TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms

Piotr Bański[1], Jack Bowers[2], Tomaž Erjavec[3]

[1] IDS Mannheim, Mannheim, Germany

[2] Austrian Academy, Vienna, Austria

[3] Jožef Stefan Institute, Ljubljana, Slovenia

E-mail: banski@ids-mannheim.de, iljackb@gmail.com, tomaz.erjavec@ijs.si

Abstract

The paper reviews results of work done in the context of TEI-Lex0, a joint ENeL / DARIAH / PARTHENOS initiative aimed at formulating guidelines for the encoding of retro-digitized dictionaries by streamlining and simplifying the recommendations of the "Print Dictionaries" chapter of the TEI Guidelines. TEI-Lex0 work is performed by teams concentrating on each of the main components of dictionary entries. The work presented here concerns proposals for constraining TEI-based encoding of orthographic, phonetic, and grammatical information on written and spoken forms of the lemma (headword), including auxiliary inflected forms. We also adduce examples of handling various types of orthographic and phonetic variants, as well as examples of handling the representation of inflectional paradigms, which have received less attention in the TEI Guidelines but which are nonetheless essential for properly exposing data content to the various uses that digitized lexica may have.

Keywords: dictionary encoding; TEI XML; TEI-Lex0

1. Introduction

The Text Encoding Initiative (TEI) Guidelines (TEI Consortium, 2016) are the chief deliverable of a project running since the early 1990s and aiming at equipping the scholar with markup suitable for describing the majority of textual forms and analytic approaches and providing extension capabilities to encompass new or infrequently found phenomena. Being a complex toolbox and aimed at being able to encode any existing work, the Guidelines provide multiple encoding solutions and have frequently been criticized on this account. The standard response to such criticism and a recommendation for the purpose of ensuring interoperability has been to fully utilize the TEI's modelling and documentation format, ODD ("One document does it all", cf. TEI Consortium, 2013). However, given that tools with the capacity to parse and semantically analyze ODD descriptions are still being developed, a common-sense strategy to secure interoperability is to come up with a lean, transparent format that may not be able to handle all the potential variation, but will instead address "90% of phenomena, 90% of the time". This is the goal of TEI-Lex0, a joint ENeL / DARIAH / PARTHENOS initiative aimed at formulating guidelines for the encoding of retro-digitized dictionaries by streamlining and simplifying the "Print Dictionaries" chapter of the TEI Guidelines and the module defined therein.

The result is not meant to replace that chapter, but rather to serve as baseline encoding against which existing dictionaries can be compared and which could serve as a pivot format for generic querying or visualization tools.

TEI-Lex0 work is performed by teams concentrating on each of the main components of dictionary entries. The main focus of the present paper is on the `form` element, designed to contain orthographic, phonetic, and grammatical information on written and spoken forms of the lemma (headword), including its inflected forms that are sometimes – depending on the source language and established lexicographic practices – used as auxiliary information for the purpose of identifying the entry, or which illustrate inflectional patterns by means of partial or complete paradigms.

Below, we first present the assumptions that underlie the work of TEI-Lex0, and then proceed to review our proposals for constraining the `form` element and its contents. At each point, an illustration is provided, frequently going beyond use types covered by the TEI Guidelines.

1. General Assumptions

This section presents the basic TEI-Lex0 assumptions relevant to the phenomena described in the remainder of the article.

1.1 Abstract Models and Serialization

A fundamental principle that TEI-Lex0, or virtually any TEI-based dictionary-modelling enterprise enterprise must rely on concerns the nature of the mapping of the physical or "near-physical" (OCR-ed) dictionary structure onto the abstract model of dictionary structure, and the mapping from said model onto its TEI XML serialization.

This is because the TEI vocabulary is heavily restricted and also influenced by some unsystematic historical decisions. The restriction is partially due to the fact that the TEI uses the same elements of the abstract model to serve many kinds of text-modelling tasks, and standardly employs 'features' or 'facets' of these elements to signal differences among them (the features in question are expressed in the XML serialization in the form of attributes, such as, e.g., `@type`). The structural context of these elements often matters as well. The fact that some elements of the serialization have names closely corresponding to what we can customarily find in the dictionary model is more or less a lucky coincidence – it is not a pattern to be expected. A lexicographer coming from outside of the TEI should not, therefore, expect their customary terms (names of dictionary objects in the dictionary model) to be straightforwardly reflected in the TEI vocabulary names.

A good illustration is provided by the elements `form` and `sense`, which might be expected to contain information about form (of the headword and related items) and about the sense, respectively. And they do, except they do it in several ways:

```
<entry>
  <form>
    <orth>bray</orth>
    <pron>brei</pron>
  </form>
...
```

Example 1.

Above, the `form` element behaves as expected, but – as exemplified in Section 2.4 below, it can also nest other `form` elements, and then the outer `form` becomes merely a "box" for form-related information. Similarly with `sense`:

```
<sense>
  <def>cry of an ass; sound of a trumpet</def>
</sense>
```

Example 2.

Above, the element `sense` contains a single definition, but it can also nest other `sense` elements, and then the outer `sense` becomes a "box" for sense-related information within the entry, and its internal structure may reflect the dictionary author's convictions or observations about the relatedness of subsenses, while the ordering of `sense` elements, whether nested or top-level, may express information about the frequency of the given subsense in the base corpus of data (we treat the term "corpus" here to mean the body of data that the lexicographer takes into consideration when creating the dictionary).[1]

The differences in the interpretation of elements such as `form` and other recursive elements make it necessary to adopt in TEI-Lex0 a rule that they may never appear without an accompanying `@type` attribute. Section 3 provides some examples.

1.2 Grammatical Information

In order to determine the complete set of properties of an element constituting a part of a hierarchy of lexicographic objects, onto which a dictionary entry can be mapped, the principle of default inheritance is assumed (cf. Ide et al. (2000) and Erjavec et al. (2000)). According to this principle, grammatical properties of a form are determined by collecting the sibling `gramGrp` of the ancestor-or-self of the focus element, where the superordinate grammatical properties can be overwritten by the lower-level properties. This principle is relatively straightforward in the case of grammatical properties, but more complex for the word paradigm, especially for variant forms.

1 Another relevant example, to which much discussion in the TEI-Lex0 group was devoted, is the `cit` element. Originally, its name derives from "citation", but its semantics has got generalized over time to the point where a more suitable name could be "container-inside-text", given the range of uses and contexts, for and in which it is now applicable.

The *modus operandi* assumed in the TEI-Lex0 is reductionist: from among the variety of means of encoding the relevant information offered by the TEI, precise guidelines for the placement and content of the `form` and `gramGrp` elements are proposed, extending to finer-grained elements of the former such as `orth` for orthography and `pron` for pronunciation, and, in the case of the latter, to various subtypes of the the `gram` element.

2. Recommendations for the Encoding of <form>

This section reviews most of the TEI-Lex0 recommendations for the treatment of `form` and dependent elements, including the treatment of `gramGrp`.

2.1 Grammatical Information

Grammatical properties of lexical entries should be specified in `entry/gramGrp`.[2] This element will typically specify at least the part-of-speech of the entry, sometimes with some further specifications, such as, for example, transitivity for verbs or gender for nouns. While the TEI has defined a number of specialized elements within `gramGrp`, TEI-Lex0 takes a more generic route in this respect, for reasons of uniformity and sustainability. The former criterion makes it possible to simplify the processing tools and unify the representation. The latter makes the format more resilient to future modifications of the TEI: if, for example, at some point in the future, the TEI defines an element `voice` for grammatical voice, the TEI-Lex0 guidelines will not need to be adjusted – all that will be necessary will be another mapping between, say, `<voice>active</voice>` in the target dictionary and `<gram type="voice">active</gram>` in TEI-Lex0. This last point is also a reminder that TEI-Lex0 is not meant as production format, but rather as the base layer for retro-digitization, and possibly a pivot format to mediate between particular implementations of the "Print Dictionaries" chapter of the TEI Guidelines.

```
<entry xml:lang="en">
  <form type="lemma"><orth>on</orth></form>
  <gramGrp><gram type="pos">prep</gram></gramGrp>
  ...
</entry>
```

Example 3.

Because the part-of-speech property is a property of the entire entry, by the principle of default inheritance mentioned in Section 2.2, it is mandatory to encode it as a direct child of the `entry` element (recall that it is inherited by the `form` element, in the absence of a conflicting specification). In cases reviewed in the following sections,

2 A `gramGrp` element that is a child of an `entry` element. The TEI format is an application of XML, and as such, it follows all the practices, conventions and restrictions that govern XML representations. For the sake of explicitness, we utilize the XPath conventions for referencing fragments of XML structure, and thus "a `gramGrp` element that is contained inside a `form` element bearing an attribute `@type` with the value 'lemma', which in turn is contained within the element `entry`" is concisely expressed as `entry/form[@type="lemma"]/gramGrp`.

where grammatical properties pertain to the headword alone or to its various inflections, the `gramGrp` element with appropriate content is placed as a child of `form[@type="lemma"]`, etc.

By the same token, in cases where headwords are distinguished only on the basis of their orthography (e.g. in dictionaries of English which treat conversion pairs of nouns and verbs, such as *run*, as belonging in single entries), `entry/gramGrp` should not be used, because its role is taken over by the individual `sense/gramGrp` elements, which either further specify grammatical properties of the individual sense or override those that pertain to the entire entry.

2.2 Representation of the Lemma

The `form` element should always be qualified by its `@type` attribute set to one of the recommended values. The lemma (i.e., headword) should be under `form[@type="lemma"]`. This is illustrated in Example 3 above.

If it is necessary to specify the grammatical properties of the lemma form itself (as opposed to the grammatical properties of entire the entry), the relevant `gramGrp` element should be a child of `form[@type="lemma"]`. This may occur in languages such as Hebrew, where verbs are lemmatized as help 3rd Person Masculine (simple) Perfect, or Greek, where verbs are lemmatized as 1st Person Singular (Active Indicative). In such cases, however, the relevant grammatical information is encoded mostly for the purpose of machine interpretation rather than for direct human consumption, and various project-dependent choices may regulate its actual placement. We will therefore not dwell on such issues here.

2.3 Representation of the Inflected Forms

Dictionaries often include additional forms next to the lemma. These forms in many cases specify irregular inflectional forms, such as *corpus* / *corpora* or *take* / *took,* while in inflectionally rich languages they enable the user to determine the correct paradigm of the word (e.g., *krava* / *-e* in Slovene or *amo* / *amare* in Latin).

Such inflected forms should be encoded in `entry/form[@type="inflected"]`, e.g.:

```
<entry>
  <form type="lemma"><orth>go</orth></form>
  <form type="inflected">
     <orth>went</orth>
  <gramGrp><gram type="tense">past</gram></gramGrp>
  </form>
...
```

Example 4.

2.4 Paradigms

When several inflected forms can be present next to the lemma, these can be embedded in an `entry/form[@type="paradigm"]` element. The decision on whether to use this extra element depends on the particular dictionary and language.

The other use case for paradigms is when the full inflectional paradigm of the word is embedded in the entry, i.e. when the dictionary also includes all the word-forms of the words covered, which can be useful for example for machine processing.

An `entry` may contain several paradigms, for example a partial one for humans and a full one for machines, or one for each stem of a verb. Each paradigm type should be distinguished by the `form/@subtype` attribute.

```
<entry xml:id="perder" xml:lang="es">
  <form type="lemma">
    <orth>perder</orth>
  </form>
  <gramGrp><gram type="pos">verb</gram></gramGrp>
  <form type="paradigm" subtype="present">
    <form type="inflected">
      <orth>pierdo</orth>
      <gramGrp>
        <gram type="person">1</gram>
        <gram type="number">sg</gram>
        <gram type="mood">indicative</gram>
        <gram type="voice">active</gram>
      </gramGrp>
    </form>
   <!-- other inflected forms (of present indicative) here -->
    <gramGrp><gram type="tense">present</gram></gramGrp>
  </form>
  <form type="paradigm" subtype="preterite">
    <form type="inflected">
      <orth>perdí</orth>
      <gramGrp>
        <gram type="person">1</gram>
        <gram type="number">sg</gram>
        <gram type="mood">indicative</gram>
        <gram type="voice">active</gram>
      </gramGrp>
    </form>
    <gramGrp><gram type="tense">preterite</gram></gramGrp>
  </form>
...
</entry>
```

Example 5.

2.5 Representation of Variants

The representation of variation within a form is highly dependant upon the specifics of what exactly varies, and how. As a general principle, variation may be encoded as

`form[@type="variant"]` and embedded within the parent element for which a subordinate feature exhibits variation. Variation within the form can occur with regard to the orthographic representation or the phonetic realization of a given form.

2.5.1 Orthographic Variation

Several kinds of orthographic variation may be distinguished. Below, we present some of the options with the corresponding examples.

The first example addresses spelling variation due to change in a language's orthography conventions.

```
<entry xml:id="Flussschifffahrt" xml:lang="de" type="compound">
  <form type="lemma">
     <orth>Flussschifffahrt</orth>
  <form type="variant">
     <orth>Fluss-Schifffahrt</orth>
  </form>
  <form type="variant">
     <orth notAfter="1996">Flußschifffahrt</orth>
     <usg type="time">Vor 1996 Rechtschreibung Reform</usg>
  </form>
  <gramGrp><gram type="pos">noun</gram></gramGrp>
....
</entry>
```

Example 6.

In the following example, the Hebrew word אֹמֶץ 'courage' can be represented by either the 'dotted' ('vowelized') spelling, or by the full spelling, where vowels are marked as separate characters.

```
<entry xml:id="courage-heb" xml:lang="heb">
  <form type="lemma">
    <form type="variant">
      <orth notation="menukad">אֹמֶץ</orth> <!-- 'dotted' spelling -->
     </form>
     <form type="variant">
      <orth notation="male">אומץ</orth> <!--full spelling -->
     </form>
     <pron notation="ipa">ˈomet͡s</pron>
  </form>
  <gramGrp><gram type="pos">noun</gram></gramGrp>
  <sense> .... </sense>
</entry>
```

Example 7.

Note that in Example 7, the phonetic representation is provided as well, according to the conventions of the International Phonetic Alphabet. The above encoding proposal might be opposed on the grounds of verbosity. However, TEI-Lex0 is primarily meant to be a derived representation format for the purpose of exchange or processing, and the primary stress is on explicitness. A project-internal representation might express

the variation simply by putting two `orth` elements next to one another, within a single `form`. In TEI-Lex0, the otherwise potentially spurious additional `form[@type="variant"]` is a matter of coherence and explicitness.

The next example illustrates a fragment of an American English dictionary in which, due to the lack of official conventions for transliteration of Arabic orthography to the English (Latin) script, the initial vowel in the surname 'Osama Bin Laden' varies between 'O' and 'U'.

```
<form type="lemma">
  <pron notation="ipa">
     <seg xml:id="ousma" corresp="#usma #osma">ow."sa.ma</seg>
     bɪnˈlaːdn̩</pron>
  <form type="variant">
     <orth type="transliterated">
        <seg xml:id="osma" corresp="#usma #ousma">Osama</seg>
        Bin Laden</orth>
  </form>
  <form type="variant">
     <orth type="transliterated">
       <seg xml:id="usma" corresp="#osma #ousma">Usama</seg>
       Bin Laden</orth>
  </form>
</form>
```

Example 8.

Note that the `seg` element is used for the purpose of providing an anchor for linking and at the same time it provides a place for the `@corresp` attribute, used to express the relevant correspondence.

2.5.2 Phonetic Variation

The example entry below contains a single orthographic form as well as phonetic transcriptions of the two roughly equally used variant pronunciations of the word 'caramel' in American English. Since all this information pertains to the lemma, it is contained within a single `form[@type="lemma"]` element.

```
<entry xml:id="caramel-en" xml:lang="en-US">
  <form type="lemma">
     <orth>caramel</orth>
     <form type="variant">
        <pron notation="ipa">'keɹə"mɛl</pron>
     </form>
     <form type="variant">
        <pron notation="ipa">'kaɹmɬ</pron>
     </form>
  </form>
  <gramGrp><gram type="pos">noun</gram></gramGrp>
...
</entry>
```

Example 9.

2.5.3 Regional and Dialectal Variation

In the following example from Mixtepec-Mixtec, there is variation in the form of the word for the city of Oaxaca between speakers from the village of Yucanany and the rest of the speakers. Since the Yucanany variety makes up only a small portion of the speakers of the language, this case of variation is represented as an embedded `form[@type="variant"]` within the lemma. Note the use of `usg[@type="geo"]/placeName` to explicitly specify this feature in addition to the use of the private language subtag "mix-x-YCNY" as per BCP 47 (Phillips and Davis, 2009).

```
<entry xml:id="Oaxaca-MIX" xml:lang="mix" type="compound">
  <form type="lemma">
     <orth>Ñuu Ntua</orth>
     <pron notation="ipa">ɲùùndùá</pron>
     <form type="variant" xml:lang="mix-x-YCNY">
        <orth>Ntua</orth>
        <pron notation="ipa">ndùá</pron>
        <usg type="geo">
           <placeName>Yucanany</placeName>
        </usg>
     </form>
  </form>
  <gramGrp>
     <gram type="pos">locationNoun</gram>
  </gramGrp>
...
</entry>
```

Example 10.

3. Summary

TEI-Lex0 focuses on staking a certain consistent path across the variety of choices offered by the TEI Guidelines, with an eye to establishing recommendations for a baseline encoding of the products of retro-digitization and at the same time a certain pivot format that may be further uniformly processed and queried. In this very paper, we concentrated on presenting a glimpse of the TEI-Lex0 effort pertaining to encoding information on the parts of entries that specify formal and grammatical features.

We have adduced examples of how orthographic and phonetic variants can be handled, and looked at the representation of inflectional paradigms, which have not received much attention in the TEI Guidelines but which are nonetheless essential for properly exposing data content to the various uses that digitized lexica can have.

4. Acknowledgements

The results presented here have been formulated in the course of work of the “Berlin task force” of the ENeL-DARIAH-PARTHENOS TEI-Lex0 initiative. We are

thankful to our colleagues for extensive discussions and brainstorming over the past months. We also stress that the ideas presented here are coloured by our subjective views, and that TEI-Lex0 is not yet a finished set of coherent guidelines. Some differences between this paper and the final TEI-Lex0 deliverable may be expected, and the blame for any deviations from the eventual specification is solely ours.

5. References

Erjavec, T., Evans, R., Ide, N., Kilgarriff, A. (2000). The CONCEDE Model for Lexical Databases.. Proceedings of the Second Language Resources and Evaluation Conference (LREC), Athens, Greece, 355-62. Available at: http://www.lrec-conf.org/proceedings/lrec2000/html/summary/335.htm

Ide, N., Kilgarriff, A., Romary, L. (2000). A Formal Model of Dictionary Structure and Content. Proceedings of Euralex 2000, Stuttgart, 113-126. Available at: https://www.kilgarriff.co.uk/Publications/2000-IdeKilgRomary-Euralex.pdf

Phillips, A. and M. Davis (eds). 2009. Tags for Identyfying Languages. BCP 47, RFC 5646. IETF. Available at: https://tools.ietf.org/html/bcp47

TEI Consortium, eds. (2013). Getting Started with P5 ODDs. Available at http://www.tei-c.org/Guidelines/Customization/odds.xml

TEI Consortium, eds. (2016). TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Version 3.1.0]. [Last updated on 15th December 2016]. TEI Consortium. Available at: http://www.tei-c.org/Guidelines/P5/ ([accessed on February 13th, 2017]).