TEI-Lex0 Etym -towards terse(r) recommendations for the encoding of etymological information

Jack Bowers, Axel Herold, Laurent Romary

TEI-Lex0 Etym – towards terse(r) recommendations for the encoding of etymological information

Jack Bowers [1,3,4]

Axel Herold [2,3,4]

Laurent Romary [2,3]

1 - Austrian Academy of Sciences (ÖAW)- Austrian Center for Digital Humanities (ACDH)
2 - Berlin Brandenburgische Akademie der Wissenschaften (BBAW)
3 - Inria - Team ALMAnaCH, Paris
4 - École Pratique des Hauts Études, Paris

TEI Conference and members meeting Tokyo, Japan - Sept 2018

Intro to TEI-Lex0

- Joint initiative of the COST action ENeL, the research infrastructure DARIAH, and EU project PARTHENOS
- It is aimed at formulating streamlined guidelines for the **TEI Dictionaries module / chapter**
- The goal: simplify recommendations for most common components of TEI dictionary encoding so to: (i) reduce the learning curve for new adopters; and (ii) to increase interoperability
- Meant to serve as baseline encoding against which existing TEI dictionaries can be compared and could serve as a pivot format for generic querying or visualization tools
- ISO 24613 - LMF Revision: TEI Serialization (part 4)

https://github.com/DARIAH-ERIC/lexicalresources

Intro to TEI-Lex0 Etym

- Builds off of recent efforts to address etymology in TEI (Bowers & Romary, 2016; Sagot, 2017)
- Defines a more restrained, set of options for encoding any given single phenomena
- Addresses issues left out of previous efforts
- Designed to be able to handle born-digital and retro-digitized print sources
- ISO 24613 - within LMF Revision: TEI Lex0-Etym will be the target format counterpart to the (LMF part 3: Diachrony-Etymology Extension)

Bowers, J., & Romary, L. (2016). Deep Encoding of Etymological Information in TEI. Journal of the Text Encoding Initiative, (Issue 10). https://doi.org/10.4000/jtei.1643
Sagot, B. (2017). Extracting an Etymological Database from Wiktionary. In Electronic Lexicography in the 21st century (eLex 2017) (pp. 716–728).

Intro to TEI-Lex0 Etym

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, etc.
- 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

I. Components of an Etymology

Structure of an Etymological Entry

Bowers & Romary (2016) describe three options for the placement of an <etym> in an <entry>:

a) as a child of <entry>

b) as a child of <sense> for sense based changes

c) (in conjunction with one of the above) embedded (0..n) times w/in another <etym> to represent multiple ordered processes in sequence

In retrodigitization of printed sources, it is feasible however that a source may place etymological information within a sense field, in which case option (b) should be used.

The attribute @type can be used on <etym> to specify etymological process, if that process itself has subtypes, @subtype can also be used

Etymology structures

Basic Minimal Flat Structure

```
<entry>
  ....
  <etym>
   <!-- text to be further marked up -->
     inherited from Middle English X
     from Old English Y
     borrowed from Latin Z
     which was from the Proto Italic Q
     from Proto Indo-European ʊ
  </etym>
</entry>
```

Etymology structures

Embedded <etym> For Sense Change

```
<entry>
  ....
  <sense>
    <etym>
     <!-- text to be further marked up -->
      Metaphorical extension of X
    </etym>
  </sense>
</entry>
```

Etymology structures

XML structure should align with the chronology of the etymology as:

Most recent = highest level

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

```
<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 ʊ
        </etym>
      </etym>
    </etym>
  </etym>
</etym>
```

Etymology structures

Embedded stages: source ordered (most > least recent)

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

<etym>

Within <etym> (*XPath:* “//etym/*”) the following elements can occur any number of times:

- <seg type=”desc”> *(for prose)*
- <lbl>
- <date>
- <cit @type>
- <bibl>
- <ref type=”bibl”> *(if bibliography is previously declared or externally stored)*
- <note> *(for editorial notes not part of the actual description in <seg type=”desc”>)*

No more <mentioned>!

<date>, <bibl> can also occur within <cit>

<etym>

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>

   <!--other stuff here -->
  <seg type="desc" part="M">zurückgeführt evtl. auch auf</seg>

   <!--other stuff here -->
  <seg type="desc" part="F">zurückführbar</seg>
   <bibl>
     <title>Pfeifer</title> <date>2014</date>
      <citedRange>329</citedRange>
   </bibl>
</etym>
```

<ref typ="bibl"> w/ @target should be used if external bibliography is provided in header or linked in project

Basic components of etymology entry: *etymons, etc.*

The rest of the most important components of an etymology, are encoded with a typed <cit> element.

(Etymological) **<cit>** can contain:

- <lang> *(0..1) for where the language of the cited form is given*
- <date>
- <form> *(w/in which <orth> and/or <pron>)*
- <sense> and/or <def> or <gloss> *(0...n) which function as a parallel to <sense>*
- <usg> (*e.g.* geo, domain,...)(*See TEI-Lex0 <usg> chapter*)
- <xr> (*for semantic relations; e.g. 'meronymOf'...)*
- <gramGrp> *(0..1) (and all grammatical sub-elements)*
- <ref>
- <bibl>

Etymons

The basic component of an etymology is an *etymon*, which is typically a form and often will include other information typical of any lexical entry e.g. language, grammatical, usage, sense/def, bibliographic source.

Etymons are encoded in <cit type=”etymon”> and is often used in parallel to <entry>

```
<entry xml:id="ntuchi">
  <form type="lemma">
    <orth xml:lang="mix">ntuchi</orth>
    <pron xml:lang="mix" notation="ipa">ndùʧí</pron>
  </form>
  <gramGrp>
    <pos>noun</pos>
  </gramGrp>
  <!-- sense: (translations, domain, etc.)-->
    ….
</entry>
```

```
<cit type="etymon">
    <!-- <lang> if needed-->
  <form>
    <orth xml:lang="mix">ntuchi</orth>
    <pron xml:lang="mix" notation="ipa">ndùʧí</pron>
  </form>
  <gramGrp>
    <pos>noun</pos>
  </gramGrp>
  <!-- sense info: (gloss/def, domain, etc.)-->
   ....
</cit>
```

Etymons

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>
 <!--other stuff here -->
  <cit type="etymon">
    <form>
      <orth xml:lang="de">fickfacken</orth>
    </form>
    <def xml:lang="de">hin- und herlaufen</def>
  </cit>
 <!--other stuff here -->
 <cit type="etymon">
    <lang>fnhd.</lang>
    <form>
      <orth xml:lang="nds-x-FNHD">fatzen</orth>
    </form>
    <def xml:lang="de">spotten, Zum Narren halten</def>
  </cit>
 <!--other stuff here -->
</etym>
```

Specific types of etymon structures

Etymons structure can vary in certain ways according to the specifics of the data/purposes, a few examples are:

- If based in external source
- If expressing a sense change (but not a form change)
- If expressing provenance but no form in source language is given

Specific types of etymon structures:

with pointer

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

Specific types of etymon structures:

sense change only (no form)

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

Specific types of etymon structures:

provenance only (no form)

```
<etym type="borrowing">
   <seg type="desc">aus</seg>
   <cit type="etymon" xml:lang="sl">
      <lang>slow.</lang>
    </cit>
</etym>

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

@xml:lang on <cit>

Other types of etymological forms: Cognates

Main entry
(Mixtepec-Mixtec)
Xiní
ʃiní

(Chalcatongo Mixtec - San Miguel El Grande)
šinì
(Macaulay, 1996);

(Ayutla Mixtec)
shīhih
(Hills, 1990);

(San Martín Duraznos Mixtec)
ʃīɲī
(Padget, 2017);

```
<cit type="cognate">
  <lang>Chalcatongo Mixtec</lang>
  <usg type="geo">
    <placeName>San Miguel El Grande</placeName>
  </usg>
  <form><pron notation="trans-macaulay-mig" xml:lang="mig">šinì</pron></form>
  <ref type="bibl" target="#Macaulay-ChalcatongoMixtec-1996">
     (Macaulay, 1996)</ref>
</cit>
<cit type="cognate">
  <lang>Ayutla Mixtec</lang>
  <form><pron notation="trans-hill-1990-miy" xml:lang="miy">shīhih</pron></form>
  <ref type="bibl" target="#Hills-AyutlaMixtec-1990">(Hills, 1990)</ref>
</cit>
<cit type="cognate">
  <lang>San Martín Duraznos Mixtec</lang>
  <form><pron notation="ipa" xml:lang="smd">ʃīɲī</pron></form>
  <ref type="bibl" target="#Padgett-2017">(Padget, 2017)</ref>
</cit>
```

Note: can also occur as <cit type=”cognateSet”> if inheriting a common bibliographic source

Other types of etymological forms: variants

In etymological sources it is very common to have variant forms of etymons, cognates, etc. These can be encoded in a parallel manner to how synchronic form variants according to TEI-Lex0 Forms (Banski et al., 2017)

Bański, P., Bowers, J., & Erjavec, T. (2017). TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms (pp. 485 – 494). Brno, Czech Republic: Lexical Computing CZ s.r.o. Retrieved from http://nbn-resolving.de/urn:nbn:de:bsz:mh39-64273

Other types of etymological forms: variants

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

```
<cit type="etymon">
  <lang>mhd.</lang>
  <form type="variant">
    <orth xml:lang="gmh">vreten</orth>
  </form>
  <form type="variant">
    <orth xml:lang="gmh">vretten</orth>
  </form>
  <form type="variant">
    <orth xml:lang="gmh">vraten</orth>
  </form>
  <def>entzünden; wundreiben; herumziehen; quälen; plagen</def>
</cit>
```

Other types of etymological forms: Derivatives

amārus 'bitter' [adj. o/ā] (Pl.+)

Derivatives:
***amārilūdō* 'bitterness' (Varro+),**
***āmāror* [m.] 'bitter taste' (Lucr.+).**

PIt. **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';.....

Other types of etymological forms: Derivatives

In TEI-Lex0 the <re> is replaced by embedded entries…

These occur as a direct child of <entry>, not <etym>

Derivatives:
***amārilūdō* 'bitterness' (Varro+),**
***āmāror* [m.] 'bitter taste' (Lucr.+).**

```
<entry>
 ….
 <lbl>Derivatives</lbl><pc>:</pc>
  <entry type="derivative">
    <form><orth xml:lang="la">amārilūdō</orth></form>
    <sense><def>'bitterness'</def></sense>
    <ref type="bibl"><bibl>(Varro+)</bibl></ref>
  </entry><pc>,</pc>

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

Minimal TEI-Lex0 Etym Encoding

Eingang m.
mhd. īnganc,
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 eingegengenen Geschäftssachen, Mannschaften’ usw. Vgl. Zugang. (Kluge, 1975)p159

```
<etym>
  <lang>mhd.</lang>
  <cit type="etymon">
    <form><orth xml:lang="gmh">īnganc</orth></form>
  </cit>
  <cit type="cognate">
    <lang>dän.</lang>
    <form><orth xml:lang="da">indgang</orth></form>
  </cit>
  <cit type="cognate">
    <lang>schwed.</lang>
    <form><orth xml:lang="sv">ingång</orth></form>
  </cit>
  <seg type="desc">Lehnübersetzung des</seg>
  <cit type="etymon">
    <lang expand="Latin">lat.</lang>
    <form><orth xml:lang="la">introitus</orth></form>
  </cit>
  <seg type="desc">Aus dem ‘Hineingehen’ als Handlung ist die ‘Stelle, an der man ins
    Haus, in den Saal geht’ geworden, neuerdings auch die ‘Gesamtheit der
    eingegengenen Geschäftssachen, Mannschaften’ usw. Vgl. <ref>Zugang</ref>.
      <ref type="bibl">(Kluge, 1975) p159</ref> </seg>
</etym>
```

II. Complex markup and typed etymologies

Extended Encoding:
Typing Etymological Processes

Eingang m.
mhd. īnganc,
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 eingegengenen Geschäftssachen, Mannschaften’ usw.
Vgl. Zugang. (Kluge, 1975)p159

```
<etym type="inheritance">
  <lang>mhd.</lang>
  <cit type="etymon">
    <form>
      <orth xml:lang="gmh">īnganc</orth>
    </form>
  </cit>
  <!--cognates here -->
 <etym type="borrowing" subtype="calque">
  <seg type="desc">Lehnübersetzung des</seg>
  <cit type="etymon">
   <lang expand="Latin">lat.</lang>
    <form>
      <orth xml:lang="la">introitus</orth>
    </form>
  </cit>
  <seg type="desc"><!--desc text here --></seg>
 </etym>
</etym>
```

Types of etymologies: Derivation

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

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

Sagot (2017) added several types of etymological descriptions to Bowers & Romary (2016) including derivation

```
<etym type="suffixalDerivation">
  <pc>(</pc>
  <seg type="desc">De</seg>
  <cit type="etymon">
    <form>
      <orth xml:lang="pt">humano</orth>
    </form>
  </cit>
  <pc>+</pc>
  <cit type="etymon">
    <gramGrp>
      <gram>suf.</gram>
    </gramGrp>
    <form>
      <orth extent="suff" xml:lang="pt">-al</orth>
    </form>
  </cit>
  <pc>)</pc>
</etym>
```

Types of etymologies: Conflicting, Divergent 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.

Conflicting, Divergent Etymological Accounts

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>
 <!-- main etymons -->
 <!-- cognates -->
  <etym type="borrowing">
    <seg type="desc" part="I">According to</seg> <bibl>Untermann 2000</bibl>,
    <lang>Latin</lang> <ref xml:lang="la">*all-</ref> <seg type="desc" part="M">was probably
     borrowed from</seg> <cit type="etymon" xml:lang="nds-x-sabe1249"><lang>Sabellic</lang></cit>,
    <seg type="desc" part="F">since <lang>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 type="lateralization">
   <seg type="desc" part="I">Others have proposed an etymology</seg>
    <cit type="etymon"><form><orth xml:lang="ine">*ad-arti-</orth></form></cit> with intervocalic
    <c xml:id="c1" next="#c2">d</c> becoming <c xml:id="c2" prev="#c1">l</c><pc>;</pc>
    <seg type="desc" part="F">the spelling <ref xml:lang="la">allers</ref> would then be analogical
     to <ref xml:lang="la">sollers</ref><pc>.</pc></seg>
 </etym>
</etym>
```

Conclusion

- Specific status of etymology in the TEI Lex-0 activity
  - A lot of new constructs => way beyond the current TEI guidelines
- Adapting/rewriting the section on <etym> in the guidelines
  - Main possibilities offered by the model (etymon- recursive processes)
  - Pointer to the TEI Lex-0 document
- Maintaining the TEI Lex-0 specification
  - Value lists
  - Complete set of examples