TEI Lex-0 Etym – towards terse recommendations for the encoding of etymological information

Jack Bowers, Axel Herold, Laurent Romary, Toma Tasovac

TEI Lex-0 Etym – towards terse recommendations for the encoding of etymological information

Jack Bowers, Axel Herold, Toma Tasovac, Laurent Romary

Abstract

The present paper describes the etymological component of the TEI Lex-0 initiative which aims at defining a terser subset of the TEI guidelines for the representation of etymological features in dictionary entries. Going beyond the basic provision of etymological mechanisms in the TEI guidelines, TEI Lex-0 Etym proposes a systematic representation of etymological and cognate descriptions by means of embedded constructs based on the <etym> (for etymologies) and <cit> (for etymons and cognates) elements. In particular, given that all the potential contents of etymons are highly analogous to those of dictionary entries in general, the contents presented herein heavily re-use many of the corresponding features and constraints introduced in other components of the TEI Lex-0 to the encoding of etymologies and etymons. The TEI Lex-0 Etym model is also closely aligned to ISO 24613-3 on modelling etymological data and the corresponding TEI serialisation available in ISO 24613-4.

Acknowledgements

The work described in this paper has benefited from a lot of interactions with all members of the DARIAH working group on lexical data and various colleagues from the EU project Elexis.

Introduction

Building off of recent efforts addressing etymology in TEI (Bowers and Romary 2017; Salmon-Alt 2006) in combination with the work carried out in ISO (ISO project 24613-3[1]), TEI Lex-0 Etym defines a more restrained set of options for encoding any given single etymological phenomenon. The recommendations herein are designed to be able to handle born-digital as well as retro-digitized print sources, for which more flexible representation mechanisms may be needed. The present proposal is named after “TEI Lex-0” (Romary and Tasovac 2018), an initiative launched in 2016 under the auspices of the DARIAH working group on lexical resources, which aims at defining a pivot format for the integration and query of heterogeneous TEI-based lexical resources[2].

The scope of our proposal covers the usage of the following concepts central to etymological description:

- Structuring etymologies through ordering and (optionally) recursivity
- Typology of etymological processes
- Etymons, their forms, senses, and additional characterising information
- Related forms (cognates, derivatives, and others)
- Temporality of etymological processes
- Bibliographical references in etymologies
- Prose description of etymological process and content
- Provenance, opinion, conflicting/divergent etymological accounts

[1] https://www.iso.org/standard/75410.html
[2] See the corresponding GitHub project under https://github.com/DARIAH-ERIC/lexicalresources

Additionally, where the content is overlapping, the recommendations in this paper are intended to reuse those which are defined in other sections of TEI Lex-0. Any exceptions to this are due to specific needs of a given feature.

I. Overview of TEI elements considered in this paper

<etym>

The basic element within which all etymological content should be described is the etymology element <etym>. With a few modifications which are described below, the basis for the use of the <etym> element is the same as described in (Bowers and Romary 2017) which features three options for its placement in an entry:

- as a child of <entry>, when describing the history of the lexical entry as a whole
- as a child of <sense> for sense-based changes[3]
- (in conjunction with one of the above) embedded (0..n) times within another <etym> to represent multiple ordered processes in sequence

Another key feature is that the attribute @type can be used on <etym> to specify etymological processes explicitly. If the process has subtypes, @subtype can also be used (see section *Nested and typed structure*). Finally, the certainty attribute @cert can be used in cases where the etymological description may not be certain. The default values of @cert are “high”, “medium”, “low” or “unknown”, editors could use any combination of these as needed.

Subcomponents of <etym>

Within <etym> the following elements can occur any number of times (some occur within each other):

- <cit @type> for complex descriptions of linguistic signs and their properties. The two possible usages in the context of etymogical representation are etymons (cit/@type=’etymon’) and cognates (cit/@type=’cognate’). We follow here the recent developments related to the ISO project 24613-3, see (Khan and Bowers 2020). The element may have the appropriate xml:lang attribute to indicate the actual language of the etymon or cognate;
- <lang>, @expand (full name of language for documentation purpose), @norm (an encoded value according to BCP 47)[4] for marking up references to languages mentioned in the etymological process
- <date>[5] for dating information (complementarity to <lang>)
- <bibl> and <biblStruct> for (complete) bibliographical references presented inline

[3] Another case in which <etym> may be embedded in <sense> is in the case of where such content is present in a legacy print source.
[4] https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry
[5] <date> and <bibl> can also occur within <cit>

- <ref type="bibl"> for pointers to bibliographical entries stored elsewhere
- As an option depending on editorial practices, <seg type="desc"> for spans of prose that do not represent any of the information types described above
- <lbl> to mark up short intertwining descriptive or connecting markers, particularly in cases of cross references (eg. *cf., see, etc.*)
- <note> for editorial notes that are not part of the actual etymological description in <seg type="desc">
- <xr> (combined with <lbl> and <ref> as specified in TEI Lex-0) for cross-references to other lexical entries, forms or senses, typically reflecting lexical-semantic relations
- In the case where explicit etymological links have to be expressed, an additional <link> element may be used to link etymons or cognates as described in ISO 24613-3 and ISO 24613-4

Basic Components of Etymons, Related Forms and Other Components of Etymologies

Other than <seg type="desc">, <bibl>, <date>, and <note>, the rest of the most important components of an etymology, which are described in the following sections are encoded as children of a typed <cit> element for describing etymons (type="etymon") or cognates (type="cognate").

<cit> can contain:

- <form> for describing the actual form corresponding to the intended etymon or cognate with the same constraints as those that apply in the general TEI Lex-0 specification. In specific cases, when the information is not provided at the etymon level, the element may have the appropriate xml:lang attribute to indicate the actual language of the form;
- <gramGrp> for providing the grammatical properties associated with the etymon or cognate (as in TEI Lex-0);
- <lang>[6] for mentioned names of the languages. When both a form and a language are provided, it may be appropriate to use @norm on <lang> with the same value as the language indication on form/@xml:lang or cit/@xml:lang ;
- <date> for dating information (period of occurrence of the etymon, whether attested or inferred)
- <gloss> in the case of a simple equivalent or paraphrase in the working language of the dictionary
- <def> for lexicographic definitions of the etymon or cognate
- <sense> when the description of the etymon or cognate requires a structured semantic description (as in TEI Lex-0)
- <usg> for usage information (as in TEI Lex-0)
- <xr> (combined with <lbl> and <ref>) for additional lexical, etymological or semantic relations, e.g. ‘meronymOf’ (see (Crist 2005) for an in-depth discussion of such possible relations)

[6] <lang> is currently not allowed in <cit>, change in the content model is required; specifically, <lang> should be made a member of model.entryPart

- <ref type="bibl" @target> for references to bibliographic entries described elsewhere in the encompassing document, and possibly <bibl> as an alternative, when no central bibliographical management is anticipated for the current dictionary

It should be noted here that making a clear cut decision as to what is to be considered as a gloss or definition is not always straightforward, in particular in the case where the meaning is given by a paraphrase in the working language (e.g. 'hin- und herlaufen'). Still, lexicographic practice has led us to keep the two possibilities and require encoders to document precisely the differentiation criteria they have used concerning these two elements. Moreover, in keeping with the general principles of TEI Lex-0, we strongly recommend that any complex semantic description associated with an etymon be actually embedded within a container <sense> element.

II. Structuring an etymology

Minimal TEI Lex-0 Etymology Encoding (flat, non-typed)

The most fundamental requirement of any TEI encoding of etymological information should simply be to include this information inside the <etym> element. In marking up an etymological entry, there are several key structural decisions that will be up to the encoder and should conform to the source and target structure of the data itself. Minimal adherence to TEI Lex-0 Etym requires only that the data be encoded using the elements described above, i.e. that all text content be wrapped in the particular element(s) specified for their data type, with <seg type=”desc”> remaining an option depending on editorial practices. Optionally, users can include multiple layered <etym> elements which may also be typed. In the sections below we describe each basic possibility, their uses, and the specifics of their encoding.

In the example below, from the Kluge’s Etymological dictionary of German (1975), we demonstrate the minimal encoding of the entry components.

> **Eingang m**.
> mhd. īnganc, nnl. ingang, dän. indgang, schwed. ingång: Lehnübersetzung des lat. introitus.
>
> Aus dem ‘Hineingehen’ als Handlung ist die ‘Stelle, an der man ins Haus, in den Saal geht’ geworden, neuerdings auch die ‘Gesamtheit der eingegangenen Geschäftssachen, Mannschaften’ usw. Vgl. Zugang. (Kluge, 1975) p.159

```
<entry xml:id="Eingang" xml:lang="de">
        <form type="lemma"><orth>Eingang</orth></form>
        <gramGrp><gen>m.</gen></gramGrp>
        <etym>
                <cit type="etymon" xml:lang="gmh">
                        <lang expand="Mittelhochdeutsch" norm="gmh">mhd.</lang>
                        <form><orth>īnganc</orth></form>
                </cit>
                <cit type="cognate" xml:lang="nl">
                        <lang expand="Neuniederländisch" norm="nl">nnl.</lang>
```

```
                <form><orth>ingang</orth></form>
            </cit>
            <cit type="cognate" xml:lang="da">
                <lang expand="Dänisch" norm="da">dän.</lang>
                <form><orth>indgang</orth></form>
            </cit>
            <cit type="cognate" xml:lang="sv">
                <lang expand="Schwedisch" norm="sv">schwed.</lang>
                <form><orth>ingång</orth></form>
            </cit>
            <lbl>Lehnübersetzung des</lbl>
            <cit type="etymon" xml:lang="la">
                <lang expand="Latein" norm="la">lat.</lang>
                <form><orth>introitus</orth></form>
            </cit>
            <note>Aus dem ‘Hineingehen’ als Handlung ist die ‘Stelle, an der man ins Haus, in den
            Saal geht’ geworden, neuerdings auch die ‘Gesamtheit der eingegangenen
            Geschäftssachen, Mannschaften’ usw. Vgl. <xr type="related"> <ref
            type="entry">Zugang</ref>. <ref type="bibl">(Kluge, 1975) p. 159</ref></xr></note>
        </etym>
</entry>
```

Example 1: Minimal encoding of an etymological description (source: Kluge, 1975)

Note that the use of etymons, and cognates will be discussed in detail in the next section *Etymons and other forms*.

In this example all the main data components are tagged in the same relative location as in the printed source. Note also that even though no explicit typology is used here (i.e. absence of etym/@type), this encoding still contains a significant amount of machine retrievable information pertaining to the etymological processes involved, i.e. in <lbl> there is the word “Lehnübersetzung” *loan translation* (aka “calque”) and “des lat. introitus” *of the Latin introitus* which contains the source language. Additionally, the presence of the Middle High German language (<lang>mhd.</lang>) would enable researchers to infer the process of inheritance into Modern German.

In any given project where terminology is consistent[7], and where the proper references to parent language stages are present, the presence of such information will enable a certain degree of machine retrievability even without adding any additional structure in the TEI encoding.

Ordering of embedded etymologies to encode chronology

[7] In the case of datasets (original or legacy) that do not use consistent terminology variation in the terminology should be normalized to allow for maximally systematic search and retrieval possibilities.

Where an entry has an etymology with multiple stages, the embedded <etym> elements should be ordered so that the element at the highest position in the hierarchy represents the most recent stage and the one at the lowest position represents the oldest stage.

```
<etym>Inherited from Middle English X
        <etym>from Old English Y
                <etym>which was borrowed from Latin Z
                        <etym>which was from the Proto Italic Q
                                <etym>from Proto Indo-European ʊ̆</etym>
                        </etym>
                </etym>
        </etym>
</etym>
```

Example 2: Embedded <etym> stages: source ordered (most to least recent)

While the structure of the source in the example above is the ideal case in that the ordering of the contents is also from the most to least recent, there may be data sources that present the etymology in the reverse order. In such cases, this structure can nonetheless be maintained in the XML hierarchy as shown in example 2 below:

```
<etym>
  <etym>
    <etym>
      <etym>ultimately from Proto Indo-European ʊ̆</etym>
      which was from the Proto Italic Q
    </etym>
    borrowed from Latin Z
  </etym>
  inherited from Middle English X
</etym>
```

Example 3: Embedded stages: source ordered (most to least recent)

Nested and typed structure

The nested <etym> structure was introduced in Bowers & Romary (2016) and allows for the recursion of an <etym> for the purposes of encoding multiple stages of an etymology and/or where an etymological change is complex and is inherently comprised of multiple interacting processes. Typing can of course be done using the @type and if a project's taxonomy/ontology of etymological processes has subtypes (i.e *calques* or *loan translations* are a subtype of the process *borrowing*), the @subtype attribute can also be used.

Re-examining example 1 above from Kluge (1975), we can see that it actually has two potential etymological layers which can be further structured. The entry implicitly states: that a) the word is *inherited* from Middle High German '*īnganc*'; and b) a common ancestor of both German '*Eingang*' as well as the cognate forms in other Germanic languages share the same source; and it explicitly states that the word in each of those languages is a loan translation

'Lehnübersetzung', i.e. a *calque* from the Latin *'introitus'.* In this alternate encoding of the etymology portion of this entry, the chronological ordering of the etymological processes is represented in the structure as described in the previous section, from most recent (i.e. inheritance from Middle High German) on the uppermost <etym> to least recent (i.e. loan translation/calque from Latin) on the lowermost (embedded) <etym>.

```
<etym type="inheritance">
  <cit type="etymon" xml:lang="gmh">
    <lang expand="Mittelhochdeutsch" norm="gmh">mhd.</lang>
    <form><orth>īnganc</orth></form>
  </cit>
  <!-- cognates here -->
  <etym type="borrowing" subtype="calque">
    <lbl>Lehnübersetzung des</lbl>
    <cit type="etymon" xml:lang="la">
      <lang expand="Latein" norm="la">lat.</lang>
      <form><orth>introitus</orth></form>
    </cit>
  </etym>
  <note>Aus dem 'Hineingehen' als Handlung ist die 'Stelle, an der man ins Haus, in den Saal geht' geworden, neuerdings auch die 'Gesamtheit der eingegangenen Geschäftssachen, Mannschaften' usw. Vgl. <xr type="related"> <ref type="entry">Zugang</ref>. <ref type="bibl">(Kluge, 1975) p. 159</ref></xr></note>
</etym>
```

Example 4: Alternate encoding of etymology of 'Eingang' from Example 1 (source: Kluge, 1975)

This nested structuring is not always possible, especially in cases of retro-digitized print sources as the ordering in which the forms and data are presented may limit or prohibit the use of nesting in any kind of systematic way. See (Bowers and Romary 2017) for more in-depth discussion of embedding <etym>.

Descriptions and Prose

Prose descriptions of the etymological components and processes can be represented in several different ways according to the editorial endeavour. They could be left alone untagged (e.g. text can be placed directly in <etym>), specifically annotated with <lbl> or <note> elements when appropriate, or, when one wants to uniformly embed all linguistic descriptions at the same encoding level in the XML tree, with the systematic use of a <seg type="desc"> element. In some cases the descriptions found in print dictionaries may occur in multiple discontinuous parts interrupted by examples or other structured content. In such cases, where desirable by the editors, the attribute @part can be used with the values of ("I" *initial* | "M" *medial* | "F" *final*)[8], where the value "M" may be used for any number of times as needed. The example below

[8] @part does also have the option of the values "Y" *yes* and "N" *no,* however these are redundant to the *initial/medial/final* options and serve no additional value thus they are not recommended.

shows this and some of the other primary descriptive elements in an etymology listed above. To the left is the print source with the TEI encoding on the right.

Etymologie Seit dem 18. Jh. belegt, auf *fickfacken* ‘hin- und herlaufen’ zurückgeführt: evtl. Auch auf fnhd. *Fatzen* ‘spotten, zum Narren halten’ zurückführbar (vgl. Pfeifer 2014:329)	<etym> <lbl>Etymologie</lbl> <date>Seit dem 18. Jh.</date> <seg type="desc" part="I">belegt, auf</seg> ... <seg type="desc" part="M">zurückgeführt: evtl. Auch auf</seg> ... <seg type="desc" part="F">zurückführbar</seg> <xr><pc>(</pc><lbl>vgl.</lbl><ref>Pfeifer 2014:329</ref><pc>)</pc></xr> </etym>

Example 5: Use of <lbl>, <date>, <bibl>, and <seg type="desc"> with discontinuous prose (from Kluge, 1975)

III. Etymons and other forms

The basic component of an etymology is an *etymon*, which is often represented by a form and which may include other information typical of any lexical entry e.g. a language name, grammatical properties, usage descriptions, semantic descriptions, or bibliographic sources. Etymons are encoded in <cit type="etymon"> and are used analogously to the organisation of <entry> both conceptually and structurally.

<entry xml:id="ntuchi" xml:lang="mix"> <form type="lemma"> <orth>ntuchi</orth> <pron notation="ipa">nduʧí</pron> </form> <gramGrp> <pos>noun</pos> </gramGrp> ... </entry>	<cit type="etymon" xml:lang="mix"> <form> <orth corresp="#bean">ntuchi</orth> </form> <gloss xml:lang="en">bean</gloss> <gloss xml:lang="es">frijol</gloss> </cit>

Example 6: Side by side comparison of entry structure and contents and basic etymon (source Bowers, 2020)

Specific types of etymon structures

The data structure of etymons can vary in certain ways according to the specifics of the conceptual content, purpose and/or sources. A few examples are:

- if based in external sources
- if expressing a semantic change (*but not a form change*)

- if expressing provenance while no form attested in the source language is provided

Below we demonstrate such scenarios and their encoding.

Linking forms to external references

If an encoder wants to link a form to an existing external resource, this can be done using the @corresp on the <form> element.

```
<cit type="etymon" xml:lang="und-x-pie">
  <form corresp="http://example.org/uekw.htm">
    <pron>u̯ekʷ-</pron>
  </form>
</cit>
```

Example 7: linking the form associated to an etymon with an external reference

Etymons in semantic changes and polysemy

In certain cases (such as in an etymology describing a semantic change, resulting in *polysemy*), the etymon may only consist of a semantic description with no form. This is possible as in cases of polysemy, the form of the new meaning/lexical item remains the same as the headword of the entry. In this encoding (Example 8), the @corresp on the <cit type="etymon"> points to the @xml:id value of the source sense.

```
<cit type="etymon" corresp="#face-PRIME">
  <sense>
    <usg type="domain">AnatomicalStructure</usg>
    <gloss xml:lang="en">face</gloss>
    <xr type="meronymy">
      <lbl xml:lang="en">as in:</lbl>
      <ref  type="sense"  target="#body-face" xml:lang="en">part of the body</ref>
    </xr>
  </sense>
</cit>
```

Example 8: Etymon with only sense change

Etymons with provenance information

In other cases there may just be a language expressing provenance (in a loanword or inheritance) or possibly simply a date corresponding to the time or first attested usage. Note, as discussed above, typing on <etym> is optional, however it is shown here to compliment the particular etymons. The use of the @xml:lang attribute on the <cit> level defines the etymon as being from the given language, despite not having a form therein. The inclusion of @xml:lang on <lang> therein distinguishes the working language (which is for the reader and is not actually part of the etymon) from the language of the etymon.

Borrowing

```
<etym type="borrowing">
   <lbl>aus</lbl>
   <cit type="etymon" xml:lang="sl">
      <lang xml:lang="de" expand="Slowenisch"
norm="sl">slow.</lang>
    </cit>
</etym>
```

Inheritance

```
<etym type="inheritance">
   <cit type="etymon" xml:lang="gmh">
      <lang xml:lang="de"
expand="Mittelhochdeutsch"
norm="gmh">mhd.</lang>
         <ref type="bibl">Lexer Wb. III 324</ref>
    </cit>
 </etym>
```

Example 9: Etymon with only provenance without forms (from Kluge, 1975)

Etymons with simple dating information

The information about etymons can sometimes be so reduced as to contain no source form information. This is the case in Example 10, in which the author of the etymological description just wanted to record the actual period or date of the first occurrence of the etymon with limited etymological background apart from the temporal information itself (or because the form might be obvious for the reader).

»ins engl. nach Trench 24 im 16. jh. mit dem hauptbegriffe des geschmückten gekommen; der ursprung ist sehr zweifelhaft;« (according to Trench 24, entered English in the 16th c. with the prominent meaning of st. decorated origin highly uncertain); s.v. /brave/ in Mueller (1878)	<cit type="etymon">ins engl. <xr><lbl>nach</lbl> <ref>Trench 24</ref> </xr><date>im 16. jh.</date> <def>mit dem hauptbegriffe des geschmückten gekommen; der ursprung ist sehr zweifelhaft</def><pc>;</pc> <xr>; s.v. /brave/ in Mueller (1878)</xr></cit>

Example 10: Etymon with date but not form information (from Kluge, 1975)

Variants of etymons and/or other forms

As mentioned, in encoding any type of forms in etymologies, <form> and its sub-elements behave the same way as when they occur on the level of the main entry. This structure is necessary for encoders of etymological dictionaries as it is common to find multiple variants and/or inflected forms of the same etymon which cannot be listed separately as they correspond to the same definition and/or other key pieces of information. Thus variants of etymons or other forms (i.e. cognates, see next section) in an etymology should be represented in accordance with the recommendation of the TEI Lex-0 Forms section (Banski, Bowers, and Erjavec 2017). Example 11 shows two such examples from print dictionaries.

Etymologie
mhd. vreten, vretten, vraten 'entzünden; wundreiben; herumziehen; quälen; plagen' (vgl. Lexer 1878 III: 502)

```
<etym>
  <cit type="etymon" xml:lang="gmh">
    <lang>mhd.</lang>
```

```
            <form type="variant">
              <orth>vreten</orth>
            </form>
            <form type="variant">
              <orth>vretten</orth>
            </form>
            <form type="variant">
              <orth>vraten</orth>
            </form>
            <gloss>entzünden</gloss>; <gloss>wundreiben</glos>; <gloss>herumziehen</gloss>;
<gloss>quälen</gloss>; <gloss>plagen</gloss>
          </cit>
          <bibl><lbl>vgl.</lbl> <title>LEXER</title> <date>1878</date> <edition>III</edition>
<citedRange>502</citedRange></bibl>
        </etym>
```

Example 11: Variants of the Middle High German etymon “vreten, vretten, vraten”

Cognates and other types of relevant forms

Cognates are forms asserted as being related in some way to the lexical entry and/or the etymon and are a ubiquitous feature of etymological dictionaries. Cognates are essentially lexical items in a language which share an etymological source. The structure of a basic representation of cognates mirrors that of etymons and uses the same <cit> structure, with the difference being the value of @type should be “cognate”. Note that in the example below, <ref type="bibl"> is used instead of <bibl> as in the project from which the examples are taken, all bibliographical sources are listed in the header with @xml:id’s.

```
        <cit type="cognate" xml:lang="mig">
          <lang>Chalcatongo Mixtec</lang>
          <usg type="geographic">
            <placeName>San Miguel El Grande</placeName>
          </usg>
          <form><pron notation="trans-macaulay-mig">šinì</pron></form>
          <ref type="bibl" target="#Macaulay-ChalcatongoMixtec-1996">(Macaulay, 1996)</ref>
        </cit>
        <cit type="cognate" xml:lang="miy">
          <lang>Ayutla Mixtec</lang>
          <form><pron notation="trans-hill-1990-miy">shīhih</pron></form>
          <ref type="bibl" target="#Hills-AyutlaMixtec-1990">(Hills, 1990)</ref>
        </cit>
        <cit type="cognate" xml:lang="miz">
          <lang>Coatzospan Mixtec</lang>
          <form><pron notation="trans-smll-miz">rkɨ</pron></form>
          <ref  type="bibl" target="#Small-CoatzospanMix-1990">(Small, 1990)</ref>
        </cit>
        <cit type="cognate" xml:lang="smd">
          <lang>San Martín Duraznos</lang>
```

```
  <form><pron notation="ipa">ʃĩɲĩ</pron></form>
  <ref type="bibl" target="#Padgett-2017">(Padget, 2017)</ref>
</cit>
```

Example 12: Collection of cognates taken from various external sources

Cognate Sets from a common bibliographic source

In some cases there may be a list of cognates that are in a print or born-digital source that are all from a single bibliographic source, or are at least presented in a source etymology as a set or list. Often these would have some kind of referential function word or abbreviation, e.g. *“cf, …”* In such cases it may be desirable to present the list of cognates as the source intended and thus group them in a single wrapper <cit type="cognateSet">.

In the case of the former where there is a referential function word or abbreviation that should be tagged with <lbl> and included as a child of <cit type="cognateSet">, preceding the etymons. Where there is a common bibliographic source <bibl>, or <ref type="bibl"> (if the bibliographic sources are already declared elsewhere) should be a child of the <cit type="cognateSet">, and placed after the given forms.

```
<cit type="cognateSet">
    <lbl>Cf.</lbl>

    <cit type="cognate" xml:lang="ffr">
      <form><orth>...</orth></form>
          ….
    </cit>
    <cit type="cognate" xml:lang="und-x-pom">
      <form><orth>...</orth></form>
          ….
    </cit>
    <cit type="cognate" xml:lang="und-x-opd">
      <form><orth>...</orth></form>
          ….
    </cit>
……
     <bibl>Bibl Source Here</bibl>
</cit>
```

Example 13: Template for set of cognates (cognateSet) from a single bibliographic source

Descendant and Derivative Forms

Certain etymological dictionaries may include descendant and/or derivative forms which were derived from the headword.

Whereas, in an entry where the lemma is a form which has been derived from another lexical item, in the etymology section, the typology would be “derivation” and the source term from which it was derived would be the etymon (see example 14). However, in cases such as that in the example below in which derivative forms are presented as related entries to the headword prior to the etymology section proper, these should be encoded as embedded <entry> elements according to TEI Lex-0 guidelines (section 2) and the value of the @type attribute should be “derivative”.

amārus 'bitter' [adj. o/ā] (Pl.+)
Derivatives: *amārilūdō* 'bitterness' (Varro+), *āmāror* [m.] ‘bitter taste' (Lucr.+).
Plt. **o/am-?*
PIE *h_2h_3m-ro-? IE cognates: Skt. *amlá-* ‘sour, acid', OIc. *apr* 'sharp, cold', OE *ampre* ‘sour one', MDu, *amper* ‘bitter, sour' < PGm. **am(p)ra-* ‘sour'; ? OIr. *om* 'raw', W. of possibly <**h_2h_3-emo-,* Skt. āmá- [adj.] ‘raw, uncooked', Gr. ωμός ‘raw’, Arm. *howm* <**$h_2eh_3$mo-**

```
<entry>
 ...
  <lbl>Derivatives</lbl><pc>:</pc>

  <entry type="derivative" xml:lang="la">
     <form><orth>amārilūdō</orth></form>
     <sense><gloss>'bitterness'</gloss></sense>
     <ref type="bibliography">(Varro+)</ref>
  </entry><pc>,</pc>

  <entry type="derivative" xml:lang="la">
     <form><orth>āmāror</orth></form>
     <pc>[</pc><gramGrp><gen>m.</gen></gramGrp><pc>]</pc>
     <sense><gloss>‘bitter taste'</gloss></sense>
     <ref type="bibliography">(Lucr.+)</ref>
  </entry><pc>.</pc>
 </entry>
```

Example 14: Derivatives from (de Vaan, 2008)

Cross-referenced forms

Often within etymological discussions (as in other portions of a dictionary entry) there are references to forms which are etymons, cognates or derivatives, but which in a given specific context, are not actually being posited as such. Where an editor wants to specify the particular etymological relationship to the lemma or other form, the particular information can be specified using the attributes on the cross-reference <xr> and embedded reference <ref> elements.

This section overlaps with the TEI Lex-0 section on structured lexical references (section 6) but its application in the context of etymological content requires us to clearly identify the contexts of usage, and the conceptual distinctions between the the cross-referenced forms and the primary features of etymology markup discussed above.

In an etymology, the contexts in which a cross-reference form should be used include:

- where a reference to another lemma or form from a separate entry is made;
- where a reference to the lemma (or another form) in the synchronic entry is made;
- where a reference to an etymon is made which is not the etymon for the actual lemma;
- where a reference to a sense (corresponding to either of the above) is made;

In the examples below we show the specific use cases of cross referenced forms in the context of etymological dictionary entries.

In this first case, the entry is for the Latin ***arcessō, -ere / accersō, -ere;*** within the etymology section of that entry there are references to the two lemma variants. The cross-references are encoded in <xr type="crossReference" xml:lang="la"><ref type="entry">. This format would apply also to external cross-references, in either case, the editors would also have the option of including a pointer to the given internal or external form(s) with the @target attribute on <ref>.

The stem occurs in two variants, ***accers-*** and ***arcess-,*** which suggests that...

```
...<xr type="related" xml:lang="la"><ref type="entry">accers-</ref></xr> ... <xr type="related" xml:lang="la"><ref type="entry">arcess-</ref></xr> ….
```

Example 15: Cross reference to lemma in etymology section

The following example (from the same entry as above) contains a reference to an etymon of the lemma/entry itself, but this is a supplementary instantiation of the given etymon which occurs in the context of a discussion of a particular phenomena (e.g. a phonetic change[9]). In such examples as the following, the use of <xr> encodes an important conceptual distinction in the data as it allows <cit type="etymon"> to be reserved for the form(s) in the given entry.

….Nussbaum 2007b gives two more arguments for regarding *accerso* as original: the noun

[9] Note the use of @prev and @next can be used to denote temporal sequences of referenced forms.

dorsum → ***dossum*** shows a phonetic change of...

...the noun <xr type="related" subtype="etymon" xml:id="etym-dorsum" xml:lang="la"><ref type="entry">dorsum</ref></xr> <lbl>→</lbl> <xr type="related" prev="#etym-dorsum" xml:lang="la"><ref type="entry">dossum</ref></xr> shows a phonetic change of...

Example 16: Cross-referenced forms for etymons not pertaining to the lemma

Finally we have an example in which there is a cross-reference to a sense of an external entry. This is also encoded as <xr type="crossReference"> but differs in the fact that it is a reference to the sense of a given entry, thus the use of <ref @type="sense"> and the embedded <gloss> within, which needs to have the @xml:lang to distinguish from the value declared at the <xr> level. In this case there is also a form (included herein a separate <ref type="sense">) however it is possible also that a cross-reference to a sense could occur without an accompanying form.

...a verb in ***-cesso*** meaning **'go get'** would be favoured by its semantic neighbours**...**

….a verb in <xr type="related" xml:lang="la"><ref type="sense">-cessō</ref> <lbl>meaning</lbl> <pc>'</pc><gloss xml:lang="en">go get</gloss><pc>'</pc></xr> would be favoured by its semantic neighbours...

Example 17: Cross-reference of a sense from an external entry

Conclusion

In this paper, we have provided a comprehensive view of the core proposals from the TEI Lex-0 Etym initiative with the expectation that its powerful and structured representations will serve to both provide guidance for the encoding of etymological content in TEI for which little precedent is available and to facilitate the interoperable encoding of a vast variety of potential etymological features found both in print (for retro-digitization) and born-digital contents. Still, as we know from the TEI guidelines, this can only be work in progress since future users of the TEI Lex-0 guidelines are likely to come up with new issues and change proposals. The online management of the whole TEI Lex-0 initiative as an open source project makes it possible for anyone to contribute and help improve the specification and documentation, which are already in use in the EU project Elexis as the default pivot format of lexical data integration.
Finally, additional work may indeed be necessary to achieve more precise and stable ontologies for typing etymological processes, as well as for qualifying cross-references in etymological contexts.

References

Banski, Piotr, Jack Bowers, and Tomaz Erjavec. 2017. “TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms.” In *Electronic Lexicography in the 21st Century: Proceedings of ELex 2017 Conference*, 485–494. Lexical Computing. https://hal.inria.fr/hal-01757108.

Bowers, Jack, and Laurent Romary. 2017. “Deep Encoding of Etymological Information in TEI.” *Journal of the Text Encoding Initiative*, no. 10. https://doi.org/10.4000/jtei.1643.

Crist, Sean. 2005. “Toward a Formal Markup Standard for Etymological Data.” In *LSA Annual Meeting*.

Khan, Fahad, and Jack Bowers. 2020. “Towards a Lexical Standard for the Representation of Etymological Data.” *Convegno Annuale Dell’Associazione per l’Informatica Umanistica e La Cultura Digitale*.

Romary, Laurent, and Toma Tasovac. 2018. “TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources.” *JADH 2018*, 274.

Salmon-Alt, Susanne. 2006. “Data Structures for Etymology: Towards an Etymological Lexical Network.” *Bulletin de Linguistique Appliquée et Générale* 31: 1–12.

Cited sources

Bowers, J. (2020) *Mixtepec-Mixtec Digital Dictionary* (https://github.com/iljackb/Mixtepec_Mixtec/blob/master/MIX-Lexicon-TEI-Dict.xml)

Kluge, F. (1975). *Etymologisches Wörterbuch der deutschen Sprache*. Berlin. New York, 21.

de Vaan, M. (2008). *Etymological Dictionary of Latin and the other Italic Languages* (edited by Alexander Lubotsky as Volume 7 of the Leiden Indo-European Etymological Dictionary Series). Leiden* Boston.

Annex: Fully Encoded Examples

In what follows, we have tried to illustrate interesting cases of etymological processes that may show how TEI Lex-0 Etym can seamlessly take into account a variety of situations. All examples have been validated and included in the TEI Lex-0 GitHub environment.

Embedded Senses, Metaphor and Compounding

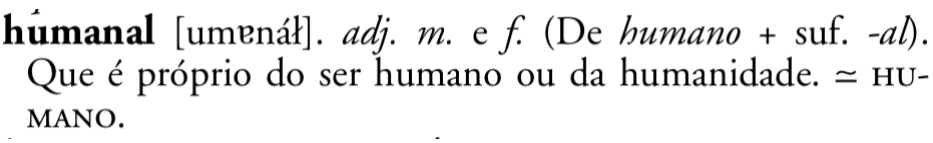
The example below shows a case of an embedded sense from the Mixtepec-Mixtec TEI dictionary (source: Bowers, 2020) in which the lemma form *xini ve’e* is a compound in which one of the components is metaphorical in nature. The portion of the etymology that is metaphorical (<etym type="metaphor">) is embedded within that of the type *compounding*, and as it is relevant to the process of metaphor, within that section, there is the domain (<usg type="domain">).

```
<sense>
  <usg type="domain">Architecture</usg>
  <cit type="translationEquivalent" xml:lang="en">
    <form>
      <orth>ceiling</orth>
    </form>
  </cit>
```

```
  <cit type="translationEquivalent" xml:lang="es">
    <form>
      <orth>techo</orth>
    </form>
  </cit>
  <etym type="compounding">
    <etym type="metaphor">
      <cit type="etymon" xml:lang="mix">
        <form  type="lemma" corresp="#body-head">
          <orth>xiní</orth>
        </form>
        <gloss xml:lang="en">head</gloss>
        <gloss xml:lang="es">cabeza</gloss>
        <usg type="domain">Anatomy</usg>
      </cit>
    </etym>
    <cit type="etymon" xml:lang="mix">
      <form type="lemma" corresp="#house">
        <orth>ve'e</orth>
      </form>
      <gloss xml:lang="en">house</gloss>
      <gloss xml:lang="es">casa</gloss>
    </cit>
  </etym>
</sense>
```

Derivational/Morphological Analysis

In the following example (source: de Vaan, 2008) for the Portuguese entry *húmanal,* we have a case of derivation (labeled in @type) in which the suffix *-al* is attached to the noun *humano* to create the attributive adjective. Given that derivation can occur in a wide variety of morphological ways, the specific type is labeled in @subtype).



húmanal [umɐnáł]. *adj. m.* e *f.* (De *humano* + suf. *-al*). Que é próprio do ser humano ou da humanidade. ≃ HUMANO.

```
    <entry xml:lang="pt">
      <form type="lemma">
        <orth>húmanal</orth>
        <pron notation="ipa">umɐnáł</pron>
      </form>
      <gramGrp>
        <pos>adj.</pos>
        <gen>m.</gen>
        <lbl>e</lbl>
        <gen>f.</gen>
      </gramGrp>
```

```
<etym subtype="derivation" subtype="suffixalDerivation">
  <pc>(</pc>
  <seg type="desc">De</seg>
  <cit type="etymon" xml:lang="pt">
    <form>
      <orth>humano</orth>
    </form>
  </cit>
   <lbl>+</lbl>
   <cit type="etymon" xml:lang="pt">
     <gramGrp>
       <pos expand="suffix">suf.</pos>
     </gramGrp>
     <form>
       <orth extent="suff">-al</orth>
     </form>
  </cit>
  <pc>)</pc>
</etym>
<!--  rest of entry here -->
</entry>
```

Phonological changes

It is very common in etymological dictionaries to have discussions about sequences of sound changes. These most often take place in the context of running prose. As described above, prose can be represented using <seg type="desc">, and where interrupted with etymons (or other content), the @part attribute can be used. The phonetic or phonological units described as having undergone particular changes are represented the same way as full word forms using <cit type="etymon">. In order to attribute the particular place in the sequence of sound changes, the @prev and @next attributes can be used to point to the @xml:id of the previous or next form in the diachrony. The following example sourced from de Vaan, 2008 illustrates these mechanisms.

Others have proposed an etymology **ad-arti-* with intervocalic *****d*** becoming ***l***; the spelling *allers* would then be analogical to *sollers*

```
<etym>
  <seg type="desc" part="I">Others have proposed an etymology</seg>
  <cit type="etymon" xml:id="ad-arti-" xml:lang="und-x-pie"><form><orth>*ad-arti-</orth></form></cit>
   <etym corresp="#ad-arti-">
     <seg type="desc" part="I">with intervocalic</seg>
     <cit type="etymon" xml:id="c1" next="#c2"><form><orth>d</orth></form></cit>
     <seg type="desc" part="F">becoming</seg> <cit type="etymon" xml:id="c2"
prev="#c1"><form><orth>l</orth></form></cit>
      <pc>;</pc>
   </etym>
```

```
   <seg type="desc" part="F">the spelling <xr type="crossReference" xml:lang="la"><ref
type="entry">allers</ref></xr> would then be <xr type="crossReference" xml:lang="la"><lbl>analogical to</lbl> <ref
type="entry">sollers</ref></xr><pc>.</pc></seg> </etym>
</etym>
```

Note also that this entry contains an embedded etymology, and it is distinguished in the data structure according to the portion that belongs directly to the author, and the portion which he is ascribing to “others”.

Multiple and/or Conflicting Etymological Accounts

In many sources there can be multiple, sometimes conflicting accounts for an etymology. In these cases nested etymologies should be used, the top layer being reserved for the editorial descriptions, and any number of separate <etym>’s can be included therein.

According to Untermann 2000, Latin **all-* was probably borrowed from Sabellic, since Latin does not have this word in its lexicon. For a word only occurring in glosses, this is of course possible.

Others have proposed an etymology **ad-arti-* with intervocalic *d* becoming *l*; the spelling *allers* would then be analogical to *sollers.*

```
        <etym>
          <!— PIt, PIE etymons-->
          <!-- cognates -->
          <etym cert="medium">
            <!-- Lat. sollers < *soti-arti- to sollus 'entire'; al(l)ers < *all-arti- to O. alio- 'entire'. -->
            <seg type="desc" part="I">According to</seg> <ref type="bibliography">Untermann 2000</ref>,
            </seg>
           <xr type="crossReference"><lang>Latin</lang> <ref xml:lang="la">*all-</ref></xr>
           <seg type="desc" part="M">was probably borrowed from</seg> <cit type="etymon" xml:lang="und-x-
sabe1249"><lang  norm="und-x-sabe1249">Sabellic</lang></cit>, <seg type="desc" part="F">since <lang
norm="la">Latin</lang> does not have this word in its lexicon. For a word only occurring in glosses, this is of course
possible.</seg>
          </etym>

          <etym cert="medium"><seg type="desc" part="I">Others have proposed an etymology</seg>
            <cit type="etymon" xml:id="ad-arti-" xml:lang="und-x-pie"><form><orth>*ad-arti-</orth></form></cit>

            <etym corresp="#ad-arti-">
              <seg type="desc" part="I">with intervocalic </seg><cit type="etymon" xml:id="c1"
next="#c2"><form><orth>d</orth></form></cit>

              <seg type="desc" part="F">becoming</seg> <cit type="etymon" xml:id="c2"
prev="#c1"><form><orth>l</orth></form></cit>
              <pc>;</pc>
            </etym>

            <seg type="desc" part="F">the spelling <xr type="related"><ref xml:lang="la">allers</ref></xr> would then
be <xr type="crossReference"><lbl>analogical to</lbl> <ref
xml:lang="la">sollers</ref></xr><pc>.</pc></seg></etym>
```

</etym>