LDK 2023

ISO LMF 24613-6: A Revised Syntax Semantics Module for the Lexical Markup Framework

Francesca Frontini, Anas Fahad Khan
Cnr-Istituto di Linguistica
Computazionale "Antonio Zampolli"
Pisa, Italy
name.surname@ilc.cnr.it

Laurent Romary
Inria - Scientific Information
and Culture Directorate
Paris, France
laurent.romary@inria.fr

Abstract

The Lexical Markup Framework (LMF) is a meta-model for representing data in monolingual and multilingual lexical databases with a view to its use in computer applications. The "new LMF" replaces the old LMF standard, ISO 24613:2008, and is being published as a multi-part standard. This short paper introduces one of these new parts, ISO 24613-6, namely the Syntax and Semantics (SynSem) module. The SynSem module allows for the description of syntactic and semantic properties of lexemes, as well as the complex interactions between them. While the new standard remains faithful to (and backwards compatible with) the syntax and semantics coverage of the previous model, the new standard clarifies and simplifies it in a few places, which will be illustrated.

1 Introduction

The Lexical Markup Framework (LMF) is undoubtedly one of the most influential lexical standards of the last two decades. First published in 2008 by the International Standards Organization (ISO) as **ISO standard 24613:2008** it was intended as a "standardized framework for the construction of computational lexicons" (Francopoulo, 2013). LMF was developed with a special focus on two different kinds of lexicon, namely, digital born electronic lexicons specifically intended for use by Natural Language Processing applications, so called *NLP dictionaries*, as well as for electronic versions of print dictionaries, or more generally lexicons primarily intended for human consumption, so called *Machine Readable Dictionaries* (MRD). The original LMF, ISO 24613:2008, contained, two modules for syntax and semantics, respectively, whose scope, taken together, was to provide means of representing the syntactic and semantic argument structure of individual lexical entries. The approach taken by the original committee tasked with drafting LMF was a theory agnostic one which identified a nucleus of elements that were generic enough to allow for the modelling of syntax, semantics and their interface without any particular theoretical bias. After its publication in 2008, LMF came to be used by a variety of different organisations and in a number of national and international projects[1]. In particular, the syntactic and semantics models were extensively used in projects such as PAROLE and SIMPLE (Ruimy et al.; Lenci et al.) as well as being the basis for other models of the syntax/semantics interface in lexical resources, such as the W3C OntoLex Syntax and Semantics Module[2].

After a detailed review of the original standard, however, the decision was made in 2015 to revise LMF and, what's more, to make it a multi-part standard with each part being published separately (as distinguished from the old LMF standard which was published in a single part but which contained separate modules as sub-parts). This new multipart version of LMF is currently being developed within the standardisation sub-committee ISO TC 37/SC 4/WG 4 (to which the authors of the current article are all contributing), with the first five parts of the new version having already been published, and other parts at an advanced stage of completion. The current paper is dedicated to ISO 24613-6, a soon-to-be published part of the revised LMF standard dealing with Syntax and Semantics (henceforth **SynSem**), two areas which as we mentioned above were previously covered by separate modules in the old LMF. SynSem stays true to the overall approach of ISO 24613:2008, but some simplifications/modifications were introduced. In what follows, we shall begin by placing SynSem in the context of the new multipart LMF, and providing an update as to its current status. Then we shall describe the constituent parts of the standard: Syntax,

[1] Searching for "LMF" in the CLARIN Virtual Language Observatory gives a good indication on resources and also tools using the 2008 model (https://vlo.clarin.eu/?1&q=lmf.

[2] https://www.w3.org/community/ontolex/wiki/Syntax_and_Semantics_Module

Semantics, and SynSem interface. Finally we shall provide some details as to its serialisation.

2 An Overview of the New Multipart LMF

Following (Romary et al., 2019) we provide a list of the new LMF parts in the present section along with their current status.

ISO 24613-1:2019 Language resource management — Lexical markup framework (LMF) — Part 1: Core model: This module defines the basic classes required to model a baseline lexicon and is a pre-requisite for the use of the other classes. **Status**: *Published in 2019 it is now being further revised to make it even easier to use.*

ISO 24613-2:2020 Language resource management — Lexical markup framework (LMF) — Part 2: Machine-readable dictionary (MRD) model: Contains components providing a deeper specification of lexical description encapsulated within the core model. **Status**: *Published in 2020*.

ISO 24613-3:2021 Language resource management — Lexical markup framework (LMF) — Part 3: Etymological extension: A completely new addition to the LMF meta-model covering etymological and diachronic information. This part makes etymologies, etymological links and etymons first class citizens. See (Khan and Bowers, 2020) for more details. **Status**: *Published in 2021*.

ISO 24613-4:2021 Language resource management — Lexical markup framework (LMF) — Part 4: TEI serialization: A TEI serialisation of the other parts of the model which aims to make both TEI and LMF fully compatible and which leverages the knowledge and makes use of the established practices of the TEI community in dealing with lexicographic resources. **Status**: *Published in 2021*.

ISO 24613-5:2022 Language resource management — Lexical markup framework (LMF) — Part 5: Lexical base exchange (LBX) serialization: Another XML serialisation. **Status**: *Published in 2022*.

ISO/CD 24613-6 Language resource management — Lexical markup framework (LMF) — Part 6: Syntax and Semantics. **Status**: *A candidate for an ISO Draft International Standard (DIS) ballot*.

3 The New SynSem module

Figure 1 gives the SynSem class diagram. The classes in white (*Lexical Entry*, *Sense* and *SenseRelation*) are inherited from the LMF core (Part 1), while the salmon-pink coloured classes are newly defined in Part 6. Notably, Part 6 introduces two important new classes which provide the means to describe both the *Syntactic Behaviour* of entries and the *Predicative Representation* of senses as well as allowing for the specification of connections between the two. The main difference with respect to ISO 24613:2008 is the absence of the previously defined *Synset* class. Indeed the semantic module of the prior version of LMF contained elements that were entirely dedicated to the modelling of WordNet-like lexicons. However, this was not judged to be necessary in the current standard since the *Sense* and *SenseRelation* classes can be used instead.

Another crucial difference with respect to the former version of LMF is the lack of a *feat* class, formerly used to make up for specific elements which a lexicographer may want to introduce but which were not generic enough to be included in the model. In the old model, class arguments could be specified as pairs of attributes of the specific tag *feat*: *att* would contain the name of the attribute, and *val* the value. In the new model, attributes can be added as needed; in Figure 3 for example a *SemanticArgument* can be specified in terms of *type* and *restriction*. Generally speaking – and here guided by the same principle already introduced for other parts – only the core features of the syntax and semantics interface are described in the present UML based standardisation, however the user can extend the model to add other features.

Regarding the modelling of syntax in Figure 1, a *LexicalEntry* may have one or more instances of *SyntacticBehaviour*, associated with separate *SubcategorizationFrame* instances, each described with *SyntacticArgument*. As for the modelling of semantics, it applies to senses. The *Sense* class, which is specified in the core package, is aggregated in the *LexicalEntry* class. A *PredicativeRepresentation* serves to connect a *Sense* with one or more instances of *SemanticPredicate*, which are described in terms of *SemanticArgument* instances. Linking between syntax and semantics is done by the *SynSemArgMap* component, which links a *SemanticArgument* with a *SyntacticArgument*.

In modelling semantics, allowance is made in

Part 6 for drawing from other relevant standards. In particular **ISO 24617-4:2014 (en) - Language resource management — Semantic annotation framework (SemAF) — Part 4: Semantic roles (SemAF-SR)** provides a background terminology and methodology for designing a semantic role scheme in a coherent way, based upon the work carried out in the LIRICS projects ((Petukhova and Bunt, 2008). The examples provided in this paper illustrate the use of such roles without providing a normative list thereof.

3.1 Examples

In this section we will illustrate Part 6 in more detail by means of an exhaustive example (Figure 3), drawn from the Parole Simple CLIPS Italian lexicon [3]. The example contains two lexical entries, the Italian verb *costruire* ('to build') and the deverbal noun *costruzione* ('a building'). For simplicity's sake, in this example each entry has just one sense (though many are possible), each linked to a separate *PerdicativeRepresentation*, but these are in turn linked to just one *SemanticPredicate* (PREDcostruire-1). The predicate is described with its two arguments to which are added semantic *roles*, and *restrictions* (the latter represented by types in the SIMPLE ontology (Del Gratta et al., 2015)). From the syntactic point of view, a *SyntacticBehaviour* element links the *LexicalEntry* to a *SubcategorizationFrame SCFtxa*, representing the transitive construction, which is in turn described by its two syntactic arguments (subject and object). A *SynSemCorrespondence* component (of type *ISObivalent*) allows for a mapping between each pair of syntactic/semantic arguments. In this rather straightforward case, the subject maps onto the agent and the object onto the patient. Finally a further diagram (Figure 2) illustrates how syntactic alternations can be represented. In the example, which represents the anti-causative syntactic alternation, a *SubcategorizationFrameSet* has been created to connect two *SubcategorizationFrames* that can be subject to alternation, as in the case of the transitive and intransitive in verbs such as *bollire* ('boil')[4]. The *SynArgMap* class can also be used to represent the link between syntactic arguments: in this case the representation tells us that the object in the transitive construction becomes the subject of the intransitive one.

3.2 Serialisation

We designed the serialisation of ISO LMF 24613-6 as an extension of the TEI guidelines[5]. In doing so, we wanted to achieve the following objectives:

- Maintain coherence with the overall serialisation framework for LMF which has already set out a dedicated TEI subset covering parts 1, 2 and 3 within the ISO LMF 24613-4 standard;
- Benefit from the TEI specification language ODD ("One Document Does it all") which provides a flexible framework compatible with literate programming principles and which allows for the generation of both schemas (DtD, RelaxNG, W3C) and documentation from a single specification document;
- Integrate the specific development of LMF syntax and semantic descriptions within a broader lexicographic landscape in which the the TEI guidelines have been widely adopted (also within the framework of the TEI Lex 0 initiative[6]) for maintaining sustainable lexical resources, which are thus FAIR by construction.

More precisely, we integrated SynSem components at three specific places within the standard structure of a TEI lexical entry:

- We added a <syntacticBehaviour> element to the possible grammatical descriptions associated with a lemma (within the TEI <gramGrp> element) that points to a sub-categorisation frame (see below);
- The content of the TEI <sense> element was expanded to contain a <predicativeRepresentation> element with references to a semantic predicate and possible syntactic-semantic correspondences;
- We extended the general intermediate of a TEI document to allow <subcategorizationFrame>, <SemanticPredicate> and <SynSemCorrespondence> elements to occur freely and be referred to from <syntacticBehaviour> and <predicativeRepresentation> within entries.

[3] http://hdl.handle.net/20.500.11752/ILC-88

[4] This example works in English ("I boil the water/The water boils")

[5] https://tei-c.org/guidelines/

[6] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

All content relating to the serialisation of 24613-6 is available from the DARIAH WG on lexical resources[7].

4 Conclusion

The new ISO LMF 24613-6 will soon be available as a published standard. Resources encoded in the previous model are easily converted to the new one, which remains overall backward compatible. Another crucial task will involve developing user-friendly conversion methodologies for other commonly used formats, particularly OntolexLemon, by defining convenient crosswalks. This would, among other things, provide an easy way to go from tree based TEI-XML representations to RDF-based graph-like representations, thus potentially contributing to the extension of the Linguistic Linked Open Data Cloud.

Acknowledgements

The work described in this paper was carried out as part of the activities of the CLARIN-IT national consortium; aspects concerning the link between the ISO standard and other formats were also explored as part of the activities of the COST Action NexusLinguarum – "European network for Web-centered linguistic data science" (CA18209), supported by COST (European Cooperation in Science and Technology) www.cost.eu.

References

Riccardo Del Gratta, Francesca Frontini, Anas Fahad Khan, and Monica Monachini. 2015. SIMPLE-LOD. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Gil Francopoulo. 2013. *LMF lexical markup framework*. John Wiley & Sons.

Fahad Khan and Jack Bowers. 2020. Towards a lexical standard for the representation of etymological data. *Convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale*.

Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. SIMPLE: A General Framework For The Development Of Multilingual Lexicons. 13(4):249–263.

[7]GitHub project under https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Schemas/LMF%20SynSem%20Specification

Volha Petukhova and Harry Bunt. 2008. LIRICS semantic role annotation: Design and evaluation of a set of data categories. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Laurent Romary, Mohamed Khemakhem, Fahad Khan, Jack Bowers, Nicoletta Calzolari, Monte George, Mandy Pet, and Piotr Bański. 2019. Lmf reloaded. *arXiv preprint arXiv:1906.02136*.

Nilda Ruimy, Ornella Corazzieri, Elisabetta Gola, Antonietta Spanu, Nicoletta Calzolari, and Antonio Zampolli. LE-PAROLE Project: The Italian Syntactic Lexicon. In *EURALEX '98*.

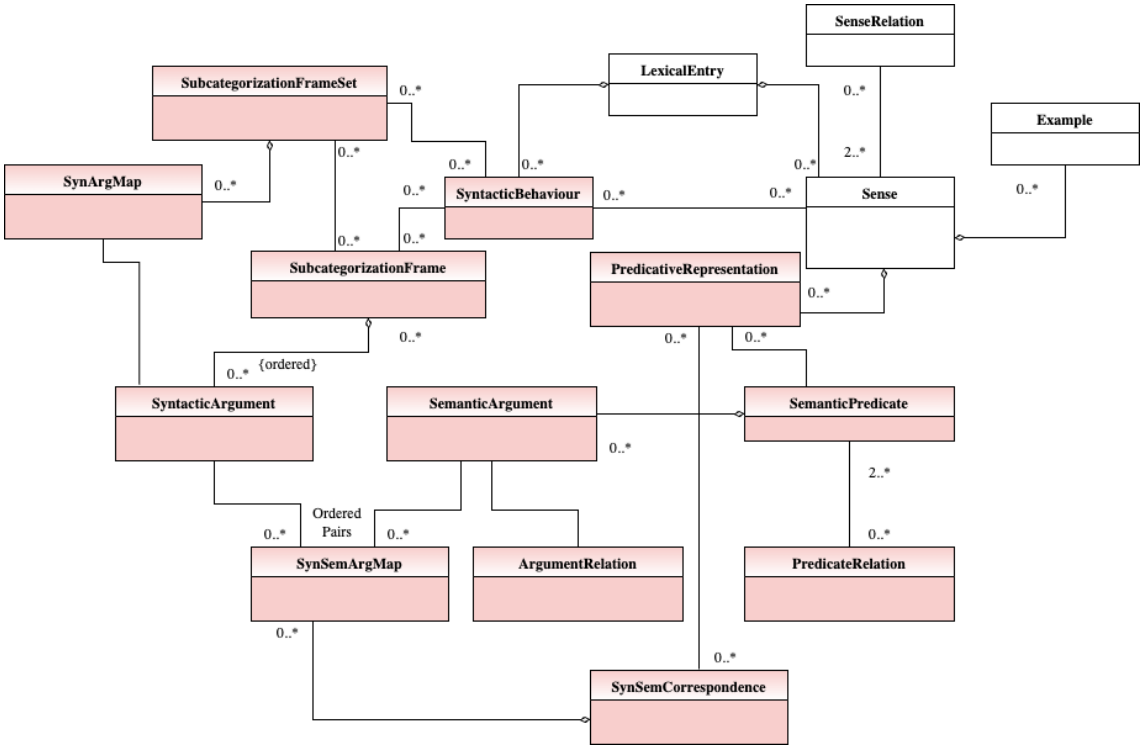

Figure 1: Synsem Module - Class diagram.

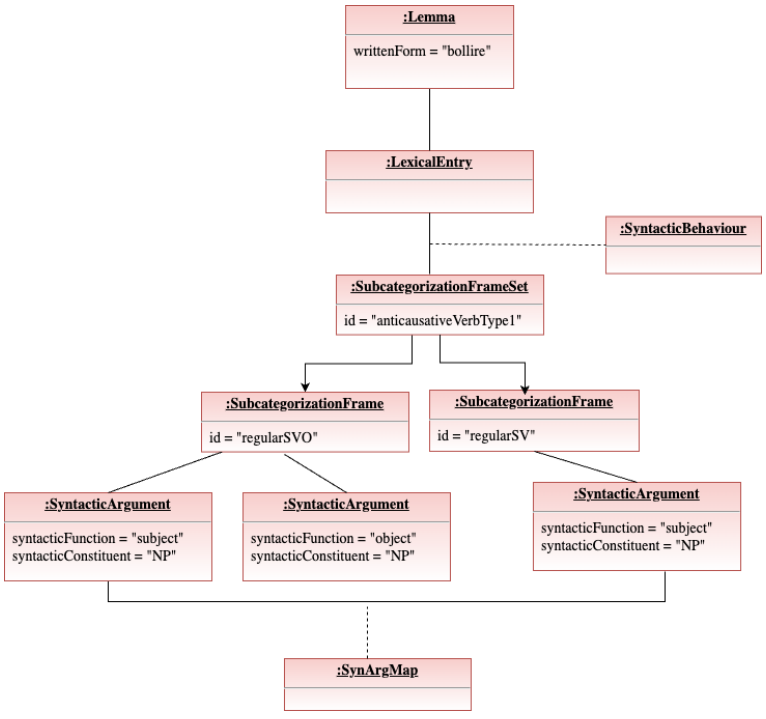

Figure 2: "Bollire" ("boil") syntactic alternation.

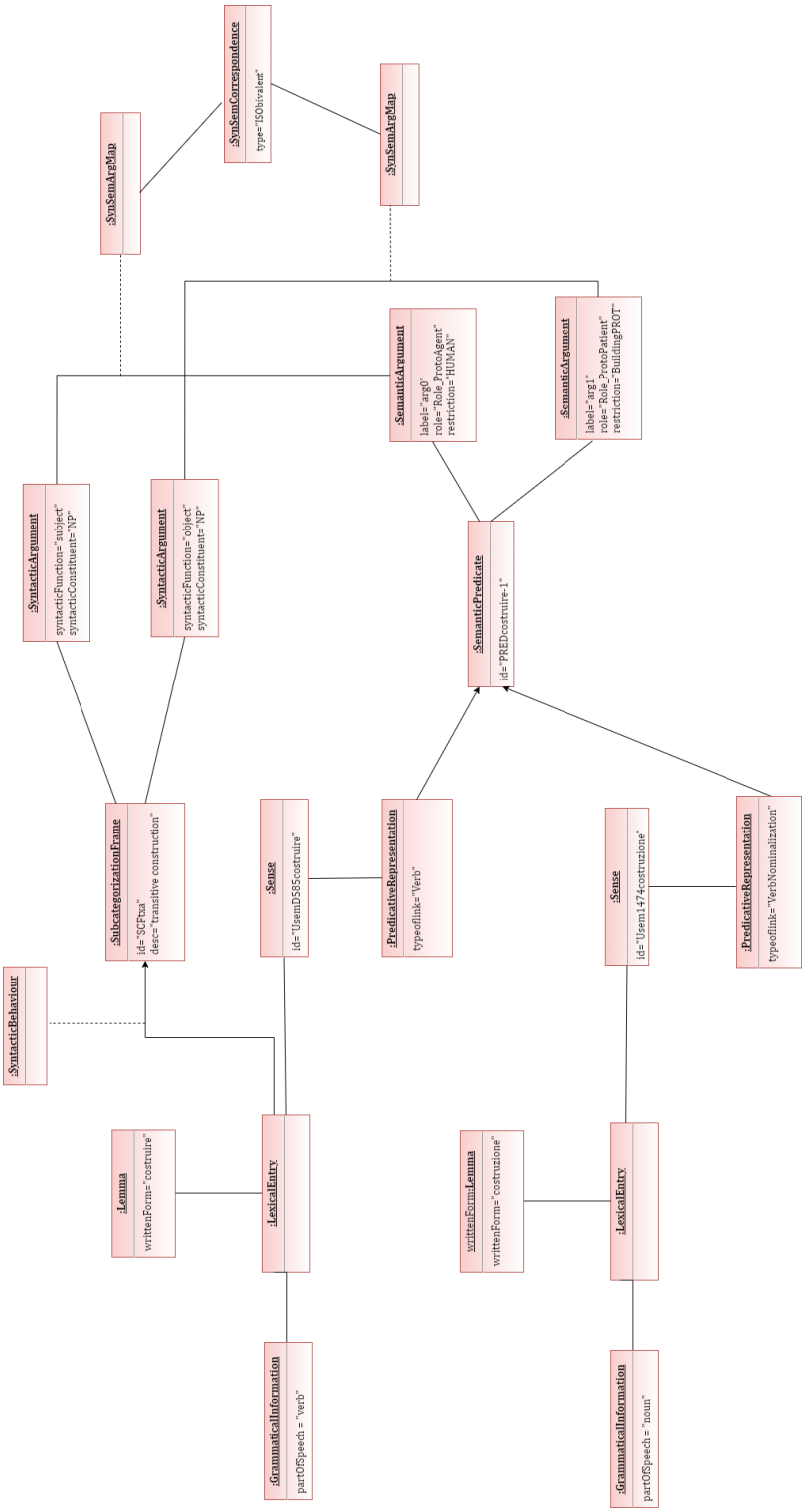

Figure 3: Costuire / costruzione (build/building) in Parole Simple CLIPS.