Lexical Markup Framework (LMF)

Gil Francopoulo[1], Monte George[2], Nicoletta Calzolari[3], Monica Monachini[4], Nuria Bel[5], Mandy Pet[6], Claudia Soria[7]

[1]INRIA-Loria: gil.francopoulo@wanadoo.fr
[2]ANSI: dracalpha@earthlink.net
[3]CNR-ILC: glottolo@ilc.cnr.it
[4]CNR-ILC: monica.monachini@ilc.cnr.it
[5]UPF: nuria.bel@upf.edu
[6]MITRE: mpet@mitre.org
[7]CNR-ILC: claudia.soria@ilc.cnr.it

Abstract

Optimizing the production, maintenance and extension of lexical resources is one the crucial aspects impacting Natural Language Processing (NLP). A second aspect involves optimizing the process leading to their integration in applications. With this respect, we believe that the production of a consensual specification on lexicons can be a useful aid for the various NLP actors. Within ISO, the purpose of LMF is to define a standard for lexicons.

LMF is a model that provides a common standardized framework for the construction of NLP lexicons. The goals of LMF are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources.

In this paper, we describe the work in progress within the sub-group ISO-TC37/SC4/WG4. Various experts from a lot of countries have been consulted in order to take into account best practices in a lot of languages for (we hope) all kinds of NLP lexicons.

1. Introduction

Optimizing the production, maintenance and extension of lexical resources is one the crucial aspects impacting Natural Language Processing (NLP). A second aspect involves optimizing the process leading to their integration in applications. With this respect, we believe that the production of a consensual specification on lexicons can be a useful aid for the various NLP actors. Within ISO, the purpose of LMF is to define a standard for lexicons.

2. History and current context

In the past, this subject has been studied and developed by a series of projects like GENELEX, EAGLES, MULTEXT, PAROLE, SIMPLE and ISLE. More recently within ISO (www.iso.org) the standard for terminology management has been successfully elaborated by the sub-committee ISO-TC37 and published under the name "Terminology Markup Framework" (TMF) with the ISO-16642 reference. Afterwards, the ISO-TC37 National delegations decided to address standards dedicated to NLP. These standards are currently elaborated as high level specifications and deal with word segmentation (ISO 24614), annotations (ISO 24611, 24612 and 24615), feature structures (ISO 24610), and lexicons (ISO 24613) with this latest one being the focus of the current paper. These standards are based on low level specifications dedicated to constants, namely data categories (revision of ISO 12620), language codes (ISO 639), scripts codes (ISO 15924), country codes (ISO 3166) and Unicode (ISO 10646).

This work is in progress. The two level organization will form a coherent family of standards with the following common and simple rule:

1) the high level specifications provide structural elements that are **adorned by the standardized constants**;
2) the low level specifications provide standardized constants.

3. Scope of LMF

LMF is a model that provides a common standardized framework for the construction of NLP lexicons. The goals of LMF are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources.

Types of individual instantiations of LMF can include monolingual, bilingual or multilingual lexical resources. The same specifications are to be used for both small and large lexicons. The descriptions range from morphology, syntax, semantic to translation. The covered languages are not restricted to European languages but cover all natural languages. The range of targeted NLP applications is not restricted. LMF is also used to model machine readable dictionaries (MRD), which are not within the scope of this paper.

4. Key standards used by LMF

LMF utilizes Unicode in order to represent the scripts and orthographies used in lexical entries regardless of language. The linguistics constants like /feminine/ or /transitive/ are not defined within LMF but are specified in the Data Category Registry (DCR) that is maintained as a global resource by ISO TC37 in compliance with ISO/IEC 11179-3:2003. And these constants are used to adorn the high level structural elements.

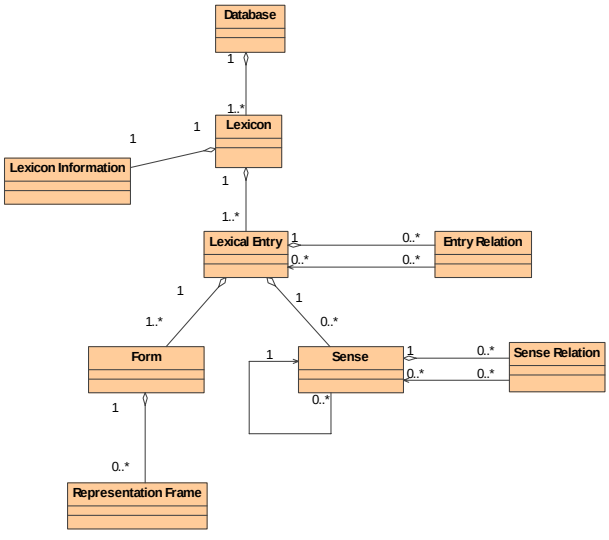
The LMF specification complies with the modeling principles of UML [Rumbaugh] as defined by OMG (www.omg.org). A model is specified by a UML class diagram (the class name is not underlined). Examples of word description are represented by UML instance diagrams (the class name is underlined).

5. Structure

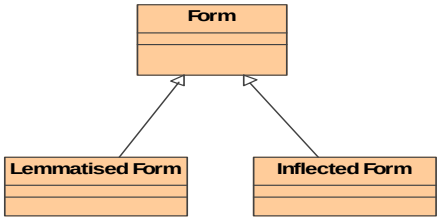
LMF is comprised of the following components:
1) **The core package** which is the structural skeleton which describes the basic hierarchy of information in a lexical entry.
2) **Extensions of the core package** which are expressed in a framework that describes the reuse of the core components in conjunction with the additional components required for a specific lexical resource.

6. Core package specification

One class called *Database* represents the entire resource and is a container for one or more lexicons. The *Lexicon* class is the container for all the lexical entries of the same language within the database. The *Lexicon Information* class contains administrative information and other general attributes. The *Lexical Entry* is a container for managing the top level language components. As a consequence, the number of single words, multi-word expressions and affixes of the lexicon is equal to the number of lexical entries in a given lexicon. The *Form* and *Sense* classes are parts of the *Lexical Entry*. The form consists of a text string that represents the word. The sense specifies or disambiguates the meaning and context of a form. Therefore, the *Lexical Entry* manages the relationship between sets of related forms and their senses. If there is more than one orthography for the word form (e.g. transliteration) the *Form* class may be associated with one to many *Representation Frames*, each of which contains a specific orthography and one to many data categories that describe the attributes of that orthography.
The core package classes are linked by the relations as defined in the following UML class diagram:

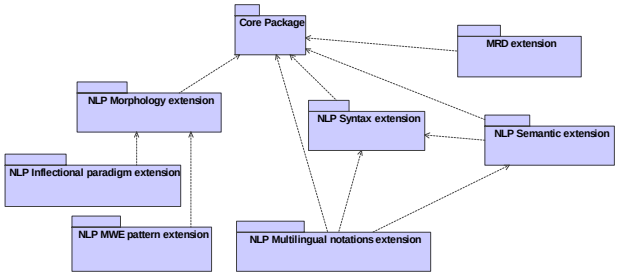


Form class can be sub-classed into *Lemmatised Form* and *Inflected Form* class as follows:



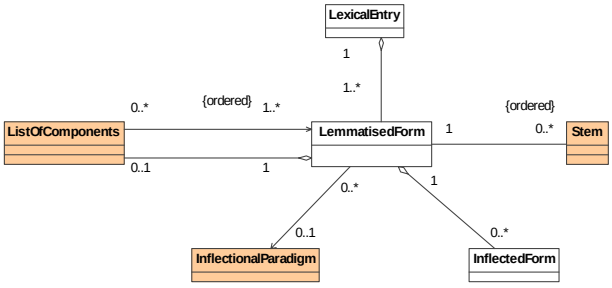
7. Extensions

The current LMF extensions are described in the annexes of the ISO document [ISO 24613] as UML packages. Extensions deal with MRD and NLP lexicons but due to the lack of space, only a sketch of the NLP extensions is given. Creators of lexicons should select the subsets of the possible extensions that are relevant to their needs. All extensions conform to the LMF core model in the sense that some of the core package classes are extended. An extension cannot be used to represent lexical data regardless of the core package. The package dependencies are presented in the following diagram:

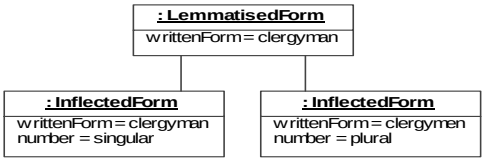


8. Morphology

The morphological extension is defined in the following class package diagram with the convention that the classes of the package are colored and the other classes are not:



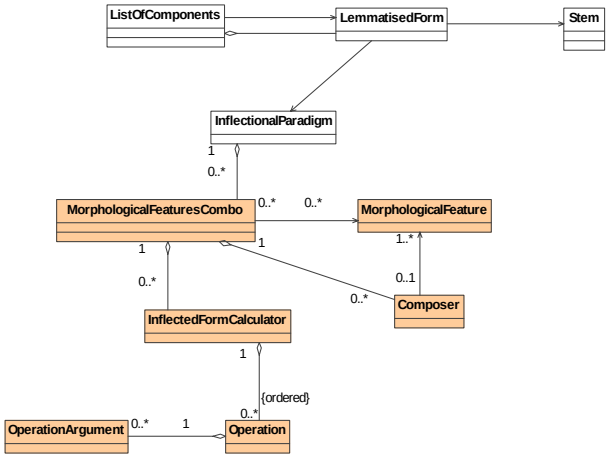
There are two possible strategies to describe the morphology of a word. The first one is to represent explicitly all inflected forms as presented in the following UML instance diagram:



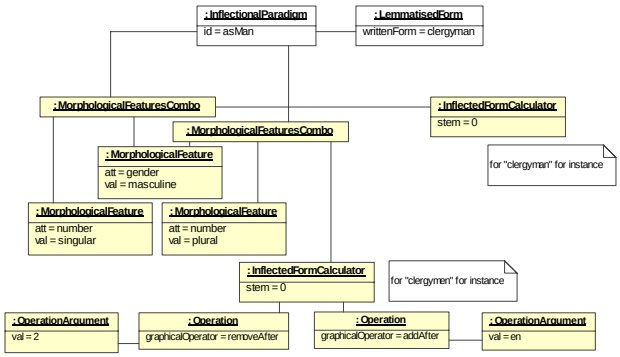
And let's recall, as explained in chapter-4, that the constants like attribute names are not defined in LMF but are taken from the DCR.

The second strategy is to use an inflectional paradigm as in the following UML instance diagram:

Then, concerning the paradigm itself, there are two sub-strategies: the first one is to declare that the inflection of the word is "asMan" (as in this example) and to stop there. This is called the "underspecified inflection". The second sub-strategy is to use the "Inflectional Paradigm" extension and to refine the inflectional paradigm description. This process seems at first view a little bit complex, but this is done only once for a given language. All the words that behave like "man" will share this description. Let's note that for a language with simple morphology like English, such an effort does not necessarily worth the value but for more complex languages like German or Hungarian, the situation is completely different. The extension is as follows:



For instance, an inflectional paradigm could be declared as:



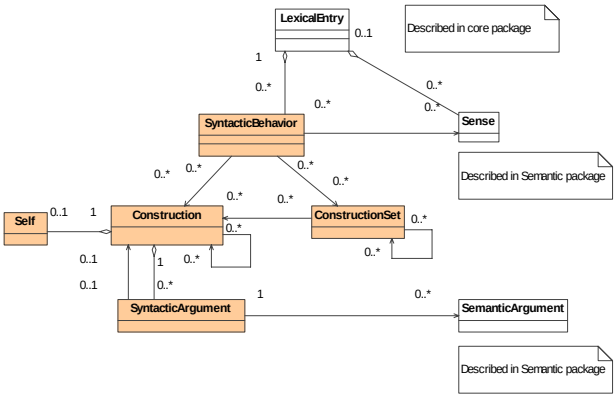
Two letters are removed and two letters are added. The English morphology is relatively simple, so the representation is simple, which means it is not necessary to manage any stem and a reference to the lemmatised form can be used. Thus, the value for the stem attribute is zero. When applied to the entry “clergyman”, the singular gives “clergyman” and the plural gives “clergymen”.

So, to summarize, the lexicon manager has the following options:
1) represent in extension all the inflected forms;
2) connect the lemmatised form to an inflectional paradigm with two sub-options:
- use an underspecified inflection,
- or use a fully specified inflection paradigm.

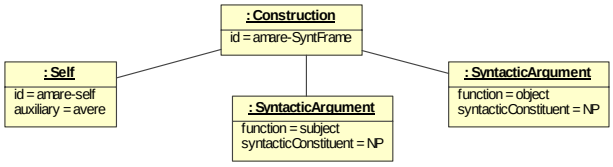
Let's add that the inflectional paradigm package can be used for frozen multiword expressions (MWEs) including agglutinating words. And a specific package (called MWE pattern package) is dedicated to semi-fixed and syntactically flexible MWEs.

9. Syntax

The extension for syntax holds five classes as specified in the following class diagram:



Let's see an example taken from the Parole/CLIPS lexicon (www.ilc.cnr.it). In this example, only syntactic structures are used. This is a rather simple construction in Italian where both the subject and the direct object are Noun Phrase. The self object describes a verb that takes the auxiliary “avere”. A typical example of such a construction is “Gianni ama Maria”.



10. Semantics

The purpose is to describe one sense and its relations with other senses belonging to the same language. Due to the intricacies of syntax and semantics in most languages, the section on semantics comprises also the connection to syntax. The linkage of senses belonging to different languages is to be described by using the multilingual notation package.
LMF does not impose any depth in the semantic description. And different degree of richness can cohabit within the same lexicon. For instance, a technical and specialized lexicon can have a rich description for its technical words and very shallow (or non-existent) semantic information attached to general words or senses.

Various descriptive mechanisms are proposed like synsets, predicates, relations or linkage with syntax. LMF does not impose any exclusive usage of these mechanisms. For instance, a user can manage a multilingual database with synsets in English and predicates in another language. Of course, these mechanisms can be combined within the same language, temporarily or in a permanent manner.
The most important classes are *Sense*, *SemanticPredicate* and *SynSet*.

Sense

Sense element is described in the core package. *Sense* element being contained in the *Lexical Entry* element, a sense is not shared among two different entries. *Sense* may have attributes like dating, style, frequency, geography or animacy.

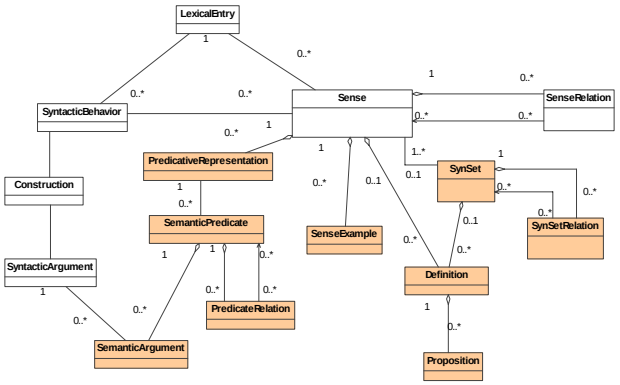
SemanticPredicate

Semantic Predicate is an element that describes an abstract meaning together with the association with *Semantic Arguments*. A semantic predicate may be used to represent the common meaning between different senses that are not necessarily fully synonyms. These senses may be linked to lexical entries whose parts of speech are different. For instance, a verb and the name of the action of the verb may share the same predicate. *Semantic Predicate* may have attributes like: name, type, definition, view.

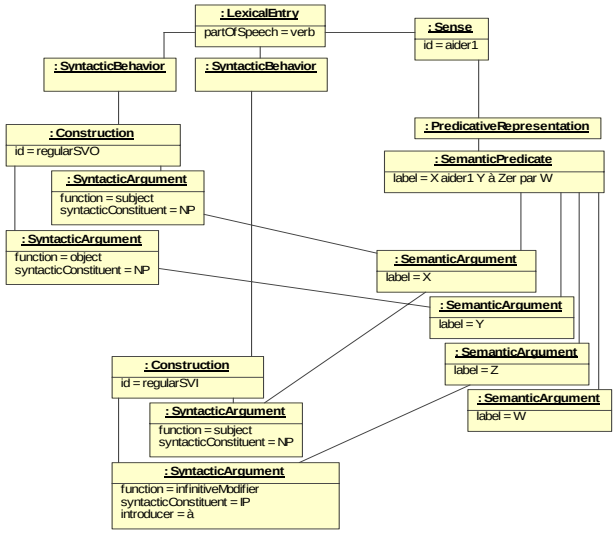
SynSet

Synset links synonyms. *Synset* is an element that describes a common and shared meaning within the same language. *Synset* may link senses of two different lexical entries with the same part of speech. Synset may have attributes like: label, source.

Class diagram is as follows:



Let's see an example in French taken from the “Dictionnaire Explicatif et Combinatoire” [Mel'cuk & al, page 120]. “Aider1” is linked to the semantic actants: “X aide Y à Z-er par W”. This example yields eight different syntactic constructions. We supply the representation for the two first ones: "La Grande-Bretagne aide ses voisins" and "La Grande-Bretagne a aidé à créer l'ONU" with a special focus on syntactic and semantic representation linking. The two constructions are related to a common semantic predicate. This predicate has its semantic arguments (X, Y, Z and W) which are shown to be related to particular syntactic arguments in the different constructions of the verb. That is, the constructions are not linked directly to the predicate, but a particular syntactic argument in each construction is linked to a particular semantic argument.



11. XML specifications

Up until now, the ISO group focused on the conceptual model by the mean of a UML specification. In Summer 2005, it has been decided to add an XML specification as an informative annex of the standard. This will be added during 2006.

12. Comparison

A serious comparison with previously existing models is not possible in this current paper due to the lack of space. We advice the interested colleague to consult the technical report "Extended examples of lexicons using LMF" located at: "http://lirics.loria.fr" in the document area. The report explains how to use LMF in order to represent OLIF-2, Parole/Clips, LC-Star, WordNet, FrameNet and BDéf.

13. Acknowledgements

This work is supported in part by the EU (eContent project 22236 LIRICS, see http://lirics.loria.fr) and the French Technolangue program (www.technolangue.net).

14. Conclusion

In this paper, we describe the work in progress within the sub-group ISO-TC37/SC4/WG4.
Various experts from a lot of countries have been consulted in order to take into account best practices in a lot of languages for (we hope) all kinds of NLP lexicons.
The target is to publish an ISO standard in 2007.

15. References

ISO 24613 Language resource management - Lexical markup framework. ISO Geneva 2005.

Mel'cuk I., Clas A., Polguère A. Introduction à la lexicologie explicative et combinatoire. Duculot Bruxelles 1995.

Rumbaugh J., Jacobson I., Booch G. The Unified Modeling language reference manual, 2nd ed, Addison Wesley 2005.