Modeling dictionary structures

TEI and TEI Lex-0 for interoperability across dictionaries

Axel Herold

Berlin-Brandenburgische Akademie der Wissenschaften

February 17th 2023

Outline

The Text Encoding Initiative

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Brief overview

- operating since the (late) 1980s
- common standard for text representation in the DH for all sorts of printed or written documents, currently based on open X* standards
  - prose, verse, drama
  - dictionaries
  - list, accounting data
  - scholary editions
  - born-digital communication
  - ...
- *lots* of active projects rely on the TEI
- open collaboration and development through special interest groups
- `http://www.tei-c.org`, `https://github.com/TEIC`

Traditional entry model(s) in TEI

Lexical modeling

- modeling $\approx$ mapping of objects and their properties (and relations) onto symbolic representations (generally also abstraction)
- modeling lexical data (esp. in digitization of printed resources) is multi-layered modeling:
  - printed characters $\longrightarrow$ codepoints (e. g. Unicode)
  - spacial relation of characters $\longrightarrow$ words (tokens)
  - typographical properties $\longrightarrow$ (hints as to) functions of words (tokens)
  - ...
- every level relies on interpretation and may introduce uncertainty
- alternative and even incompatible interpretations (and therefore conflicting models) are possible

Traditional entry model(s) in TEI

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Lexical modeling in TEI

different "views" on lexical data:

typographical "the two-dimensional printed page, including information about line and page breaks and other features of layout"

editorial "the one-dimensional sequence of tokens which can be seen as the input to the typesetting process . . . "

lexicographic ". . . the underlying information represented in a dictionary, without concern for its exact textual form"

(TEI Guidelines, chapter 9)

Traditional entry model(s) in TEI

Lexical modeling in TEI

different “views” on lexical data:

- in print production: lexical $\longrightarrow$ typographical
- in (retro-)digitization: typographical $\longrightarrow$ lexical
- often desirable to retain several views
- good practice:
  - keep literal values as character data in elements
  - add annotations, normalizations as attribute values

Traditional entry model(s) in TEI

Lexical modeling, example entry

Flusspat, von nhd. *Flußspat,* so genannt, weil das mineral als zusatz beim schmelzen verwandt wurde, um die masse *in Fluß* zu bringen. Hierfür holl. *vloeispaath,* engl. *fluor* und *fluor-spar* (vgl. *feltspat*). Zugrunde liegt mlat. *fluor,* eigentlich „das fließen". Siehe *spat* I.

Falk/Torp (1910)

Traditional entry model(s) in TEI

Typographical view

```
<lb/><p><hi rendition="#b">Flusspat,</hi> von
    nhd. <hi rendition="#i">Flußspat</hi>,
    so genannt, weil das mineral als
    <lb/>zusatz beim schmelzen verwandt wurde,
    um die masse <hi rendition="#i">in</hi>
    <hi rendition="#i">Fluß</hi> zu
    <lb/>bringen. Hierfür holl.
    <hi rendition="#i">vloeispaath</hi>,
    engl. <hi rendition="#i">fluor</hi> und
    <hi rendition="#i">fluor-spar</hi> (vgl.
    <lb/><hi rendition="#g #i">feltspat</hi>).
    Zugrunde liegt mlat.
    <hi rendition="#i">fluor</hi>,
    eigentlich „das fließen“.
    <lb/>Siehe <hi rendition="#g #i">spat</hi>
    I.</p>
```

Traditional entry model(s) in TEI

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Editorial view

```
<p><hi rendition="#b">Flusspat,</hi> von
    nhd. <hi rendition="#i">Flußspat</hi>,
    so genannt, weil das mineral als
    zusatz beim schmelzen verwandt wurde,
    um die masse <hi rendition="#i">in</hi>
    <hi rendition="#i">Fluß</hi> zu bringen.
    Hierfür holl.
    <hi rendition="#i">vloeispaath</hi>,
    engl. <hi rendition="#i">fluor</hi> und
    <hi rendition="#i">fluor-spar</hi> (vgl.
    <hi rendition="#g #i">feltspat</hi>).
    Zugrunde liegt mlat.
    <hi rendition="#i">fluor</hi>,
    eigentlich „das fließen“. Siehe
    <hi rendition="#g #i">spat</hi> I.</p>
```

Traditional entry model(s) in TEI

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Lexical view

```
<entry>
    <form><orth>Flusspat,</orth></form>
    <etym>von <lang>nhd.</lang>
    <mentioned>Flußspat</mentioned>, so
    genannt, weil das mineral als zusatz
    beim schmelzen verwandt wurde, um die
    masse <mentioned>in Fluß</mentioned>
    zu bringen. Hierfür <lang>holl.</lang>
    <mentioned>vloeispaath</mentioned>,
    <lang>engl.</lang>
    <mentioned>fluor</mentioned> und
    <mentioned>fluor-spar</mentioned>
    (vgl. <ref>feltspat</ref>).
    Zugrunde liegt <lang>mlat.</lang>
        <!-- ... --> </etym>
</entry>
```

Traditional entry model(s) in TEI

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Lexical view, alternative (with metadata annotations)

```
<entry type="main">
    <form type="headword">
        <orth>Flusspat,</orth>
        <gramGrp><pos value="NN"/>
        </gramGrp></form>
    <etym>von <lang>nhd.</lang>
    <mentioned
      xml:lang="de">Flußspat</mentioned>,
    so genannt, weil das mineral als
    zusatz beim schmelzen verwandt wurde,
    um die masse <mentioned
        xml:lang="de">in Fluß</mentioned>
    zu bringen. Hierfür <lang>holl.</lang>
        <!-- ... --> </etym>
</entry>
```

Traditional entry model(s) in TEI

Problems with vanilla TEI

(specifically for dictionary modeling)

- sometimes several similar models
- often several ways to encode the same abstract model
- some abstract models do not have direct equivalents in TEI
- difficult to encode alternative models
- often open or semi-fixed vocabulary for annotations

Some of these problems will be discussed in this talk.

TEI Lex-0

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

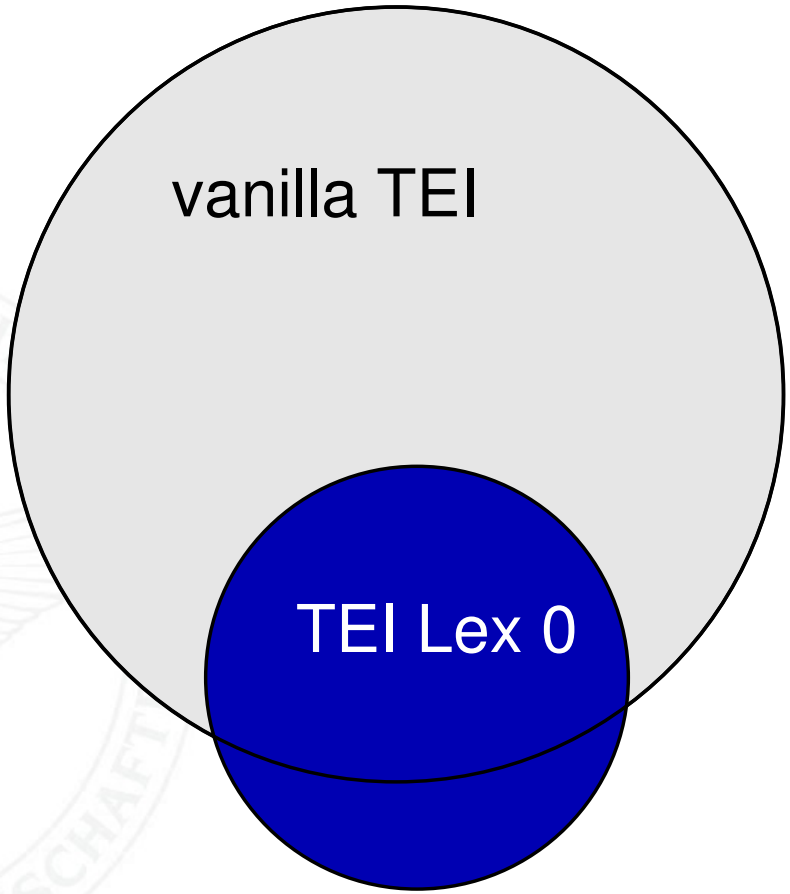
Background

- international working group with strong ties to the TEI
- work started in 2016, supported by ENeL, Dariah, independent research institutes (currently cooperating with Elexis)
- general use-case: mapping typographic structures onto lexical structures
- not a chapter 9 replacement, rather a proposed baseline by way of a TEI customization
- aims at interoperability by
  - restricting some alternatives
  - streamlining some content models
  - closing and fixing vocabulary
  - ...
- quite some proposals went upstream already

TEI Lex-0

Relation between vanilla TEI and TEI Lex-0

“not a chapter 9 replacement, rather a proposed baseline” for lexicographic information



TEI Lex-0

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Entries

(“sometimes several similar models”)

- in TEI, several models for entries (and entry-like entities):

entry	“contains a single structured entry . . . ”
entryFree	“contains a single unstructured entry . . . ”
superEntry	“groups a sequence of entries . . . ”
hom	“groups information relating to one homograph within an entry.”
re	“contains a dictionary entry for a lexical item related to the headword . . . ”

- all models with slightly different content models
- in TEI Lex-0: only **`entry`**, but recursively nestable

TEI Lex-0

Entries

	Leder	nn	leather
	leder	n	of leather; leathern, leathery, tough
ab	leder	n	wipe with chamois skin
Ober	leder	nn	upper leather of shoe
Unter	leder	nn	sole leather

Keller (1978)

TEI Lex-0

Entries

	Leder	nn	leather
	leder n		of leather; leathern, leathery, tough
ab	leder n		wipe with chamois skin
Ober	leder	nn	upper leather of shoe
Unter	leder	nn	sole leather

Keller (1978)

```
<entry type="word-family">
  <entry type="word">
    <form type="lemma" xml:lang="de">
      <orth>Leder</orth></form>
    <gramGrp><gram type="pos">nn.</gram></gramGrp>
    <sense><def xml:lang="en">leather</def></sense>
  </entry>
  <entry type="word"> <!-- ledern [vb.] --> </entry>
  <entry type="word"> <!-- abledern [vb.] --> </entry>
  <!-- ... -->
</entry>
```

TEI Lex-0

Forms and grammatical information

("often several ways to encode the same abstract model")

- `gramGrp` may appear on entry, form, sense, ... in vanilla TEI
- TEI Lex-0 restricts this and relies on enheritance as expressed in the XML structure:
  - entry-level grammatical information on `entry`
  - sense-level grammatical information on `sense`
  - form specific grammatical information on `form`
  - (with narrow exceptions)
- mandatory `form/@type`, e. g. lemma, inflected, paradigm, variant in Lex-0

TEI Lex-0

Forms and grammatical information

grunt vb. ME. *grunte gronte* OE. *grunnettan*; ident. w. G. *grunzen*, DAN. *grynte*, SW. *grynta* A more primit. stem appears in OE. *grunian* 'grunt'. The √*grun* is imitation of sound; cp. LAT *grunnire*.

Kluge/Lutz (1898)

```
<entry>
  <form type="lemma"><orth>grunt</orth></form>
  <gramGrp><gram type="pos">vb.</gram></gramGrp>
  <etym>
    <!-- ... -->
  </etym>
</entry>
```

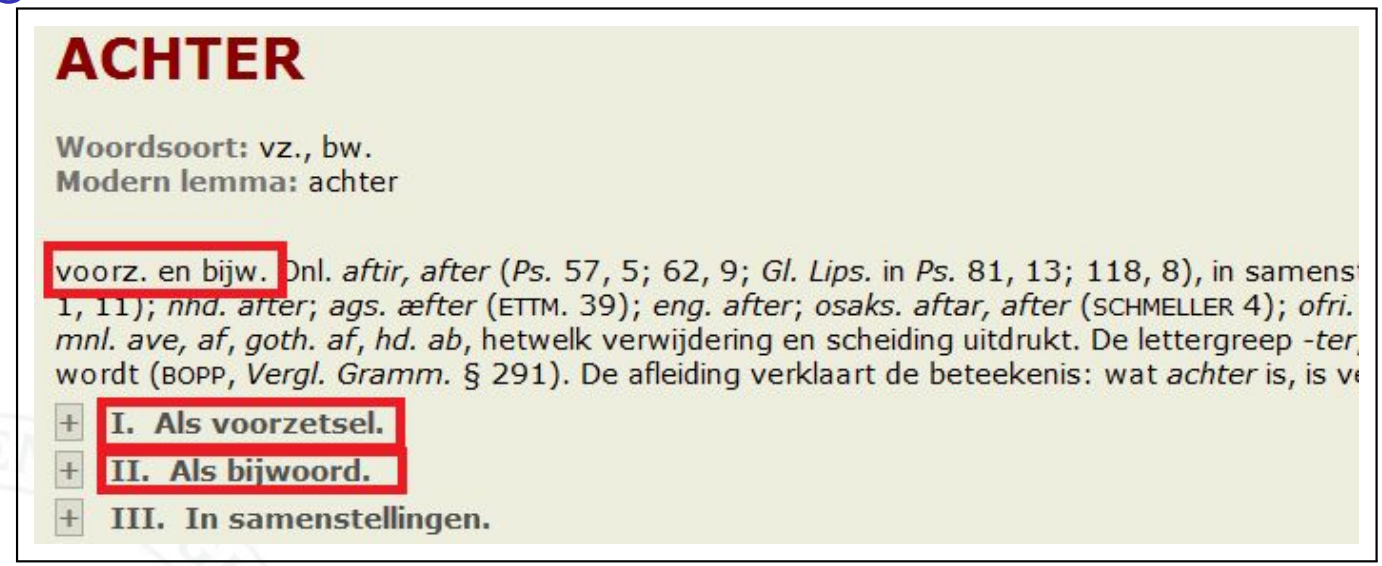
Forms and grammatical information

aid /eɪd/ *noun* **1.** help, especially money, food or other gifts given to people living in difficult conditions ○ *aid to the earthquake zone* ○ *an aid worker* (NOTE: This meaning of **aid** has no plural.) □ **in aid of** in order to help ○ *We give money in aid of the Red Cross.* ○ *They are collecting money in aid of refugees.* **2.** something which helps you to do something ○ *kitchen aids* ■ *verb* **1.** to help something to happen **2.** to help someone

```
<entry>
  <form type="lemma">
    <orth>aid</orth>
    <pron>e&#305;d</pron>
  </form>
  <entry>
    <gramGrp><gram type="pos">noun</gram></gramGrp>
    <!-- ... -->
  </entry>
  <entry>
    <gramGrp><gram type="pos">verb</gram></gramGrp>
  </entry>
</entry>
```

TEI Lex-0

Forms and grammatical information



ACHTER

Woordsoort: vz., bw.
Modern lemma: achter

voorz. en bijw. Onl. *aftir, after* (*Ps.* 57, 5; 62, 9; *Gl. Lips.* in *Ps.* 81, 13; 118, 8), in samens
1, 11); *nhd. after*; *ags. æfter* (ETTM. 39); *eng. after*; *osaks. aftar, after* (SCHMELLER 4); *ofri.*
mnl. ave, af, goth. af, hd. ab, hetwelk verwijdering en scheiding uitdrukt. De lettergreep *-ter*
wordt (BOPP, *Vergl. Gramm.* § 291). De afleiding verklaart de beteekenis: wat *achter* is, is ve

+ **I. Als voorzetsel.**
+ **II. Als bijwoord.**
+ **III. In samenstellingen.**

```
<entry>
  <form type="lemma"><orth>ACHTER</orth></form>
  <gramGrp><gram type="pos">voorz. en bijw</gram>
      </gramGrp><etym> ... </etym>
  <sense n="I">
    <gramGrp><gram type="pos">Als voorzetsel.</gram>
        </gramGrp> ... </sense>
  <sense n="II">
    <gramGrp><gram type="pos">Als bijwoord.</gram>
        </gramGrp> ... </sense> ... </entry>
```

Forms and grammatical information, inflected forms

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

We really want inheritance to work!

TEI Lex-0

Usage labels

("often open or semi-fixed vocabulary for annotations")

- `usg` covers multiple dimensions:
  - geo(graphic)
  - time
  - dom(ain)
  - register, style
  - plev (preference level)
  - lang(uage)
  - gram(matical)
  - syn(onym), hyp(ernym)
  - colloc(ation), comp(lement), obj(ect), subj(ect), verb
  - hint
- many dimensions have their own (better) model in TEI
- TEI Lex-0 tries to streamline this vocabulary (ongoing discussion)

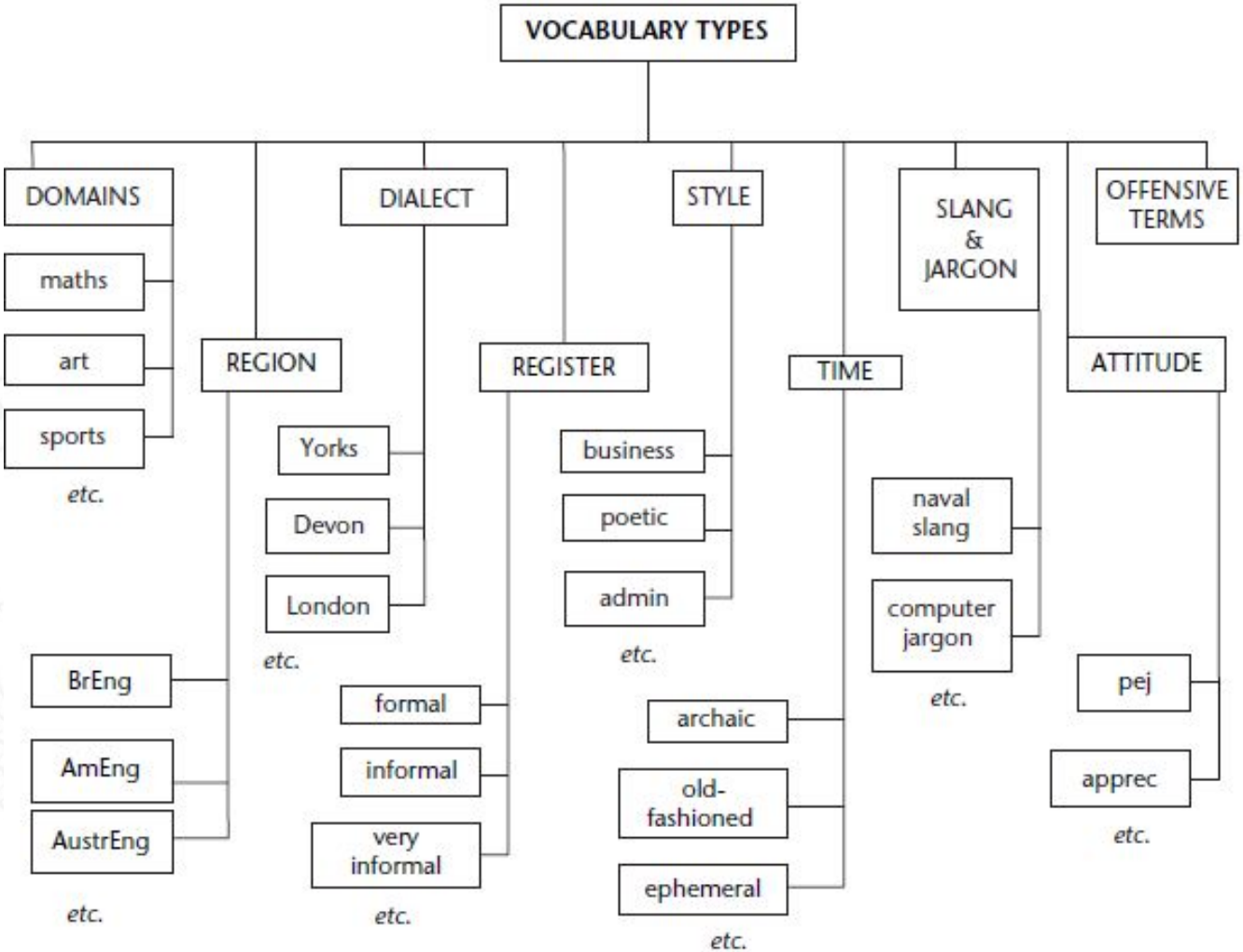
Usage labels

Criterion	Type of marking	Unmarked centre	Marked periphery	Examples of labels
Time	diachronic	contemporary language	archaism – neologism	*arch, dated, old use*
Place	diatopic	standard language	regionalism, dialect word	*AmE, Scot., dial.*
Nationality	diaintegrative	native word	foreign word	*Lat., Fr.*
Medium	diamedial	neutral	spoken – written	*colloq., spoken*
Socio-cultural	diastratic	neutral	sociolects	*pop., slang, vulgar*
Formality	diaphasic	neutral	formal – informal	*fml, infml*
Text type	diatextual	neutral	poetic, literary, journalese	*poet., lit.*
Technicality	diatechnical	general language	technical language	*Geogr., Mil., Biol., Mus.*
Frequency	diafrequential	common	rare	*rare, occas.*
Attitude	diaevaluative	neutral	connoted	*derog., iron., euphem.*
Normativity	dianormative	correct	incorrect	*non-standard*

Svensén (2009) after Hausmann (1989)

TEI Lex-0

Usage labels



Atkins/Rundell (2008)

Etymology

(“sometimes abstract models have no direct equivalent in TEI”)

etymological prose . . .

- often is exactly this: prose
  (i. e. not necessarily rigidly structured)
- outlines complex linguistic entities
- outlines complex (historical) relations among them

$\longrightarrow$ Can we formalize this more deeply than vanilla TEI?

Etymology

grunt vb. ME. *grunte gronte* OE. *grunnettan*; ident. w. G. *grunzen*, DAN. *grynte*, SW. *grynta* A more primit. stem appears in OE. *grunian* 'grunt'. The √*grun* is imitation of sound; cp. LAT *grunnire*.

- non-etymological information is covered in TEI
- etymological information not so much . . .
- essentially, we need a model for complex *mentioned* forms (such as etymons) and their relations
- need a model of the temporality of etymological processes

TEI Lex-0

Etymology

grunt vb. ME. *grunte gronte* OE. *grunnettan*; ident. w. G. *grunzen*, DAN. *grynte*, SW. *grynta* A more primit. stem appears in OE. *grunian* 'grunt'. The √*grun* is imitation of sound; cp. LAT *grunnire*.

```
<entry>
  <form type="lemma"><orth>grunt</orth></form>
  <gramGrp><gram type="pos">vb.</gram></gramGrp>
  <etym>
    <!-- ... -->
  </etym>
</entry>
```

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Etymology

grunt vb. ME. *grunte gronte* OE. *grunnettan*; ident. w. G. *grunzen*, DAN. *grynte*, SW. *grynta* A more primit. stem appears in OE. *grunian* 'grunt'. The √*grun* is imitation of sound; cp. LAT *grunnire*.

grunt

↑

grunte, gronte

↑

grunnetan

Problem in vanilla TEI: make (chained) relations explicit

TEI Lex-0

Etymology

grunt vb. ME. *grunte gronte* OE. *grunnettan*; ident. w. G. *grunzen*, DAN. *grynte*, SW. *grynta* A more primit. stem appears in OE. *grunian* 'grunt'. The √*grun* is imitation of sound; cp. LAT *grunnire*.

```
<etym>
  <lang rendition="sc">me.</lang>
  <mentioned>grunte</mentioned>
  <mentioned>gronte</mentioned>
  <lang rendition="sc">oe.</lang>
  <mentioned>grunnetan</mentioned>;
  <!-- ... -->
</etym>
```

Problem in vanilla TEI: associate `lang` and `mentioned`

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Etymology

grunt vb. ME. *grunte gronte* OE. *grunnettan*; ident. w. G. *grunzen*, DAN. *grynte*, SW. *grynta* A more primit. stem appears in OE. *grunian* 'grunt'. The √*grun* is imitation of sound; cp. LAT *grunnire*.

```
<cit type="etymon">
  <lang rendition="sc">me.</lang>
  <form type="lemma" xml:lang="enm">
    <orth>grunte</orth>
    <orth>gronte</orth>
  </form>
</cit>
```

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Etymology

grunt vb. ME. *grunte gronte* OE. *grunnettan*; ident. w. G. *grunzen*, DAN. *grynte*, SW. *grynta* A more primit. stem appears in OE. *grunian* 'grunt'. The √*grun* is imitation of sound; cp. LAT *grunnire*.

```
<cit type="etymon">
  <lang rendition="sc">oe.</lang>
  <form type="lemma" xml:lang="ang">
    <orth>grunian</orth>
  </form>
  <def>'grunt'</def>
</cit>
```

TEI Lex-0

Etymology

grunt vb. ME. *grunte gronte* OE. *grunnettan*; ident. w. G. *grunzen*, DAN. *grynte*, SW. *grynta* A more primit. stem appears in OE. *grunian* 'grunt'. The √*grun* is imitation of sound; cp. LAT *grunnire*.

```
<cit type="cognate">
  <lang rendition="sc">dan.</lang>
  <form type="lemma" xml:lang="da">
    <orth>grynte</orth>
  </form>
</cit>
```

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Etymology

grunt vb. ME. *grunte gronte* OE. *grunnettan*; ident. w. G. *grunzen*, DAN. *grynte*, SW. *grynta* A more primit. stem appears in OE. *grunian* 'grunt'. The √*grun* is imitation of sound; cp. LAT *grunnire*.

```
<etym type="inheritance">
  <cit type="etymon">
    <lang rendition="sc">me.</lang>
    <form type="lemma" xml:lang="enm">...</form>
  </cit>
  <cit type="etymon">
    <lang rendition="sc">oe.</lang>
    <form type="lemma" xml:lang="ang">...</form>
  </cit>
</etym>
```

Etymology

complex descriptions of linguistic signs via `cit[@type="etymon"]`, `cit[@type="cognate"]`:

- in a way, `cit[@type="..."]` is very close to `entry`
- may contain
  - lang (not in vanilla TEI)
  - date
  - form
  - def/gloss (even sense?)
  - usg
  - xr
  - gramGrp
  - ref
  - bibl

TEI Lex-0

Etymology

complex relations among cits via
`etym[@type="..."]`:

- types may include borrowing, inheritance, compounding, derivation, metaphor, ...
- typing may be expensive, therefore optional
- etyms contain mostly cits, maybe segs
  (for unmarked stretches of prose)
- conflicting etymologies can be siblings
  (and may get indications of responsibility)
- Bowers/Herold/Tasovac/Romary (2022): TEI Lex-0 Etym: Toward Terse Recommendations for the Encoding of Etymological Information. In: Journal of the TEI (rolling issue), `https://doi.org/10.4000/jtei.4300`

Outline

Outlook

- more areas of work not covered today,
  e. g. cross-references, bilingual dictionaries, ...
- things to come: lexical metadata, linking `cit` with corpora
- frequent group meetings
- open collaboration on GitHub:
  `https://github.com/DARIAH-ERIC/lexicalresources`
  (take a look!)
- close collaboration with the TEI consortium
- TEI Lex-0 is still work in progress (version 0.9.1)

Thank you for listening!