Lexical Relations and Domain Knowledge: The Bio-Lexicon Meets the Qualia Structure

Monica MONACHINI
ILC-CNR
Via Moruzzi 1
Pisa, Italy, 56124
monica.monachini@ilc.cnr.it

Valeria QUOCHI
ILC-CNR
Via Moruzzi 1
Pisa, Italy, 56124
valeria.quochi@ilc.cnr.it

Nilda RUIMY
ILC-CNR
Via Moruzzi 1
Pisa, Italy, 56124
nilda.ruimy@ilc.cnr.it

Nicoletta CALZOLARI
ILC-CNR
Via Moruzzi 1
Pisa, Italy, 56124
nicoletta.calzolari@ilc.cnr.it

Abstract

This paper assumes that the linguistic side of terminologies is necessarily partially informed by the knowledge of the specific domain and claims that semantic relations, especially those accounting for the syntagmatic relations of words in context, are crucial for the representation of this kind of information. This paper also argues that the privileged representational device for encoding these relations is the set of Qualia Structure. This will be shown by presenting a new lexical resource, the BioLexicon, especially designed for the representation of all linguistically relevant aspects of a lexicon for information extraction in the bio domain. The BioLexicon presents several novelties: it follows the most well-known international lexical standards and its semantic layer is inspired by the SIMPLE model which is based on a novel theory of lexical semantics, the Generative Lexicon. We will focus on how the Extended Qualia Structure is particularly suitable and efficient in a resource where lexical and conceptual representation are strongly intertwined, with no neat boundary between the two. We will show how the set of qualia relations can be further augmented and tuned to cope with domain specific semantic information. Bulding up, for the biomedical domain, such a comprehensive terminological resource, which follows lexical and ontological standards and links concepts to lexical items in a linguistic lexicon (containing morphological and syntactic descriptions of the terms) is a huge scientific challenge.

1 Introduction

As the Generative Lexicon theory builds on the idea that lexical semantics is partially informed by world knowledge, this paper starts from the assumption that the linguistic side of domain lexicons is necessarily partially informed by the knowledge of the specific domain and, consequently, semantic relations – especially those accounting for the syntagmatic relations of words in context – appear to be crucial for the representation of this information. The set of Qualia Relations (Pustejovsky, 1995) constitutes the privileged representational device for encoding this kind of notions. This will be argued by describing the BioLexicon, which is being built in the framework of BOOTStrep[1], an EU project for the development of resources and NLP tools for text-based knowledge harvesting in order to support information extraction and text mining in the biomedical domain. The BioLexicon is a lexical resource with interesting features of novelty: it integrates domain terms with lexical information typically contained in a computational lexicon and allows the alignement of term-related syntactic and semantic information with concepts of the BioOntology, the ontological resource of the project[2]. Together, these tightly interconnected resources will constitute a terminological backbone, which is the major innovation of BOOTStrep. Such a combination of lexical and ontological information, indeed, is largely missing from major bio linguistic resources. By linking both types of resources, the project aims at developing, for the bio domain, an innovative integrated resource that is, at the same time, semantically interoperable over knowledge repositories and useful for information extraction purposes.

The main claim of this paper is that the Qualia Structure is particularly suitable and efficient in a resource where lexicon and conceptual

[1] www.bootstrep.eu

[2] The BioOntology is a separate resource: it is a formal ontology designed and represented in OWL. It will be linked to the BioLexicon through an intermediate conceptual layer (the Lexicon Ontology) and together these two resources will become part of the BioKnowlegde Store.

representation are strongly intertwined, with no neat boundary between the two. As a matter of fact, event concepts in the ontology can be mapped onto the lexical semantic level and accommodated in the lexicon through qualia roles, represented as relations among term senses.

The rest of the paper is organized as follows. The next section provides a brief account of the key requirements for a complex lexical resource that should combine different needs from different communities. Section 2 briefly presents the state-of-the-art of existing lexical models, in particular the SIMPLE lexicon model and the LMF standard, the abstract meta-model for the representation of computational lexicons. Section 3 describes the BioLexicon, focussing on its semantic layer and the power of qualia structure. A case-study is presented to show how a specific set of qualia relations can be defined to account for bio-events typically represented in the ontology, and accommodate them in the lexicon. Finally, section 4 presents the conclusions that can be drawn from this experiment.

1.1 Specific Requirements for a Lexical Resource in the Biomedical Field

Given the specific domain of application of the project, i.e. biomedical domain and gene regulation, in particular, there are some special requirements that the BioLexicon model should satisfy. Some of the most important ones involve the type of knowledge which is crucial to represent and the way the lexicon is equipped in view of information extraction. The BioLexicon should be a comprehensive resource that combines together properties of either computational open-domain lexicons or terminologies, and above all should contain most of information pertaining to the semantic dimension of lexical items/terms: in particular, semantic variants of terms, biological processes and events together with their typical participants (argument structure and selectional restrictions on arguments of relevant verbs and event nouns), and various types of relations between entities, e.g. functional relations between proteins. Of particular interest in the present work are those semantic relations relevant for the domain at hand.

Exploring the literature in the field, i.e. the most used gene and protein databases (like UniProt, or ENTREZ Gene) or the best known ontologies (like OBO and GENA), it appears that most of the concepts which are extensively used to describe biochemical terms are in fact types of semantic relations among terms. A few examples are: synonymy and hyponymy/hyperonymy relations; or other semantic relations like `produced_by, phenotype_of, molecular_function_of, has_cellular_components, is_ instance _of, belongs / applies_to_species, applies_to_tissue,` etc.

2 Existing Linguistic Resources and Models

The linguistic resources that can be considered the state-of-the-art in the biological field, namely the Specialist Lexicon (Browne *et al*. 2000), the BioThesaurus (Liu *et al.* 2006), Termino (Harkema *et al.* 2004) present some shortcomings. In a nutshell, the BioThesaurus completely lacks linguistic types of information, which are instead fundamental for text mining and information extraction purposes; the Specialist Lexicon, instead, has no particular focus on molecular biology and moreover, just like as Termino, is not a standardized and reusable resource.

As to lexical models, the SIMPLE model (Lenci *et al.* 2000) and the ISO Lexical Markup Framework (Francopoulo *et al.* 2006a) have been taken into consideration for the design of the BOOTStrep lexical resource.

2.1 SIMPLE Lexicon Model

The SIMPLE model, grounded on the Generative Lexicon theory, was designed for the semantic classification and uniform representation of the content of word senses as well as of the semantic characteristics of their context. In SIMPLE-based lexical resources, the lexicon is structured in terms of a semantic type hierarchy. The *SIMPLE Ontology* is a multidimensional type system, based on both hierarchical and non-hierarchical conceptual relations, and which is tailored to account for the different degrees of internal complexity of word meaning, making explicit its componential and relational nature. In the type system, multidimensionality is captured by *qualia roles*, as conceived in GL theory, that define the distinctive properties of semantic types and differentiate their internal semantic constituency. The SIMPLE ontology distinguishes between *simple* (one-dimensional) and *unified* (multi-dimensional) semantic types, the latter implementing the principle of *orthogonal inheritance* (Pustejovsky and Boguraev, 1993). Each sense of a lemma is encoded as a *Semantic Unit* and is assigned a semantic type, as well as a wide range of additional fine-grained, structured information. Among these, a revised version of the traditional Qualia Structure, the *Extended Qualia Structure* (henceforth EQS), where each of the four *qualia* roles (Formal, Constitutive, Agentive, Telic) is the top of a hierarchy of semantic relations (60 in total). Extended Qualia relations

allow modeling the different meaning dimensions of a word sense and systematically structuring and specifying its relationships to other lexical units, on both the paradigmatic and syntagmatic axes.

2.2 Lexical Markup Framework

A recent activity of the ISO community aims at creating an abstract meta-model for the construction/description of computational lexicons. This activity draws inspiration from the SIMPLE and ISLE/MILE (Calzolari *et al.* 2003) experiences and integrates many of their representational devices. We will not enter in the details of the MILE metamodel here, and will only briefly present the most recent ISO standard.

The ISO meta-model, named Lexical Markup Framework (LMF), provides a common, shared representation of lexical objects that allows the encoding of rich linguistic information, including morphological, syntactic, and semantic aspects.

Its main goal is to allow the mapping of differently conceived lexical and terminological resources, in order to allow for interoperability among even very different resources.

Besides, LMF encoding of linguistic information enables reusability in different applications and for different tasks. LMF is organized in several different packages. The mandatory core package describes the basic hierarchy of information included in a lexical entry. This core package is then enriched with resources that are part of the definition of LMF. These resources include specific data categories used to adorn both the core model and its extensions.

3 The BOOTStrep Lexical Resource

The BioLexicon will be a semantically-enriched lexicon for the biological domain, usable by text analysis and knowledge-capture systems to establish, for a string in a text, a direct link from the lexical to the conceptual layer, i.e. the BioOntology. These two resources, BioLexicon and BioOntology, will then serve as the terminological backbone for harvesting information from documents. Such a linkage is exactly one of the main outcomes that BOOTStrep intends to provide to the biomedical domain.

The BioLexicon is designed to be a reusable and flexible enough resource to adapt to different application needs. Moreover, it is automatically populated both with terms and term related information gathered from already existing electronic resources, and with terms extracted from texts by text mining applications [3]. For these reasons, the lexicon model has to be as extendable and expressive as possible. Since we aim at semantic interoperability in the biomedical community, and by virtue of the characteristic features that have just been highlighted, the ISO Lexical Markup Framework was chosen as the reference metamodel for the structure of the BioLexicon. In its final form, it will account for (English) lexical forms, through lemmatization, and will contain morphological, syntactic and lexical semantic properties of terms from the biology domain. The lexicon will therefore provide those linguistic pieces of information that domain ontologies typically lack, and which instead are crucial to improve text and knowledge mining results.

3.1 The BioLexicon Model: The Semantic Layer

In this section we briefly present the first version of the BioLexicon structure, focusing in particular on the semantic layer.

The BioLexicon model consists of a number of lexical objects (or lexical classes), a set of relations among such objects, and a set of data categories, i.e. attribute-value pairs, some of which drawn from standard repositories, others specifically designed for the project. The BioLexicon is modeled according to the LMF core model plus three NLP extensions for the representation of morphological, syntactic and lexical semantics aspects of terms.

The fundamental lexical objects are: *LexicalEntry*, *Lemma*, *Sense*, and *Syntactic Behaviour*. In the present paper, only those objects that pertain to the semantic level of representation will be described (Fig. 1).

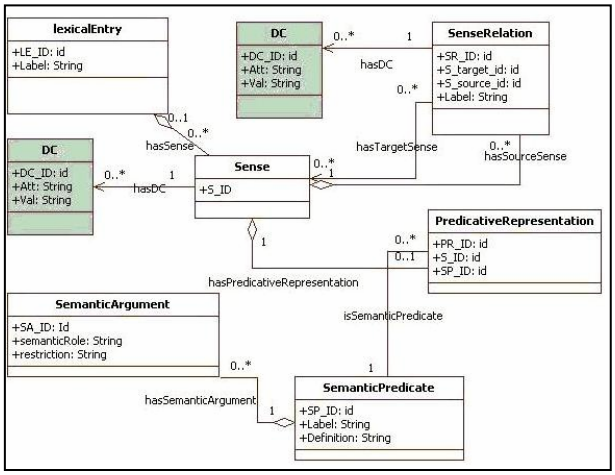

Fig. 1: The BioLexicon Semantic Extension

[3] The gathering of terms from existing resources is performed and curated in particular by one the project partners, EBI; while term extraction from texts will be done by MIB.

The *Lexical Entry* class is used to represent the lexemes of our domain language, i.e. the abstract units of vocabulary, at all three levels of description: i.e. morphology, syntax and semantics. The three levels are accounted for in separate objects (i.e. *Lemma, SyntacticBehaviour, Sense*) independently linked to the LexicalEntry. The basic information units at the semantic level are (word/term) senses. All different information types that are used to discriminate senses either monolingually or multilingually are considered as semantic basic notions. *Sense* is the class used for the representation of the lexical meanings of a word/term; it is inspired by the SIMPLE Semantic Unit and the way it is represented and articulated in the SIMPLE Semantic layer (see 2.1. above). Each *Sense* instance represents and describes one meaning.

3.2 Semantic Relations and Qualia Structure

The notion of semantic relation is fundamental for a full representation of the meaning and contextual behaviour of terms; also they are of much use in structuring lexical data. In the SIMPLE model, they in fact intervene right from the ontological level where qualia roles – which coherently structure the different meaning dimensions that inherently characterize a semantic type – are expressed in templates by qualia semantic relations[4].

Relations may express either (semantic) paradigmatic similarity, typically represented through taxonomical hyperonymy and synonymy links, or (semantic) syntagmatic relatedness, e.g. co-occurrence relations between a predicate and its arguments, and be expressed by non-taxonomical relations such as `is_the_activity_of`, `used_for`, etc. The latter type of semantic links informs on the contextual environment of the term at hand and enables the interpretation of syntagmatic relations of various types, including collocational ones. Such collocation patterns are intimately connected to the term's predicative structure and may be used for different disambiguation purposes, e.g. for resolving structural ambiguities such as prepositional phrase attachment and the relation between constituents in nominal compounds.

In the BioLexicon model, semantic relatedness among terms is expressed through the *SenseRelation* class, which encodes (lexical) semantic relationships among instances of the Sense class (i.e. between meanings of terms). The BioLexicon semantic relations build on the 60 *Extended Qualia relations* of the SIMPLE model: some of them are just a subset of the EQS[5], whereas others are specifically designed for the biological domain. Most of these relations, also shared by well-known ontologies of the biomedical domain, prove to be perfectly suitable to define interrelationships among terms of the domain of interest. To give but a few examples, it is the case of the `is_a` hyperonymic relation of the formal role, e.g. (`is_a` *Interleukin7, protein*); the meronymic/holonymic `part-of` relations `is_a_part_of/has_as_part`, e.g. `is_part_of` (*Cytoplasm, cell*), `has_as_part` (*cell, plasma membrane*), but also of the contains, `typical_of` (*Cox2*, *rat*) or `is_in` constitutive relations; of the agentive `derived_from` and of the telic `is_the_activity_of`. Such relations can therefore be straightforwardly imported into the BioLexicon model and used to relate terms to each others. Clearly, with respect to this relation set, more specific relations, crucial to the domain knowledge, are needed. The flexible design of the EQS, with four hierarchies of relations headed by qualia roles, enables a rather straightforward customization process. On the one hand, it is possible – without corrupting its structure – to draw from the EQS set what is valuable for our enterprise, discarding those few SIMPLE relations[6] which seem of no relevance for the domain of interest. On the other hand, once the knowledge areas of primary importance to the domain are identified, it is possible to create specific, fine-grained semantic relations to account for properties/restrictions of terms and their relationships in order to fully capture the conceptual organization of the biological domain. Taking the OBO semantic relations[7] as reference point, we envisage:

- the creation of new relations in the four hierarchies, e.g. regulates and the inverse relation `is_regulated_by` would more precisely account for the specific relationships holding among terms in utterances such as *'IL2 negatively regulates IL7' / 'Transcription factor regulates cell division'*, while three different links would be necessary using the

[4] SIMPLE templates are schematic structures containing clusters of structured, language-independent information corresponding to the semantic content of ontological types which were proposed to the lexicographers in order to guide the encoding process.

[5] The EQS relations have been submitted as candidates for inclusion in the official ISO Data Category Registry and are currently under evaluation.

[6] e.g. `is_a_member_of`

[7] for a list of relations used in all OBO ontologies http://obo.cvs.sourceforge.net/*checkout*/obo/obo/ontology/OBO_REL/ro.obo

present SIMPLE EQS, namely `is_the_activity_of` (*IL2/ Transcription factor,regulate*); `object_of_the_activity` (*IL7 / cell division, regulate*); `related_to` (*IL2, IL7 / Transcription factor, cell division*).

- The addition of further subtypes to existing relations, e.g. the more granular `has_integral_part`, `has_proper_part` under the constitutive relation `has_as_part`.

Semantic Relations in the BioLexicon are represented as Data Categories drawn from the Data Category Selection specifically defined to meet the needs of the bio-domain and of the BOOTStrep project. Data categories (Wright 2004) represent the main *building blocks* which, in combination with the structures of the lexicon data-model, allow for the creation of many different lexical entries as instances of the same abstract schema (Francopoulo *et al.* 2006b). In fact, they provide the attributes and values that are used to *adorn the objects* [8]. A Data Category Selection is built by gathering part of the standard DCs taken from the Data Category Registry (Ide and Romary 2004) and by defining a set of specific DCs needed for the representation of the domain terminolgy, whenever missing from standard repositories.

3.3 The case of Protein Regulation

The descriptive power of the EQS appears particularly suited for representing the kind of knowledge which is especially crucial for information extraction. In the domain targeted by BOOTStrep, i.e. gene regulation, it offers a workable representation device for recognizing bio-events and functional relations between proteins, e.g. *regulate*, *inhibit*, *block*, typically stored in the ontology. It offers also a means for representing them in the lexicon as relations holding between bio-terms. This is precisely the kind of knowledge that is required in NLP applications, e.g. for the process of anaphora resolution or coreference found in texts (*10-4 M amiloride **inhibits** the regulation of ... This **inhibitor** ...)*.

Several interesting issues should be considered in identifying what goes into the BioOntology, what into the BioLexicon (or in both) and how information should be represented in order for the two resources to be interfaced.

Let the example "*NF-AT positively regulates IL2*" be a sentence expressing a typical mechanism of transcription regulation in gene expression.

Ontologically speaking (Fig. 2), each piece of information found in the text, has a corresponding concept. ProteinPositiveRegulation constitutes an event concept, subclass of the more general Regulation event. The concepts Protein and TranscriptionFactor, in their turn, are semantic types which account for the two event participants, IL2 and NF-AT.

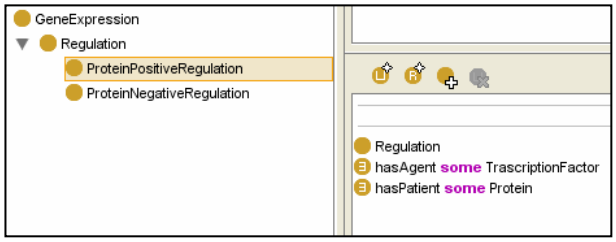

Fig. 2: Protein Regulation in the Ontology

Given the above representation of nodes in the BioOntology, the focus is now on the lexicon and on how this information is mapped onto its semantic layer and translated into the BioLexicon *parlance*. The event Regulation can be automatically imported into the lexical semantic layer, and give rise to an instance of the class SemanticPredicate (Fig. 1).

In the case at hand, Regulation gives rise to the `PredRegulate` [9], whose semantic arguments are instantiated on the basis of the object properties hasAgent and hasPatient provided by the ontology. They correspond to `Arg0Regulate` and `Arg1Regulate`, which are imposed roles and selectional restrictions as indicated from the constraints in the ontology: `Arg0-Agent-TranscriptionFactor`, `Arg1-Patient-Protein`. The predicate `PredRegulate` informs about its potential argument structure in that all lexical instances of the classes TranscriptionFactor and Protein are licensed as possible fillers of, respectively, agent and patient arguments. NF-AT and IL2 (found in the example above) give rise to two semantic entries linked to TranscriptionFactor and Protein respectively, by means of an `is_a` relation. The `PredRegulate`, at the lexical level, is pointed to and shared by the semantic entries "regulate" and "regulation" (Fig. 3), which are proto-typical lexical instances of the Regulation class.

[8] More precisely, a Data Category is *a linguistic constant*, representing the basic linguistic notions: it is either an attribute name like e.g. /semanticRelation/ or a value for the attribute, e.g. /is_the_activity_of/.

[9] Note that for inherent polar verbs, e.g. *up-regulate* and *down-regulate*, a polarity semantic feature would be added to the semantc entry of the term.

```
<LexicalEntry id="LEregulate1">
    <DC att="partOfSpeech" val="verb"/>
        <Sense id="Sregulate1">
            <PredicativeRepresentation predicate="PredRegulate1">
                <SemanticPredicate id="PredRegulate1">
                    <SemanticArgument id="arg0">
                        <DC att="role" val="Agent"/>
                        <DC att="restriction" val="TranscripFactor"/>
                    </SemanticArgument>
                    <SemanticArgument id="arg1">
                        <DC att="role" val="Patient"/>
                        <DC att="restriction" val="Protein"/>
                    </SemanticArgument>
            </Sense>
```

Fig. 3: Assigment of `PredRegulate` to *regulate*

The definition and assigment of predicates to lexical entries, in the BioLexicon, is both inspired by the strategy implemented in SIMPLE, and follows the LMF model. As evidenced by the above description, the predicate constitutes the first connection between the ontology and the lexicon.

In the BioLexicon, however, the mapping between the conceptual and lexical levels will be pushed further: a set of new domain-specific *qualia* relations will be automatically derived from event concepts (of the BioOntology) and constrained to a predicate, allowing to link explicitly noun terms to the event(s) they participate in. In our example, the qualia relations `regulates` and `is_regulated_by` (sub-typed under the Telic role and linked to the predicate `PredRegulate`) are used to connect TranscriptionFactor and Protein-typed term entries (NF-AT and IL2, respectively) involved in gene regulation (Fig. 4). Thanks to the explicit information about source and target semantic entries, they reveal the actual fillers of the regulator and regulatee roles.

```
<LexicalEntry ID="LE_nuclear_factor_activated_T_cell">
    <RepresentationFrame ID="RF_NF-AT">
        <Variant DC writtenForm="NF-AT"
VariantType="Achronym"/>
    </RepresentationFrame>
    <Sense id="S_nuclear_factor_activated_T_cell">
        <SenseRelation targets="S_interleukin-2">
            <DC att="TelicSemanticRelation" val="regulates"/>
        </SenseRelation>
    </Sense>
</LexicalEntry>

<LexicalEntry ID="LE_interleukin-2">
    <RepresentationFrame ID="RF_IL2">
        <Variant DC writtenForm="IL2" VariantType="Achronym"/>
    <Sense id="S_interleukin7"> </sense>
```

Fig. 4: `regulate` relation links *NF-AT and IL2*

Besides creating a dense network of information encoded in both resources, the reinforcement of lexicon-ontology interconnections (Fig. 5) allows to determine event representations, interaction patterns and constraints between biological entities.

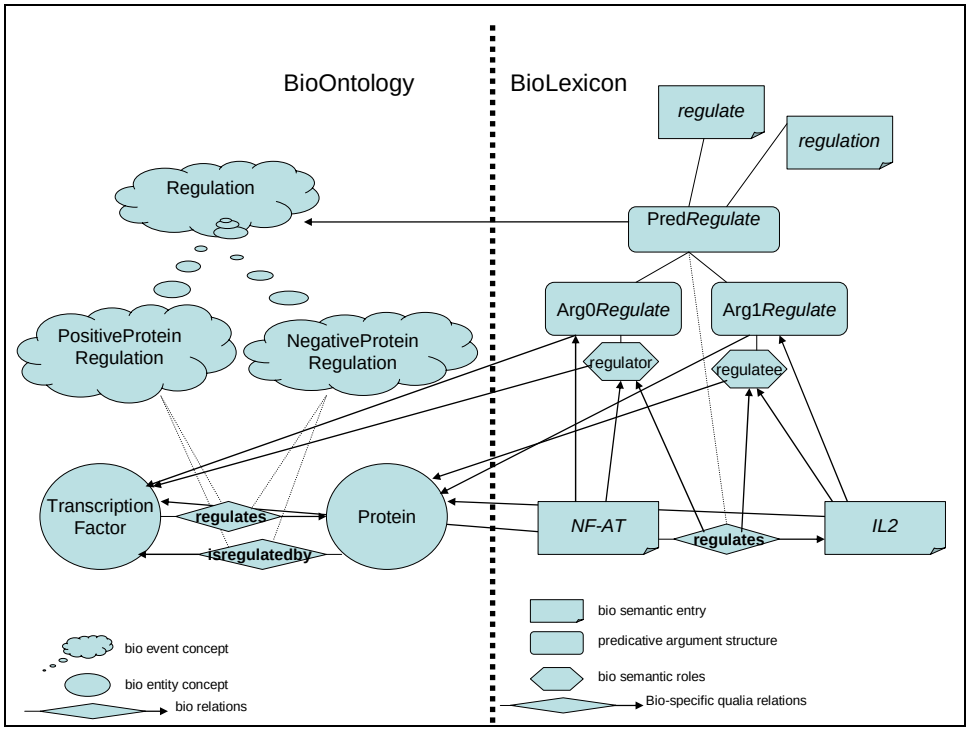

Fig. 5: Lexicon – Ontology Interconnections

The apparent redundancy of semantic information in the two resources guarantees, in fact, a better efficiency for text mining and information extraction: the bio-event-specific semantic relations used throughout the lexicon, enable to harvest, immediately and explicitly, those bio-entities which are actual fillers of agent and patient roles and to constrain them to the relevant bio-event, without needing to go to and from the two resources.

4 Conclusion

Since in the biology/biochemistry domain new papers and written material are continuously being produced, there is an increasing demand for comprehensive terminological resources that allow for semantic interoperability and especially that support text-based knowledge harvesting, through NLP techniques. Existing bio-ontologies and bio-terminologies lack linguistic information about terms, which is instead fundamental for text mining applications. Open-domain computational lexicons, on their turn, provide little information related to the biological domain.

The BioLexicon, being specifically designed to target biomedical domain requirements, is a resource that integrates features of both terminologies and lexicons and constitutes one of the BOOTStrep innovations with respect to related works. The BioLexicon is also an extendable resource, able to be incremented with entries and lexical properties for bio-terms and bio-events automatically extracted from texts. The other novelty is the alignment of term-related syntactic and semantic information with concepts from the BioOntology.

In this paper, we have shown how the Extended Qualia representation device is particularly suited for making the connections between the BioLexicon and the BioOntology possible, and, in particular, for explicitly representing the domain-specific roles of event participants. We described how a set of specific qualia semantic relations, which allow to capture the lexical fillers for the argument slots, can be derived from an event concept in the ontology. This way, it is possible to explicitly gather lexical fillers of predicate semantic (and also syntactic) arguments, without having to look-up the semantic type hierarchy in the ontology.

Concluding, by linking both the lexical and ontological resources, BOOTStrep aims at creating an innovative integrated resource suitable for the tasks of information extraction in the bio domain. Building up such a comprehensive terminological resource in a way that follows both lexical and ontological standards constitutes a scientific challenge.

As a matter of fact, the BioLexicon, which adheres to the most innovative theoretical approaches in lexical semantics and the most recent lexical standards, shows that the state-of-the-art both in the bio-domain and in NLP lexicons is mature enough for us to aim at creating a terminological-lexical resource which is candidate to become "*the*" standard in this domain.

5 Acknowledgements

This research was funded by the EC´s 6th Framework Programme (4th call) and conducted within the BOOTStrep consortium under grant FP6-028099.

References

Browne A.C., Mc Cray A.T., Srinivasan S. 2000. The SPECIALIST Lexicon, National Library of Medicine, Bethesda, Maryland.

Calzolari N., Bertagna F., Lenci A., Monachini M. (eds) 2003. Standards and best Practice for Multilingual Computational Lexicons. MILE (The Multilingual ISLE Lexical Entry). ISLE CLWG Deliverable D2.2 & 3.2 Pisa.

EntrezGene http://www.ncbi.nlm.nih.gov/entrez/

Francopulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006a. Lexical Markup Framework (LMF) In Proceedings of the LREC2006, Genova, Italy.

Francopulo G., Monachini M., Declerck T., Romary L., 2006b. Morpho-syntactic Profile in the ISO-TC37/SC4 Data Category Registry. In Proceedings of the LREC2006, Genova, Italy.

Harkema H., Gaizauskas R., Hepple M., Angus R., Roberts I., Davis N., Guo Y. 2004. A Large Scale Teminology Resource for Biomedical Text Processing. HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Database, Boston, Massachusetts, USA.

Ide N. and Romary L. 2004. A registry of standard data categories for linguistic annotation. In Proceeding of the LREC04, Lisbon, Portugal.

ISO-12620 2006. Terminology and other content language resources – Data Categories – Specifications of data categories and management of a Data Category Registry for language resources. ISO/TC37/SC3/WG4.

Lenci A., Busa F., Ruimy N., Gola E., Monachini M., Calzolari N., Zampolli A., Pustejovsky J. et al. 2000. Linguistic Specifications. SIMPLE Deliverable D2.1. Pisa: ILC-CNR.

Pustejovsky J. 1995. *The Generative Lexicon* Cambridge Mass: MIT Press.

Pustejovsky J. and Boguraev B. 1993. Lexical Knowledge Representation and Natural Language Processing. *Artificial Intelligence* 63: 193-223.

Liu, H., Hu, Z.Z., Zhang, J., and Wu, C. 2006. BioThesaurus: a web-based thesaurus of protein and gene names. In *Bioinformatics*, 22(1): 103-5.

UniProt http://www.pir.uniprot.org/

Wright S.E. 2004. A global data category registry for interoperable language resources. In *Proceedings of LREC04*, Lisbon, Portugal.