LMF for multilingual, specialized lexicons

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria

LMF for multilingual, specialized lexicons

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 (ISO-24613) is to define a standard for lexicons that covers multilingual and specialized data.

1. Introduction

Lexical Markup Framework (LMF) is a model that provides a common standardized framework for the construction of Natural Language Processing (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 a 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 description range from morphology, syntax, semantic to translation information organized as different extensions of an obligatory core package. The model is being developed to 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.

2. History and current context

In the past, this subject has been studied and developed by a series of projects like GENELEX [Antoni-Lay], EAGLES, MULTEXT, PAROLE, SIMPLE , ISLE and MILE [Bertagna]. More recently within ISO[1] 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), dates (ISO 8601) and Unicode (ISO 10646).

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

1) **low level specifications** provide standardized constants;

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

3. Scope and challenges

The task of designing a lexicon model that satisfies every user is not an easy task. But all the efforts are directed to elaborate a proposal that fits the major needs of most existing models.

In order to summarise the objectives, let's see what is in the scope and what is not.

LMF addresses the following difficult challenges:

1. Represent words in languages where multiple orthographies (native or transliterations) are possible, e.g. some Asian languages.
2. Represent the morphology of languages where a description in extension of all inflected forms is not manageable (e.g. Hungarian). In this case, representation in intension is the only manageable issue.
3. Easily associate written forms and spoken forms for all languages.
4. Represent complex compound words (like in German, Dutch among other languages)
5. Represent fixed, semi-fixed and flexible multiword expressions.
6. Represent specific syntactic behaviors (as recommended in Eagles).
7. Allow complex argument mapping between syntactic and semantic descriptions (as recommended in Eagles).
8. Allow a semantic organization based on SynSets (like in WordNet) or on semantic predicates (like in FrameNet).

[1] www.iso.org

9. Represent large scale multilingual resources based on interlingual pivots or on transfer linking.

LMF does not address the following topics:
1. General sentence grammar of a language
2. World knowledge representation

In other terms, LMF is mainly focused on lexical linguistic information representation.

4. Key standards used by LMF

LMF utilizes Unicode in order to represent the scripts and orthographies used in lexical entries regardless of language.

Linguistic 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.

The LMF specification complies with the modeling principles of Unified Modeling Language (UML) as defined by OMG[2] [Rumbaugh]. A model is specified by a UML class diagram within a UML package: the class name is not underlined. The various examples of word description are represented by UML instance diagrams: the class name is underlined.

5. Structure and core package

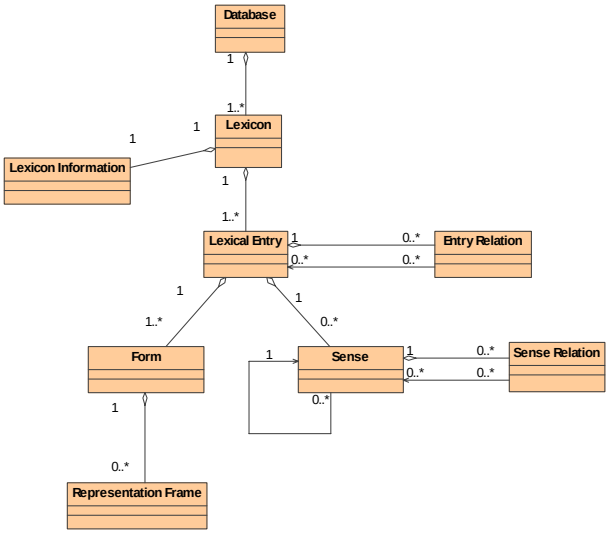
LMF is comprised of two components:

1) **The core package** which is the structural skeleton which describes the basic hierarchy of information in a lexical entry.

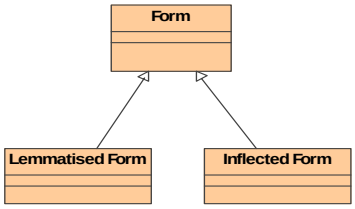
2) **Extensions to the core package**, which are expressed in a framework that describes the re-use of the core components in conjunction with these additional components required for the description of the contents of a specific lexical resource.

In the core package, 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* class is a container for managing the top level language components. As a consequence, the number of representatives of single words, multiword 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*. Form consists of a text string that represents the word. *Sense* specifies or identifies the meaning and context of the related 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 subclassed into *Lemmatised Form* and *Inflected Form* class as follows:



A subset of the core package classes are extended to cover different kinds of linguistic data. All extensions conform to the LMF core package and cannot be used to represent lexical data independently of the core package. From the point of view of UML, an extension is a UML package. Current extensions for NLP dictionaries are: NLP Morphology, NLP inflectional paradigm, NLP Multiword Expression pattern, NLP Syntax, NLP Semantic and Multilingual notations, which is the focus of this paper. Extensions for Morphology, Syntax and Semantic extensions are described in [Francopoulo]. All extensions are described in [LMF 2006].

6. NLP Multilingual extension

The NLP multilingual notation extension is dedicated to the description of the mapping between two or more languages in a LMF database. The model is based on the notion of *Axis* that links the notions of *Sense*, *Syntactic Behavior* and *Example* pertaining to different languages. "Axis" is a term taken from the Papillon project[3] [Sérasset]. Axis can be organized at the lexicon manager convenience in order to link directly or indirectly objects of different languages.

6.1. Considerations for standardizing multilingual data

The simplest configuration of multilingual data is a bilingual lexicon where a single link is used to represent

[2] www.omg.org

[3] www.papillon-dictionary.org

the translation of a given form/sense pair from one language into another. But a survey of actual practices clearly reveals other requirements that make the model more complex. Consequently, LMF has focused on the following ones:

(i) Cases where the relation 1-to-1 is impossible because of lexical differences among languages. An example is the case of English word “river” that relates to French words “rivière” and “fleuve”, where this last one is used for specifying that the referent is a river that flows into the sea. The bilingual lexicon should specify how these units relate.

(ii) The bilingual lexicon approach should be optimized to allow the easiest management of large databases for real multilingual scenarios. In order to reduce the explosion of links in a multibilingual scenario, translation equivalence can be managed through an intermediate "Axis". This object can be shared in order to contain the number of links in manageable proportions.

(iii) The model should cover both *transfer* and *pivot* approaches to translation, taking also into account hybrid approaches. In LMF, the pivot approach is implemented by a “Sense Axis”. The transfer approach is implemented by a “Transfer Axis”.

(iv) A situation that is not very easy to deal with is how to represent translations to languages that are similar. The problem arises for instance when the task is to represent translations from English to European Portuguese and Brazilian. The difference between the two last languages is not very important: a certain number of words are different and the syntax of pronouns is different. Instead of managing two distinct copies, it is more effective to distinguish variations through a limited number of specific Axis, the vast majority of Axis being shared.

(v) The model should allow for representing the information that restricts or conditions the translations. The representation of tests that combine logical operations upon syntactic and semantic features must be covered.

6.2. Structure

The model is based on the notion of Axis that link Senses, Syntactic Behavior and examples pertaining to different languages. Axis can be organized at the lexicon manager convenience in order to link directly or indirectly objects of different languages. A direct link is implemented by a single axis. An indirect link is implemented by several axis and one or several relations.

The model is based on three main classes: Sense Axis, Transfer Axis, Example Axis.

6.3. Sense Axis

Sense Axis is used to link closely related senses in different languages, under the same assumptions of the interlingual pivot approach, and, optionally, it can also be used to refer to one or several external knowledge representation systems.

The use of the Sense Axis facilitates the representation of the translation of words that do not necessarily have the same valence or morphological form in one language than in another. For example, in a language, we can have a single word that will be translated by a compound word into another language: English “wheelchair” to Spanish “silla de ruedas”. Sense Axis may have the following attributes: a label, the name of an external descriptive system, a reference to a specific node inside an external description.

6.4. Sense Axis Relation

Sense Axis Relation permits to describe the linking between two different Sense Axis. The element may have attributes like label, view, etc.

The label enables the coding of simple interlingual relations like the specialization of “fleuve” compared to “rivière” and “river”. It is not, however, the goal of this strategy to code a complex system for knowledge representation, which ideally should be structured as a complete coherent system designed specifically for that purpose.

6.5. Transfer Axis

Transfer Axis is designed to represent multilingual transfer approach. Here, linkage refers to information contained in syntax. For example, this approach enables the representation of syntactic actants involving inversion, such as (1):

(1) fra:“elle me manque” => eng:“I miss her”

Due to the fact that a lexical entry can be a support verb, it is possible to represent translations that start from a plain verb to a support verb like (2):

(2) fra:“Marie rêve” => jpn:"Marie wa yume wo miru"
(Mary dreams)

6.6. Transfer Axis Relation

Transfer Axis Relation links two Transfer Axis. The element may have attributes like: label, variation.

6.7. Source Test and Target Test

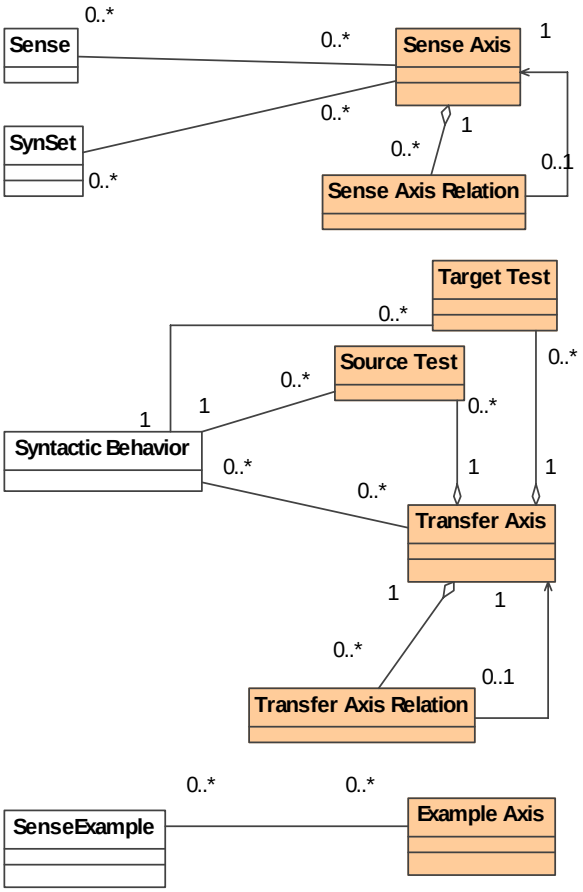
Source Test permits to express a condition on the translation on the source language side while Target Test does it on the target language side. Both elements may have attributes like: text and comment.

6.8. Example Axis

Example Axis supplies documentation for sample translations. The purpose is not to record large scale multilingual corpora. The goal is to link a Lexical Entry with a typical example of translation. The element may have attributes like: comment, source.

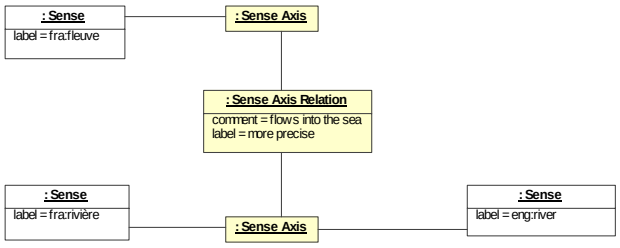
6.9. Class Model Diagram

The UML class model diagram for multilingual notations is as follows:

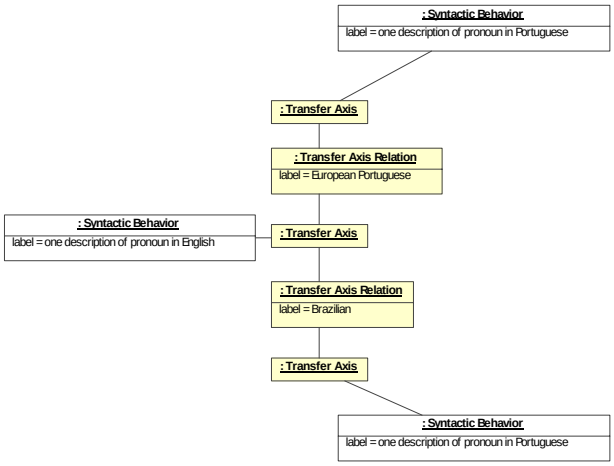


7. Three examples

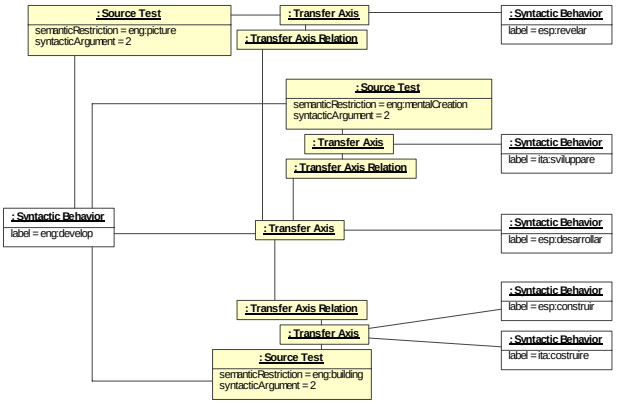
The first example is about the interlingual approach with two axis to represent a near match between "fleuve" in French and "river" in English. The axis on the top is not linked directly to any English sense because this notion does not exist in English. In the diagram, French is located on the left side and English on the right side.



Let's see now an example about the transfer approach about slight variations between similar languages. The example is about English on one side and European Portuguese and Brazilian on the other side. Due to the fact that these two last languages have a very similar syntax, but with some local exceptions, the goal is to avoid a full and dummy duplication in order to ease maintenance of both languages. The transfer axis relations hold a label to distinguish which axis to use depending on the target language.



A third example shows how to use the Transfer Axis relation to relate different information in a multilingual transfer lexicon. It represents the translation of the English "develop" into Italian and Spanish. Recall that the more general sense links "eng:develop" and "esp:desarrollar". Both Spanish and Italian have restrictions that should be tested in the source language: if the second argument of the construction refers to certain elements (picture, mentalCreation, building) it should be translated into specific verbs.



8. LMF for specialized lexicons

LMF, that has not specially been conceived and tested on specialized lexicons, can be used for all kinds of lexicons included the specialized ones.

Compared to general NLP lexicons, specialized lexicons have the following properties:

1. High number of multiword expressions
2. High number of orthographic variants including abbreviations and acronyms
3. Inclusion of domain specific information: terminological definitions, particular codes (like in UMLS).
4. Domain (and sub-domain) marks are needed in the two following situations:
   - when the domain is subdivided into several subdomains
   - when the lexicon is a mix of general and specialized words.

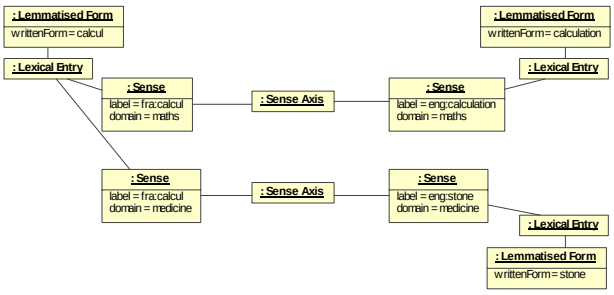
LMF offers for these cases different solutions which are mostly in line with the recommendations for general language lexica [LMF 2006].

The first case is for the encoding of multiword expressions which can be referred to as a unique element because of, for instance, translation equivalences. This is the case for Italian “cervello terminale” which must be translated into English as “cerebrum” and into Spanish as “encéfalo”.

The second case: variation can take the form of orthographic variation, as in the case of “gonadotropin” vs. “gonadotrophin”. But it can also be two entries linked by a synonym relation: take the case of the English medical terms “hypophysis” and “pituitary gland”.

Concerning the two last cases (i.e. domain specific information and domain marks), every LMF element can be adorned by an attribute/value pair. In a multilingual perspective, these marks can be used to condition a translation.

Let's see for instance, the translation of the French word "calcul" into English. There are two senses in French: one in Maths and the other one in Medicine. The translations into English give two different senses and two different lexical entries, as follows:



9. LMF in XML

During the last three years, the ISO group focused on the conceptual model by the mean of a UML specification. In the last version of the LMF document [LMF 2006] a DTD has been provided as an informative annex. Concerning UML to XML conversion, the following conventions are adopted:

1. each UML attribute is transcoded as a DC element
2. each UML class is transcoded as an XML element
3. UML aggregations are transcoded as content inclusion
4. UML shared associations (i.e. associations that are not aggregations) are transcoded as IDREF(S)

An example of entries is the following XML tag structure, where three senses are shown: a French entry "gonadotrophine" is linked both to a Spanish entry "gonadotrofina" and to an English entry "gonadotropin". The Spanish fragment shows two orthographic variants "gonadotrofina" and "gonadotropina". The English fragment shows also two variants.

```
<Database languageCode="ISO-639-2">
<!—                                French section -->
<Lexicon>
<LexiconInformation>
      <DC att="name" val=”French Extract”/>
      <DC att="language" val="fra"/>
</LexiconInformation>
<LexicalEntry
      <DC att=”partOfSpeech” val="noun"/>
      <LemmatisedForm>
            <DC att=”writtenForm” val="gonadotrophine"/>
      </LemmatisedForm>
      <Sense id="fra#gonadotrophine">
            <DC att="domain" val="medicine"/>
      <SemanticDefinition>
            <DC att=”text” val="Lycoprotéine d'un poids moléculaire d'environ 43 000 daltons produite par le syncytiotrophoblaste"/>
            <DC att=”source” val="Wikipedia"/>
      </SemanticDefinition>
      </Sense>
</LexicalEntry>
</Lexicon>
<!—                                Spanish section -->
<LexiconInformation>
     <DC att="name" val=”Spanish Extract”/>
     <DC att="language" val="esp"/>
</LexiconInformation>
<LexicalEntry
     <DC att=”partOfSpeech” val="noun"/>
     <LemmatisedForm>
            <DC att=”writtenForm” val="gonadotrofina"/>
     </LemmatisedForm>
     <LemmatisedForm>
            <DC att=”writtenForm” val="gonadotropina"/>
     </LemmatisedForm>
     <Sense id="esp#gonadotrofina">
            <DC att="domain" val="medicine"/>
     <SemanticDefinition>
            <DC att=”text” val="Cada una de las hormonas secretadas mayoritariamente por la hipófisis"/>
            <DC att=”source” val="UPF-Term"/>
      </SemanticDefinition>
      </Sense>
</LexicalEntry>
</Lexicon>
<!—                                Multilingual section -->
<SenseAxis id="A1" senses="fra#gonadotrophine esp#gonadotrofina eng#gonadotropin">
</SenseAxis>
<!—-                               English section -->
<LexiconInformation>
     <DC att="name" val=”English Extract”/>
     <DC att="language" val="eng"/>
</LexiconInformation>
<LexicalEntry
     <DC att=”partOfSpeech” val="noun"/>
     <LemmatisedForm>
            <DC att=”writtenForm” val="gonadotropin"/>
     </LemmatisedForm>
     <LemmatisedForm>
            <DC att=”writtenForm” val="gonadotrophin"/>
     </LemmatisedForm>
     <Sense id="eng#gonadotropin">
            <DC att="domain"  val="medicine"/>
     <SemanticDefinition>
```

```
        <DC att=”text” val="a hormone (eg, follicle-stimulating hormone) that acts on the gonads to promote their growth and function"/>
        <DC att=”source” val=”www.aegis.com"/>
        <DC att=”UMLS code” val=”E0030121” />
      </SemanticDefinition>
      </Sense>
</LexicalEntry>
</Lexicon> </Database>
```

10. Conclusion

In this paper we presented the results of the ongoing research activity of the LMF ISO standard. The design of a common and standardized framework for multilingual lexical databases will contribute to the optimization of the use of lexical resources, specially their reusability for different applications and tasks. Interoperability is the condition of a effective deployment of usable lexical resources.

In order to reach a consensus, the work done has paid attention to the similarities and differences of existing lexicons and the models behind them.

Acknowledgements

The work presented here is partially funded by the EU eContent-22236 LIRICS project[4], partially by the French TECHNOLANGUE[5] + OUTILEX[6] programs.

References

Antoni-Lay M-H., Francopoulo G., Zaysser L. 1994 A generic model for reusable lexicons: the GENELEX project. Literary and linguistic computing 9(1) 47-54

Bertagna F., Lenci A., Monachini M., Calzolari N. 2004 Content interoperability of lexical resources, open issues and MILE perspectives LREC Lisbon

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 Lexical Markup Framework LREC Genoa

LMF 2006 Lexical Markup Framework ISO-CD24613-revision-9, ISO Geneva

Rumbaugh J., Jacobson I., Booch G. 2004 The unified modeling language reference manual, second edition, Addison Wesley

Sérasset G., Mangeot-Lerebours M. 2001 Papillon Lexical Database project: monolingual dictionaries & interlingual links NLPRS Tokyo

[4] http://lirics.loria.fr
[5] www.technolangue.net
[6] www.at-lci.com/outilex/outilex.html