Lexical markup framework: an ISO standard for electronic lexicons and its implications for Asian languages

Gil Francopoulo · Chu-Ren Huang

Received: 7 May 2014 / Accepted: 10 May 2014 / Published online: 17 June 2014
© Springer Berlin Heidelberg 2014

Abstract Lexical markup framework (LMF) is the ISO standard for representing machine-readable dictionaries (MRD) and natural language-processing lexicons. The formal specification has been officially published in 2008 under the reference ISO-24613:2008 after a 5-year study and series of meetings gathering 60 lexicon managers and linguists coming from various cultures and languages. The ISO document contains a formal specification under the form of a Unified Modeling Language accompanied with a selection of examples of word description in Asian, European, Semitic and Turkish languages. Afterward, the model has been applied to a couple of African languages. In this current text, after a brief introduction of LMF, we present some difficult challenges which are required to represent a selection of Asian languages, especially in the context of dictionaries in general and MRD in particular.

Keywords Machine-readable dictionaries · Natural language processing · Asian languages · ISO · UML

1 Introduction

Lexical markup framework (LMF) is the ISO¹ standard for representing machine-readable dictionaries (MRD) and natural language-processing (NLP) lexicons. The

¹ See www.iso.org for ISO, and <http://www.lexicalmarkupframework.org/> for LMF.

Communicated by Yukio Tono.

G. Francopoulo
IMMI-CNRS + Tagmatica, Paris, France

C.-R. Huang (✉)
The Hong Kong Polytechnic University, Kowloon, Hong Kong
e-mail: churen.huang@polyu.edu.hk

formal specification has been officially published in 2008 under the reference ISO-24613:2008 after a 5-year study and series of meetings gathering 60 lexicon managers and linguists coming from various cultures and languages. The ISO document contains a formal specification under the form of a Unified Modeling Language (UML) accompanied with a selection of examples of word description in Asian, European, Semitic and Turkish languages. Afterward, the model has been applied to a couple of African languages.

In this current text, after a brief introduction of LMF, we present some difficult challenges which are required to represent a selection of Asian languages, especially in the context of dictionaries in general and MRD in particular.

2 Historical context

The historical root of the specification comes from the 1990s with the *Acquilex* (Boguraev et al. 1988) and *Genelex* (Antoni-Lay et al. 1994) models which were European research projects funded by the European Commission with the aim of boosting what was (and still is) called “the transnational integration” among this mosaic of languages that constitutes Europe. Technically speaking, the linguistic perimeter did not really cover all the European languages but was limited to a core subset comprising only English, French, German, Spanish and Italian. Based on the entity–relation model and a SGML DTD² defined within the *Genelex* project, some years later, the *Eagles* model was defined in a broader perspective to cover all European languages in collaboration with a group of Asian experts.³

The LMF work began, in 2003, as an agreement between the US association for standardization, namely ANSI, and the French association for standardization, namely AFNOR, for a joint effort to define an international standard for MRD’s and NLP lexicons. Very early in the process, Asian lexicon managers took part in the specification design in order to establish the broad lines of the formal model. Nowadays, both the model and the terminology defined in the ISO specification are applied in different continents, and in the light of these applications and experimentations, it may be the time now to try to answer to the question: Is the model suited for Asian language lexicons?

3 LMF specifications

The full LMF specification is published by ISO in Geneva. A book (Francopoulo 2013) and series of scientific papers have been published afterwards. However,

² Let’s note that, in the period 1989–1994, the use of SGML as a physical model was particularly innovative, thanks to the IBM’s French scientific center that suggested this idea, as a member of *Genelex*.

³ For a detailed history, see (Calzolari et al. 2013).

aside from a short presentation, they mainly deal with application of LMF for lexicons. The LMF full specification comprises four parts: A glossary which is important with respect to the damaging fuzziness of certain linguistic terms, a UML conceptual specification, a set of documented examples and finally, a physical XML specification. The first two parts constitute the normative portion of LMF, and the third and fourth parts (i.e., the examples and the XML portion) are defined as informative in order to document the purpose of the various mechanisms and also to help users who do not want to define their own XML schema.

The UML conceptual specification can be understood with respect to two different directions, corresponding to two different notions. The first notion deals with the degree of specialization of the represented material: There is a mandatory core UML package, called the “core model”, and a series of optional packages, called “extensions” in order to reduce the complexity of the resulting structure. For instance, when a lexicon manager wants to represent a morphological monolingual lexicon, this person will pick the core model and the morphological extension without being burdened by syntactic or translation mechanisms. The second direction or notion concerns each structure and deals with the level of detail, the lexicon manager wants to have at hand. Each UML package (being the core model or an extension) is defined as a set of organized classes bound with specific links, and each class may be adorned by a vast range of attributes. For instance, for a language with morphological cases, the lexicon manager may adorn a specific UML class with a case value to indicate that the case of a particular inflected form is dative. The UML class is valid for all languages and is clearly defined within LMF, but the adornment is specific to a particular language and/or decision made by the lexicon manager. The adornment is a pair “attribute/value”, in our example: “case/dative” to be taken outside LMF from an external ISO registry called “isocat”⁴ and whose content may evolve, in a strictly incremental manner, over time. That means that it is possible to have two lexicons with the same structure but, due to the fact that they represent two different languages, their respective adornments could be different.

The core model is shown in Fig. 1, with the UML graphical convention that a relation with a diamond means the composition (i.e., the relation “part of”), the diagram specifies that a *Lexicon Resources* instance contains a certain number of *Lexicon* instances. In turn, a *Lexicon* contains a certain number of *Lexical Entry* instances in a “Russian dolls” manner. On the right side, the diagram specifies that a *Lexical Entry* may be connected to some *Sense* instances in order to represent polysemy. Each *Sense* may have one or several definitions, possibly in different languages, and in turn, each definition could be subdivided into fragments of definitions called *Statement* instances. On the left side, the box name is italicized, and, again following UML conventions, that means that the class is abstract with a full definition in the following morphological package. On the other hand, to the right hand side, we can

⁴ www.isocat.org.

Fig. 1 Core model

see that to a *Lexical Entry* instance is required to be associated with one or more forms. In other words, strict synonyms with different lexical forms are allowed, but a *Lexical Entry* instance without a lexical form is not allowed. In lexicographical terms, this definition requires that a lexical entry to be a form-meaning pair, where lexical form in non-vacuous, yet multiple forms or senses (including zero sense) are allowed. We believe that this formulation captures the range of decisions that a lexicographer can make in defining a lexical entry in various language dictionaries.

With the morphological extension, a *Lexical Entry*, shown in Fig. 2, contains a *Lemma* class with a strict cardinality of one. That means that there is always one (and only one) lemma for each *Lexical Entry* instance. It should be noted, and we

Fig. 2 Morphological extension with new classes marked as colored

will see this point further that this does not mean that a lemma should be connected to a single textual representation. Going back to the lexical Entry, this instance may be associated to a certain number of word forms in order to describe a lexicon with the full set of inflected forms explicitly in the lexicon, as opposed to a description in intension by means of a specific extension call “NLP morphological patterns extension” which will be detailed further. On the right side, the classes *List Of Components* and *Component* are specified for the representation of multiword expressions, such as “the White House” or “take advantage of”.

The extensions are more or less independent from each other. Some extensions are usable directly from the core model, but some others require other extensions. More precisely, the dependence is shown in Fig. 3 below.

4 Practical considerations

The UML specification may be a bit too abstract for the lexicon managers who are not used to formalize data structures with such a tool. Let us see now how the UML classes may be used to organize and define an XML physical representation. This XML file may be used for instance as an interchange format or as a dump for backup operations. We will see that the mapping between the UML specification and the XML schema is easy to understand.

A very simple lexicon in English may be as follows, with a *Lexicon* container and only one entry, with a lemma and two word forms, one for “clergyman” and one for “clergymen”, as in Fig. 4. These two word forms are explicitly recorded in the lexicon.

Fig. 3 Dependency bindings between UML packages

Fig. 4 Explicit word forms for an English example

The same data can be expressed by the following XML fragment:

```
<LexicalResource dtdVersion="17">
  <GlobalInformation>
    <feat att="languageCoding" val="ISO 639-3"/>
  </GlobalInformation>
  <Lexicon>
    <feat att="language" val="eng"/>
    <LexicalEntry>
      <feat att="partOfSpeech" val="commonNoun"/>
      <Lemma>
        <feat att="writtenForm" val="clergyman"/>
      </Lemma>
      <WordForm>
        <feat att="writtenForm" val="clergyman"/>
        <feat att="grammaticalNumber" val="singular"/>
      </WordForm>
      <WordForm>
        <feat att="writtenForm" val="clergymen"/>
        <feat att="grammaticalNumber" val="plural"/>
      </WordForm>
    </LexicalEntry>
  </Lexicon>
</LexicalResource>
```

The mapping rules are simple:

- Every UML class is rendered as an XML element with the same name and associated with a very small set of mandatory attributes like the DTD version,
- All XML elements may be adorned with a pair combining a DCR (data category registry) constant (like “grammatical Number”) with a value which may be another DCR constant (like “singular”) or a free string value (like “clergymen”).

5 Application to some Asian languages

As mentioned earlier, Asian colleagues were active in contributing to the design and drafting of LMF. Their contribution and study of applying various stages of LMF proposals to Asian languages were reported in Tokunaga et al. (2006, 2009, 2013), Chung et al. (2007), Shirai et al. (2008), Lee et al. (2009) and Yu et al. (2009). These studies cover Bangla, Chinese, Japanese, Malay, Taiwanese and Thai.

Subsequent studies also extended LMF to Tagalog as well as several Indian languages. Although studies up-to-date are limited to MRD's and NLP lexicons, we believe that the same issues explored could shed light on the compilation of dictionaries for human reader, especially in the context of electronic dictionaries. In this section, we will focus on few linguistics characteristics which are common and features of Asian languages which are not typically dealt with in a western lexicon. Apparently, the Asian languages do not require any specific mechanisms from the syntactic, semantic, multilingual notations, MRD and constraint expression packages when compared with other languages. Let us add also, that due to space limitations, the variation of forms presented here concerns the mainly the written system, but it is easy to understand how the formal UML mechanism can be applied to the phonetic descriptions.

The required properties are the following ones:

- To store and document different representations for the same lemma and the same word forms.
- To document honorifics, like for instance for Thai and Japanese.
- To organize words according to semantic classifiers, like for instance for Chinese and Japanese.
- To explicitly represent and document the inflection or the derivation process, thus, a powerful mechanism is needed to record one or several generic morphological patterns within the lexicon.

6 Different representations

For certain languages, there is the need to store and document different representations for the same lemma and the same word forms. An example of such languages is Japanese, in order to represent the same lexical entry under a hiragana script form or under a katakana script form. The number of *Form Representation* instances may be more important because four kinds of writing systems coexist and combine: hiragana, katakana, kanji and their romanization. A set of variants with the same script name may be combined as in the following example representing *curly hair*. In this example, the whole lexicon is marked as being a lexicon for Japanese, but the script name is not global to the lexicon, it is local to each entry. Some entries may have a single representation, but some entries may be written with different values for the *written attribute* corresponding in parallel with a different script value. For most complex situations, concerning the Latin script name for instance, the script value may be subcategorized by means of an *orthography name* attribute as shown in the right side of the example diagram, as in Fig. 5.

It should be added that there is the same challenge for Chinese. The Chinese writing system changed over the years. The strategy of simplification involved a reduction in the number of strokes of commonly used characters, and at the moment, the two variants are in use. According to ISO-15924, script code is *Hans* for simplified variant and *Hant* for traditional variant. It is obviously possible to

Fig. 5 Multiple scripts and variants

represent only simplified entries in a single lexicon: In this configuration, the script mark *Hans* should be set on the *Lexicon* instance. But, it is also very easy to mix simplified and traditional forms in the lexicon. In addition, with increasing globalization of the society as the ubiquity of mobile devices with western keyboards, there are now a few hundred so-called alphabetic words in Chinese which can only be represented by English alphabets, or a mixture of Chinese characters and English alphabets, such as X 线/X 光 “X-ray” or PK “to play each other in an elimination game,” etc. In this configuration, the script attribute should not be set on the *Lexicon* instance but should be set on each *Form Representation* instance with the possibly of code-mixing, as in the previous Japanese example.

7 Honorifics

Many Asian languages have some level of distinction at the lexical level representing the differences between members of a conversation based on their social level, in terms of a superior/inferior hierarchy (Tokunaga et al. 2013). Three studies have been conducted on different languages.

- In Thai, where a developed honorific system is in use, based on social status, seniority and formal/informal relationships for commercial and social links. There are four types of honorific words:
 - special diction for the King and the royal family,
 - special diction for religious figures,
 - respectful forms,
 - polite forms.
- In Japanese, the honorific system has five forms:
 - respectful for those in higher positions, e.g., a boss at work, a customer,

- humble form to show respect to others, but it is achieved by the speakers abasing themselves. The system is subcategorized into humble forms concerning third persons and humble forms concerning the hearer.
- Polite forms show politeness without differentiating social levels,
- beautification,
- special diction for the Imperial Family.
- In Chinese, honorifics are lexicalized and implemented at rhetoric level, that is, the use of the second person pronoun 您 “you-honorific”, referential prefix 贵/ 貴 “honorable/esteemed”, or deferential prefix 敝 “humble”. Since these morphemes form productive compounds, honorifics in Chinese can be dealt with the regular morphological process.

For LMF, the Thai/Japanese type honorific requires that Honorific be included as a DCR constant and be encoded as an attribute of the wordform which have values (of DCR constants) as defined according to levels in the honorific system of a particular language. In languages like Chinese, honorific values will be directly assigned to its sense.

8 Semantic classifiers

Many Asian languages do not distinguish singularity and plurality of nouns, but instead use numerative classifiers to denote the number of objects. In addition, semantic agreement between classifiers and nouns should be taken into account. This agreement is not as simple as number and gender agreement as in European languages; it is rather similar to a selectional restriction on arguments of predicates. Detailed and complex linguistically motivated ontologies have been studied and constructed for Chinese and Japanese in order to mark lemmas within LMF lexicons, see (Shirai et al. 2008). A dictionary, like the corpus-based collocational dictionary for classifiers and nouns in Chinese by Huang et al. (1997), can explicitly represent such lexical knowledge. For LMF, the solution is to have Classifier as a DCR constant which is an attribute of wordform of nouns, and allowed to have values or one of more classifiers which are represented as DCR constants.

9 Morphological paradigm patterns

9.1 Naïve example

For a language with a simple morphology like English, it is conceivable to record explicitly all the word forms, like those in Fig. 4, given earlier in Sect. 4 dedicated to practical considerations.

This configuration is conceivable for English because the number of word forms for a single lemma is not so high, i.e., limited to two or three forms, aside for a very small number of cases where a variant needs to be recorded. This strategy is called “extensional description”. But when the number of forms is high

or when the lexicon manager wants to describe the linguistic transformation that links the lemma to the word forms, another sort of UML mechanism is available, this is called the morphological pattern. In contrast, this strategy is called “intensional description”.

9.2 Model specification

This system is not limited to inflection but may be applied to agglutination, derivation and compounding as well. More precisely, the inflected forms (and resp. derived, etc.) are not explicitly listed within the lexicon, but the *Lexical Entry* instance is associated with a shared *Morphological Pattern* instance. This instance is shared in the sense that it is common to a group of words. The forms referred to by the Morphological Pattern instance could be the lemma or a series of stems, which are specific to a lexical entry. In contrast, character strings, affixes and different kinds of conditions are located in the *Paradigm Pattern* instances and are generic. The UML class diagram is as follows in Fig. 6:

Fig. 6 Morphological pattern model

This system is complex but powerful. It has been applied to numerous languages with complex morphology and allows different theoretical approaches. Due to space limitation, we are not going to develop all the possible configurations but we will show three difficult examples taken from Tagalog, Bangla and Thai, and we will try to explain the range of possibilities along with the examples.

9.3 Example in Tagalog

For Tagalog (Hocker 1954), a pure item and arrangement approach is impossible. It is not possible to consider the computation of a verbal word form as an additive arrangement of a root with a morph. The following example shows how to form the future tense by taking the first consonant, adding the first vowel and adding these letters to the left side of the lemma, as in Fig. 7.

The *Morphological Pattern* instance is directly contained at the lexicon level, said in other terms, from the point of view of the lexicon manager, this instance is available to be attached to a given lexical entry. For a given language, it is possible, for instance, to have 50 patterns for verbs, 10 for nouns and 10 for adjectives.

9.4 Example in Bangla

For complex inflectional languages like Bangla, when the lexicon manager wants to represent and document the association of affixes and stems, a powerful mechanism is needed to record one or several generic morphological patterns within the lexicon.

Fig. 7 Verbal form example in Tagalog

The following example, illustrated with Fig. 8, uses an item and arrangement approach to implement the verb conjugation using an *Affix Template* to manage a pattern of inflectional affixes. The *Affix Template* representing participial tenses references two *Template Slot* instances, one managing the perfect participial suffix and the second managing a set of verb suffixes in the perfective tenses. In Bangla, a subset of the simple verb suffixes also serves as components in the perfective verb tenses. In order to reduce the presence of redundant *Affix* objects in an implementation, the *Template Slot* instances manages the affixes indirectly through *Affix Slot* instances. The *Affix Slot* instances, in turn, reference shared *Affix* objects. The purpose of this design option is illustrated through the case of the first person present imperfect suffix. This suffix is used for the imperfect tense and as a component in the present perfect tense. Because the affix allomorphs in the different tenses have different phonetic environments, the condition instances may reference sets of *Grammatical Features* instances as relevant constraints.

9.5 Example in Thai

The lexicon manager may need to represent the derivation process when this is a frequent mechanism like in Thai, this is again a situation where the morphological

Fig. 8 Complex affix template example

pattern may be of great help. Indeed, it is time-consuming to record a separate entry for the derived form. Generic reduplication is thus used to modify the sense of a lexeme by some operation to repeat the sound of the lemma. The types of derivation using reduplication are:

- AA type. The form of reduplication is generated by attaching a character symbol “Mai Yamok” ((๑)) to produce a reduplicated sound of the lemma. For instance, the lemma “ดำ” (to be pronounced “dam0” and that means “black”) can be modified to give “ดำ๑” (to be pronounced “dam0-dam0” and that means “blackish”) in order to express a generalization.
- A’A type (tone change in the first syllable), for instance, “ด้าดำ” (to be pronounced “dam3-dam0” and that means “extremely black”) for intensification.
- AA’A type (triplication), for instance, “กินกินกิน” (to be pronounced “kin0-kin4-kin0” and that means “eat like a horse”) for intensification.
- More complex mechanisms like AABB or AB’AB types.

It should be noted, that in the following example, given in Fig. 9, in contrast to the other examples, two different morphological pattern instances are attached to the given lexical entry instance. In this configuration, the marks “generalization” and “intensity” are considered as grammatical features and should be treated as such when reading the lexicon.

It is worth noting that reduplication is not specific to derivation but appears also in languages like Indonesian in order to express plural forms.

Fig. 9 complex affix template example

10 Conclusion

Our study did not cover all Asian languages; thus, it is peremptory to state that LMF suits for all Asian languages. We studied different mechanisms reported in the ISO meetings as being challenging linguistic features of Tagalog, Bangla, Thai, Indonesian, Korean, Chinese and Japanese and designed UML specification in order to cover these configurations, and then we tested these examples. Let us add that due to space limitations, only a subset of these challenges was reported in the current contribution. We hope that the examples given in our paper demonstrate LMF does have the versatility and flexibility to allow for computational representation of necessary information for lexical entries in various Asian languages. It is hoped that by introducing LMF to the community of lexicographers as a standard machine-readable platform for representing and sharing lexicographical information, joint efforts can be directed toward a collaborative and synergic common platform for developing mono- and multilingual lexica in Asia and elsewhere.

References

- Antoni-Lay, M. H., G. Francopoulo, and L. Zaysser. 1994. A generic model for reusable lexicons: The Genelex Project. In *Literary and linguistic computing*, eds. N. Ostler, A. Zampolli, 9(1): 47–54.
- Boguraev, B., E. J. Briscoe, C. Calzolari, A. Cater, W. Meijs, and A. Zampolli. 1988. *Acquisition of lexical knowledge for natural language processing systems (ACQUILEX)*, Proposal for ESPRIT Basic Research Actions No. 3030. Cambridge (UK).
- Calzolari, N., M. Monachini, and C. Soria. 2013. LMF—historical context and perspectives. In *LMF—Lexical Markup Framework*, ed. G. Francopoulo. London: ISTE/Wiley.
- Chung, S., T. Jiang, K. Hasan, S. Lee, I. Su, L. Prevot, and C. Huang. 2007. Extending an international lexical framework for Asian languages, the case of Mandarin, Taiwanese, Cantonese, Bangla and Malay. In *Proceedings of the first international workshop on intercultural collaboration (IWIC)*. Kyoto: Kyoto University, January 24–26.
- Francopoulo G. ed. 2013. *LMF—Lexical Markup Framework*. London: ISTE/Wiley.
- Hocker, C. 1954. Two models of grammatical description. *Word* 10: 210–234.
- Huang, C., K. Chen, and C. Lai. eds. 1997. *Mandarin Daily Dictionary of Chinese Classifiers*. (國語日報量詞典) Taipei: Mandarin Daily Press.
- Lee, L., S. Hsieh, and C. Huang. 2009. Cwn-Lmf: Chinese Wordnet in the Lexical Markup Framework. In *Presented at the 7th Workshop on Asian Language Resources (ALR7), ACL-IJCNLP 2009*. Singapore, August 2–9.
- Shirai, K., T. Takunaga, T. Huang, S. Hsieh, L. Huo, V. Sornlertlamvanich, and T. Charoenporn. 2008. Constructing Taxonomy of Numerative Classifiers for Asian Languages. In *Proceeding of the 3rd international joint conference on natural language processing (IJCNLP)*, Hyderabad, India, 2008.
- Tokunaga, T., V. Sornlertlamvanich, T. Chareonporn, N. Calzolari, M. Monachini, C. Soria, C. Huang, Y. Xia, H. Yu, L. Prevot, and K. Shirai. 2006. Infrastructure for standardization of Asian language resources. In *Presented at the 2006 COLING/ACL Joint Conference*. Sydney, Australia. July 17–21.
- Tokunaga, T., D. Kaplan, N. Calzolari, M. Monachini, C. Soria, V. Sornlertlamvanich, T. Charoenporn, Y. Xia, C. Huang, S. Hsieh, and K. Shirai. 2009. Query expansion using Lmf-compliant lexical resources. In *Presented at the 7th Workshop on Asian Language Resources (ALR7), ACL-IJCNLP 2009*. Singapore, August 2–9.
- Tokunaga, T., S. Y. M. Lee, V. Sornlertlamvanich, K. Shirai, S. Hsieh, and C. Huang. 2013. LMF and its implementation in some Asian languages, In *LMF—Lexical Markup Framework*, ed. G. Francopoulo. London: ISTE/Wiley.
- Yu, Y., L. Lee, S. Hsieh, and C. Huang. 2009. Chinese word sense distinction in the Lexical Markup Framework: A study in environmental domain. In *Presented at Chinese Lexical Semantics Workshop (CLSWS)* 2009. Yantai, China, July 27–31.