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 occurs naturally when one observes the history of both initiatives, 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 by 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 on consensus on the strategy to define a reliable and stable XML serialisation has forced the ISO working group on LMF to confine it as 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, it prevents it from being properly maintained, in order to either make corrections on identified weak points or bugs, or to add step by step additional features;
- It is only defined as a DTD, 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 property of LMF to be an extensible model, as it does not contain any dedicated mechanism for customization;
- A more intrinsic weak aspect of the suggested LMF serialisation, in 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 pure 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 which is similar to that of ISO

[1] Not even compliant with ISO standard 16642 (TMF) which defined such an element before ISO 24610 was in place.

1951, that is, creating 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 consensus between comprehensiveness, simplicity and technical adequacy. The topic of our paper is indeed to contribute to improve LMF by considering it getting closer to the TEI, an initiative that has itself gone through many years of fruitful iterations.

TEI as a data modelling environment

Even if 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 now.

The scope of the TEI covers all documents whose main content can be seen as 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 600 or more elements of the TEI guidelines are all defined in a specification language based on the TEI vocabulary itself. In a way, such 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) and 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, with the W3C using it for it ITS recommendation or ISO committee 37 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.

Within 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 (name, gloss, definition, examples and

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

remarks) and technical information (content model based on 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. These are called *model classes*;
- *Attribute classes* are also available to factorise attributes that are used uniformly by several elements;
- Elements may also be grouped together as *modules.*

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.

[Tools Roma, OxGarage, SourceForge]

TEI as a quasi LMF compliant framework

Now that the motivations and general context for our approach has 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), there existed a specific module inspired from these two trends and extensively covering all aspects of typed feature structures, with mechanisms for declaring constraints on them. In 2006, following an agreement between the TEI consortium and ISO, the module became an ISO standard (ISO 2461-1) and is now the reference XML representation for feature structures.

Applying the ISO-TEI feature structure format for representing data 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 simple features with a symbolic value (mapped onto a simple data category in ISOCat).

The controlling of mappings between features and feature values with ISOCat entries can be made either by eliciting the association within a feature structure definition (FSD), or even when describing a feature library to factorise the information expressed within lexical entries. These mechanisms, related to the use of the dcs attributes are based upon the technical description provided in (Menzo, this volume) and will not be elaborated further here.

[Unless present in Menzo's paper, I would put an Annex with an FSD excerpt]

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.

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

Even if one does not want to go as far as using full-pledged feature structures but keep at least the general principles of the LMF serialisation skeleton (element named according to their equivalent component in the meta model), it is still possible to use the ISO TEI feature syntax for the corresponding descriptors in an LMF representation[3]. One possible advantage, beyond a better convergence across standardisation initiatives is that it allows, like was alluded to before, a simple declaration of the corresponding feature in connexion to ISOCat. This is illustrated below with the again "clergyman" example expressed according to the suggested mixed-approach:

```
<LexicalResource xmlns:tei="http://www.tei-c.org/ns/1.0">
   <GlobalInformation>
      <tei:f name="languageCoding" val="ISO 639-3"/>
   </GlobalInformation>
   <Lexicon>
      <tei:f name="language" val="eng"/>
```

[3] 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/

```
            <LexicalEntry>
               <tei:f name="partOfSpeech" val="commonNoun"/>
               <Lemma>
                  <tei:f name="writtenForm" val="clergyman"/>
               </Lemma>
               <WordForm>
                  <tei:f name="writtenForm" val="clergyman"/>
                  <tei:f name="grammaticalNumber" val="singular"/>
               </WordForm>
               <WordForm>
                 <tei:f name="writtenForm" val="clergymen"/>
                 <tei:f name="grammaticalNumber" val="plural"/>
               </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 kind of verbosity, makes it more difficult to maintain by human lexicographers. When this becomes an issues, it is than 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. 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 so well the LMF core package[4]. Still, it is important to keep in mind that the initial 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) does contain much more constructs intended for the representation of human oriented features (typically the etymology of a word) and b) it offers specific "flat" representations intended to cover the early steps of the digitization process.

Whereas we will provide concrete crosswalks examples between the LMF model and theTEO *Dictionaries* chapter of the TEI 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 XX 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 to an entry;

[4] It is all the less a surprise that the TEI principles informed the first ISO meeting in Korea (February 2004) where the LMF ideas have been initially put together (cf. Romary et alii, 2004)

- <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.

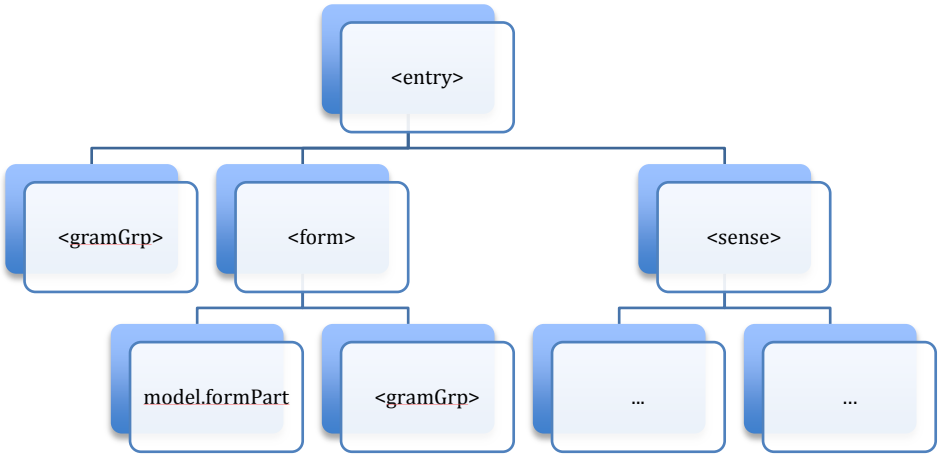


Figure XX: The structural skeleton of the TEI *Dictionaries* chapter

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 one and single representation means for a given phenomenon. Even if the same critic could be addressed, even more fiercely to the LMF standard itself, it is true that the experience gained over the year 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 make sometimes 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 too specific elements for representing a lexical entry in a less structured manner, namely, <entryFree> to allow any kind of combination and order of dictionary components, and <dictScrap>, which lets one have parts of a dictionary entry left unencoded. These alternatives are indeed intended to deal with the specific scenarios of legacy human dictionaries, especially ancient ones, whose entries may not straightforwardly organised (<entryFree>) or in the case of a multi-step scenario (<dictScrap>) whereby an initially OCRed dictionary is manually encoded step by step. All in all, the target scheme, remains a systematic use of <entry>, which warrants the production of LMF compliant data.

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 directly within <entry>. This was 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. Like 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 be definitely used even as external resource to LMF compliant representations[5] that are not expressed using the TEI guidelines proper.

[Example here to illustrate this point]

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 component and identify, a) how the available TEI elements for form description can be matched onto it and b) what perspective it brings about for representing an essential type of lexical objects in various language technology domains, namely, full form dictionaries.

From an LMF point of view, the description of form information within a lexical entry (see figure ZZZ) is based upon 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.

[5] 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.

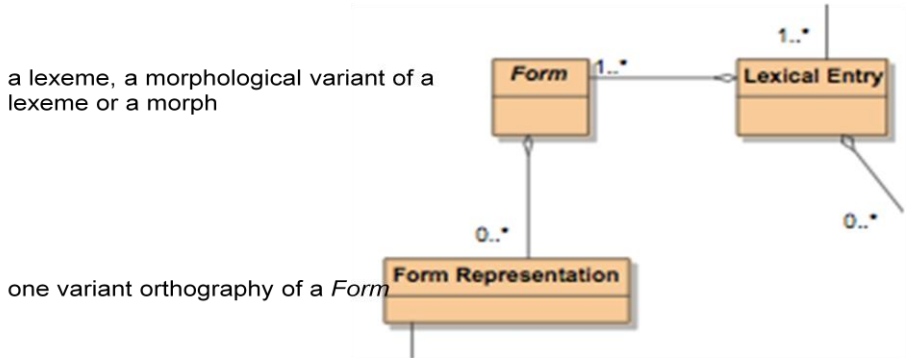


Figure ZZZ: the Form and Form Representation components of the LMF core package

The two-level structure representation is an essential aspect to gain "form autonomy"[6] within a lexical entry. The canonical use of such a construct is typically when a form may occur in several written forms according to the script or transliteration mode being used. For instance, the Hangeul representation of the verb "chida" (en:"to hit") can associated with its Romanized transliteration as sketched below.

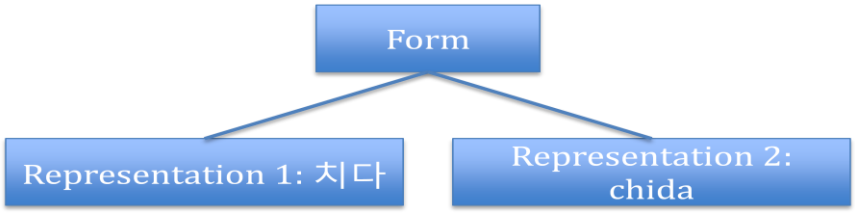
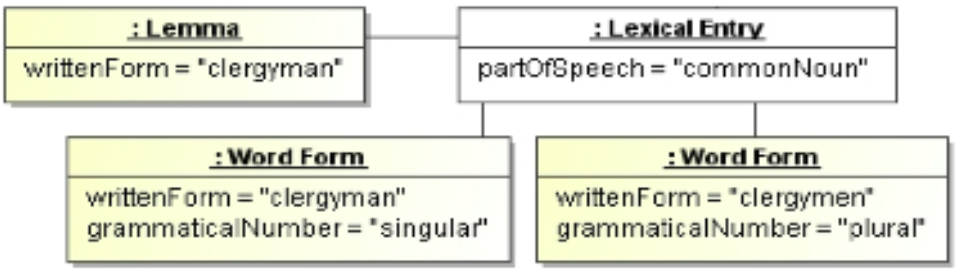


Figure TTT: multiple scripting of the Korean verb "chida"

Given the canonical mapping that exists between the Form-Form representations components in LMF and the form-model.formPart 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, Hangeul vs. Romanized) being used.

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

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



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

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

Table TTT: 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 string basis for the implementation of any kind of full form lexica[7].

```
<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 morph-osyntactic features typically needed for full form lexica. Still, there are several issues 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 systemic 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 one single grammatical feature (<number>) appears in the

[7] See also the first experiments done on the Morphalou dictionary (Romary et alii, 2004)

<gramGrp>, a coherent representation with other word categories (for instance verbs) or other languages, requires that the latter should not be omitted. 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 the 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" val="commonNoun"/>
     </fs>
   </form>
   <form type="inflected">
     <orth>clergyman</orth>
     <fs type="grammar">
       <f name="number" val="singular"/>
     </fs>
   </form>
   <form type="inflected">
     <orth>clergymen</orth>
     <fs type="grammar">
       <f name="number" val="plural"/>
    </fs>
   </form>
 </entry>
```

Finally, we should address here the issue of linguistic coverage, with the possibility to constrain 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 still necessary to describe morpho-syntactic constraints in other languages. To this purpose, the TEI comprises 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 by far recommended, when one has such representational needs to design an *ad hoc* element in one's ODD specification and relate this specification to ISOCat by eans of either the <equiv> construct or the appropriate DCS attributes [to be discussed with Menzo].

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 does hardly cover 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 one to have the missing LMF constructs.

To illustrate this concretely, we take here an experiment carried out on the Korean Wordnet lexicon (REF) and more particularly the verb frame structure associated to verb entries, whose structure corresponds to the syntactic extension of LMF.

•verb_concept means what concept is assigned to each verb. In this file, you can see "chida" (치다). This verb behaves like "take" in English. The number means the concept number. We recently assigned our concept number to PWN synsets.

•verb_frame means a kind of predicate-argument structure. In my sample, it is also for the verb "chida" (치다). Every Korean verb frame has its corresponding Japanese verb in my dictionary.

•noun_concept means what concept is assigned to each noun.

•But the whole vocabulary size is limited to the most frequent 50,000 words that was selected from KAIST corpus. --

Senses

Sub-senses

```
<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>
```

치다 3 vi

(1) 12222112#생기, 12231211#날씨

① N1이/가 치다

눈보라 [12231214#눈] ふぶく

비바람 [12222#비<기상<기상/천체현상>>] 吹きつける

(2) 12222112#생기, 12231211#날씨

① N1이/가 치다

번개 [1223121B1#천둥] する

벼락 [1223121B1#천둥] 鳴る, 打つ

(4) 12222112#생기, 122224142#흔듦(비의태), 12222416#동요

① N1이/가 치다

파도 [12231219#파도] 打つ

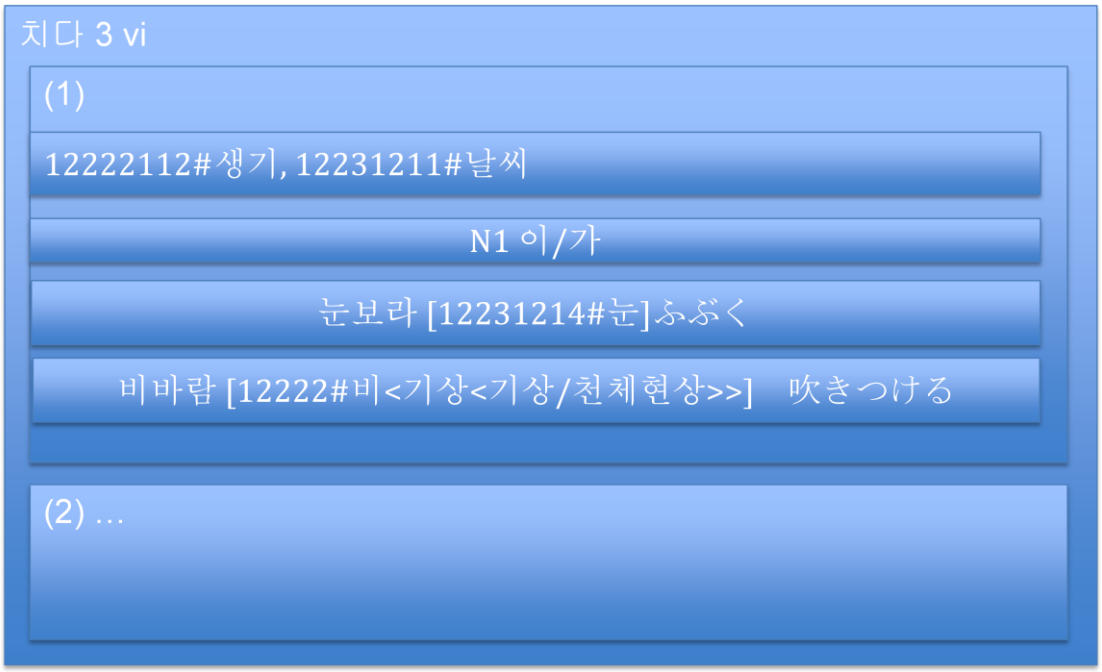
치다 3 vt

(1) 1221282691#치기

① N1이/가 N2을/를 치다

[11111#인간] 박수 [122126341#칭찬] 打つ

손바닥 [1131123132#손바닥] 打つ



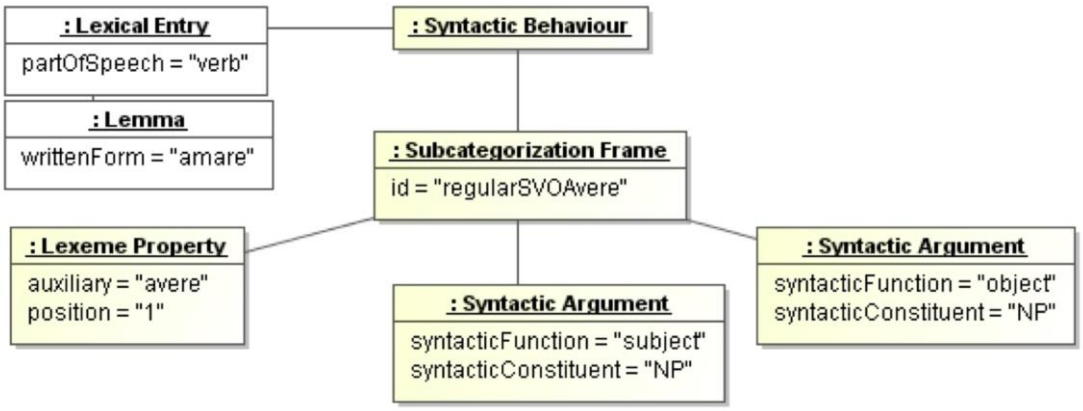
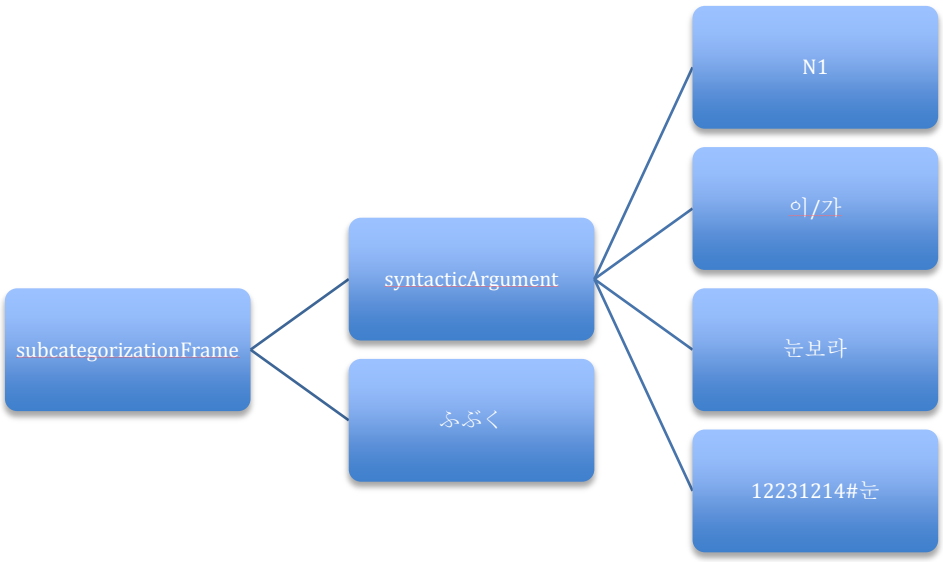


Figure WWW: AN instance of the LMF syntactic extension

Tranposed on our example from the Korean wordnet, this structure easily maps onto the frame description we have there, as depicted in figure ZZZ.



Contributing to the LMF packages: linguistic quotations

We now address the opposite case as 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 define 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: *quotations structures*.

Quotations in a lexical database are all linguistic segments which illustrate the use of the headword either as a constructed example, the citation of an external source or through the embedding of excerpts that have been automatically extracted, or possibly selected, from a corpus. In some lexicographic projects (cf. e.g. Kilgarriff and

Tugwell, 2001 or Sinclair 1987) such quotations may even be 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[8]:

> **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[9]:

> **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 definition, 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[10] that merged the various specific constructs that existed for examples (<eg> element in the

---

[8] 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)

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

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

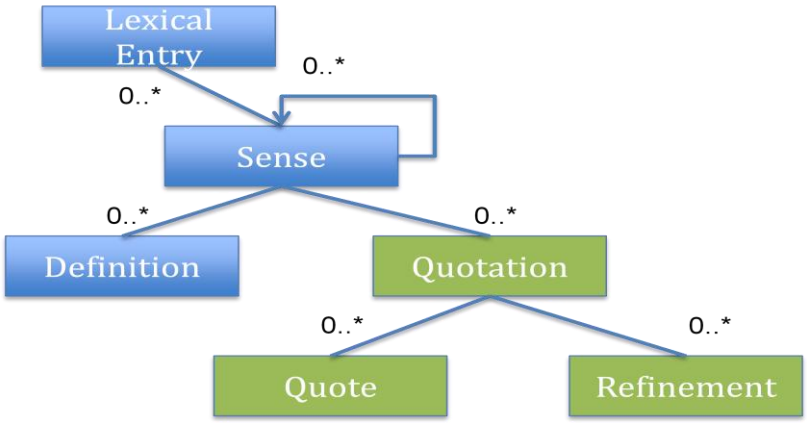
P4 edition of the TEI guidelines) or translations (<tr> element in P4). This construct can be characterised as follows:

- 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 can get a @type attribute to further constraint the nature of the quotation construct, for instance “example” or “translation”.

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

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



- exclusively related to sense
- Quote
- Refinement

- Provision of a source

–Authorship

–Bibliographical reference

- Morphosyntactic refinement

–Orthographical variant, pronunciation

–Grammatical constraints

- Sense refinement

–Usage

Implementation in TEI,

Note that a refinement can also be a quotation :

```
<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>
```

Or even a definition:

```
<cit type="translation" xml:lang="en”>
   <quote>OAS</quote>
   <def>illegal military organization supporting French rule
of Algeria</def>
</cit>
```

More complex entry (LMF compliant with extension)

```
<entry>
   <form>
      <orth>dresser</orth>
   </form>
   <sense n="a">
      <sense>
         <usg type="dom">Theater</usg>
         <cit type="translation" xml:lang="fr">
            <quote>habilleur</quote>
            <gramGrp>
               <gen>m</gen>
            </gramGrp>
         </cit>
      </sense>
      …
   </sense>
   …
</entry>
```

Highly deterministic structure (more constraints than in the TEI, cf. Romary & Wegstein), in the spirit of Ide and Veronis

Included in a wider vision of quotations in text. Cf. humanities papers:

```
Lexicography has shown little sign of being affected by the
work of followers of J.R. Firth, probably best summarized in
his slogan, <cit> <quote>You shall know a word by the company
it keeps.</quote> <ref>(Firth, 1957)</ref></cit>
```

Towards more convergence between initiatives: a roadmap

Cf. TMF-TBX missing chapter

Providing a real customization platform to LMF

Cf. FS

Main danger: oppose communities. Lack of vision. “It is not because we don’t know them that they do not exist.”

References

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.

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, 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 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., 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.

Kiyong Lee, 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. and Pierrel J.-M. (1989) "The Use of the Dempster-Shafer Rule in the Lexical Component of a Man-Machine Oral Dialog System", *Speech Communication* 8, 2 159-176

Romary L., Salmon-Alt S., Francopoulo G. (2004). Standards going concrete : from LMF to Morphalou. Workshop on Electronic Dictionaries, Coling 2004, Geneva, Switzerland.

Romary L. and W. Wegstein (2012), "Consistent modelling of heterogeneous lexical structures", Journal of the Text Encoding Initiative, Issue 4.

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.

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.