agentive, constitutive and telic qualia roles (*AGENTIVE*, *CONSTITUTIVE* and *TELIC*). These three nodes are designed to include semantic units definable only in terms of qualia dimensions. The direct subtypes of the node *ENTITY* are the semantic types *CONCRETE_ENTITY*, *PROPERTY*, *ABSTRACT_ENTITY*, *REPRESENTATION* and *EVENT*.

Table 1: Relations encoded in the template for the semantic type *INSTRUMENT*

Relation	Qualia role	Constraint value
Isa	Formal	Yes
Hasaspart	Constitutive	RecNo
Madeof	Constitutive	RecNo
Usedfor	Telic	RecYes

Each semantic type is associated with one template. These act as *blueprints* for any given type in the ontology and provide the conditions of well-formedness and constraints for lexical items belonging to that type. The template structure is built like a schema that works as an interface between the lexicon and the ontology: it imposes conditions for the belonging of a given semantic unit to a semantic type. The template is then a help and a guide for the encoding of information referring to the ontology.

Relations and features are the elements that allow to assign properties to the semantic units. They can be applied as constraints within templates, in this case they act as type-defining (prototypical) for the semantic units included in these templates and can take one of the following values:

- Yes. The information is mandatory. I.e. every semantic unit that belongs to the semantic type should initialise this property.
- RecYes. The information is mandatory and the cardinality can be higher than one. I.e. a semantic unit can be linked to more than one element via this property.
- No. The information is optional.
- RecNo. The information is optional and the cardinality can be higher than one.

Table 1 provides an example on the encoding of relation constraints in templates. It presents the relations included in the definition of the template *INSTRUMENT* and for each of them the qualia type and the constraint value.

Features are used to characterise those attributes for which a closed range of values can be specified (e.g. edible = yes, sex = male, female). Features are useful to connect nodes across the ontology that share a given aspect and that otherwise would remain isolated. Relations, differently, link pairs of semantic units. There are different kinds of relations; there exist four types for the corresponding top roles of the extended qualia structure (formal, constitutive, agentive and telic), and others of non-qualia nature (e.g. synonymy). E.g. the semantic unit *bisturi* (*scalpel*) that belongs to the semantic type *INSTRUMENT* is linked to the semantic unit *incidere* (*engrave*) by the telic relation *Usedfor*.

3 Ontology modelling

This section describes our on-going research to formalise and enrich the ontology of PAROLE-SIMPLE-

Table 2: Modelling cardinality restrictions

Original Value	Cardinality restriction
Yes	min 1, max 1
RecYes	min 1
No	min 0, max 1
RecNo	min 0

CLIPS. Subsection 3.1 deals with the aspects regarding the formalisation in OWL while subsection 3.2 introduces the approach related to the semantic enrichment.

3.1 Formalisation

The elements of PAROLE-SIMPLE-CLIPS which have been considered to be modelled in OWL are those used to define the original ontology, i.e. the semantic types, the qualia relations and the features that apply to the templates to which the semantic types are associated. Further details on how each of these elements has been modelled follow.

Semantic types, as aforementioned, are the nodes that constitute the ontology. Therefore, they are modelled in OWL as classes. All sibling classes across the OWL ontology are made disjoint.

Relations are modelled as object properties. For qualia relations the domain and the range is made up of the classes *ENTITY* and the class that corresponds to the type of qualia relation (*AGENTIVE*, *CONSTITUTIVE* or *TELIC*). On the other hand, for non-qualia relations both the domain and the range are set to the top node of the ontology.

The application of relations to semantic types, as represented in the templates (see Table 1), is modelled with cardinality restrictions. To each value corresponds a different cardinality restriction, as shown in table 2.

The multidimensional nature of some semantic types is preserved in the OWL ontology by the inclusion of restrictions on qualia relations. The latter are in fact the elements that in PAROLE-SIMPLE-CLIPS allow to have multidimensional semantic types (also called unified types). If an ontology class contains a mandatory cardinality restriction on a qualia relation, then this class has as an additional defining dimension the corresponding qualia type. E.g. The class *INSTRUMENT* has a mandatory restriction on the constitutive qualia relation *Madeof*. Therefore, *INSTRUMENT* has as an additional defining dimension the constitutive one.

Figure 1 provides an example on the assignment of the relation constraints to the classes, the establishment of cardinalities and the role of inheritance (*INSTRUMENT* is a daughter of *ARTIFACT*). This figure is a snapshot of Protégé, a software that supports the edition of OWL ontologies [4], which shows the asserted conditions for the type *INSTRUMENT* in the formalised ontology.

Finally, features are modelled as *DataType* properties. Their domain is the union of the classes that share the feature (e.g. *FOOD*, *VEGETABLE*, etc. for the feature *PLUS_EDIBLE*).

3.2 Bottom-up lexicon-driven enrichment

Besides formalising the ontology in OWL, we enrich it by following a bottom-up approach that extracts semantic information from the word senses of the lexicon by using the qualia structure as a generative device. This initial research enriches the ontology with constraints on relations and features extracted from the lexicon. On-going research will provide further enrichment of the ontology by extracting predicates and subclasses.

Each class (semantic type) of the ontology is enriched with additional constraints on relations and features which are extracted by exploring the word senses (semantic units) that belong to it. The procedure extracts all the relations and features that are defined for the word senses that belong to the semantic type. Afterwards, from these relations and features, it selects those that are considered to be representative of the class and proposes them to be modelled in the class definition as cardinality restrictions.

As the objective is to extract those relations and features that are relevant for the class definition, we consider discriminating by frequency of appearance, i.e. the percentage of word senses (semantic units) that belong to a class for which the given relation or feature is defined. As an initial experiment, we have established a threshold for each class to be the frequency of the least frequent relation/feature that is defined in the template of the class. Thus, those relations/features not defined in the template but whose frequency is higher than that of the threshold are proposed to be considered in the class definition.

The outlined procedure finds 218 relations and 229 features that are not considered in the class definitions but that, according to our hypothesis, could be included in the ontology as cardinality restrictions because they convey information that characterises the semantic units that belong to the semantic types. In order to make more comprehensible this matter, we examine some relations and features that are extracted to enrich the ontology.

Beginning with relations, let's consider the semantic type *INSTRUMENT*. The agentive relation *Createdby* is not included in the template definition but due to its high frequency of appearance is proposed to enrich the ontology by including it as a defining relation in this template. Clearly, an instrument is an artificial entity and therefore the relation *Createdby* applies and so is included in the node definition¹.

Regarding features, we take *PLUS_HUMAN*. This feature is not defined for any class of the ontology. However, it is applied to a high percentage of semantic units across several nodes of the ontology: *PROFESSION*, *HUMAN_GROUP*, *HUMAN*, *PEOPLE*, etc. It is clear that any semantic unit that belongs to any of these semantic types would be of a human nature. Therefore, this feature is promoted to be type-defining in the aforementioned classes.

Another feature that could serve as an example is *PLUS_EDIBLE*. This feature is included in the tem-

¹ As it can be seen in figure 1 this relation is already present in the formalised ontology as it is inherited from the class *ARTIFACT*.

Asserted Conditions	
NECESSARY & SUFFICIENT	
NECESSARY	
p1:Artifact	E
p1:hasHasaspart min 0	E
p1:hasMadeof min 0	E
INHERITED	
p1:hasCreatedby min 1	[from p1:Artifact] E
p1:hasSynonym min 0	[from p1:Entity] E
p1:hasUsedfor min 1	[from p1:Artifact] E

Fig. 1: Asserted Conditions for the class *INSTRUMENT*

plate of several semantic types such as *FOOD* or *VEGETABLE*. However, although having a high frequency of appearance in the type *SUBSTANCE-FOOD*, it is not included in the definition of this class. The bottom-up procedure incorporates this feature as a type-defining element for this node.

4 Application to the biological domain

This section introduces the application of the presented procedures to the BioLexicon, a lexicon for the biological domain designed in the framework of the BOOTStrep project² which is inspired by the Generative theory and to a wide extent builds on top of the structures introduced by the PSC model [8]. This lexicon, together with the BioOntology, constitute a terminological backbone by combining lexical and ontological information thus becoming an innovative integrated resource suitable for NLP tasks in the bio domain.

The BioLexicon semantic relations build on the 60 Extended Qualia Relations of the SIMPLE model. The Extended Qualia Relations allow modelling different meaning dimensions of a word sense and specifying its relations to other lexical units (either paradigmatic or syntagmatic). Most of these relations, also shared by well-known ontologies of the biological domain, prove to be suitable for the domain of interest and therefore are imported into the BioLexicon model. Clearly, there are relations not considered in the Qualia Structure that however are relevant for this domain. We have studied the Open Biomedical Ontologies (OBO) Relations, an ontology of core relations for the biomedical domain, in order to find relevant relations not present in the Qualia Structure. Each of these has been added to the BioLexicon model, some of them as new relations whereas some others as subtypes of existing qualia relations.

The BioLexicon and BioOntology have been separately designed and constructed. The BioLexicon has been automatically populated with terms gathered from available bio terminologies and augmented with linguistic information about terms extracted from

texts [2]; the BioOntology has been built integrating different ontological resources of the domain. This is why we hypothesise that the procedures introduced in this paper might be useful in this case in order to synchronise the information present in lexicon and in the ontology:

- From the data present in the lexicon, we can generalise constraints from instantiated relations and check whether or not they have been included in the ontology definition. In other words, the procedures can be useful in order to find definition gaps in the ontology as its design has been done separately of the lexicon population.
- On the other hand, the procedures can easily be used to guarantee that the data encoded in the lexicon is consistent with the constraints that are present in the BioOntology.

5 Conclusions

This paper has presented a proposal to formalise and enrich the ontology of a GL resource. The approach followed has proved to success to formalise the GL ontology in the standard OWL format. Moreover, we have applied a bottom-up procedure in order to enrich the converted ontology with further semantic information obtained from the lexicon.

The formalisation allows the ontology to be processed and checked by standard reasoners. This can be useful for building semantic applications as well as to enhance the quality of the resource by validating it (through reasoning we can look for inconsistencies or conflicts).

Besides, the paper has studied the feasibility of the procedures to be applied to a GL-based domain specific resource. Also in this case, the ontology can be enriched with additional semantic information and the resource can be checked and thus consistency be guaranteed.

The possible uses of the resulting formalised and enriched ontology are twofold. First, as it is an OWL ontology it could be used in Semantic Web applications. Second, as it is a semantically rich resource, it could be applied to semantic NLP tasks. In fact, we

² www.bootstrep.eu

plan to use it for semantic Information Extraction and Knowledge Acquisition.

As for future work, some aspects regarding the modelling are to be considered. On one hand, we plan to research on enriching the ontology with semantic predicates, an additional kind of semantic information which is encoded in the lexicon and which could play an important role when using the ontology for NLP purposes. Once this is done, we will investigate regarding the atomisation of the formalisation and the enrichment with all the considered information.

Acknowledgements

This research is part of an European Ph.D. program. It has been partially funded by a research grant of the ILC-CNR and by the ECs 6th Framework Programme (4th call), conducted within the BOOTStrep consortium under grant FP6-028099. We also would like to thank Riccardo del Gratta for his ideas and valuable comments on the application of the procedures presented to the BioLexicon.

References

- [1] M. Dean and G. Schreiber. OWL web ontology language reference. W3C recommendation, W3C, February 2004.
- [2] R. del Gratta, V. Quochi, E. Sassolini, M. Monachini, and N. Calzolari. Toward a Standard Lexical Resource in the Bio Domain (to appear). In *3rd Language and Technology Conference*, Poznan, Poland, October 2007.
- [3] C. Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.
- [4] H. Knublauch, R. W. Ferguson, N. F. Noy, and M. A. Musen. The protege owl plugin: An open development environment for semantic web applications. In *Proceedings of the Third International Semantic Web Conference*, 2004.
- [5] D. Lenat. *From 2001 to 2001: Common sense and the mind of HAL*, pages 193–208. MIT Press, Cambridge, MA, 1998.
- [6] A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, and A. Zampolli. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263, 2000.
- [7] D. L. Mcguinness and F. van Harmelen. Owl web ontology language overview, February 2004.
- [8] M. Monachini, V. Quochi, N. Ruimy, and N. Calzolari. Lexical Relations and Domain Knowledge: The Bio-Lexicon Meets the Qualia Structure. In *Fourth International Workshop on Generative Approaches to the Lexicon*, Paris, France, May 2007.
- [9] J. Pustejovsky. The generative lexicon. *Computational Linguistics*, 17(4):409–441, December 1991.
- [10] N. Ruimy, O. Corazzari, E. Gola, A. Spanu, N. Calzolari, and A. Zampolli. The european le-parole project: The italian syntactic lexicon. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain, 1998.
- [11] N. Ruimy, M. Monachini, R. Distanto, E. Guazzini, S. Molino, M. Olivieri, N. Calzolari, and A. Zampolli. Clips, a multi-level italian computational lexicon: A glimpse to data. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas de Gran Canaria, Spain, 2002.
- [12] L. F. Soualmia, C. Golbreich, and S. J. Darmoni. Representing the mesh in owl: Towards a semi-automatic migration. In U. Hahn, editor, *KR-MED*, volume 102 of *CEUR Workshop Proceedings*, pages 81–87. CEUR-WS.org, 2004.
- [13] M. van Assem, A. Gangemi, and G. Schreiber. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006.
- [14] M. van Assem, M. R. Menken, G. Schreiber, J. Wielemaker, and B. Wielinga. A method for converting thesauri to rdf/owl. In *Proceedings of the Third International Semantic Web Conference (ISWC'04)*, number 3298 in *Lecture Notes in Computer Science*, pages 17–31, Hiroshima, Japan, November 2004.
- [15] P. Vossen. Introduction to EuroWordNet. *Computers and the Humanities*, 32:73–89, 1998.