berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

TEI für Wörterbücher

TEI, TEI Lex-0 und Interoperabilität zwischen Wörterbüchern

Axel Herold[1,2]

[1]Berlin-Brandenburgische Akademie der Wissenschaften
[2]École Pratique des Hautes Études

Oktober 4th, 2018

Überblick

TEI-Hintergrund

Modellierung nach TEI

- Modellierung lexikografischer Daten
- Typografische Perspektive
- Editorische Perspektive
- Lexikografische Perspektive
- Probleme mit „reinem“ TEI

TEI-Lex-0

- Artikel
- Forms and grammatical information
- Usage labels
- Etymology

Outlook

TEI-Hintergrund

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

TEI

- Community, Consortium
- Format
- Guidelines
- „Ökosystem“
- seit Ende der 1980er Jahre aktiv
- faktisch weit verbreiteter DH-Standard
- basiert auf X*-Standards

TEI-Hintergrund

TEI für Wörterbücher

- eigenes Kapitel schon seit P1 (1990)
- separates Modul *dictionaries* (andere textsortenspezifische Module bspw. *drama*, *verse*, . . .)
- in vielen Projekten Austausch- oder Archivformat
  - Wörterbuchnetz (Trier)
  - WDG, EtymWB im DWDS (Berlin)
  - MWB (Mainz)
  - AWB (Leipzig)
- *keine* SIG „Dictionaries“
- internationale Initiative TEI-Lex-0
- `http://www.tei-c.org`, `https://github.com/TEIC`

Modellierung nach TEI

Modellierung lexikografischer Daten

- Modellierung $\approx$ Abbildung von Objekten und Eigenschaften (und deren Relationen) auf Symbole (typischerweise bei gleichzeitiger Abstraktion)
- lexikografische Daten werden auf verschiedenen Ebenen modelliert:
  - gedruckte Zeichen $\longrightarrow$ Codepoints (z. b. Unicode)
  - räumliche Beziehungen zwischen Zeichen $\longrightarrow$ Wörter (Token)
  - typografische Eigenschaften $\longrightarrow$ Funktion(en) von Wörtern (Token)
  - ...
- jede Ebene: Interpretation, möglicherweise Unsicherheit
- alternative und / oder inkompatible Interpretationen (und Modelle) möglich

Modellierung nach TEI

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Modellierung lexikografischer Daten nach TEI

verschiedene Perspektiven („views“) auf die Daten:

typografisch „the two-dimensional printed page, including information about line and page breaks and other features of layout“

editorisch „the one-dimensional sequence of tokens which can be seen as the input to the typesetting process . . . “

lexikografisch „. . . the underlying information represented in a dictionary, without concern for its exact textual form“

(TEI-Guidelines, Kapitel 9)

Modellierung lexikografischer Daten nach TEI

verschiedene Perspektiven („views“) auf die Daten:

- Wörterbuchherstellung: lexikografisch $\longrightarrow$ typografisch
- (Retro)digitilisierung: typografisch $\longrightarrow$ lexikografisch
- mehrere Perspektiven gleichzeitig behalten?
- typische (und empfehlenswerte) Entscheidung:
  - (Druck)zeichen als *character data* in Elementen
  - Annotationen, Normalisierungen als Attributwerte
  - typografische Eigenschaften als Attributwerte

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Modellierung lexikografischer Daten, Beispielartikel

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)

Modellierung nach TEI

Typografische Perspektive

```
<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>
```

Modellierung nach TEI

Editorische Perspektive

```
<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>
```

Modellierung nach TEI

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Lexikografische Perspektive

```
<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>
```

Modellierung nach TEI

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Lexikografische Perspektive, alternative

```
<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>
```

Modellierung nach TEI

Probleme mit „reinem“ TEI

(in Bezug auf Wörterbuchmodellierung)

- verschiendene sehr ähnliche Modelle
- verschiedene Möglichkeiten, ein Modell zu kodieren
- manche Modelle können nicht in TEI beschrieben werden
- Alternative Annotationen sind schwierig
- Vokabular für Annotationen ist manchmal „unscharf“

Um einige dieser Probleme geht es im Folgenden.

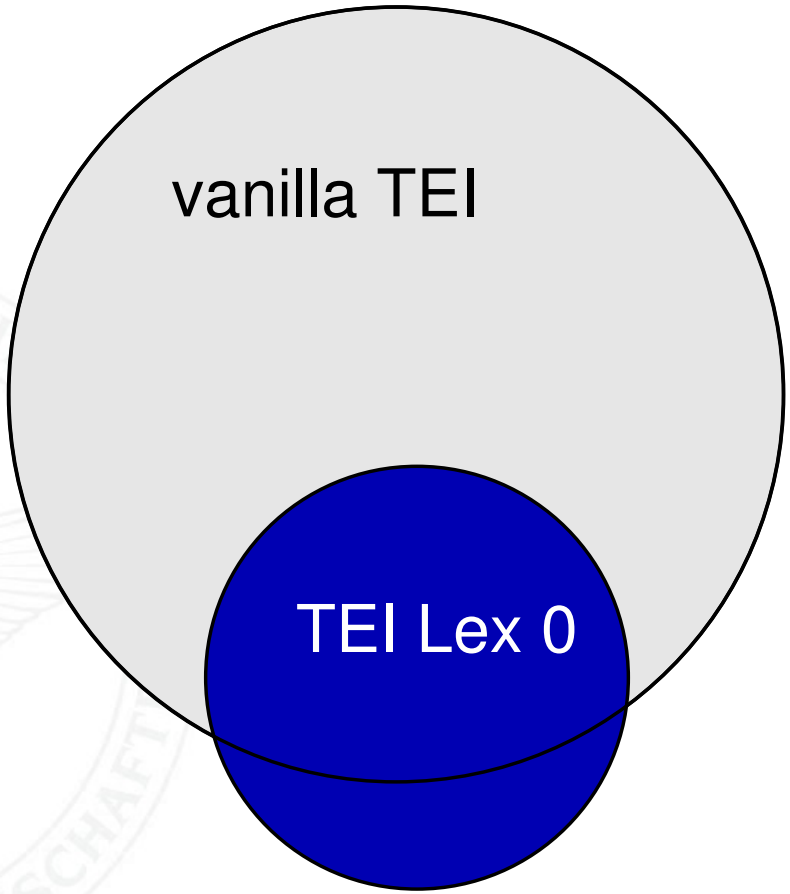
Hintergrund

- Arbeitsgruppe seit 2016, unterstützt von ENeL, Dariah, verschiedenen Forschungsinstituten
- internationale Gruppe mit engen Verbindungen zur TEI
- Anwendungsfall: typografisch $\longrightarrow$ lexikografisch
- kein Ersatz für Kapitel 9, eher eine Best-Practice-Anleitung (mit Ergänzungen, also eine *customization*)
- Hauptziel: Interoperabilität
  - Alternativen beschränken
  - Inhaltsmodelle beschränken
  - Vokabulare „reparieren“
  - …
- Änderungen zum Teil schon in TEI P5 (3.4.0) eingeflossen

Verhältnis zwischen TEI und TEI-Lex-0

„kein Ersatz für Kapitel 9, eher eine Best-Practice-Anleitung“

. . . für lexikografische Daten



Wichtige Nebenbemerkung: What's a TEI name, anyway?

```
<form>
  <orth>tree</ort>
</form>
```

```
<form>
  <orth>tree</ort>
  <form>
    <orth>trees</orth>
    <gramGrp><number>plur.</number></gramGrp>
  </form>
  <gramGrp><pos>noun</pos></gramGrp>
</form>
```

- „Formangabe“ oder „Container für form-bezogene Informationen“
- konsistentere Typisierungen sind nötig

TEI-Lex-0

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Artikel

(„verschiendene sehr ähnliche Modelle“)

- 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

Artikel

	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

Artikel

	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><pos>nn.</pos></gramGrp>
    <sense><def xml:lang="en">leather</def></sense>
  </entry>
  <entry type="word"> <!-- ledern [Adj.] --> </entry>
  <entry type="word"> <!-- abledern [vb.] --> </entry>
  <!-- ... -->
</entry>
```

Forms and grammatical information

(„verschiedene Möglichkeiten, ein Modell zu kodieren“)

- `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

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

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)

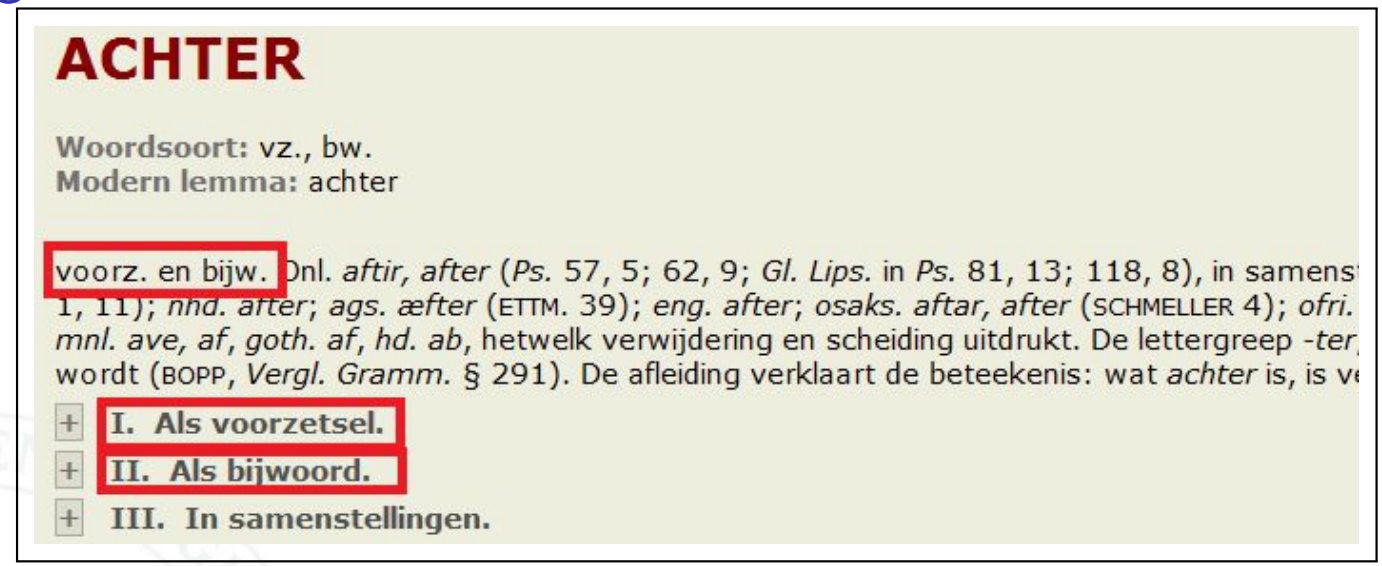
```
<entry>
  <form type="lemma"><orth>grunt</orth></form>
  <gramGrp><pos>vb.</pos></gramGrp>
  <etym>
    <!-- ... -->
  </etym>
</entry>
```

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><pos>noun</pos></gramGrp>
    <!-- ... -->
  </entry>
  <entry>
    <gramGrp><pos>verb</pos></gramGrp>
  </entry>
</entry>
```

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><pos>voorz. en bijw</pos></gramGrp>
  <etym> <!-- --> </etym>
  <sense n="I">
    <gramGrp><pos>Als voorzetsel.</pos></gramGrp>
    <!-- ... --></sense>
  <sense n="II">
    <gramGrp><pos>Als bijwoord.</pos></gramGrp>
    <!-- ... --></sense> <!-- ... --></entry>
```

Forms and grammatical information, exception

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

We really want inheritance to work!

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!

Usage labels

(„Vokabular für Annotationen ist manchmal ‚unscharf‘“)

- `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)

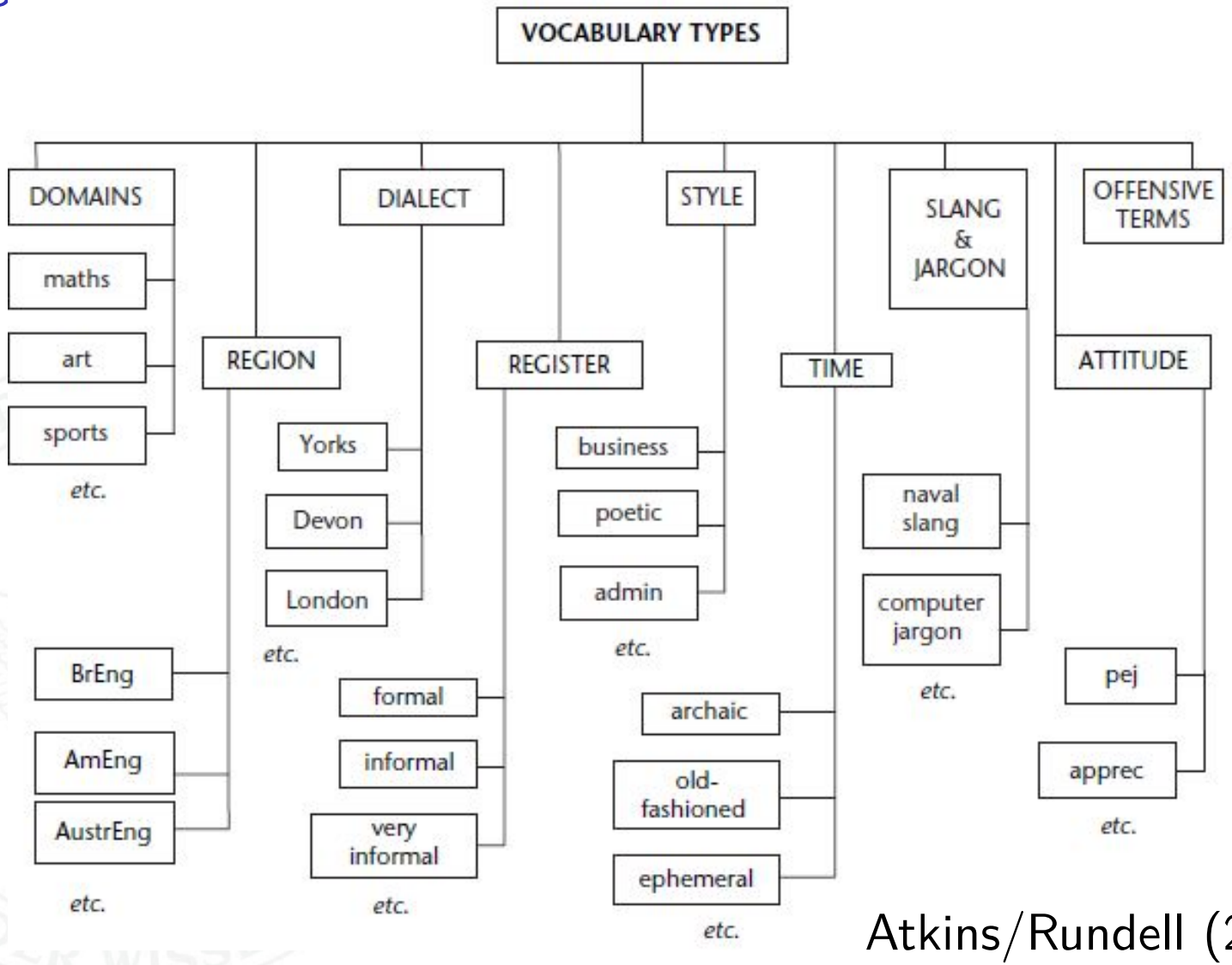
Usage labels

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

(„manche Modelle können nicht in TEI beschrieben werden“)

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 device to model complex *mentioned* forms (such as etymons) and their relations
- need a model of the temporality of etymological processes

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><pos>vb.</pos></gramGrp>
  <etym>
    <!-- ... -->
  </etym>
</entry>
```

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`

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>
```

Remeber: What's a TEI name, anyway ...

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>
```

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

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

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=ëtymon"]`, `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

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)
- upcoming paper by Bowers/Herold/Romary

Outline

Outlook

- more areas of work not covered here, e. g. for cross-references, bilingual dictionaries, . . .
- frequent group meetings
- open collaboration on GitHub (take a look!)
- close collaboration with the TEI consortium
- TEI Lex-0 is still work in progress

Thank you for listening!