TEI and LMF crosswalks

Laurent Romary

TEI and LMF crosswalks

Laurent Romary, Inria & HUB-IDSL

The intimate relationship between the TEI and the LMF standards

This chapter is about a simple thesis: the TEI framework could be the optimal serialisation background for the LMF standard, since it provides both an ideal XML specification platform and a representation vocabulary that can be easily tuned (or *customized*) to cover the various LMF packages and components. This thesis does not come out of the blue but arises naturally when one observes the history of both initiatives, and their current impacts in various communities in the humanities and in computational linguistics, but also when one ponders on the relevance of having an LMF-specific serialisation when lexical data are in essence to be interconnected with various other types of linguistic resources.

As a matter of fact, the current XML serialisation of LMF suffers from both generic and specific problems that have prevented it from being widely used by the various communities interested in digital lexical resources. Right from the onset, the lack of consensus on the strategy to define a reliable and stable XML serialisation has forced the ISO working group on LMF to confine it to an informative annex, with the following main shortcomings:

- Being carved in stone within the ISO standard, rather than being pointed to as an external and stable online resource, prevents it from being properly maintained, in order to either make corrections on identified weak points or bugs, or to add additional features;
- It is only defined as a DTD, a schema format which hardly any XML developer currently uses anymore and which deeply limits its capacity to express constraints on types or to factorise global attributes. For the sake of simplicity (and this can be easily understood when one has to finalise a text for an ISO standard) no parallel definition of a RelaxNG or W3C schema was provided;
- It does not reflect the intrinsic extensibility of LMF, as it does not contain any dedicated mechanism for customization, for instance when the developer of a new lexical model would like to add her own extensions;
- A more intrinsic weakness of the suggested LMF serialisation is that it hardly takes up any existing vocabulary that could be reused to express either the macro- or micro-structure of a lexical entry. From a purely technical point of view, basic representation objects such as @xml:id, which are standard practice in XML design, are redefined locally. At a low level, it misses using ISO 24610 for the representation of feature structures and redefines its own <feat> object[1]. As a whole, it suffers from a syndrome similar to that of the

[1] The LMF <feat> object is not even compliant with ISO standard 16642 (TMF) which defined such an element before ISO 24610 was in place.

unfortunate ISO standard 1951: it creates a specific silo that shows as little reuse of other initiatives as possible.

Let us be clear here that such infelicities are usually the characteristics of standards that are in many other respects ahead of their time (think of ISO 8879:1986, SGML!) and which require further years of ripening before they reach the best balance between comprehensiveness, simplicity and technical adequacy. The topic of our paper is indeed to contribute to improving LMF by considering bringing it closer to the TEI, an initiative that has itself gone through many years of fruitful iterations.

TEI as a data modelling environment

Although the Text Encoding Initiative started quite some years ago in 1987, with its establishment as a consortium some 15 years ago, we will focus here on its current technical characteristics, knowing that the maintenance mechanisms we will describe have contributed to its being the existing powerful infrastructure we know today.

The scope of the TEI covers all documents whose content can be seen as mainly textual. This encompasses many types of possible objects such as manuscripts (Burghart & Rehbein, 2012), scholarly papers (Holmes & Romary, 2010) or spoken data (Schmidt, 2011). As we shall see lexical data are part of the covered domains but at this stage the most important feature to stress is that the almost 600 elements of the TEI guidelines are all defined in a specification language based on the TEI vocabulary itself. In a way, as was the case for Lisp in the good old days, the TEI is expressed in its own language.

More fundamentally, the specification principles of the TEI infrastructure, reflected in the so-called ODD (One Document Does it all) vocabulary, are based upon the concept of literate programming introduced by Knuth (1984), which advocates an integrated process through which technical specifications and prose descriptions are intimately linked with one another, so that one can easily work with one while having direct access to the equivalent object in the other. From the point of view of the TEI, this means that out of the ODD specification one can generate various schema formats (DTD, RelaxNG schemas, W3C schemas) as well as the documentation in any kind of possible format (pdf, docx, ePub, etc.).

Beyond the fact that the TEI is itself specified in ODD, the language is generic enough to be applicable to non-TEI environments. This has indeed been the case for several initiatives in the standardisation domain, the W3C using it for its ITS recommendation, and ISO committee 37 using it for drafting several of its standards[2]. Moreover, ODD is well designed to combine heterogeneous vocabularies, like integrating CALS tables or MathML formulae within a TEI document. This is particularly important for the reuse of components (typically ISO-TEI feature structures) within a newly designed document model.

Without providing too many technical details here, we can describe the main aspects that give ODD its strength and flexibility:

- The core declarative object is naturally the XML element, which can be associated with various descriptive properties (name, gloss, definition, examples and remarks) and technical information (content model based on

[2] ISO 24611, ISO 24616, ISO 24617-1, and ongoing revision of ISO 16642

RelaxNG snippets, further constraints (e.g. Schematron), attribute declarations);

- In complement to element, the ODD language allows the definition of classes, which are grouping objects for elements having a similar semantics or occurring in the same syntactical context (for example all grammatical features). These are called *model classes*;
- *Attribute classes* are also available to factorise attributes that are used uniformly by several elements (for instance all attributes providing additional temporal constraints to an element);
- Elements may also be grouped together as *modules* (ex.: *drama*, *transcription of speech* and … *dictionaries*).

As described in Burnard and Rahtz (2004) these various components provides a wealth of customization facilities, with for instance the possibility to easily add to or remove an element from a content model by changing its belonging to a given class in the TEI infrastructure. This specification and customization platform also paves the way to the description of coherent XML substructures (or *crystals*, see Romary and Wegstein, 2012), that are essential for a component based data modelling and, as we shall see, correspond to the kind of granularity needed to implement LMF packages.

Finally, all these mechanisms are actually maintained and implemented as an open source portfolio of specifications[3] and tools[4] that facilitate their adoption by a wide range of users.

TEI as a quasi-LMF-compliant framework

Now that the motivations and general context for our approach have been set, we can focus on the actual representational tools that the TEI offers to deal with LMF compliant lexical structures. There are indeed two main approaches that one can consider here:

- Considering lexical structures as feature structures and using the corresponding ISO-TEI joint vocabulary to this end;
- Taking the XML vocabulary available from the TEI chapter for dictionaries.

The baseline – feature structures

The idea of representing lexical entries as feature structures has come to light in conjunction with the necessity of providing a structured representation of lexical data in the context of formal linguistic theories (e.g. Pollard & Sag, 1994; see also Haddar et alii, 2012 for an LMF proposal in this respect) but also to account for the deterministic representation and access to legacy dictionary data (Véronis & Ide, 1992). As a matter of fact, since the early days of the TEI guidelines (See Langendoen & Simons, 1995 and Lee et alii, 2004), there existed a specific module[5] inspired by

[3] http://tei.sourceforge.net/

[4] For instance, Roma (http://www.tei-c.org/Roma/startroma.php) for the online design of customization, or Oxgarage for the transformation of TEI documents from and to various possible formats or schema languages.

[5] Chapter 18 in TEI P5 - http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html

these two trends and extensively covering all aspects of typed feature structures, with mechanisms for declaring constraints on them[6]. In 2006, following an agreement between the TEI consortium and ISO, the module became an ISO standard (ISO 24610-1) and is now the reference XML representation for feature structures.

Applying the ISO-TEI feature structure format for representing data in a way compliant to the LMF meta-model can be achieved quite straightforwardly by mapping LMF concepts as follows:

- Components are implemented as features whose value is a complex feature structure;
- Elementary descriptors (i.e. which correspond to complex data categories in ISOcat) are implemented as elementary features with a symbolic value (mapped onto a simple data category in ISOcat).

Mappings between features and feature values with ISOcat entries can be controlled either by eliciting the association within an FSD, or even by describing a feature library to factorise the information expressed within lexical entries. These mechanisms, related to the use of the dcr attributes (see Windhouwer and Wright, 2012), are based upon the technical description provided in (Aristar-Dry et alii, 2012) and will not be elaborated further here.

To visualize what such an LMF compliant representation could look like, we provide below a verbatim representation of the "clergyman" example from the LMF standard according to the principles stated above[7].

```
<fs type="Lexicon" xmlns="http://www.tei-c.org/ns/1.0">
   <f name="language"/>en</f>
   <f name="LexicalEntry">
      <fs>
         <f name="partOfSpeech">commonNoun</f>
         <f name="Lemma">
            <fs>
               <f name="writtenForm">clergyman</f>
            </fs>
         </f>
         <f name="WordForm">
            <fs>
               <f name="writtenForm">clergyman</f>
               <f name="grammaticalNumber">singular</f>
            </fs>
         </f>
         <f name="WordForm">
            <fs>
               <f name="writtenForm">clergymen</f>
               <f name="grammaticalNumber"/>plural</f>
            </fs>
         </f>
      </fs>
   </f>
</fs>
```

Even if one does not want to go as far as using fully-fledged feature structures but limits oneself to keeping at least the general principles of the LMF serialisation skeleton (elements named according to their equivalent component in the meta

[6] FSD – Feature Structure Declarations

[7] In all our examples, we will use the simplified (untyped) form for feature values as plain text content of the <f> element. More elaborate implementations should distinguish specific subtypes as specified in the ISO-TEI specification.

model), it is still possible to use the ISO TEI feature syntax for the corresponding descriptors in an LMF representation[8]. One possible advantage, beyond a better convergence across standardisation initiatives is that it allows, as was alluded to before, a simple declaration of the corresponding feature in connection to ISOcat. The suggested mixed-approach is illustrated below with again the "clergyman" example:

```
<LexicalResource xmlns:tei="http://www.tei-c.org/ns/1.0">
   <GlobalInformation>
      <tei:f name="languageCoding">ISO 639-3</f>
   </GlobalInformation>
   <Lexicon>
      <tei:f name="language">eng</f>
      <LexicalEntry>
         <tei:f name="partOfSpeech">commonNoun</f>
         <Lemma>
            <tei:f name="writtenForm"/>clergyman</f>
         </Lemma>
         <WordForm>
            <tei:f name="writtenForm">clergyman</f>
            <tei:f name="grammaticalNumber">singular</f>
         </WordForm>
         <WordForm>
           <tei:f name="writtenForm">clergymen</f>
           <tei:f name="grammaticalNumber">plural</f>
         </WordForm>
      </LexicalEntry>
   </Lexicon>
</LexicalResource>
```

All in all, the feature structure module of the TEI offers several possibilities to work within an LMF friendly environment, with the advantage of being based on a strong formalism where data validation is actually built-in. On the weak side, the generic character of feature structures, which comes with some degree of verbosity, makes it more difficult to maintain by human lexicographers. When this becomes an issue, it is reasonable to turn to a more lexical oriented format.

The TEI *Dictionaries* chapter

The TEI guidelines actually come with a quite elaborate XML vocabulary for the description of electronic dictionaries[9]. Conceived initially on the basis of an underlying formal model of the hierarchical nature of a lexical entry (Ide & Véronis, 1995), and based upon previous theoretical (Véronis & Ide, 1992) and descriptive (Ide et alii, 1992) works anticipating the idea of a solid structural skeleton further decorated by means of a variety of descriptors, it is not a surprise that the TEI model matches the LMF core package so well[10]. Still, it is important to keep in mind that the original chapter of the TEI guidelines, then named "Print dictionaries", was strongly oriented towards the representation of digitized material rather than on the creation of born digital lexical data. This had basically two consequences: a) it contains many more constructs intended for the representation of human oriented features (typically the etymology of a word, cf. Alt, 2006 and Salmon-Alt et alii-b 2005) and b) it offers

---

[8] A very similar approach has indeed been developed by Menzo Windhouwer in the context of the Relish project, see http://tla.mpi.nl/relish/lmf/ and (Aristar-Dry et al., 2012)

[9] see http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

[10] It is less surprising given that the TEI principles informed the first ISO meeting in Korea (February 2004) where the LMF ideas were initially put together (cf. Romary et alii, 2004)

specific “flat” representations intended to cover the early steps of the digitization process, and that are outside the scope of the structured view we consider in this paper.

Whereas we will provide concrete crosswalks examples between the LMF model and the TEI *Dictionaries* chapter in the following section, we focus here on the description of the main elements that form the basis of the TEI descriptive toolbox for dictionaries.

The main structural elements of the TEI *Dictionaries* chapter are presented below and schematised in Figure 1 to illustrate their structural relationships:

- <entry> is the basic structuring element of a lexicon (in the LMF sense) and groups together form information, grammatical information (cf. comments in the following section), sense information and related entries;
- <form> can be used to describe one or several forms associated with an entry;
- <gramGrp> groups together all grammatical features that may be attached to the entry as a whole, to a specific form or even as constraint on one of the senses of a word;
- <sense> brings together all sense related information, i.e. definitions, examples, usage information and additional notes. It matches the Form component of the Sense component of the LMF standard.

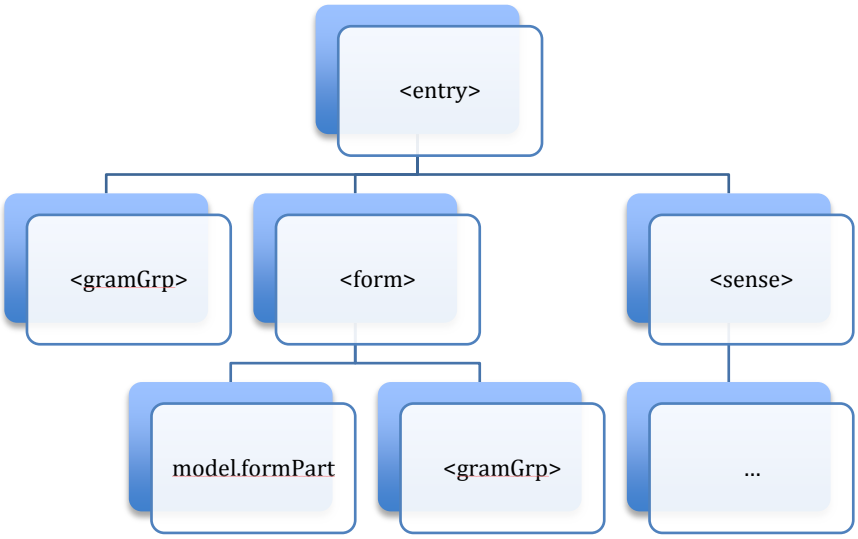

Figure 1: The structural skeleton of the TEI *Dictionaries* chapter

The richness of the TEI descriptive toolbox has at times had the paradoxical effect that one could get deterred from using it simply because it does not come as a ready made module offering a single method of representing a given phenomenon. Although the same criticism could be addressed even more fiercely to the LMF standard itself, it is true that the experience gained over the years with the representation of lexical databases based on the TEI guidelines suggests that it is necessary to introduce more constraints, or at least some precise recommendation to make lexical representations more interoperable (cf. for instance Romary & Wegstein, 2012).

Among the core issues that sometimes make dictionary designers ponder upon which descriptive object to use is the variety of alternative elements that the TEI offers to <entry> proper. Apart from the possibility to group together homonyms (<hom>) or

homographs (<superEntry>), the TEI has two specific elements for representing a lexical entry in a less structured manner:, <entryFree> to allow any kind of combination and order of dictionary components, and <dictScrap>, which allows parts of a dictionary entry to be left unencoded. These alternatives are indeed intended to deal with the specific scenarios of legacy human dictionaries, especially ancient ones, whose entries may not be straightforwardly organised (<entryFree>) or in the case of a multi-step scenario (<dictScrap>) whereby an initially OCRed dictionary is manually encoded step by step.

In the perspective of identifying the optimal customisation of the TEI guidelines which might implement the LMF model, we consider these various alternative constructs as transient objects that are part of specific workflows. For the purpose of disseminating LMF compliant data, we will thus from now onwards only consider <entry> as a proper implementation of the *LexicalEntry* component.

Another typical case of representational ambiguity results from the fact that the core sense related sub-elements (<cit>, <def> or <usg>, with the ambivalent case of <gramGrp>) can actually occur freely as children of the <entry> element. This was initially intended to simplify representations where only one sense is being recorded and the encoder wants to avoid the supposedly superfluous <sense> element around such information. But at the end of the day, the resulting representations are not interoperable with one another and, in the context of our current argument, some of them are not even LMF compliant. It is thus essential for the TEI community (or the LMF standard in one of its further revisions) to identify which subset of the TEI guidelines can be set as the reference LMF compliant one. As elicited in Romary & Wegstein (2012), such a customization should make <sense> mandatory for the representation of semantic content in <entry>, even if there is indeed only one sense.

Finally, on a more positive note, it can be observed that the TEI brings a lot of potential elements, which, in complement to the basic lexical encoding mechanisms provided by LMF can be useful for the encoding of deep textual features with text fields. Typically, names, dates, foreign expressions in definitions or examples are not part of the LMF ontology. Still, they are usually important for the proper traversal or cross-linking of lexical material. Whether they are manually or automatically detected, the corresponding TEI vocabulary can definitely be used even as an external resource to LMF compliant representations[11] that are not expressed using the TEI guidelines proper. Typically a location can be tagged within a definition as in the following example:

```
<def>Orchidée épiphyte, originaire d'<geogName>Amérique
tropicale</geogName>, et dont l'espèce la plus connue est très
recherchée pour l'élégance de ses fleurs mauves à grand labelle
en cornet onduleux.</def>
```

A canonical match: form representation in TEI

As we mentioned earlier, the *TEI Dictionaries* chapter already contains most of the basic constructs needed to implement the various components of the LMF core package. In this section, we would like to focus more specifically on the Form

[11] see for instance the chapter "Names, Dates, People, and Places" (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html) for the encoding of basic name entities.

component and identify, a) how the available TEI elements for form description can be matched to the LMF specification and b) what perspective it brings about for the representation of full-form dictionaries, which we will take as an typical example of the type of lexical objects that are needed in the language technology domain.

From an LMF point of view, the description of form information within a lexical entry (see figure 2) consists of a very simple, yet extremely expressive, structure based upon two components:

- a Form component, which can be iterated within a lexical entry and unites all descriptions associated to what is considered as a single and coherent morphological object associated to the entry;
- a Form Representation component, which allows one to provide as many descriptive views as needed for a given form.

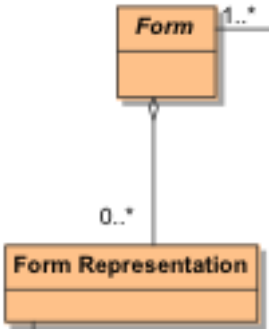

Figure 2: the Form and Form Representation components of the LMF core package

The two-level structure representation is an essential aspect to gain "form autonomy"[12] within a lexical entry. The canonical use of such a construct is typically when a word may occur in several written forms according to the script or transliteration mode being used. For instance, the Hangul representation of the verb "chida" (en: "to hit") can be associated with its Romanized transliteration as sketched below.

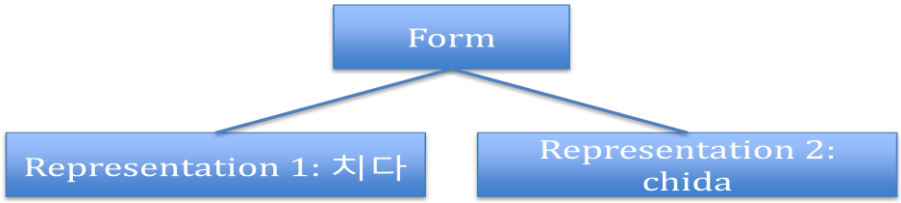

Figure 3: multiple scripting of the Korean verb "chida"

Given the canonical mapping that exists between the Form - Form Representation components in LMF and the <form> element - model.formPart model class in the TEI guidelines, this excerpt can be simply represented in TEI: as follows, where the @type attribute is used to characterize the orthographical methods (here, Hangul vs. Romanized) being used.

```
<form>
   <orth type="hangul">치다</orth>
   <orth type="romanized">chida</orth>
</form>
```

[12] Like we have the term autonomy principle in terminology

If we now move to the slightly more elaborate "clergyman" example depicted in figure 4, the situation is hardly more complex and can be summarized by mean of the mapping table 1.

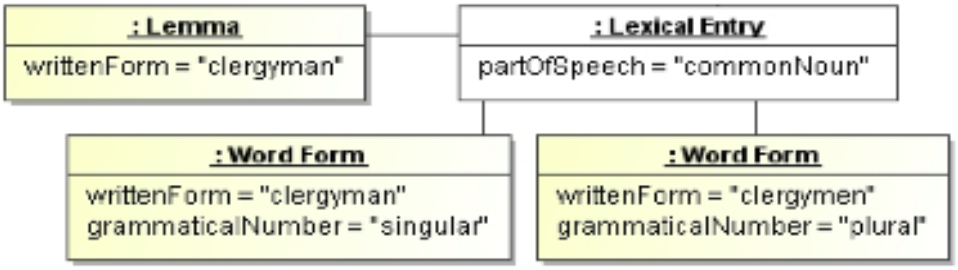

Figure 4: Schematic representation for the entry "Clergyman" (source: LMF standard)

LMF component	*TEI representation*
LexicalEntry	<entry>
Lemma	<form type="lemma">
Word Form	<form type="inflected">
writtenForm	<orth>
partOfSpeech	<pos>
grammaticalNumber	<number>

Table 1: Mapping between LMF components and corresponding TEI elements

The resulting representation, presented below, corresponds to a strict one-to-one mapping to the corresponding LMF model, which indeed can make it a strong basis for the implementation of any kind of full form lexica[13].

```
<entry>
   <gramGrp>
      <pos>commonNoun</pos>
   </gramGrp>
   <form type="lemma">
     <orth>clergyman</orth>
   </form>
   <form type="inflected">
      <orth>clergyman</orth>
      <gramGrp>
         <number>singular</number>
      </gramGrp>
   </form>
   <form type="inflected">
      <orth>clergymen</orth>
      <gramGrp>
         <number>plural</number>
      </gramGrp>
   </form>
</entry>
```

As can be seen, the TEI guidelines provide quite a good coverage of the morpho-syntactic features typically needed for full form lexica. Still, there are several issues

[13] See also the first experiments done on the Morphalou dictionary (Romary et alii, 2004) or for the Arabic language (Salmon-Alt et alii, 2005-a)

that have to be considered before one can systematically represent such lexica in an interoperable way for a variety of languages.

From a pure TEI point of view, we already tackled the issue of representational ambiguity, which can make encoders use different constructs to represent the same phenomenon. In the case of inflected forms, both the coherence of their representation and the necessity to remain compliant with LMF requires a systematic use of <form> and <gramGrp> to embed form and grammatical related information respectively, even if in both cases it may be seen as redundant. In the preceding example for instance, even if only a single grammatical feature (<number>) appears in the <gramGrp>, a coherent representation with other word categories (for instance verbs) or other languages, requires that the latter should not be omitted[14]. This allows for instance that a search for the various grammatical constraints used in a lexicon can be made with <gramGrp> as an entry point.

From a data model perspective, this also ensures, as demonstrated in the previous section, a coherent and strict equivalence of <gramGrp> with a feature structure in case one wants to use this generic representation means in place of <gramGrp> within <form>. For instance, the previous example can be reformulated as:

```
<entry>
   <form type="lemma">
     <orth>clergyman</orth>
     <fs type="grammar">
       <f name="pos">commonNoun</f>
     </fs>
   </form>
   <form type="inflected">
     <orth>clergyman</orth>
     <fs type="grammar">
       <f name="number">singular</f>
     </fs>
   </form>
   <form type="inflected">
     <orth>clergymen</orth>
     <fs type="grammar">
       <f name="number">plural</f>
    </fs>
   </form>
 </entry>
```

Finally, we should address here the issue of linguistic coverage, with the possibility of constraining the semantics of the grammatical features used in such representations, and furthermore to add features that may not be part of the core grammatical elements of the TEI, but which are still necessary to describe morpho-syntactic constraints in other languages. For this purpose, the TEI provides a generic <gram> element, which, coupled with the appropriate value for its @type attribute, can theoretically mark any kind of grammatical feature. Still, it is strongly recommended, when one has such a representational need, to design an *ad hoc* element in one's ODD specification and relate this specification to ISOcat by means of either the <equiv> construct or the appropriate DCS attributes[15].

---

[14] Following a personal communication made by Martin Holmes, I mention that In the case where the is no grammatical information available, it is preferable not to have a <gramGrp> element at all. Indeed, it is important to keep to the general encoding principle that void elements should be avoided in the absence of further information.

[15] Namely: dcr:datcat and dcr:valueDatcat

Adding components to the TEI framework: the syntactic case

Since the *TEI Dictionaries* chapter was initially conceived to account for the kind of information that appears in machine-readable dictionaries, it only sparsely covers features related to language processing and in particular does not propose any specific element for representing syntactic or semantic structures. When one looks at the various additional packages of LMF on the one hand and at the customisation facilities of the TEI infrastructure on the other, it appears to be relatively easy to define extensions that actually allow TEI based customisation to include the missing LMF constructs.

In this section we present the basic principles to be applied to create such a customization that extends the TEI guidelines by means of an ODD specification for the syntactic package of LMF. This presentation will be carried out by going through ta specific example, namely the encoding of verbal structures in CoreNet, the Korean Wordnet.

CoreNet, the Korean Wordnet lexicon (also known as CoreNet, see Choi et alii, 2003 and Choi et alii, 2004) has been put together as a deep semantic and syntactic encoding of a selection of the 50 000 Korean most frequent words (mainly nouns and verbs). Looking at verbs proper, their representation is based upon a double filing system of a) *verb concepts*, associating a concept number (and therefore a Wordnet entry, via a specific conceptual mapping) to the various senses and b) *verb frames*, associating each sense with one or several predicate-argument structure.

Figure 5: An entry from the verb concept section of CoreNet

As illustrated in figure 5 for the verb "chida" (치다), the verb concept structure is organised in senses and sub-senses, to which are attached both a Wordnet reference and a gloss. This two level semasiological representation is indeed entirely construable as a standard TEI <entry> structure as illustrated below:

```
<entry>
   <form>
      <orth type="한글">치다</orth>
```

```
            <orth type="Romanization">chida</orth>
        </form>
        …
        <sense n="3">
            <gramGrp>
                <subc>vt</subc>
            </gramGrp>
            <sense n="1">
                <ref type="wordnet">
                    <idno>1221282691</idno>
                    <gloss>치기</gloss>
                </ref>
            </sense>
            <sense n="2">…
            </sense>
        </sense>
    </entry>
```

The verb frame structure is in turn illustrated in figure 6, where one can see that a complementary semasiological structure is being used, grouping together senses from the verb concept structure (represented here by a combination of concept number and gloss) and associating such groups to one or several predicate argument representations. An additional Japanese gloss is provided for each semantic group, on the basis of the actual semantic restriction introduced for the corresponding arguments.

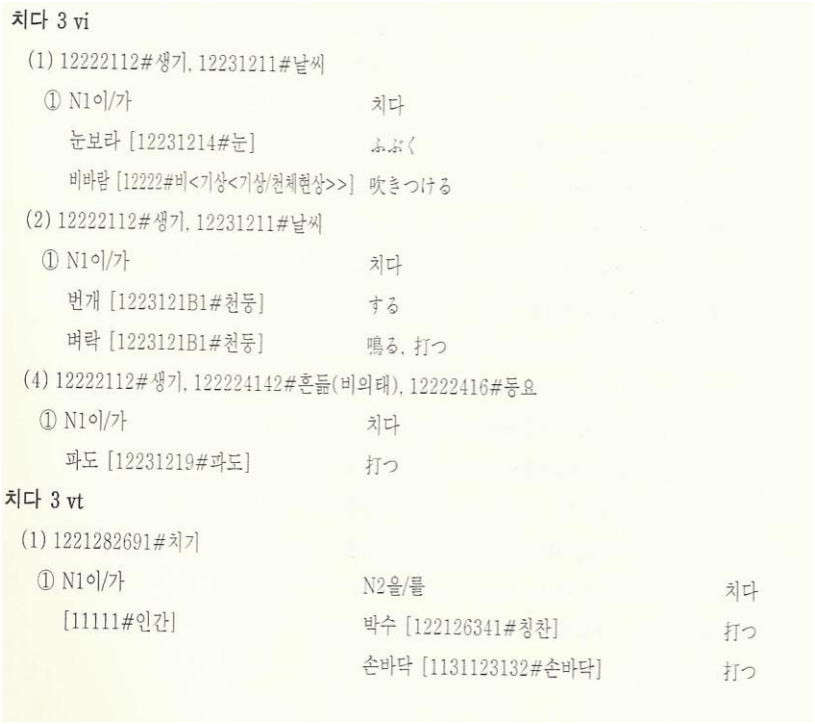

Figure 6: Two entries from the verb frame section of CoreNet

This predicate argument structure is indeed a good instance to the syntactic extension of LMF that is based on the notion of a sub-categorisation frame (component: Subcategorisation Frame), which is then linked to various syntactic arguments (component: Syntactic Argument). Figure 7, which takes up an Italian example from the LMF standard, illustrates this core structure and shows how it is directly anchored on the Lexical Entry component.

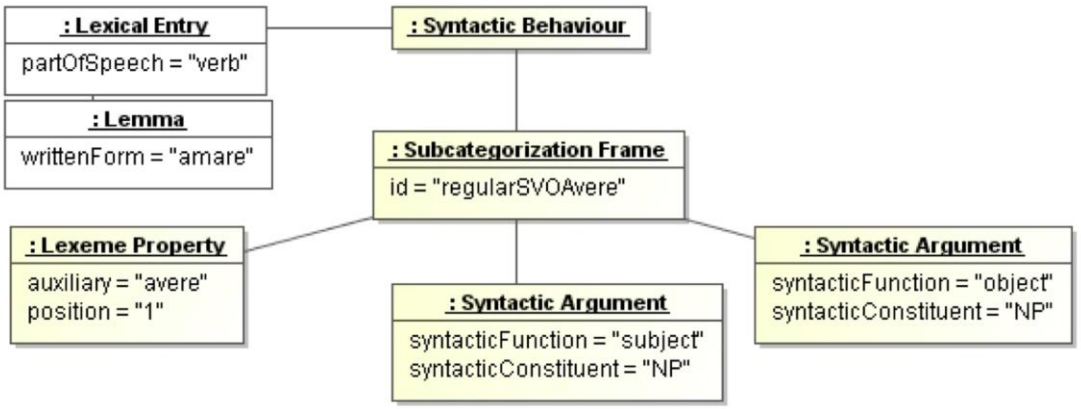

Figure 7: An instance of the LMF syntactic extension (source ISO 24613)

When transposing this model to our KoreNet example, we can actually embed the syntactic description within the sense level of the lexical entry[16]. This leads to a possible TEI extended construct that may look as follows:

```
<tei:sense>
   <tei:gloss xml:lang="jp">ふぶく</tei:gloss>
   <lmf:syntacticBehaviour>
      <lmf:subcategorizationFrame>
         <lmf:syntacticArgument>
            <lmf:syntacticFunction>N1</lmf:syntacticFunction>
            <tei:colloc type="particle">이/가</tei:colloc>
            <tei:gloss>눈보라</tei:gloss>
            <tei:ref type="wordnet">
               <tei:idno>12231214</tei:idno>
               <tei:gloss>눈</tei:gloss>
            </tei:ref>
         </lmf:syntacticArgument>
      </lmf:subcategorizationFrame>
   <lmf:syntacticBehaviour>
</tei:sense>
```

In this representation, we applied the following core specification principles, which, to our view, should be systematically applied for any further TEI based LMF extension:

- Limit the introduction of specific element to those for which there are no equivalent construct in the TEI infrastructure
- Keep new elements within their own namespace. This is a general principle for TEI customization, but it allows here a clear management of the heterogeneous mix-up of elements that we suggest here at all levels of the representation
- Avoid introducing new LMF elements within existing TEI constructs apart from the clear anchoring of the LMF syntax crystal within the <sense> element. This principle is essential at this stage to facilitate the future integration of our proposal as an official extension to the TEI guidelines, where unintended side-effects should be avoided

[16] The full LMF package for syntax is (rightly) intended to allow the factorisation of syntactic construct across several entries. We simplify here the representation to make out point clearer. The full ODD specification should indeed implement both views.

As an aside note, we can see the interesting case of the various usages of the TEI <gloss> element in this representation. Depending on the context, it can be applied in a systematic way to mark any kind of equivalent wording in the various object or working languages of the dictionary.

The actual implementation of such an extension is rather straight forward. Following the general principles outlined in $$REF for implementing a TEI customisation in ODD, we only give here the essential aspects of the proposed syntax extension to the TEI Dictionaries chapter[17].

The first step is to create a background customisation comprising the core modules of the TEI guidelines together with the Dictionaries module as follows:

```
<schemaSpec ident="LMFSyntax">
   <moduleRef key="core"/>
   <moduleRef key="tei"/>
   <moduleRef key="header"/>
   <moduleRef key="textstructure"/>
   <moduleRef key="dictionaries"/>
</schemaSpec>
```

The second step is to create specifications for all new elements within a specific LMF namespace. When such elements have a complex content model, an associated element class is created so that the content model is easy to customise further. For instance, a simplified specification for the <syntacticArgument> element may look as follows:

```
<elementSpec ident="syntacticArgument" module="Syntax"
   ns="http://www.iso.org/ns/LMF">
   <classes>
      <memberOf key="model.subcategorizationFramePart"/>
   </classes>
   <content>
      <rng:oneOrMore>
         <rng:ref name="model.syntacticArgumentPart"/>
      </rng:oneOrMore>
   </content>
</elementSpec>
```

Finally, as seen also in the preceding example each element is made a member of the appropriate classes to appear in the intended content models.

The resulting specification is all in all quite simple and allows one to edit syntactic lexica right away, while remaining within the TEI realm. Moreover, it shows that implementing similar extensions for some additional packages would definitely be an easy tasks that would not take too much time for a minimally TEI minded person.

Contributing to the LMF packages: linguistic quotations

We now address the opposite case to the one we have just seen, namely when some existing constructs in the TEI infrastructure do not have any counterpart in the LMF standard and can thus contribute to defining additional packages. There are indeed several such interesting cases in the TEI guidelines (one may think in particular of all etymological related aspects), but in order to make the point clear we will focus on a simple yet essential type of information: *quotation structures*.

Quotations in a lexical database are linguistic segments that illustrate the use of the headword either as a constructed example, as the citation of an external source or

[17] The complete customisation is available under http://hal.inria.fr/hal-00762664

through the embedding of excerpts that have been automatically extracted and selected from a corpus. In some lexicographic projects (cf. e.g. Kilgarriff and Tugwell, 2001 or Sinclair 1987) such quotations have even beeen the organising principle of the whole lexical matter.

In their simplest form, quotations appear as a textual sequence embedded within other descriptive information of the word, for instance[18]:

> **ain't (eInt)** *Not standard. contraction of* am not, is not, are not, have not or has not: *I ain't seen it*.

When the quotation is actually taken from a known source, it is usually accompanied by an explicit (usually abbreviated) reference to it, as in[19]:

> **valeur** … n. f. … 2. Vx. Vaillance, bravoure (spécial., au combat). 'La valeur n'attend pas le nombre des années' (Corneille).

In the case of multilingual dictionaries, we can extend the notion of quotations to the provision of a translation, possibly accompanied by additional contextualising information. This falls indeed within our earlier definition of a quotation, since such translations actually illustrate the intended meaning in the target language. In the following example we see for instance how such a translation can in turn be refined by an explicit gloss for the corresponding meaning:

> **rémoulade [Remulad]** nf remoulade, rémoulade (*dressing containing mustard and herbs*).

Further types of quotation refinements can be observed in existing dictionaries and indeed, any kind of morpho-syntactic, syntactic or semantic information may be associated with quotations, as long as it provides a qualification for the corresponding usage. Taking again the case of multilingual dictionaries, it is indeed standard practice to refine a translation by means of gender information as in the following excerpt:

> **dresser** … (a) (Theat) habilleur m, -euse f; (Comm: window ~) étalagiste mf. she's a stylish ~ elle s'habille avec chic; V hair. (b) (tool) (for wood) raboteuse f; (for stone) rabotin m.

In this example, we see various types of refinements, with a simple marking of gender for the translation (*habilleur m*), to a combination of morpho-syntactic and semantic constraints (*(for wood) raboteuse f*).

As can be seen, quotation structures are a strong component of the organisation of lexical entries in senses. We are used to observing these in traditional print dictionaries, but indeed, it is easy to foresee a generic mechanism that applies to any lexical database where illustrative text (examples or translations) are to be integrated.

In this respect, the TEI has taken this issue very seriously by introducing in its recent editions (from P5 onwards), a single construct based on the <cit> element[20] that merged the various specific constructs that existed for examples (the <eg> element in

---

[18] Source: TEI P5, chapter "Dictionaries", http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html (original source: *Collins English Dictionary*. London: Collins)

[19] *ibid.* (original source: Guerard, Françoise. *Le Dictionnaire de Notre Temps*, ed. Paris: Hachette, 1990)

[20] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-cit.html

the P4 edition of the TEI guidelines) or translations (the <tr> element in P4). This construct can be characterised as follows:

- it is based upon a very generic two level structure where the <cit> element is the entry point and comprises a language excerpt expressed by means of a <quote> (occasionally a <q>) element;
- the <cit> element may have a @type attribute to further constrain the nature of the quotation construct, for instance “example” or “translation”.

In the simplest case, when no further constraint or bibliographic reference is needed, the <cit> construct boils down to something as simple as the following example representing a translation:

```
<cit type="translation" xml:lang=”fr”>
   <quote>horrifier</quote>
</cit>
```

When further refinements are expressed in relation to the quotation, these are added to the actual quoted sequence, using the usual descriptive vocabulary available from the TEI guidelines. For instance, the provision of the gender for the French equivalent to the headword “dresser” in English would be expressed as follows:

```
<cit type="translation" xml:lang="fr">
   <quote>habilleur</quote>
   <gramGrp>
      <gen>m</gen>
   </gramGrp>
</cit>
```

Finally, an important feature of the <cit> element is its recursivity where for instance the actual translation for a example is also provided, as in the following example:

```
<cit type="example”>
   <quote>she was horrified at the expense.</quote>
   <cit type="translation" xml:lang=”fr”>
      <quote>elle était horrifiée par la dépense.</quote>
   </cit>
</cit>
```

The LMF standard does not have a real equivalent to the <cit> crystal and the only similar structure that appears in LMF may be the possibility to associate a statement in a definition (€€REF). We thus propose to define an optional extension to the LMF core package, anchored on the sense component and schematized in figure 8.

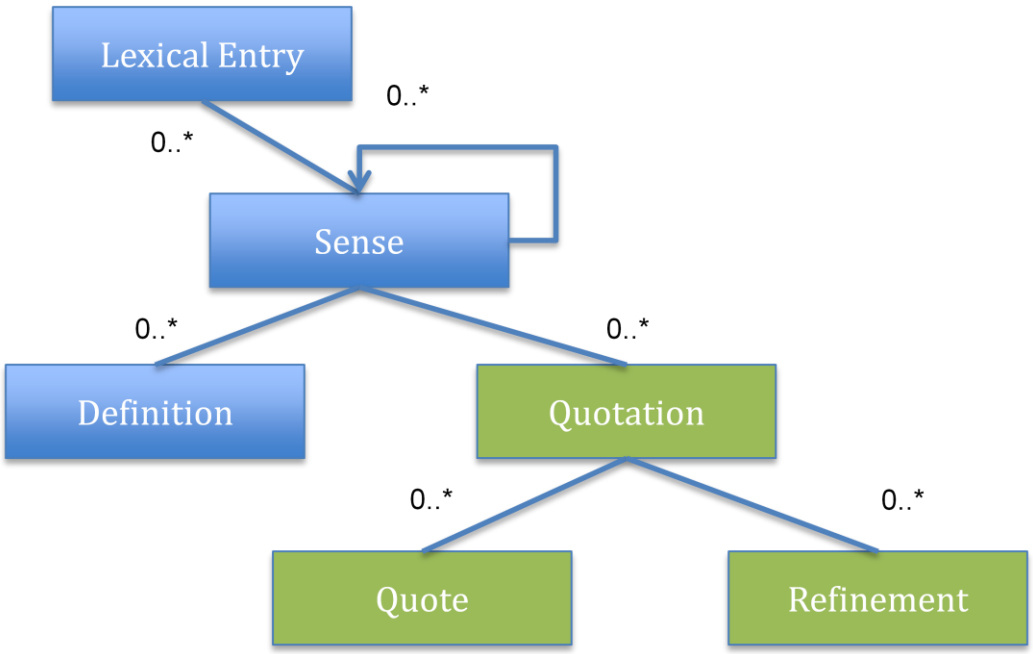

Figure 8: A sketch for a Quotation package in LMF

As we can see, the package is directly part of the Sense component aggregation and further defined as a combination of a Quote (an instance of the Text Representation component in LMF) and a Refinement component.

A further specification process, which should be carried out in concertation with the community of lexical databases developers and users, should clarify what should pertain to the Refinement component in this model. As we have seen, we have here a wide spectrum of possibilities, ranging from authorship or bibliographical information to morpho-syntactic constraints and comprising various alternative forms (pronunciation, variants, translations) or usage information (subject, definition, gloss). Of course, a possible instance of a Requirement may also be a Quotation.

Towards more convergence between initiatives: a roadmap

One of the underlying ails of this paper is to demonstrate that there are some good possibilities ot work towards a better convergence between the LMF and the TEI initiatives in the domain of lexical structures, and in particular take full benefit of each side's strengths. Indeed, whereas the ISO perspective brings stability and an international validation, it should not be neglected how large the current TEI community is. With this perspective in mind, the project of having an LMF serialisation entirely expressed as a TEI customisation can be seen as one of the most important endeavour to offer a common and strong basis for any kind of lexical work both in the language technology and the digital humanities domains. This will also provide LMF with a real customisation platform that will facilitate the work of defining project specific subset within a coherent framework that guaranties compliance to the underlying reference standard.

References

Alt Susanne (2006) "Data structures for etymology: towards an etymological lexical network", BULAG 31 1-12 — http://hal.archives-ouvertes.fr/hal-00110971

Aristar-Dry, H., Drude, S., Windhouwer, M., Gippert, J., and Nevskaya, I. (2012). „Rendering Endangered Lexicons Interoperable through Standards Harmonization": The RELISH Project. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, May 23rd-25th, 2012 (pp. 766-770).

Burghart M. and M. Rehbein, "The Present and Future of the TEI Community for Manuscript Encoding", *Journal of the Text Encoding Initiative* [Online], Issue 2 | February 2012, Online since 03 February 2012, connection on 12 October 2012. URL : http://jtei.revues.org/372 ; DOI : 10.4000/jtei.372

Burnard L. and S. Rahtz (2004) "RelaxNG with Son of ODD". Extreme Markup Languages conference.

Choi Key-Sun, Guseong-dong and Yuseong-gu (2003) "CoreNet: Chinese-Japanese-Korean wordnet with shared semantic hierarchy", Proc. IEEE International Conference on Natural Language Processing and Knowledge Engineering - NLP-KE.

Choi, Key-Sun , Hee-Sook Bae, Wonseok Kang, Juho Lee, Eunhe Kim, Hekyeong Kim, Donghee Kim, Youngbin Song and Hyosik Shin (2004) "Korean-Chinese-Japanese Multilingual Wordnet with Shared Semantic Hierarchy", Proc. LREC 2004, Lisbon, Portugal, 26 May - 28 May 2004

Haddar Kais, Héla Fehri and Laurent Romary (2012) "A prototype for projecting HPSG syntactic lexica towards LMF", *Journal of Language Technology and Computational Linguistics*, 27(1), 21-46; http://hal.inria.fr/hal-00719954.

Ide N. and J. Véronis, (1995). Encoding dictionaries. In Ide, N., Veronis, J. (Eds.) *The Text Encoding Initiative: Background and Context.* Dordrecht: Kluwer Academic Publishers, 167-80.

Ide N., J. Veronis, S. Warwick-Armstrong and N. Calzolari (1992) Principles for encoding machine readable dictionaries, *EURALEX'92 Proceedings*, H. Tommola, K. Varantola, T. Salmi-Tolonen, Y. Schopp, eds., in *Studia Translatologica*, Ser. a, 2, Tampere, Finland, 239-246. Available from: http://www.cs.vassar.edu/~ide/papers/Euralex92.pdf

ISO 1951:2007 Presentation/representation of entries in dictionaries -- Requirements, recommendations and information

ISO 24610-1:2006 Language resource management -- Feature structures -- Part 1: Feature structure representation

ISO 24613:2008 Language resource management - Lexical markup framework (LMF)

Kilgarriff A. and D. Tugwell (2001) "WORD SKETCH: Extraction and display of significant collocations for lexicography." Proc Collocations workshop, ACL 2001

Knuth D. (1984) "Literate Programming " in Literate Programming. CSLI, 1992, pg. 99.

Holmes M. and Romary L. "Encoding models for scholarly literature", in *Publishing and digital libraries: Legal and organizational issues*, Ioannis Iglezakis, Tatiana-Eleni Synodinou, Sarantos Kapidakis (Ed.) (2010) 88-110 - http://hal.archives-ouvertes.fr/hal-00390966

Langendoen, D. Terence and Gary F. Simons, (1995) "A rationale for the TEI recommendations for feature-structure markup." *Computers and the Humanities* 29: 191-209.

Lee Kiyong , Lou Burnard, Laurent Romary, Eric De La Clergerie , Thierry Declerck, Syd Bauman, Harry Bunt, Lionel Clement, Tomaz Erjavec, Azim Roussanaly, Claude Roux (2004) “Towards an international standard on feature structures representation” 4th International Conference on Language Resources and Evaluation - LREC'04 373-376

Pollard C. and I. A. Sag (1994): Head-Driven Phrase Structure Grammar. Chicago: University of Chicago Press.

Romary L., Salmon-Alt S., Francopoulo G. (2004) “Standards going concrete: from LMF to Morphalou”, Workshop on Electronic Dictionaries, Coling 2004, Geneva, Switzerland — http://hal.inria.fr/inria-00121489

Romary L. and W. Wegstein (2012), “Consistent modelling of heterogeneous lexical structures”, *Journal of the Text Encoding Initiative*, Issue 3 | November 2012, Online since 15 October 2012, connection on 09 January 2013. URL : http://jtei.revues.org/540 ; DOI : 10.4000/jtei.540 — http://hal.inria.fr/hal-00704511

Salmon-Alt S., Akrout A., Romary L. (2005). Proposals for a normalized representation of Standard Arabic full form lexica. Second International Conference on Machine Intelligence (ACIDCA-ICMI 2005), Tozeur, Tunisia — http://halshs.archives-ouvertes.fr/halshs-00004541

Salmon-Alt S., L. Romary, E. Buchi (2005). “Modeling Diachrony in Dictionaries”. ACH-ALLC 2005, Vancouver, Canada.

Schmidt T., “A TEI-based Approach to Standardising Spoken Language Transcription”, *Journal of the Text Encoding Initiative* [Online], Issue 1 | June 2011, Online since 08 June 2011, connection on 12 October 2012. URL : http://jtei.revues.org/142 ; DOI : 10.4000/jtei.142

Sinclair J. M. (ed.) 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Collins, London.

Véronis, J. and N. Ide (1992). A feature-based model for lexical databases. *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92*,Nantes, France, 588-594.

Windhouwer, M., and Wright, S. E. (2012). “Linking to linguistic data categories in ISOcat”. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), *Linked data in linguistics: Representing and connecting language data and language metadata* (pp. 99-107). Berlin: Springer.