Projekt *Die Schule von Salamanca*

Kontakt: info.salamanca@adwmainz.de

Allgemeine Erfassungsanweisungen für Werke des Projekts *Die Schule von Salamanca*

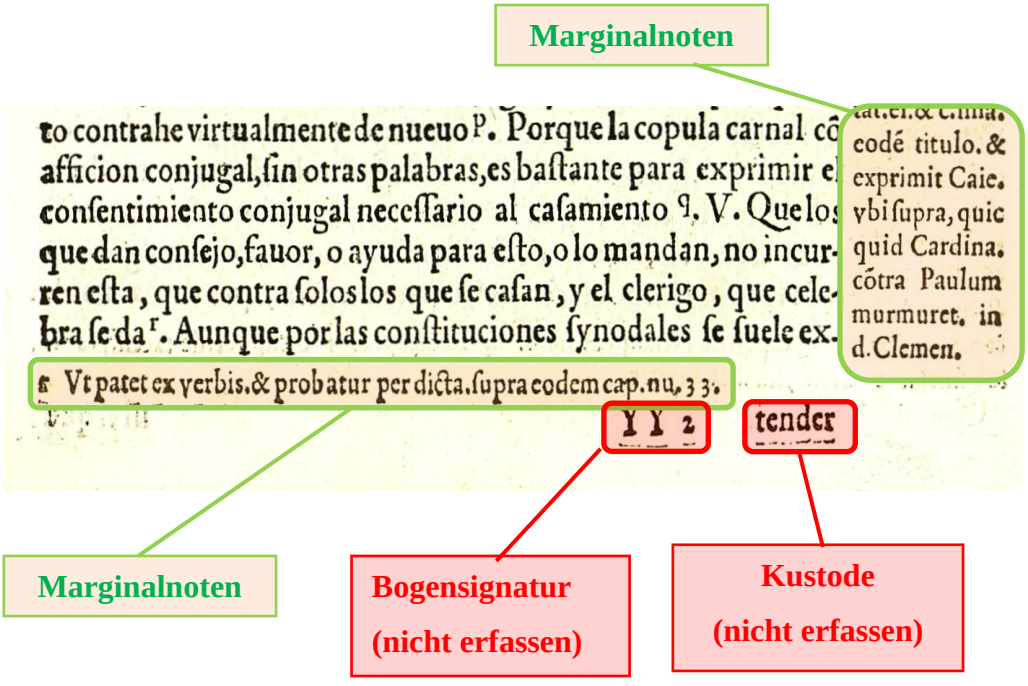
Stand: 20.04.2018

I. Allgemeine Definitionen: Was (nicht) zu erfassen ist

Hauptbereich und Marginalbereich des Textes

Der 'Hauptbereich' ist der innerste (d.h. selbst nicht wieder einen anderen Textbereich umfließende), ggf. mehrspaltige Textblock; er wird komplett erfasst. Der 'Marginalbereich' ist der Bereich vom äußeren Seitenrand bis zum Haupttext. Er umfließt den Hauptbereich oben und unten sowie innen und außen und beinhaltet Seitenzahlen, Kustoden, Noten, Kolumnentitel und Marginalnoten. Davon werden nicht erfasst: Seiten-/Foliozahlen, Kopfzeilen, Kolumnentitel, Bogensignaturen, Kustoden. Erfasst werden die Marginalnoten, die in der Regel an den Seiten und unterhalb des Textes stehen.

Hier als generelles Beispiel ein Seitenende (nicht werkspezifisch):



Handschriftliche Elemente

Handschriftliche Notizen, Anmerkungen, Symbole, Zeichnungen und Anstreichungen werden nicht erfasst. Für den Umgang mit (möglicherweise handschriftlich) durchgestrichenen Passagen siehe Abschnitt III.4.a „Nicht oder schwer lesbare Stellen“.

II. Zeichenkodierung

Allgemeines

Textzeichen werden grundsätzlich und so weit wie möglich als Unicode-Zeichen erfasst. Dies gilt auch für Digraphen und Zeichen mit diakritischen Marken (z.B. „æ“, „œ“; „é“, „ñ“, „ä“, „ç“) sowie für nicht-lateinische Zeichen (etwa griechische oder hebräische Zeichen, siehe hierzu auch Punkt III.3.m). Sind Zeichen nicht als Unicode-Zeichen darstellbar, so sollen sie in NCR-Notation (d.h. in der hexadezimalen Schreibweise von xml/html entities) ausgedrückt werden; wie (besondere) Zeichen erfasst werden, kann dabei generell der Transkriptionsübersicht der ungewöhnlichen Zeichen entnommen werden.

Kann ein Zeichen nicht eindeutig mithilfe der „Transkriptionsübersicht“ erfasst werden, so wird anstelle des entsprechenden Zeichens das leere TEI-Element <g sameAs="..."/> gesetzt, wobei der Wert von @sameAs auf einen entsprechenden Eintrag in einer externen Sonderzeichen-Dokumentationsdatei (XML) verweist, in welcher ggf. entsprechende Symbolgrafiken verlinkt sind. Sollte sich bereits vor oder im Laufe der Erfassung eines Textes abzeichnen, dass ein darin häufig benutztes Zeichen nicht in der „Transkriptionsübersicht“ vorhanden ist, soll das Projekt bezüglich der richtigen Kodierung und der Anpassung der „Transkriptionsübersicht“ kontaktiert werden.

Umgang mit besonderen Buchstaben und Zeichen

„S“ und Ligaturen

Rundes S („S“/„s“), Schaft-S („ſ“) und die ſ-s-Ligatur („ß“) werden als solche erfasst, unabhängig von benachbarten Zeichen. Ligaturen mit „s“ oder „ſ“ werden dementsprechend in aufgelöster Form erfasst (siehe auch folgenden Absatz).

Nur die Ligaturen „Æ“/„æ“ und „Œ“, „œ“ werden als solche erfasst. Andere Ligaturen (z.B. „ij“, „ﬅ“, „ﬆ“ „ﬂ“) werden in aufgelöster Form erfasst (z.B. „ij“, „ſt“, „st“ „fl“).

Zur Kodierung von Ligaturen (und Zeichen generell) siehe auch die Transkriptionsübersicht der ungewöhnlichen Zeichen

Absatzzeichen

Zeichen wie  ,  ,  , die einen (sinnhaften) Absatzbeginn markieren (auch innerhalb des typographischen Absatzes), sollen als „¶“ (Latin-1 Supplement Code Block , bzw. ¶ in NCR-Notation) erfasst werden .

„I“ vs. „J“ und „V“ vs. „U“
„V“/„v“ und „I“/„i“ werden als solche erfasst (d.h. nicht etwa normalisiert als „U“/„u“ oder „J“/„j“).

Anführungsstriche
Anführungsstriche werden grundsätzlich nicht mittels (Unicode-)Zeichen wiedergegeben, sondern mithilfe des TEI-Elements <q> ausgedrückt und durch dieses ersetzt; siehe hierzu unbedingt III.3.h „Text in Anführungsstrichen“.

Leerzeichen
Die Verwendung von Leerzeichen (nach Satzzeichen) wird, soweit möglich, stillschweigend an moderne Gepflogenheiten angepasst, so dass die korrekte Zusammen- und Getrenntschreibung, anders als im Druckbild der Vorlage, deutlich werden. Fehlende Leerzeichen nach Satzzeichen werden also beispielsweise ergänzt, während überzählige Leerzeichen (sofern der erhöhte Zwischenraum nicht bedeutungstragend ist) reduziert werden.

Trennungszeichen und Gleichheitszeichen
Trennstriche (als Markierungen von Worttrennungen) werden in normalisierter Form als moderner Bindestrich „-“ erfasst (siehe hierzu in jedem Fall Punkt III.3.a „Zeilenumbrüche (und Worttrennungen)“). Dies gilt etwa auch für das rechts schräg nach oben gestellte Gleichheitszeichen, das vor allem in Fraktuschriften vorkommt; normale Gleichheitszeichen (also nicht als Trennungsstrich benutzte G., die im Frakturdruck sehr lang sein können) sollen aber als „=“ erfasst werden.

Fraktur- und Antiquaschrift
Die Benutzung von Fraktur- oder Antiquaschrift, oder der Wechsel zwischen beiden Schriftsystemen, wird in den Texten nicht als solche(r) ausgezeichnet.

III. Allgemeine Textauszeichnung

Die Textauszeichnung soll dem ***TEI Tite***-Schema folgen (siehe http://www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_tite.doc.html), einem speziell für Transkriptionen spezifizierten Schema der *Text Encoding Initiative*. Das ‚Endprodukt‘ der Transkription des Werkes sollte somit in Form von *TEI Tite*-XML vorliegen. Da *TEI Tite* keinen XML-Header erlaubt (<text> ist das Wurzelelement), sind bibliographische o.ä. Metadaten für die Transkription und Textauszeichnung irrelevant. Für die genauen Spezifikationen von *TEI Tite* kann auch dessen Dokumentation konsultiert werden. Für Fragen zur finalen Edition und zur allgemeinen Orientierung bei der Textauszeichnung können auch die Editionsrichtlinien des SvSal-Projekts hilfreich sein, die jedoch nicht als Erfassungsanweisungen (etwa im Hinblick auf Zeichenkodierung) anzusehen sind.

Auszuzeichnende Entitäten und die Arten der Auszeichnung werden im Folgenden genauer erläutert.

1. Typografische Merkmale
Für typografische Besonderheiten stehen folgende Elemente zur Verfügung:

- Fett: …
- Kursiv: …
- Tiefgestellt: _…
- Hochgestellt: […]
- Kapitälchen: <smcap>…</smcap>
- Initialen: <hi rend="init">…</hi>; ist eine Initiale nicht erkennbar, so wird ein leeres <hi rend="init"/> gesetzt
- Sperrsatz (erhöhter Abstand zwischen den Buchstaben): <hi rend="sp">...</hi>
- Recte: innerhalb eines kursiven Textes recte (“gerade” bzw. nicht kursiv) gesetzter Text wird mit <hi rend="recte">...</hi> ausgezeichnet
- größere Lücken im Text innerhalb einer Zeile (nur im Hauptbereich des Textes): <seg rend="gap"/>

2. Textausrichtung und Schriftgröße

Textblöcke, die zentriert oder rechtsbündig stehen UND keine Überschriften oder Verstext sind, sollen mittels <hi rend="center"> bzw. <hi rend="right"> entsprechend ausgezeichnet werden. Um eine Trennung zwischen Textausrichtung/Typografie und „konzeptuellen“ Elementen (z.B. Paragraphen) herzustellen, soll die Textausrichtung stets mittels <hi>-Tags erfasst werden und nicht innerhalb konzeptueller Elemente: also etwa ** <hi rend="right">...</hi> ** (anstatt ...).

Die Schriftgröße (bzw. Unterschiede in derselben im Laufe eines Texts) wird nicht erfasst (abgesehen von dem Sonderfall der Initialen, s.o.).

3. Strukturelle Elemente

a. Zeilenbeginn (und Worttrennungen)

Der Beginn jeder neuen Zeile (im Haupt- und im Marginalbereich des Textes) wird mit <lb/> ausgezeichnet. Beginnen zugleich strukturelle bzw. konzeptuelle Annotationen (; ; <table>, <row>, <cell>; <foreign>; <list>, <head>, <item>, <ref>; <lg>, <l>; <note>; <q>; <title>, <titlePage>, <titlePart>; <unclear>), so soll <lb/> möglichst innerhalb des entsprechenden strukturellen Elements gesetzt werden. Im Falle eines Spaltenformats werden die Zeilenanfänge für jede Spalte ausgezeichnet. Auch im Falle unlesbarer, zeilenübergreifender Textstellen ist die Markierung von Zeilenanfängen per <lb/> durchzuführen, wenn diese erkennbar oder genau abschätzbar sind; im Falle von grafischen Elementen wie Illustrationen oder Ornamenten wird jedoch keine Auszeichnung von (impliziten) Zeilenanfängen durchgeführt.

Geht mit dem Zeilenumbruch eine explizit, d.h. durch Trennungszeichen, markierte Worttrennung einher, so bekommt <lb/> ein Attribut @type mit dem Wert "nb" (für "no break"), also: <lb type="nb"/>; das Trennzeichen

wird in normalisierter Form (d.h. als moderner Trennstrich „-“) direkt (!) vor dem <lb type="nb"/> miterfasst. Ist die Worttrennung nicht explizit durch Trennungszeichen markiert, kann aber aufgrund grammatikalischer oder sonstiger Regeln/Muster abgeleitet werden, so wird der Zeilenbeginn ebenfalls mit <lb type="nb"/>, aber ohne Trennstrich annotiert. Geht mit dem Zeilenbeginn bei einer (impliziten oder expliziten) Worttrennung zugleich ein Spalten- (<cb/>) und/oder Seitenbeginn (<pb.../>) einher, so wird dieser nur im Element für den Zeilenbeginn (<lb.../>) in obiger Form annotiert (und nicht etwa in <cb/> oder <pb/>). Hierbei (wie auch sonst) ist die richtige Reihenfolge der Elemente zu beachten, also: 1.) <pb.../>, 2.) <cb/>, 3.) <lb .../>.

b. Spaltenbeginn und -format

Der Beginn jeder einzelnen Spalte wird durch das Element <cb/> markiert, wobei eine eventuelle Nummerierung der Spalten im (nur in diesem Fall vorhandenen) @n-Attribut des jeweiligen <cb/> angegeben wird. In jedem Fall (außer in Marginalnoten, siehe III.3.f) bekommt <cb/> eine eindeutige ID in @xml:id (deren Wert frei gewählt werden kann). Beispiel:
<cb n="2" xml:id="W0002_cb123"/>

Bei der Änderung des Spaltenformats im Lauf des Textes (auch etwa für spaltenübergreifende Überschriften) wird ein <colShift cols="…"/> gesetzt, wobei @cols die Anzahl der Spalten pro Seite im folgenden Textabschnitt angibt (es hat also mindestens den Wert “1“). Bei mehrspaltigem Format im folgenden Text wird <colShift .../> direkt vor dem ersten <cb/> gesetzt.

c. Seitenbeginn

Der Beginn jeder Seite wird durch <pb n="..." facs="..." xml:id ="..."/> markiert, wobei @n die Seitenzahl (falls es im Werk eine durchgängige Seitenpaginierung gibt) und @facs den Dateinamen des zugehörigen Faksimiles angibt. (Zu Seitenzahlen und deren Normalisierung siehe evtl. auch die besonderen Erfassungsanweisungen für das jeweilige Werk.) Zusätzlich (außer in Marginalnoten, siehe III.3.f) soll jedes <pb/> eine eindeutige ID in @xml:id haben, deren Wert sich z.B. an jenen von @facs anlehnen kann. Beispielsweise:
<pb n="41" facs="W0002-0060" xml:id="W0002-0060"/>

Nicht zu erfassende Seiten (etwa die leeren Seiten zu Beginn und Ende eines Drucks) werden jeweils als <pb/> mit leerem @n repräsentiert, also bspw. <pb n="" facs="W0096-A-0002"/>.

Für die Positionierung von Umbruchselementen (<pb/>, <colShift/>, <cb/>, <lb/>) gilt allgemein folgende Regel:
Fällt ein Umbruch mit dem Beginn „konzeptueller“ Textelemente (z.B. <head>, , <div?>, <table>, <row>, <cell>; <foreign>; <list>, <head>, <item>, <ref>; <lg>, <l>; <note>; <q>; <title>, <titlePage>, <titlePart>; <unclear>) zusammen, dann soll das Umbruchselement (<pb/>,<colShift/>,<cb/>,<lb/>) *als erstes Kindelement (child) des ersten Mixed-Content-Elements* gesetzt werden. Beispiele:

<div1><head><pb/>... , nicht: <pb/> <head>...
<table><row><cell><lb/>..., nicht: <table><row><lb/><cell>...
...</item><item><cb/>..., nicht: ...</item><cb/><item>...
<colShift cols="2"/><cb/><hi rend="init">...

d. Überschriften

Überschriften werden grundsätzlich als ... erfasst, wobei das „?“ in @rend für eine Zahl steht, die die relative Überschriftengröße angibt – analog etwa zur Repräsentation von Überschriften in HTML mit <h1>, <h2> usw. Beispielsweise werden in einem Werk Überschriften mit der größten Schrift als annotiert, die nächstkleineren mit @rend="h2“. Kann keine klare Zuordnung getroffen werden, so wird der Attributwert "hx" vergeben, also ... . Bei werkspezifischen Fragen, was in welcher Weise als Überschrift anzusehen ist, siehe auch die besonderen Erfassungsanweisungen für das jeweilige Werk.

Überschriften in Listen werden innerhalb von <list> ausgezeichnet und als <head>...</head> erfasst.

Überschriften auf Titelseiten werden mit den entsprechenden Tags erfasst (siehe hierzu Punkt j. „Titelseite(n)“).

e. Paragraphen

Die typographischen Absatz-Unterteilungen im Hauptbereich, markiert in der Quelle durch vertikale Abstände, Erstzeileneinzüge, verkürzte (also nicht bis in den Blocksatz ausgetriebene) Zeilenenden o.ä., werden als -Paragraphen ausgezeichnet.

f. Marginalnoten und Querverweise

Falls vorhanden, werden Marginalnoten (Text im Marginalbereich, s.o.) per <note>-Element mit dem Attribut @type="margin" festgehalten. Zusätzlich erhält die Note ein @xml:id. Dabei sind unterschiedliche Fälle möglich: Die Note kann in der Vorlage durch ein Zeichen (z.B. einen hochgestellten Buchstaben oder ein Sternchen) identifiziert werden; dieses kann an der genauen Position, an der die Note verankert ist, stehen, am Beginn der Note

selbst oder an beiden Positionen.

Ist die genaue Position, an der die Note im Hauptbereich des Textes verankert ist, ersichtlich (etwa ein hochgestellter Buchstabe), so wird die Note (vollständig) auch an genau dieser Stelle erfasst und unterbricht somit im TEI-XML den Haupttext. In diesem Fall wird als Wert von @n in <note> das Verankerungszeichen (z.B. ein Buchstabe) festgehalten; das Zeichen selbst (im Text) wird dann durch den <note>-Tag ersetzt (da es in note/@n "aufgelöst" wurde).

Ansonsten – falls es kein referenzierendes Zeichen im Hauptbereich gibt – wird die Note (vollständig) am Ende der Zeile erfasst, auf deren Höhe sie beginnt (oder, falls sie auf der Höhe zwischen zwei Zeilen steht, am Ende der ersten Zeile) und erhält das Attribut @rend="noRef". „Am Ende der Zeile“ heißt hierbei: direkt vor dem <lb/>, mit dem die folgende Zeile beginnt; das gilt auch für Fälle, bei denen mit dem Zeilenumbruch eine Worttrennung einher geht (<lb type="nb"/>, s.o.).

Wenn in der Note ein Seiten- oder Spaltenwechsel erfolgt, verweist dieser mit dem @sameAs-Attribut auf die xml:id des Seitenwechsels im Haupttext (<pb/> bzw. <cb/> wird in diesem Fall also sowohl im Haupttext als auch in der <note> gesetzt). Zum Beispiel (<pb> innerhalb von <note> verweist auf <pb> im Haupttext):
<pb facs="W0066-0005" sameAs="#W0066-0005"/>
Oder (<cb> innerhalb von <note> verweist auf <cb> im Haupttext):
<cb sameAs="#W0066-cb123"/>
Es ist zu beachten, dass <cb/> bzw. <pb/> hier, im Gegensatz zu <pb/> / <cb/> im Haupttext, kein Attribut @xml:id haben.

(Allgemeiner Hinweis: Wenn Text im Marginalbereich nur aus einer Ziffer oder einzelnen Symbolen besteht, handelt es sich um Absatz- oder Artikelnummern, die nicht als Marginalnoten, sondern mittels spezifischer Elemente (z.B. <milestone…/>) aufgenommen werden; siehe hierzu evtl. die besonderen Erfassungsanweisungen für das jeweilige Werk.)

g. Verstext

Verstext (meist an Einrückung und Linksbündigkeit zu erkennen) wird insgesamt in einem <lg> (pro Verstext-Block, also z.B. pro Gedicht oder Verszitat) erfasst, mögliche Strophen (falls identifizierbar) werden als einzelne <lg> (unterhalb des Gesamt-<lg> für den Verstext-Block) und die einzelnen Verszeilen jeweils als <l> annotiert.
Dabei gilt grundsätzlich jede typografische Zeile auch als Verszeile, es sei denn, die folgende Zeile ist gegenüber der aktuellen Zeile eingerückt; in letzterem Fall handelt es sich um einen Zeilenumbruch (<lb/>) innerhalb der Verszeile, also z.B.:

<l>[erster Teil der Verszeile]<lb/>[zweiter Teil der Verszeile]</l>

Die Textausrichtung ist bei Verstext nicht gesondert zu annotieren.

h. Text in Anführungsstrichen

Mit Anführungsstrichen markierter Text wird entweder mit <q>…</q> annotiert, falls Anfang und Ende des Textes eindeutig bestimmt werden können. Ist dies nicht der Fall, so werden am Ort der Anführungsstriche leere <q/> gesetzt. Im Original vorhandene Anführungsstriche entfallen dabei. In jedem Fall werden also öffnende Anführungsstriche durch <q> ersetzt (wenn es zugehörige schließende Anführungsstriche gibt); schließende Anführungsstriche werden durch </q> ersetzt (wenn es zugehörige öffnende Anführungsstriche gibt); und alle anderen Anführungsstriche werden durch <q/> ersetzt.

i. Ornamente, Grafiken und Illustrationen

Ornamente (d.h. als strukturelle Marker fungierende grafische Elemente, wie z.B. langgezogene Striche) werden jeweils durch ein leeres <ornament/> erfasst. Alle weiteren grafischen Elemente werden nicht transkribiert, sondern lediglich jeweils durch ein leeres <figure/>-Element markiert. Das Setzen von (impliziten) Zeilenanfängen als <lb .../> entfällt dabei.

j. Titelseite(n)

Spezielle Textelemente der Titelseiten können mit den dafür in TEI Tite zur Verfügung stehenden Elementen (<titlePage>, <titlePart>, <byline>, <docAuthor> etc.) ausgezeichnet werden. Ansonsten können entsprechende Elemente der Einfachheit halber auch mit den "üblichen" Elementen (v.a. <p> und <p rend="hx">) ausgezeichnet werden.

k. Grobstrukturierung

Die Titelseite(n) eines Werkes wird/werden insgesamt mit <front>...</front> umschlossen; der gesamte darauf folgende Text (bis zum Ende des Buches) wird der Einfachheit halber insgesamt mit <body><div1>...</div1></body> umschlossen. (D.h. es darf im Werk genau ein <body>-Tag und ein <div1>-Tag, in direkter Aufeinanderfolge, geben. Die genauere Unterteilung (auch mittels <back> und strukturelle Tiefenannotation mittels erfolgt in der späteren Redaktionsarbeit.)

l. Mehrbändige Werke

Bei mehrbändigen Werken wird jeder Band einzeln (d.h. wie eine Monographie) erfasst.

m. Textstellen mit nicht-lateinischen Schriftzeichen

Bei Textpassagen in anderen Sprachen als der Hauptsprache des Werkes werden diese Passagen mit einem <foreign xml:lang="...">...</foreign>

markiert, wobei der Wert von @xml:lang eine entsprechende Abkürzung der Sprache ist. Dies ist obligatorisch für Textpassagen in nicht-lateinischer Schrift, v.a. griechisch (<foreign xml:lang="grc">) und hebräisch (<foreign xml:lang="heb">), durchzuführen und für Passagen in anderen Sprachen (sofern diese ohne Weiteres identifiziert werden können) optional. Bei sehr häufigen und kleinteiligen fremdsprachlichen Textpassagen innerhalb eines Werkes ist Rücksprache mit dem Projekt über die genaue Art der Annotation zu halten.

4. Umgang mit nicht oder schwer lesbaren Stellen sowie mit Zweifelsfällen

a. Nicht oder schwer lesbare Stellen

Bei völligem Textverlust – unabhängig von der Ursache – wird der genaue Ort bzw. Bereich des Verlustes mit einem (möglicherweise leeren) <unclear>-Element markiert. Alle gesetzten <unclear>-Tags sollen grundsätzlich eine Verantwortlichkeitsangabe mittels @resp mit dem Wert "#TL" erhalten, also: <unclear resp="#TL"...>...

Bei zeilen-, spalten- oder seitenübergreifendem Textverlust muss die Strukturauszeichnung unbedingt beibehalten werden, indem innerhalb von <unclear> die entsprechenden Tags gesetzt werden. Ein Textverlust von genau drei Zeilen wäre etwa folgendermaßen auszuzeichnen: <unclear resp="#TL"><lb/><lb/><lb/></unclear>

Ist der Text noch erkennbar, aber nicht eindeutig transkribierbar, so wird er ebenfalls mit einem <unclear>-Element ausgezeichnet, wobei innerhalb des Elements der transkribierte Text steht. Auch in diesem Fall ist die Zeilenzählung per <lb/> selbstverständlich weiterzuführen.

b. Zweifelsfälle

Ergeben sich bei der Transkription bzw. Textauszeichnung Zweifelsfälle, etwa wenn ein Textphänomen nicht durch die hier beschriebenen Tags erfasst werden kann oder sich Fragen hinsichtlich der „richtigen“ Annotation ergeben, so sind sie mit einem leeren <gap resp="#TL"/> (auch wenn semantisch nicht unbedingt ganz passend) als Hinweis für die spätere Redaktionsarbeit zu markieren. Bei durchgängig oder häufig in einem Werk auftauchenden Zweifelsfällen ist das Projekt hinsichtlich der richtigen Annotationsweise zu kontaktieren.