Using LMF to Shape a Lexicon for the Biomedical Domain

Monica Monachini, Valeria Quochi, Riccardo Del Gratta, Nicoletta Calzolari

Istituto di Linguistica Computazionale – CNR, Pisa, Italy

name.surname@ilc.cnr.it

Abstract

This paper describes the design, implementation and population of the BioLexicon in the framework of BootStrep, an FP6 project. The BioLexicon (BL) is a lexical resource designed for text mining in the bio-domain. It has been conceived to meet both domain requirements and upcoming ISO standards for lexical representation. The data model and data categories are compliant to the ISO Lexical Markup Framework and the Data Category Registry. The BioLexicon integrates features of lexicons and terminologies: term entries (and variants) derived from existing resources are enriched with linguistic features, including subcategorization and predicate-argument information, extracted from texts. Thus, it is an extendable resource. Furthermore, the lexical entries will be aligned to concepts in the BioOntology, the ontological resource of the project. The BL implementation is an extensible relational database with automatic population procedures. Population relies on a dedicated input data structure allowing to upload terms and their linguistic properties and "pull-and-push" them in the database. The BioLexicon teaches that the state-of-the-art is mature enough to aim at setting up a standard in this domain. Being conformant to lexical standards, the BioLexicon is interoperable and portable to other areas.

Index Terms: domain terminologies, computational lexicons, lexical standards, lexical architectures

1. Motivation and background

Bio-literature is continuously being produced and new knowledge is continuously being developed and it is of paramount importance to share and disseminate knowledge in the biomedical domain especially for boosting and supporting discoveries of new illnesses, treatments, medicaments, and similar. The reuse of information however requires time and efforts because it needs to integrate often redundant and partial pieces of information, which are often stored in different formats.

Intensive research has being carried out to develop language technologies that provide intelligent access to such knowledge and build lexical and ontological resources targeted to fulfill special demands for the biologist community: i.e. normalized nomenclatures (see Kors et al. 2005), extensible databases for storing terminological information like Termino (Harkema et al. 2004), lexical and ontological resources like the SPECIALIST lexicon. Still, access and interoperability of biological databases is hampered, due to persistent lack of structuring and uniformity of formats. Moreover, available bio-terminologies lack information relevant to knowledge extraction, such as predicate argument structures and syntactic complementation patterns. A comprehensive and continuously growing resource where bio-terms from different sources are integrated, encoded on the basis of the most accredited standards, enriched with relevant linguist description and linked to concepts in the ontology would significantly improve text analysis and knowledge capture systems (Hahn and Markó 2001). One of the main resources of BOOTStrep knowledge core is the BioLexicon: an expected state-of-the-art lexical resource that meets both bio-domain requirements and the most recent standards for lexical representation. The BioLexicon is an integrated resource in that it is semi-automatically populated with data collected from different available biomedical sources (e.g. UniProt/ Swiss-Prot, ChEBI, BioThesaurus, NCBI taxonomy) and is further integrated with morphological, syntactic and lexical semantic features either extracted from texts and or from domain ontologies.

2. The BioLexicon

The BioLexicon is a computational lexicon for the biology domain, designed to be reusable and flexible enough to adapt to different application needs: e.g. text mining, information extraction, information retrieval. The BioLexicon accounts for (English) lemmas and terms related to the bio-domain and contain morphological, syntactic and lexical semantic properties of them.

Since one of our main aims is to foster semantic interoperability in the community, the ISO Lexical Markup Framework (Francopoulo et al. 2006a) was chosen as the reference meta-model for the structure of the BioLexicon. The Lexical Markup Framework provides a common and shared representation of lexical objects that allows for the encoding of rich linguistic information. The BioLexicon is modeled in an XML DTD according to the LMF DTD: it implements the core model plus objects taken from the NLP extensions for the representation of morphological, syntactic and lexical semantics aspects of words and terms. The model consists of a number of independent lexical objects (or classes) and a set of Data Categories (DCs), i.e. attributevalue pairs which represent the main building blocks of lexical representation. In conformity to the ISO philosophy, the Data Category Selection for the BioLexicon is partially drawn from the ISO 12620 Data Category Registry (Francopoulo et al. 2006b,c, Wright 2004), and partially defined for the specific purposes of the project and the special domain. Furthermore, in order to be able to automatically constrain and check the consistency of the DCs on each specific object most DCs have been typed.

A key innovation is that the DB comes equipped with automatic loading procedures for its population with data coming from partners. Also, the BioLexicon will be linked to the BioOntology, and the two will serve as the terminological backbone for harvesting information from documents.

3. The BioLexicon data model

The core lexical objects of the BioLexicon are: *LexicalEntry*, *Lemma*, *Sense*, and *Syntactic Behaviour*.

The *Lexical Entry* class represents the abstract units of vocabulary at three levels of description: morphology, syntax and semantics. To ensure modularity and extendibility the three levels of description are accounted for in separate lexical objects, independently linked to the *LexicalEntry*, which, thus, functions as a bridge among the *Lemma* – and their forms – its related *Sense(s)*, and *Syntactic Behavior(s)*. *Lexical Entry* bears a Part-Of-Speech DC, plus additional non mandatory attributes.

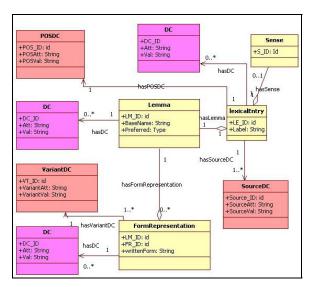


Figure 1: The BioLexicon morphology extension.

A specific requirement coming from the biology community is that the resource should keep track of the ids of the terms in other well known reference databases and ontology (see Harkema et al. 2004). External references in the BioLexicon are thus represented as typed data categories that are added as attributes to the *Lexical Entry* object. *Lemma* is used to represent the base form of lexemes plus additional grammatical properties; because it is in a one-to-one relation with the *Lexical Entry*, homonyms in the BioLexicon are represented as separate entries.

Syntactic Behaviour is dedicated to the representation of how lexical items and terms are used in context. One Syntactic Behaviour describes specific syntactic properties of an item related to one of the possible contextual behaviors of a lexical entry.

Finally, the basic information units at the semantic level are senses. *Sense* is therefore the class used for the representation of the lexical meanings of a word/term, and it is inspired by the SIMPLE Semantic Unit (Ruimy *et al.* 2003). Each *Sense* instance represents and describes one meaning of a given *Lexical Entry*, contains information on the specific (sub)domain to which the sense applies, and contains a link to the Bio-ontology.

3.1. The morphology extension

In a terminological lexicon for biology a key requirement is the representation of the different types of term variants. Variants in fact are extremely frequent and common in the biology literature (Nenadic et al. 2004). Given that linguistic information are automatically extracted from texts, in the BioLexcon we chose to distinguish only between two types of variants: variants of form and semantic variants. The

morphology extension therefore has been implemented mainly to allow for a rich and extensible representation of variants of form. The *FormRepresentation* object has in fact the function of representing multiple orthographies. The basic DC specifying the *FormRepresentation* is the *writtenform*, i.e. the string identifying the form in question. Each variant is then adorned with properties represented by specific DCs: the type of variants ("orthographic", for variants and "preferred" for baseforms), and a confidence score that the automatic extraction techniques assigned to each variant (for details on the treatment of variants see Quochi et al 2007). The *InflectedForm* class is used in the BioLexicon to represent the automatically generated inflected forms of domain-relevant verbs

3.2. The syntactic extension

As mentioned above, *Syntactic Behavior* represents one of the possible behaviors that a lexical entry shows in context. A detailed description of the syntactic behavior of a lexical entry is further defined by the *Subcategorisation Frame* object, which is the "hearth" of the syntax module.

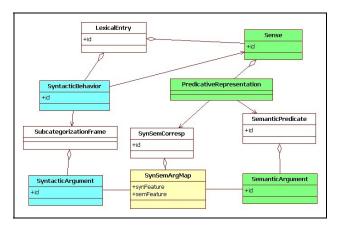


Figure 2: The BioLexicon syntactic extension.

Subcategorisation Frame is used to represent one syntactic configuration and does not depend on individual syntactic units; rather it may be shared by different units. The LMF syntax extension is adapted in view of accommodating the subcategorisation behaviors of terminological verbs automatically extracted from texts by appropriate NLP algorithms, and thus a probability score will be recorded as a property of the Syntactic Behavior belonging to a give SubcategorisationFrame.

3.3. The semantic extension

The semantic module of the lexicon is made of lexical objects related to the *Sense* class. As said above, *Sense* represents lexical items as lexical semantic units. Semantic relatedness among terms is expressed through the *SenseRelation* class, which encodes (lexical) semantic relationships among instances of the Sense class. The BioLexicon *Semantic Relations* build on the 60 *Extended Qualia relations* of the SIMPLE model and 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 (for details on bio-relations and the semantic extension in general see Monachini et al. 2007).

The Semantic Predicate class, instead, is independent from specific entries and represents an abstract meaning together with its associated semantic "arguments". It represents a meaning that may be shares by more senses that are not necessarily considered as synonyms. It is referred to by the *Predicative Representation* class, which represents the semantic behavior of lexical entries and senses in context, i.e. it describes the complete semantic argument structure of a predicative lexical item.

4. The BioLexicon data base

The software implementation of the BioLexicon consists of two modules: a relational database MySQL and a java-based loading software for the automatic population of the database. External to the DB, but fundamental for its automatic population, is an XML Interchange Format (XIF hereafter) specifically tailored to the BioLexicon structure.

4.1. Database Architecture

The database is structured into three logically distinct but strongly interconnected layers (see Figure 3). The TARGET FRAME layer contains the actual BioLexicon tables, i.e. tables that directly instantiate the lexical objects and relations designed in the conceptual model presented in the sections above and defined in a corresponding DTD.

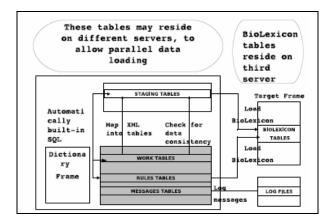


Figure 3: The BioLexicon database architecture

Each module of the BioLexicon structure (syntax, semantics, morphology) is independently accessible by queries and represents a self-consistent set of data. Each table has its internal identifier, but native identifiers, i.e. original source identifiers, are maintained as attributes of a given lexical entry (i.e. as DCs). The other two layers, DICTIONARY and STAGING, may be considered as operational layers: the DICTIONARY contains rules to populate target tables, whereas STAGING is an intermediate set of hybrid tables for the storage of volatile data: staging table columns consist of attributes of the XML Interchange Format (see below) and attributes of target tables. In addition, the staging frame is the level dedicated to data-cleaning and normalization. This neat separation between target tables (the BioLexicon proper) and operational tables allows for the optimization of the uploading of data into the BioLexicon DB and ensures extensibility both of the database and of the uploading procedures. Faced with the need to add a new table (i.e. a new lexical object) or a new attribute to an existing table, for example, it is sufficient to add only the definition of the new tables or attributes to the DB and to modify accordingly only the relevant portion of the Dictionary Frame layer to handle the novelties.

4.2. The XIF and Automatic Population

The XML Interchange Format (XIF) is designed with the main purpose of automatically populating the BioLexicon with data provided by domain experts and by lexical acquisition systems. Within the project, data are extracted and gathered through automatic procedures both from existing resources and by research papers in biology. The XIF DTD is to be considered a simplified version of the BioLexicon DTD which accommodates the needs of data providers and facilitates the automatic uploading of the DB. By means of the XIF, we therefore allow for a standardization of the data extracted from the different terminological resources and from texts. Differently from other similar lexical systems (like Termino), the XIF allows for the independency of the uploading procedures from native data formats. This way, any system/group wishing to feed new data into the BioLexicon would only need to encode this in an XML file according to the XIF DTD. The XIF DTD partially mirrors the way biological data are stored into domain knowledge databases and also accommodates the way these data are extracted from those resources. The XIF is organized in clusters of terms, i.e. in sets of coherent types of information. A cluster contains one or more synonymous entries with information related to their lemmas, parts-of-speech, inflected forms, semantic relations and external references. Such an organization, furthermore, permits the splitting of the input file by clusters, which in turn allows for a parallel uploading of the data into the DB. The XIF, therefore, has been conceived as a link between existing resources and the BioLexicon. From the implementation perspective, the XIF may be considered as the physical counterpart of the Dictionary Frame: that is to say, the loading software uses rules contained in the dictionary tables to correctly interpret the input file. Faced with the need to add a new table or to alter an existing one, it is sufficient to add new elements or new attributes to the XIF DTD and to add relevant instructions to the dictionary frame. The loading software interprets the XIF as usual and applies the new Dictionary instructions automatically inserting the new table or attribute. This property of the XIF together with the neat separation of the three layers of the DB mentioned above allows any agent (human or machine) to easily populate, update and create new objects and attributes. The data that we are currently uploading are terms gathered by Project BOOTStrep partners from existing databases with information relevant to their external source references, variants, lexical category and, to some degree, semantic relations.

5. Statistics and Validation

Since we are dealing with the building of a lexical resource within an ongoing project, no evaluation is available yet. However, some kind of content validation can be made, taking into account the input resources and the documentation so far produced. For validation of the resource we readapt templates from the ELRA Validation Manual for Lexica 2.0 (Fersøe, 2004).

The BioLexicon is a monolingual English lexical resource that represents Bio-terms as well as general lexical items relevant to the bio-medical domain. Both nouns and verbs are represented: nouns cover a wide portion of the existing biological terminologies and come from the most used databases in the sector. Mainly they represents terms denoting enzymes, chemical, species, genes and/or proteins, especially those relevant for the gene regulation topic. Verbs are also

represented, but are very limited in number: for the time being only verbs relevant to the E.Coli species have been included (see Table 1).

Table 1: The BioLexicon coverage

POS	Sem.Type	Lexical Entries	Variants
N	Enzymes	4,016	9,379
N	Genes/Proteins	841,164	1,547856
N	Species	367,565	439,336
N	Chemicals	13437	51332
POS	Sem.Type	Lexical Entries	Infl. Forms
V		489	2435

For each entry the part-of-speech is encoded together with the written form of both its lemma and its variants. Also some semantic relations are instantiated: synonymy, part-of, and a few other biological relations (see Table 2).

Table 2: Instantiated semantic relations

Relation Type	#
is_a	464,078
is_synonym_of	847,873
is_part_of	189
is_conjugate_base_of	637

The BioLexicon resource is implemented as a MySQL relational database that runs both under Linux and Windows systems. The database is shaped according to the model XML DTD, and therefore easily allows for XML outputs. So far, the DB has automatically uploaded all the input files provided by the bio-experts within the BOOTStrep project (EBI-Cambridge and MIB Manchester), which gathered and systematized biological terminology from the major online databases. Altogether the DB contains 25 million records and occupies ca. 1.7G of memory space. It consists of 1,309,496 Lexical Entries, Lemmas and Senses; 2,047,903 orthographic variants and 1,697,804 semantic relations.

6. Conclusions

The biological literature is continuously developing, which leads to the need for large-scale terminological lexicons that can support text mining and information extraction applications, which would make the life of biologists much easier. The BioLexicon, described in this paper, is designed to integrate both typical information provided by domain ontologies and typical linguistic information generally available in open-domain computational lexicons. The DB, as well as the BioLexicon model, is modular, extensible and, by means of the protocol defined through the XIF, can easily and automatically upload new data, and provide outputs by means of web services. A brief hint at internal validations of the resource has been added, based on the first data coming from our project partners. A suitable evaluation of the resource is not feasible for the moment, and will only be possible in the future, when the DB will actually be integrated in the BOOTStrep UIMA infrastructure; that is when it starts to dynamically interact with applications.

7. Acknowledgements

The work presented here has been funded by the EC's 6th Framework Programme (4th call), conducted within the BOOTStrep consortium under grant FP6-028099.

8. References

- [1] 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.
- [2] Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006a. Lexical Markup Frame-work (LMF). Proceedings of the LREC 2006, Genova, Italy.
- [3] Francopoulo G. Monachini M., Declerck T., Romary L.. 2006b. The relevance of standards for research infrastructure. Proceeding of the LREC 2006, Genoa, Italy
- [4] Francopoulo G., Monachini M., Declerck T., Romary L., 2006c. Morpho-syntactic Profile in the ISO-TC37/SC4 Data Category Registry. In Proceedings of the LREC2006, Genova, Italy.
- [5] Hahn U., Markó K. 2001. Joint Knowledge Capture for Grammars and Ontologies. Proceedings of the 1st international conference on Knowledge capture Victoria, British Columbia, Canada.
- [6] Harkema H., Gaizauskas R., Hepple M., Angus R., Roberts I., Davis N., Guo Y. et al. 2004. A Large Scale Teminology Resource for Biomedical Text Processing. HLT-NAACL 2004 Workshop: Bio-LINK 2004, Linking Biological Literature, Ontologies and Database, Boston, Massachusetts, USA.
- [7] Ide N. and Romary L. 2004. A registry of standard data categories for linguistic annotation. In Proceeding of the LREC04, Lisbon, Portugal.
- [8] 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.
- [9] Kors J. A. et al. 2005. Combination of Genetic Databases for Improving Identification of Gens and Proteins in Text, Rotterdam, Netherlands
- [10] ISO-12620. 2006. "Terminology and other content language resources – Data Categories – Specifications of data categories and management of a Data Cate-gory Registry for language resources. ISO/TC37/SC3/WG4.
- [11] Monachini M., Quochi V., Ruimy N., Calzolari N. 2007. Lexical Relations and Domain Knowledge: The BioLexicon Meets the Qualia Structure. In GL2007: 4th International Workshop on Generative Approaches to the Lexicon, 10-11 May 2007, Paris. CD-ROM.
- [12] Quochi V., Del Gratta R., Sassolini E., Monachini M., Calzolari N. 2007. Toward a Standard Lexical resource in the Bio Domain. In Vetulani (ed.), Proceedings of 3rd Language & Technology Conference. Fundacja Uniwersytetu im A. Mickiewicza, Poznań. 295-299.
- [13] Ruimy N., Monachini M., Gola E., Calzolari N., Del Fiorentino M. C., Ulivieri M., Rossi Sergio. 2002. In Linguistica Computazionale, Vol.XVIII-XIX, I.L.C. and Computational Linguistics, special issue, A. Zampolli, N. Calzolari, L. Cignoni, (Eds.), I.E.P.I., Pisa-Roma.
- [14] Wright S.E. 2004. A global data category registry for interoperable language resources. In Proceedings of LREC04 Lisbon, Portugal.