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

4. Oktober 2018

Überblick

TEI-Hintergrund

Modellierung nach TEI

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

TEI-Lex-0

- Artikel
- Formangaben und grammatische Angaben
- Gebrauchsangaben
- Etymology

Ausblick

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

Modellierung nach TEI

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Modellierung lexikografischer Daten, Interoperabilität

Einheitliche Kodierung lexikografischer Ressourcen vereinfacht:

- Werkzeugentwicklung
- (wissenschaftliche) Auswertung
- Ressourcenverknüpfung
- (Langzeit)archivierung

$\longrightarrow$ gut für Entwicklung von *Forschungsinfrastrukturen*

$\longrightarrow$ muss kein Ersatz für projektinterne Formate sein

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

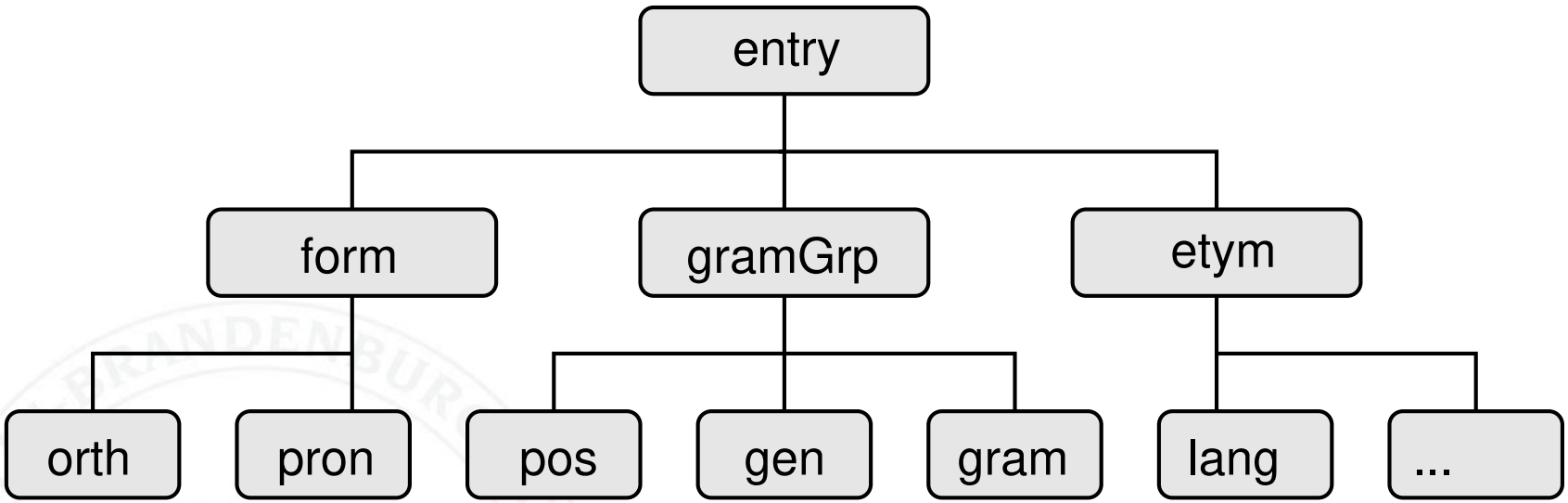
berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

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

Lexikografische Perspektive, allgemeine Struktur



- nur Auswahl an Elementen
- Artikel mit Baumstruktur vs. diskursiver Struktur

Modellierung nach TEI

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Probleme mit „reinem“ TEI

(in Bezug auf Wörterbuchmodellierung)

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

Um einige dieser Probleme geht es im Folgenden.

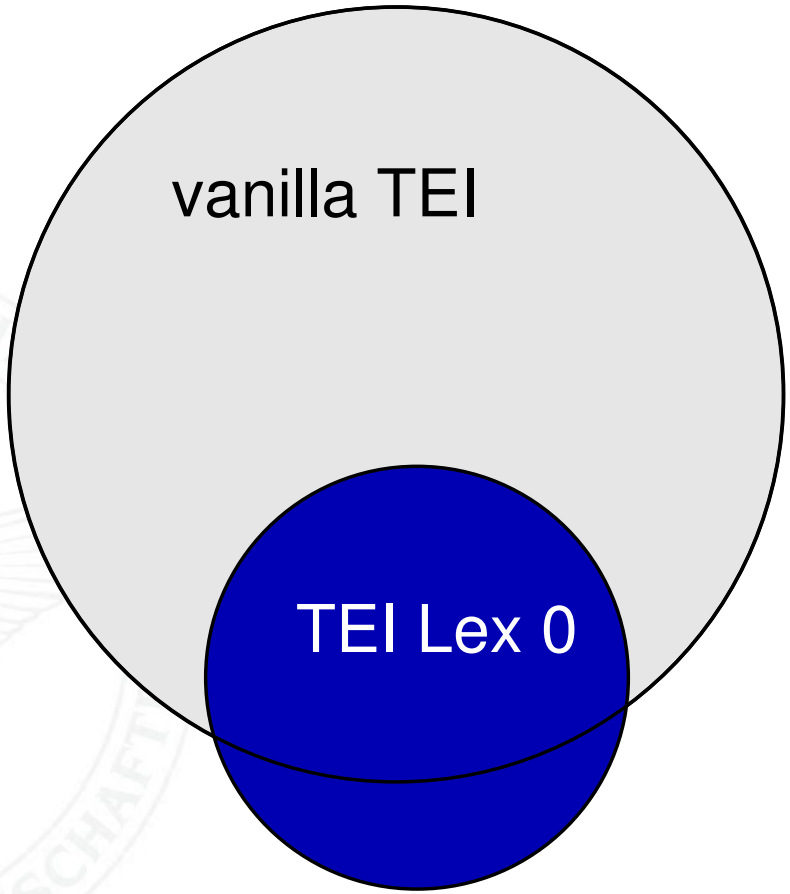
Hintergrund

- Arbeitsgruppe seit 2016, unterstützt von ENeL, Dariah-EU, elexis, 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 (zum Beispiel `form/pc`)

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, aber sehr ähnliche TEI-Modelle“)

- TEI-Modelle für Wörterbuchartikel (und ähnliche Strukturen):

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 . . . “

- alle Modelle haben minimal unterschiedliche *content models*
- in TEI Lex-0: nur **`entry`**, dafür rekursiv verwendbar

Artikel, Beispiel

	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)

Artikel, Beispiel

	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>
```

Formangaben und grammatische Angaben

(„verschiedene Möglichkeiten, ein Modell in TEI zu kodieren“)

- `entry/gramGrp`, `form/gramGrp`, `sense/gramGrp`, `...` in TEI
- in TEI-Lex-0 restriktiver und explizit mit Vererbung (entsprechend der Einbettung):
  - eintragsbezogene Angaben: `entry/gramGrp`
  - lesartenbezogene Angaben: `sense/gramGrp`
  - formbezogene Angaben: `form/gramGrp`
- in TEI-Lex-0: `form/@type` obligatorisch,
  z. B. `lemma`, `inflected`, `paradigm`, `variant`

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Formangaben und grammatische Angaben

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)

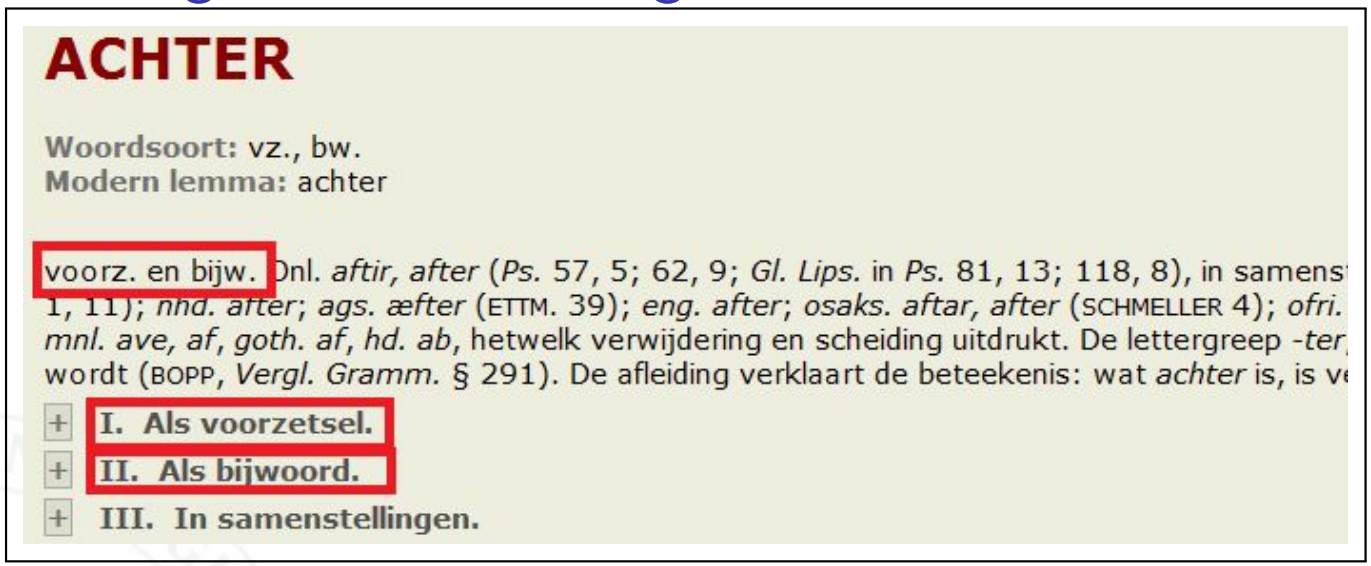
```
<entry>
  <form type="lemma"><orth>grunt</orth></form>
  <gramGrp><pos>vb.</pos></gramGrp>
  <etym>
    <!-- ... -->
  </etym>
</entry>
```

Formangaben und grammatische Angaben

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>
```

Formangaben und grammatische Angaben



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>
```

Formangaben und grammatische Angaben, flektierte Formen

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

(Bezug explizit durch Vererbung)

TEI-Lex-0

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Gebrauchsangaben

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

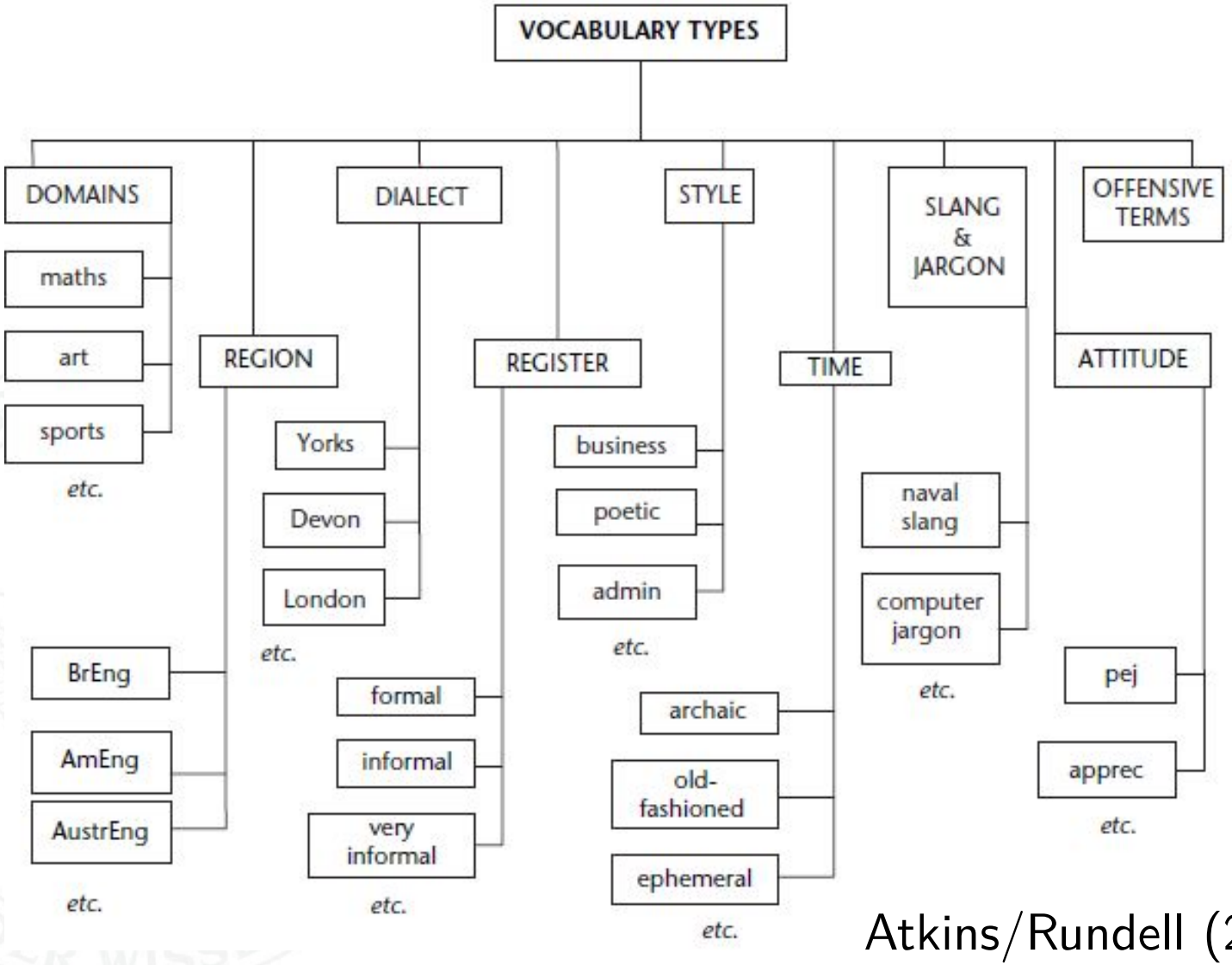
- `usg` beschreibt sehr heterogene Dimensionen:
  - geo(graphic)
  - time
  - dom(ain)
  - register, style
  - plev (preference level)
  - lang(uage)
  - gram(matical)
  - syn(onym), hyp(ernym)
  - colloc(ation), comp(lement), obj(ect), subj(ect), verb
  - hint
- viele Dimensions haben eigene (und besser) TEI-Modelle
- TEI-Lex-0 wird Vokabular überarbeiten (aktuelle Diskussion)

Gebrauchsangaben

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) nach Hausmann (1989)

Gebrauchsangaben



Atkins/Rundell (2008)

Etymology

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

etymologische Prosa . . .

- ist oft genau das: Prosa
  (also nicht notwendigerweise streng strukturiert)
- beschreibt komplexe linguistische Zeichen
- beschreibt komplexe (historische) Zusammenhänge zwischen diesen Zeichen

$\longrightarrow$ Können wir bessere Modelle entwickeln als TEI sie vorsieht?

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*.

- nicht-etymologische Information („grunt vb.“) ist leicht zu modellieren
- etymologische Information ist nicht so eindeutig in TEI modellierbar
- im Kern brauchen wir:
  - ein Modell für *komplexe* erwähnte Zeichen (z. B. Etyma) und deren Relationen
  - ein Model für die zeitliche Abfolge von Sprachwandelprozessen

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*.

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

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

Problem hier: Bezug zwischen `lang` und `mentioned` explizieren

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>
```

Wie gesagt: 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>
```

(auch weitere Angabetypen möglich)

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*.

Problem hier: verkettete Relationen explizieren

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 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>
```

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 type="cognacy">
  <cit type="cognate">
    <lang rendition="sc">g.</lang>
    <form type="lemma" xml:lang="deu">...</form>
  </cit>
  <cit type="cognate">
    <lang rendition="sc">dan.</lang>
    <form type="lemma" xml:lang="dan">...</form>
  </cit>
</etym>
```

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">
   <etym
      type="inheritance">
      <!-- Old English
      to Middle English -->
   </etym>
   <etym type="cognacy">
      <!-- German, Danish -->
   </etym>
   <cit type="root">
      <form type="lemma" xml:lang="x-pie">
         <orth>grun</orth></form>
   </cit>
</etym>
```

Etymology

komplexe Modelle sprachlicher Zeichen durch
`cit[@type="etymon"]`, `cit[@type="cognate"]`:

- `cit[@type="..."]` ist konzeptionell fast schon ein `entry`
- mögliche Angabentypen in TEI-Lex-0 (mehr in TEI):
  - lang
  - date
  - form
  - def / gloss
  - usg
  - xr
  - gramGrp
  - ref
  - bibl

Etymology

komplexe Relationen zwischen `cits` durch `etym[@type="..."]`:

- Typen: `borrowing`, `inheritance`, `compounding`, `derivation`, `metaphor`, `...`
- Typisierung optional, da (manuell) aufwändig zu annotieren
- `etym` enthält hauptsächlich `cit`,
  aber vor allem auch andere `etym`
  (in TEI jetzt auch möglich)
- alternative Etymologien als Geschwister modellieren
- Bowers/Herold/Romary (in Arbeit)

Outline

Ausblick

- TEI-Lex-0 hat noch mehr Themen, z. B. Verweise, mehrsprachige Wörterbücher, . . .
- regelmäßige Treffen der Gruppe
- Arbeitsstand öffentlich einsehbar unter `https://github.com/DARIAH-ERIC/lexicalresources`
- enge Kooperation mit TEI-Council
- TEI Lex-0 ist *work in progress*

Vielen Dank
für die Aufmerksamkeit!