Ludwig-Maximilians-Universität München
Institut für Musikwissenschaft
Lektürekurs „Digital Humanities – Digitale Methoden in der Musikwissenschaft“

Notizen zur 4. Sitzung

(30.04.2018)

Brieftranskription, Teil 2

Besprechung der Hausaufgabe

Aufgabenstellung: Auszeichnung des Briefs der Allgemeinen Musikgesellschaft Zürich an Adolf Sandberger (19.10.1899) mit den in Sitzung 3 besprochenen TEI-Tags.
Musterlösung: siehe amg_zuerich_dtb_20180430.xml

verwendete Tags für Briefabschnitte (bekannt aus Sitzung 3):

`<dateline>`	Zeile im Briefkopf mit Angabe von Datum, Ort etc.
`<salute>`	Anrede oder Schlussformel
`<p>`	normaler Absatz (Mengentext im Brief)
`<signed>`	Unterschrift

Zeilenumbrüche werden mit dem leeren Element `<lb/>` codiert. Die Zeilenumbrüche im Text der XML-Datei dürfen damit sogar entfallen; die ganze Information steckt ja jetzt in diesem Element. Steht vor dem Zeilenumbruch ein Trennstrich, bietet es sich an, diesen speziell auszuzeichnen. Dafür gibt es das Element `<pc>` („punctuation character“, Satzzeichen). Das zusätzliche Attribut `force="weak"` zeigt an, dass das Satzzeichen keine zwei Wörter trennt, sondern die Buchstaben links und rechts vom Element zum gleichen Wort gehören. Insgesamt hat man also zum Beispiel:

```
Musik<pc force="weak">-</pc><lb/>gesellschaft
```

Für den in Anführungszeichen gesetzten Text kann das `<q>`-Element verwendet werden. Bei der Übertragung von Primärquellen empfiehlt es sich jedoch, die Anführungszeichen weiterhin auch als Zeichen zu übertragen. Im Gegensatz zu Sekundärliteratur (vgl. die Codierung des Nägele-Aufsatzes in Sitzung 2 und 3) kann nämlich beispielsweise die Form der Anführungszeichen (etwa „“ oder »«) von Interesse sein.

Für die Abkürzung „gef.“ kann in der Transkription eine Auflösung angegeben werden:

```
<choice>
   <abbr>gef.</abbr>
   <expan>gefällige</expan>
</choice>
```

Für die Unterschrift steht in der gegebenen Rohtranskription lediglich der Platzhalter „[Unterschrift]“ – wohl, weil diese schwer zu entziffern ist. Für die TEI-Codierung bietet

sich hier das (leere) <gap/>-Element an. Dieses steht zunächst für eine „Lücke“ jeglicher Art: bei der Transkription übersprungene Teile, Auslassungen in Zitaten, oder eben auch eine Transkriptionslücke wegen unleserlichen Texts. Der Grund für die Auslassung kann im Attribut reason näher angegeben werden. In TEI wird die Angabe „[Unterschrift]“ somit wie folgt vollständig abgebildet:

```
<signed><gap reason="illegible"/></signed>
```

Opener und Closer

Die TEI-Richtlinien empfehlen für Brieftranskriptionen, die formalisierten Anfangs- und Schlussteile nicht nur mit Elementen wie <dateline> und <salute> auszuzeichnen, sondern die betreffenden Zeilen auf höherer Ebene zu <opener> und <closer> zusammenzufassen, zum Beispiel:

```
<opener>
   <dateline>Zürich, 19. X 99</dateline>
   <salute>Herr Dr. A Sandberger <lb/>München.</salute>
</opener>


<closer>
   <salute>namens des Vorstands <lb/>der Aktuar</salute>
   <signed><gap reason="illegible"/></signed>
</closer>
```

Der Nachteil dieser Elemente besteht jedoch darin, dass das TEI-Schema recht strenge Vorgaben zu den möglichen Kindelementen und deren Reihenfolge macht, die nicht unbedingt von jeder Briefquelle erfüllt werden.

Metadaten

Ein Punkt, der in diesem Kurs recht kurz kommen wird, ist die Erfassung von Metadaten zu einem digitalen Text. Für eine nachhaltige Archivierung sind gut gepflegte Metadaten jedoch extrem wichtig. Anhand des Briefs zum DTB-Beitritt Prinz Ludwig Ferdinands soll deshalb kurz gezeigt werden, wie zumindest die wichtigsten Arten von Metadaten im <teiHeader> erfasst werden können (vgl. Datei prinz_ludwig_dtb_20180430.xml):

Quellenbeschreibung

Die dem Text im <text>-Element zugrundeliegende(n) Quelle(n) werden in der <sourceDesc> („source description“, Quellenbeschreibung) angegeben. Gibt es keine Textvorlage, reicht eine Angabe wie Born digital. (siehe TEI-Rumpfdatei). Stammt der Text aus einer physischen Quelle, verwendet man stattdessen das Element <msDesc> („manuscript description“), das zumindest einen <msIdentifier> („manuscript identifier“) enthalten sollte, der den Aufbewahrungsort der Quelle angibt. Die Angaben in der Fußzeile der Rohtranskription ließen sich etwa so abbilden:

```
<sourceDesc>
   <msDesc>
      <msIdentifier>
         <repository>Bayerische Staatsbibliothek</repository>
         <collection>Nachlass der Gesellschaft zur Herausgabe der DTB,
            Ana 529, Schachtel 2</collection>
      </msIdentifier>
   </msDesc>
</sourceDesc>
```

`<repository>`	aufbewahrende Institution
`<collection>`	Sammlung innerhalb der Institution, zu der die Quelle gehört

So weiß ein späterer Nutzer der TEI-Datei, wo die Quelle zu finden ist – eine editorische Mindestanforderung.

Briefmetadaten

Für Briefe gibt es eine eigene Klasse von Metadaten, für die eigens das Element `<correspDesc>` („correspondence description“) eingeführt wurde, das Teil eines Elements `<profileDesc>` ist. In diesem Element sollen die verschiedenen „korrespondenztypischen“ Aktionen wie Versand und Empfang des Briefs in einzelnen `<correspAction>`-Elementen mit entsprechenden `type`-Attributen aufgelistet werden. Jede `<correspAction>` bekommt Kindelemente, die Details zu dieser bestimmten Aktion angeben: Absendername, Ort und Datum kommen in `<correspAction type=ßent»`, Empfängername, Ankunftsdatum und Zielort (soweit bekannt) in `<correspAction type="received»`.

Für den vorliegenden Brief kann folgender Block im `<teiHeader>` ergänzt werden:

```
<profileDesc>
   <correspDesc>
      <correspAction type="sent">
         <orgName>Hofsecretariat seiner kgl. Hoheit des Prinzen Ludwig
            Ferdinand von Bayern</orgName>
         <placeName>München</placeName>
         <date when="1900-06-04"/>
      </correspAction>
      <correspAction type="received">
         <name>Unknown</name>
      </correspAction>
   </correspDesc>
</profileDesc>
```

Der Absender wurde hier aus dem Stempelaufdruck hergeleitet; alternativ könnte hier auch der (unleserliche) Name des unterschreibenden Hofrats stehen. „Prinz Ludwig Ferdinand“ allein wäre ungenau, denn aus dem Brieftext geht klar hervor, dass Ludwig Ferdinand den Brief nicht selbst verfasst hat. Statt das generische `<name>` zu benutzen, kann mit den Elementen `<persName>` und `<orgName>` genauer spezifiziert werden, ob es sich um den Namen einer Person oder einer Organisation/Institution handelt.

Das <date>-Element erlaubt in seinem when-Attribut die Angabe eines maschinenlesbaren Datums im Format YYYY-MM-DD. Weitere Angaben braucht es dann nicht; das Element enthält keinen Text oder Kindelemente.

Für die Angabe, dass ein Name unbekannt ist, gibt es keinen festen Standard; die Codierung hier ist ein Notbehelf, der sich jedoch bei zahlreichen Briefeditionen eingebürgert hat. (Adolf Sandberger ist nicht Adressat des Briefs, da er nicht als „Hochwohlgeboren" angeredet würde; die DTB-Gesellschaft kann ebenfalls nicht gemeint sein – der Empfänger dieses Briefs ist derzeit in der Tat nicht bekannt.)

TEI Guidelines

Die semantische Bedeutung aller TEI-Elemente sowie die zulässigen Kombinationen von Elementen und Attributen werden in den *TEI Guidelines* spezifiziert. Die jeweils aktuelle Version der Spezifikation kann unter
`http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html`
abgerufen werden. Hiervon werden auch die von der Schemadatei tei_all.rng geprüften Syntaxregeln abgeleitet.

Da die *Guidelines* den gesamten TEI-Sprachschatz und alle möglichen Anwendungsbereiche abdecken, lohnt es sich nicht, alle 23 Hauptkapitel auf einmal durchzulesen, sondern gezielt nach den benötigten Informationen zu stöbern. Als Beispiel soll die Spezifikation des bereits bekannten <choice>-Elements betrachtet werden:
`http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-choice.html`
(auffindbar über Appendix C Elements – c – choice oder die Suchfunktion).

Die Spezifikation ist in folgende Teile gegliedert (von „leicht verständlich" bis „nur mit technischem Vorwissen lesbar"):

Description Eine Kurzbeschreibung der semantischen Bedeutung sowie Links zu relevanten Kapiteln der *Guidelines*.

Example Ein einfaches Beispiel für eine typische Anwendung, hier ein Abschnitt aus *Gulliver's Travels*, bei dem eine Jahreszahl korrigiert (<choice> – <sic> – <corr>) und zu einer britisch-englischen auch eine amerikanisch-englische Schreibweise angegeben wird (<choice> – <orig> – <reg>).

Note Weitere Verwendungshinweise:

1. Die einzelnen Kindelemente eines <choice>-Elements codieren alternative Transkriptionen, von der in der Regel eine gewählt werden muss, wenn man aus der TEI-Codierung einen linearen Text ableiten will.
2. Innerhalb eines <choice>-Elements kann ein weiteres <choice> stehen (engl. „self-nesting"). (Beispiel: Zu einem mittelhochdeutschen Wort wird mit <choice> – <orig> – <reg> die neuhochdeutsche Variante angegeben, gleichzeitig ist das mittelhochdeutsche Wort in der Originalquelle falsch geschrieben, also muss die <orig>-Variante mit <choice> – <sic> – <corr> korrigiert werden.)
3. Werden zu einem Werk mehrere Textzeugen (z. B. mehrere Quellen) transkribiert, sollen die Unterschiede nicht mit <choice>, sondern mit <app> codiert werden. <choice> kommt nur zum Einsatz, wenn sich die verschiedenen Lesarten aus *einem* Textzeugen ergeben.

May contain Elemente, die als Kindelemente von zugelassen sind. Bei `<choice>` sind das nicht viele: neben `<sic>` und `<corr>` (für Korrekturen des Transkribierers) sowie `<abbr>` und `<expan>` (Auflösung von Abkürzungen, `<am>` und `<ex>` ermöglichen eine noch feinere Codierung) gibt es noch `<orig>/<reg>` (u. a. zur Anpassung von Text an moderne Sprachgewohnheiten), `<unclear>` (zur Markierung unsicherer Transkriptionen; in Verbindung mit `<choice>` zum Anbieten mehrerer alternativer Transkriptionen bei schwer lesbarem Text), `<supplied>` (für durch den Herausgeber ergänzten Text) und `<seg>` (für ein allgemeines Textsegment ohne weitere editorische „Wertung“).

Contained by Elemente, innerhalb derer `<choice>` benutzt werden darf. Da in allen möglichen Arten von Textabschnitten und auch in Metadatenfeldern Alternativen angegeben werden dürfen, sind das für `<choice>` relativ viele. Um sich besser orientieren zu können, sind die zahlreichen TEI-Elemente in Klassen (sogenannte *Modellklassen*) untergliedert. Neben dem *core* („Kern“ der quasi überall zum Einsatz kommenden Elemente) gibt es beispielsweise *textstructure* für erweiterte Textgliederung (z. B. bei Briefen) oder *drama* für Dramentexte. *dictionaries* braucht man etwa nur, wenn man ein Wörterbuch codieren will, nicht für Quellentranskriptionen.

Attributes Erlaubte Attribute des Elements. Auch diese sind in Klassen unterteilt, um Attributdefinitionen mehrfach verwenden zu können. Siehe zum Beispiel die Attribute der Klasse `att.global.rendition`, die bei jedem Element, für das sie erlaubt sind, die gleiche Bedeutung haben (Angaben, wie das Element bei der Anzeige dargestellt werden soll).

Content model Technische Spezifikation des Elementinhalts in einer TEI-eigenen XML-Syntax. Diese kann zusätzliche Informationen enthalten, die von den anderen Kategorien nicht abgedeckt wird. Für `<choice>` ergibt sich hieraus zum Beispiel, dass dieses Element mindestens zwei Kindelemente braucht (`minOccurs="2"`), beliebig viele Kindelemente erlaubt sind (`maxOccurs=ünbounded"`; das heißt es können z. B. auch 3 alternative Lesarten angegeben werden), und die Kindelemente entweder ein weiteres `<choice>` oder Elemente der Modellklasse `model.choicePart` sein müssen (zu der genau die oben genannten Elemente gehören).

Schema Declaration Weitere technische Spezifikation der Elementsyntax, die sich an der Syntax regulärer Ausdrücke orientiert. Dass `<choice>` mindestens zwei Kindelemente braucht, ist hier jedoch nicht spezifiziert.

Die Links unter *Description* führen zu den Kapiteln der *Guidelines*, die nähere Erläuterungen zu Semantik und Verwendung des Elements liefern. Für `<choice>` werden die typischen Kombinationen von Kindelementen beschrieben, und das Element wird in den breiteren Kontext von „Simple Editorial Changes“ eingebettet: In Kapitel 3.4.3 wird etwa auch das `<gap>`-Element erwähnt, das der Kennzeichnung von Transkriptionslücken dient.

Vergleiche als weiteres Beispiel die Spezifikation von `<correspAction>`:
`http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-correspAction.html`

Das Element ist nur innerhalb von `<correspDesc>` erlaubt – es handelt sich um ein recht spezifisches Metadatenfeld. In der Liste der möglichen Attribute wird zwar auf einige universal definierte Attributklassen verwiesen, das Attribut `type` jedoch individuell spezifiziert: Damit vergleichbare Korrespondenzvorgänge von allen TEI-Anwendern einheitlich

codiert werden, gibt es eine Liste von *Suggested values*, die nach Möglichkeit verwendet werden sollen: `sent`, `received`, `transmitted`, `redirected`, `forwarded`.

Textannotation

Auch wenn der Text bereits übertragen, formatiert und mit editorischen Markierungen versehen ist, kann die Transkription des Briefs von Prinz Ludwig Ferdinand noch weiter ausgebaut werden (vgl. Datei prinz_ludwig_dtb_20180430.xml).

Der Brief enthält zwei Datumsangaben, eine in der Datumszeile („4. Juni 1900“) und eine im Text, die nicht unbedingt als solche leicht erkennbar ist (vor allem nicht durch eine schlecht trainierte maschinelle Textanalyse): „v. Mts.“ steht für „vorigen Monats“, also ist der „14. v. Mts.“ im Kontext dieses Briefs der 14. Mai 1900.

Um solche Datumsangaben in der Transkription zu markieren, benutzt man ebenfalls das `<date>`-Element (beachte beim zweiten Datum die Position von Anfangs- und End-Tag im Zusammenhang mit der Abkürzungsauflösung!):

```
<date>4. Juni 1900</date>
<date>14. <choice>
   <abbr>v. Mts.</abbr>
   <expan>vorigen Monats</expan>
</choice></date>
```

Ähnlich wie in der `<correspDesc>` kann hier auch ein maschinell lesbares Datum im `when`-Attribut hinterlegt werden; durch die Codierung in einem Attribut bleibt der ausgezeichnete Text der Quelle unangetastet:

```
<date when="1900-06-04">4. Juni 1900</date>
<date when="1900-05-14">14. <choice>
   <abbr>v. Mts.</abbr>
   <expan>vorigen Monats</expan>
</choice></date>
```

Auf diese Weise kann ein entsprechend gebautes Analyseprogramm die Datumsangaben nun leicht auffinden. Wäre die Transkription Teil einer größeren Briefsammlung, könnte das Dokument dadurch kontextualisiert werden: Welche Briefe wurden am 14. 05. 1900 noch geschrieben? Ist vielleicht die in diesem Brief erwähnte „freundliche Mittheilung“ dabei?

Eine solche sogenannte *Annotation* geht über das hinaus, was in einer normalen Textverarbeitung vorgesehen wäre, und reichert die Datei mit über den Text hinausgehenden Informationen an, die sich später wissenschaftlich „nachnutzen“ lassen.