Umwandlung proprietären Markups in TEI-konformes XML mithilfe von TUSTEP.

Ein Werkstattbericht aus dem *Parzival*-Projekt Bern

1. Textmengen im *Parzival*-Projekt

· 16 Handschriften (mehr oder weniger vollständig)
· 1 Frühdruck
· 70 Fragmente
Die Überlieferung beträgt ca. 500.000 Verse mit mehr als 2 Millionen Wörtern. Der edierte Text besteht aus 4 synoptisch dargestellten Fassungen (*D, *m, *G, *T), die zusammen ca. 100.000 Verse mit etwa 450.000 Wörtern ausmachen. Gesamtumfang des Projektmaterials: ca. **2,5 Millionen** Wörter.

2. Proprietäres Markup

Gründe für die Entwicklung eines proprietären Markups:
· Zu Beginn des Projekts (um 2000) hatte XML seinen Siegeszug noch nicht angetreten.
· Die Text Encoding Initiative (TEI) war noch kein fächerübergreifender Standard.
· Es bestanden spezielle Bedürfnisse des Projekts, die nicht über die TEI-Richtlinien abgedeckt waren (z.B. Auszeichnung von Initialen).

Daher wurde ein eigenes, an SGML/XML angelehntes und auf Tags basierendes Markup entwickelt, welches alle Anforderungen des Projekts erfüllte. Nun besteht die Aufgabe darin, alle Daten ohne grösseren Aufwand in ein standardisiertes Markup zu übertragen.
· Die projektinternen Tags werden im Gegensatz zu XML-Tags in eckige Klammern eingeschlossen. Bsp: **[sic]...[/sic]**
· Milestones (**[z]** = Zeilenumbruch) erhalten keinen abschliessenden Schrägstrich.
· Statt Attributen können zusätzliche Informationen direkt im Starttag kodiert werden. Bsp.: **[3Lrinit]** = 3-zeilige Lombarde in roter Farbe, **[uhvh]...[/hvh]** = Hervorhebung durch Unterstreichung.
· Nur die später hinzugefügten Tags für Eigennamen entsprechen den Attributregeln von XML. Bsp.: **[np Name="Parzival"]Parcifales[/np]**

3. Vollautomatische Konvertierung

Umwandlung der Tags:
Tags, die sich nur durch die Klammern und den Tagnamen von ihren Entsprechungen in TEI unterscheiden, können durch einfache Austauschanweisungen umgewandelt werden, dem ein #KOPIERE-Programm zugrunde liegt, das mit dem Parameter XX gesteuert wird:

```
          * Unsicherheiten, Lücken, Bogenhäufungen
XX   50   |\[zw\]|<unclear reason="Zweifel">|
XX   50   |\[/zw\]|</unclear>|
XX   51   |\[unl\]{1-3}:\[/unl\]|<gap reason="unleserlich"/>|
XX   52   |\[frag\]{1-3}:\[/frag\]|<gap reason="Fragmentverlust"/>|
XX   53   |\[sic\]|<sic>|
XX   53   |\[/sic\]|</sic>|
XX   54   |\[ct\]|<unclear reason="ct">|
XX   54   |\[/ct\]|</unclear>|
XX   54   |\[sf\]|<unclear reason="sf">|
XX   54   |\[/sf\]|</unclear>|
XX   55   |\[b{#}\]|<supplied reason="Bogenhäufung" extent="{+3=}" unit="Schaft">|
XX   55   |\[/b{#}\]|</supplied>|
XX   56   |\[a\]|<unclear reason="Fragliche Abkürzung">|
XX   56   |\[/a\]|</unclear>|
```

Tags wie die Initialen, die Zusatzinformationen enthalten, können mithilfe der Mustererkennung von TUSTEP in wenigen Schritten weiter verarbeitet werden. Häufigkeitsangaben, Zeichengruppen und Verweise helfen, die Anzahl der Austauschregeln gering zu halten.

```
          * Initialen
XX   20   |\[{#}xinit\]|<hi rend="{+2=}-zeilige nicht ausgeführte Initiale">|
XX   20   |\[minit\]|<hi rend="Majuskel">|
XX   20   |\[vinit\]|<hi rend="Versal">|
XX   20   |\[{#}L{0-3}{&a}init\]|<hi rend="{+2=}-zeilige Lombarde ({+4=})">|
XX   20   |\[{#}F{0-3}{&a}init\]|<hi rend="{+2=}-zeiliges Fleuronee ({+4=})">|
XX   20   |\[{#}Pinit\]|<hi rend="{+2=}-zeilige Prachtinitiale">|
XX   20   |\[/init\]|</hi>|
```

Umwandlung der Sonderzeichen:
Die Konvertierung der Sonderzeichen erfolgt ebenfalls mithilfe von Austauschoperationen. Die besonderen Unicode-Zeichen werden in der Ziel-Datei »hart« kodiert, d.h. der hexadezimale Unicode wird explizit angegeben, z.B.: ſ = 017F. Dadurch wird sichergestellt, dass die Sonderzeichen inklusive Diakritika und Superskripte in der XML-Datei korrekt angezeigt werden, auch wenn dem Benutzer die Schriftarten hierzu fehlen.

```
          * Umwandlung der Sonderzeichen
          * Einzelne Sonderzeichen (Schaft-s, Ligaturen)
XX   90   |#.s|&#x017F;|
XX   90   |#.\ä|&#x00E6;|
XX   90   |#.\Ä|&#x00C6;|
XX   90   |#.\ö|&#x0153;|
XX   90   |#.\Ö|&#x0152;|

          * Diakritika auf Sonderzeichen
XX   91   |%/&#x{4}?;|{+3-7=}&#x0301;|
XX   91   |%<&#x{4}?;|{+3-7=}&#x0302;|
XX   91   |%.&#x{4}?;|{+3-7=}&#x0307;|
XX   91   |%:&#x{4}?;|{+3-7=}&#x0308;|
XX   91   |%"&#x{4}?;|{+3-7=}&#x030B;|
XX   91   |%=&#x{4}?;|{+3-7=}&#x036F;|
XX   91   |%\{&#x{4}?;|{+3-7=}&#x0351;|

          * Diakritika auf sonstigen Buchstaben
XX   91   |%/{&a}|{+3=}&#x0301;|
XX   91   |%<{&a}|{+3=}&#x0302;|
XX   91   |%.{&a}|{+3=}&#x0307;|
XX   91   |%:{&a}|{+3=}&#x0308;|
XX   91   |%"{&a}|{+3=}&#x030B;|
XX   91   |%={&a}|{+3=}&#x036F;|
XX   91   |%\{{&a}|{+3=}&#x0351;|

          * Superskripte auf Sonderzeichen
XX   92   |#;e&#x{4}?;|{+4-8=}&#x0364;|
XX   92   |#;o&#x{4}?;|{+4-8=}&#x0366;|
XX   92   |#;u&#x{4}?;|{+4-8=}&#x0367;|
XX   92   |#;v&#x{4}?;|{+4-8=}&#x036E;|

          * Superskripte auf sonstigen Buchstaben
XX   92   |#;e{&a}|{+4=}&#x0364;|
XX   92   |#;o{&a}|{+4=}&#x0366;|
XX   92   |#;u{&a}|{+4=}&#x0367;|
XX   92   |#;v{&a}|{+4=}&#x036E;|
```

Umwandlung von Spaltenumbrüchen:
Ein besonderes Problem bereiten Umbrüche innerhalb eines Verses. Bei einem einfachen Austausch würde der zweite Bestandteil des Verses »in der Luft hängen«:

```
$<L xxx.xx>_Lorem ipsum dolor
$|F 5rb|
$sit amet
```

```
<l xml:id="...">Lorem ipsum dolor</l>
<pb xml:id="..."/>
sit amet
```

Um dieses Problem zu lösen, braucht es eines höher entwickelten Programms (Tuscript), welches Schleifen (LOOP) und IF-Abfragen unterstützt sowie Variablen und Funktionen bereitstellt (s. S. 3). Dadurch können die isolierten Bestandteile gefunden und korrekt in den Vers eingeschlossen werden.

5. Literatur

Viehhauser, Gabriel: Standardisierung und proprietäre Annotation im Berner *Parzival*-Projekt. In: Jahrbuch für Computerphilologie. Online auf:
http://computerphilologie.digital-humanities.de/jg09/viehhauser.html [25.09.2013]
Weitere Publikationen finden Sie online auf:
http://www.parzival.unibe.ch/projektpraesentationen.html

6. Kontakt

Christian Griesinger M.A.
Institut für Germanistik, Universität Bern
Länggassstrasse 49, D103
CH–3012 Bern, Schweiz
Tel.: +41 (0)31 631 34 65
Mail: Christian.Griesinger@germ.unibe.ch

Tuscript-Lösung zum Korrigieren von Spaltenumbrüchen innerhalb von Versen:

```
#- Zwischenschritt: Spaltenumbrüche innerhalb von Versen lösen
#MAKRO
$$ MODE TUSCRIPT
SET quelle = "tmp1"
SET status = OPEN (quelle, WRITE, -STD-)
SET ziel = "tmp3"
SET status = CREATE (ziel, SEQ-O, -STD-)
SET status = ERASE (ziel)
SET i = 1
ACCESS q: READ/RECORDS "{quelle}" sn.zn/un, text
ACCESS z: WRITE/ERASE/RECORDS "{ziel}" sn.zn/un, text
LOOP/999999
 READ/NEXT/EXIT q
 SELECT i
 CASE 1
   SET a = text
 CASE 2
   SET b = text
 CASE 3
   SET c = text
   IF ("{c}" .SW. "<l" .OR.
       "{c}" .SW. "<note" .OR.
       "{c}" .SW. "<cb" .OR.
       "{c}" .SW. "<pb" .OR.
       "{c}" .SW. "<milestone") THEN
   - tue nichts weiter
   ELSE
     a = EXTRACT (a, 0, -4)
     c = APPEND (c, "", "</l>")
   ENDIF
   text = a
   WRITE/ADJUST z
 DEFAULT
   SET a = b
   SET b = c
   SET c = text
   IF ("{c}" .SW. "<l" .OR.
       "{c}" .SW. "<note" .OR.
       "{c}" .SW. "<cb" .OR.
       "{c}" .SW. "<pb" .OR.
       "{c}" .SW. "<milestone") THEN
   - tue nichts weiter
   ELSE
     a = EXTRACT (a, 0, -4)
     c = APPEND (c, "", "</l>")
   ENDIF
   text = a
   WRITE z
 ENDSELECT
 SET i = i + 1
ENDLOOP
text = b
WRITE/ADJUST z
text = c
WRITE/ADJUST z
ENDACCESS/PRINT q
ENDACCESS/PRINT z
*eof
```

Beispieldreissiger vor der Konvertierung:

```
$<L 734.01>_[6Pinit]U[/init]il livte de#.s hat v[7](er)[/7]drozzen.
$<L 734.02>_den diz mære wa#.s vor be=[z]#.slozzen.
$<L 734.03>_[minit]G[/init]n#;ovge chvndenz [z] nie ervarn.
$<L 734.04>_[minit]N[/init]v wil ich daz [z] niht langer #.sparn.
$<L 734.05>_ich t#;ovnz iv chvnt mit [z] rehter #.sage.
$<L 734.06>_[minit]W[/init]and[03](e)[/03] ich in dem mvnde [z] trage.
$<L 734.07>_[minit]D[/init]az #.sloz dirre Aventivre.
$<L 734.08>_wi der #.s#;evze vnt der gehivre.
$<L 734.09>_[np Name="Anfortas"]Anfortas[/np] wart wol ge#.svnt.
$<L 734.10>_vn#.s t#;ovt div aventivre chvnt.
$<L 734.11>_wi von [no Ortsname="Pelrapeire"]Pelrapeire[/no] div kvnegin.
$<L 734.12>_ir chiv#.scen wiplichen #.sin.
$<L 734.13>_behielt vnz an ir lon#.s #.stat.
$<L 734.14>_da #.si in hohe #.sælde trat.
$<L 734.15>_[np Name="Parzival"]Parcifal[/np] daz wirbet.
$<L 734.16>_ob min chvn#.st niht verdirbet.
$<L 734.17>_ich #.sage alre#.st #.sin arbeit.
$<L 734.18>_#.swaz #.sin hant ie ge#.streit.
$<L 734.19>_daz wa#.s mit kinden her getan.
$<L 734.20>_mohte ich di#.s#.s mæres wandel han.
$<L 734.21>_vngern wolt ich in wagen.
$<L 734.22>_des chvnde #;voch mich betragen.
$<L 734.23>_[minit]N[/init]v bevilh ich #.sin gelvcke.
$<L 734.24>_#.sime hercen der #.sælden #.stvche.
$<L 734.25>_da div vrævel bi der chiv#.sce lach.
$<L 734.26>_wandez nie zagheit gepflach.
$<L 734.27>_daz m#;evze im ve#.stenvnge gebn.
$<L 734.28>_daz er behalde nv #.sin lebn.
$<L 734.29>_#.sit ez #.sich hat an den gezogt.
$<L 734.30>_in be#.stet ob allem #.strîte ein vogt.
```

Beispieldreissiger nach der Konvertierung:

```
<l xml:id="D_734.01"><hi rend="6-zeilige Prachtinitiale">U</hi>il livte de&#x017F; hat
  v<choice><am><g ref="#a7"/></am><ex>er</ex></choice>drozzen.</l>
<l xml:id="D_734.02">den diz m&#x00E6;re wa&#x017F; vor be=<lb/>&#x017F;lozzen.</l>
<l xml:id="D_734.03"><hi rend="Majuskel">G</hi>nv&#x0366;ge chvndenz <lb/> nie ervarn.</l>
<l xml:id="D_734.04"><hi rend="Majuskel">N</hi>v wil ich daz <lb/> niht langer &#x017F;parn.</l>
<l xml:id="D_734.05">ich tv&#x0366;nz iv chvnt mit <lb/> rehter &#x017F;age.</l>
<l xml:id="D_734.06"><hi rend="Majuskel">W</hi>and<choice><am><g ref="#a03"/></am><ex>e</ex></choice>
  ich in dem mvnde <lb/> trage.</l>
<l xml:id="D_734.07"><hi rend="Majuskel">D</hi>az &#x017F;loz dirre Aventivre.</l>
<l xml:id="D_734.08">wi der &#x017F;v&#x0364;ze vnt der gehivre.</l>
<l xml:id="D_734.09"><name type="Person" ref="regp:Anfortas">Anfortas</name> wart wol ge&#x017F;vnt.</l>
<l xml:id="D_734.10">vn&#x017F; tv&#x0366;t div aventivre chvnt.</l>
<l xml:id="D_734.11">wi von <name type="Ort" ref="rego:Pelrapeire">Pelrapeire</name> div kvnegin.</l>
<l xml:id="D_734.12">ir chiv&#x017F;cen wiplichen &#x017F;in.</l>
<l xml:id="D_734.13">behielt vnz an ir lon&#x017F; &#x017F;tat.</l>
<l xml:id="D_734.14">da &#x017F;i in hohe &#x017F;&#x00E6;lde trat.</l>
<l xml:id="D_734.15"><name type="Person" ref="regp:Parzival">Parcifal</name> daz wirbet.</l>
<l xml:id="D_734.16">ob min chvn&#x017F;t niht verdirbet.</l>
<l xml:id="D_734.17">ich &#x017F;age alre&#x017F;t &#x017F;in arbeit.</l>
<l xml:id="D_734.18">&#x017F;waz &#x017F;in hant ie ge&#x017F;treit.</l>
<l xml:id="D_734.19">daz wa&#x017F; mit kinden her getan.</l>
<l xml:id="D_734.20">mohte ich di&#x017F;&#x017F; m&#x00E6;res wandel han.</l>
<l xml:id="D_734.21">vngern wolt ich in wagen.</l>
<l xml:id="D_734.22">des chvnde o&#x036E;ch mich betragen.</l>
<l xml:id="D_734.23"><hi rend="Majuskel">N</hi>v bevilh ich &#x017F;in gelvcke.</l>
<l xml:id="D_734.24">&#x017F;ime hercen der &#x017F;&#x00E6;lden &#x017F;tvche.</l>
<l xml:id="D_734.25">da div vr&#x00E6;vel bi der chiv&#x017F;ce lach.</l>
<l xml:id="D_734.26">wandez nie zagheit gepflach.</l>
<l xml:id="D_734.27">daz mv&#x0364;ze im ve&#x017F;tenvnge gebn.</l>
<l xml:id="D_734.28">daz er behalde nv &#x017F;in lebn.</l>
<l xml:id="D_734.29">&#x017F;it ez &#x017F;ich hat an den gezogt.</l>
<l xml:id="D_734.30">in be&#x017F;tet ob allem &#x017F;tri&#x0302;te ein vogt.</l>
```