SGML/XML-Tags als Makros für das TUSTEP-Satzprogramm

Als erster kleiner Schritt zur leichteren Verarbeitung von SGML-konform ausgezeichneten Texten mit TUSTEP wurden in der Version November 1996 die sogenannten Spitzklammermakros im Satzprogramm eingeführt (vgl. BI 96/7+8, S. 8–9).

Mit der nächsten TUSTEP-Version (November 1997) wird ein weiterer wichtiger Schritt zur direkten Weiterverarbeitung von SGML/XML-Dateien bzw. zu deren Formatierung mit Hilfe des TUSTEP-Satzprogramms getan.

A) Programm SATZ

1. Satzmakros können jetzt zwischen den spitzen Klammern auch Blanks enthalten. Damit ist die Möglichkeit gegeben, auch SGML-Tags mit Attributen direkt als Makros für das Satzprogramm zu benutzen.

2. Bei Satzmakros, die den Start- bzw. Ende-Tags von Elementen eines nach SGML/XML kodierten Textes entsprechen, kann angegeben werden, daß ihre Stellung in der Hierarchie der Elemente bei der Auswertung berücksichtigt werden soll. Dies geschieht dadurch, daß
– dem Satzprogramm die hierarchische Struktur des Textes anhand von Makros, die den im Text vorkommenden Tags entsprechen und deren Reihenfolge mitgeteilt wird, und
– ein und demselben Makro, das mehrfach – an jeweils unterschiedlichen Stellen in der Hierarchie der Elemente – vorkommen kann, je nach seiner Stellung in dieser Hierarchie unterschiedliche Satz-Steueranweisungen zugeordnet werden. (Beispiel: Eine Zwischenüberschrift kann im Vorwort eines Buches anders dargestellt werden als im Hauptteil, und dort wieder anders als im Register, auch wenn an allen drei Stellen das selbe Tag zur Markierung der Zwischenüberschrift benutzt ist.)

Die hierarchische Struktur des Textes wird dem Satzprogramm also nicht über die zugehörige DTD (Document Type Definition) mitgeteilt, sondern über die Reihenfolge der Makros, die den Start- und Ende-Tags der Elemente entsprechen. Die Start- und Ende-Tags eines Elements, das innerhalb eines übergeordneten Elements vorkommt, stehen dabei zwischen dem Start- und dem Ende-Tag dieses übergeordneten Elements. Dies setzt gleichzeitig voraus, daß in dem zu setzenden Text keine Tag-Minimierung vorgenommen wurde: Wie in XML-kodierten Texten müssen zu allen nicht-leeren Elementen Ende-Tags im Text vorhanden sein.

Die hierarchische Struktur eines Textes bzw. die möglichen Verschachtelungen seiner Elemente können im Prinzip aus der DTD abgelesen werden; danach könnten die Satzmakros angegeben und aufgelöst werden.

Diese hierarchische Struktur kann jedoch, wenn alle Start- und Ende-Tags vorhanden sind (also keine Minimierung von Start- und/oder Ende-Tags vorgenommen wurde), auch aus dem Text selbst abgelesen werden. Neben der Tatsache, daß die DTD nicht vorliegen muß, bietet dieses Verfahren eine einfache Möglichkeit, beim Zusammenstellen der Makros nur die im Text tatsächlich vorkommenden Tags berücksichtigen zu müssen (dies kann deutlich weniger sein als das, was aufgrund der DTD möglich ist).

B) Standardmakro *TAGS

Mit dem Makro *TAGS kann eine Liste der in einer Text-Datei vorkommenden SGML- bzw. XML-Tags unter Berücksichtigung ihrer hierarchischen Ordnung erzeugt und in Parameter – MAC für »einfache Makros« bzw. MAH für »hierarchische Makros« – für das Satzprogramm verwandelt werden. Tags, die nicht paarweise vorkommen (Tags für leere Elemente, »milestones«) werden dabei zu »einfachen Makros«, paarweise vorkommende Tags (für Elemente mit Start- und Ende-Tags) zu »hierarchischen Makros«. In die so erzeugten Parameter MAC bzw. MAH müssen nur noch die für den Satz erforderlichen Steueranweisungen eingetragen werden.

Um die Übersicht über die Dokumentstruktur und die Zuordnung der typographischen Steueranweisungen zu den einzelnen Tags an den verschiedenen Stellen dieser Struktur zu erleichtern, können mit dem Makro *TAGS weitere Listen erzeugt werden:

a) eine Liste der im Text vorkommenden hierarchischen Tags, in der – im Unterschied zu den daraus erzeugten Makros – anstelle der Punkte für übergeordnete Hierarchiestufen die entsprechenden Tags ausgeschrieben sind (vgl. unten Liste 2),
b) eine Liste, die zu den in Liste a) enthaltenen Tags alle Stellen mit Seiten- und Zeilennummer aufführt, an denen diese Tags im Text vorkommen (vgl. unten Liste 3),
c) eine Liste, in der die in Liste a) enthaltenen

Tags (einschließlich der übergeordneten Tags) in der Textreihenfolge aufgeführt sind, je eine Zeile pro Tag (vgl. unten Liste 4).

Außerdem wird geprüft, ob die durch bereits verwendete Tags festgelegte Struktur auch weiterhin eingehalten wird, insbesondere ob zu jedem Start-Tag ein Ende-Tag vorhanden ist und ob die Verschachtelung der Elemente stimmt.

Ein Beispiel soll dies erläutern. Dem Beispiel liegt der nach TEI-Lite kodierte Text von Conan Doyle's »The Hound of the Baskervilles« zugrunde, der im Oxford Text Archive (http://sable.ox.ac.uk/ota/) zugänglich ist. Dieser Text hat für unseren Zweck den Vorteil, daß seiner Kodierung nach TEI-Lite eine relativ einfache Struktur zugrundegelegt wurde. Es sind nur insgesamt 10 verschiedene Elemente markiert (text, front, body, div, head, letter, p, quote, signed, s), die (weil zu »body« fälschlicherweise das Ende-Tag fehlt) zu insgesamt nur 31 Makros werden. Der Aufruf

```
#*TAGS, QUELLE=hound, ZIEL=hound.mac,
PROTOKOLL=hound.prt, REIHENFOLGE=hound.tat,
HIERARCHIE=hound.tag
```

produziert zu diesem Text folgende Listen:

1. Makros für #SATZ (Datei hound.mac)

```
MAC     <body>
        Tags mit Attributen:
        <div type=chap n="hound.01">
        <div type=chap n="Hound.02">
        <div type=chap n="Hound.03">
        <div type=chap n="Hound.04">
        <div type=chap n="Hound.05">
        <div type=chap n="Hound.06">
        <div type=chap n="Hound.07">
        <div type=chap n="Hound.08">
        <div type=chap n="Hound.09">
        <div type=chap n="Hound.10">
        <div type=chap n="Hound.11">
        <div type=chap n="Hound.12">
        <div type=chap n="Hound.14">
        <div type=chap n="Hound.15">
        <div type=chap n="Hound1.3">
        <text id=HOUND n='Hound of the
        Baskervilles'>
        hierarchische Tags:
MAH     <text >         [1]
MAH     . <front>       [2]
MAH     . . <head>      [3]
MAH     . . </head>     [4]
MAH     . </front>      [5]
MAH     . <div >        [6]
MAH     . . <head>      [7]
MAH     . . </head>     [8]
MAH     . . <p>         [9]
MAH     . . . <s>       [10]
MAH     . . . </s>      [11]
MAH     . . </p>        [12]
MAH     . . <quote>     [13]
MAH     . . . <p>       [14]
MAH     . . . . <s>     [15]
MAH     . . . . </s>    [16]
MAH     . . . </p>      [17]
MAH     . . </quote>    [18]
MAH     . . <letter>    [20]
MAH     . . . <p>       [21]
MAH     . . . . <s>     [22]
MAH     . . . . </s>    [23]
MAH     . . . </p>      [24]
MAH     . . . <signed>  [25]
MAH     . . . . <s>     [26]
MAH     . . . . </s>    [27]
MAH     . . . </signed> [28]
MAH     . . </letter>   [29]
MAH     . </div>        [19]
MAH     </text>         [30]
```

Da in der Datei das Ende-Tag zu <body> fehlt, ist das Makro <body> nicht in die Hierarchie der Tags (die in den mit MAH beginnenden Zeilen sichtbar wird) eingegliedert, sondern zum Parameter MAC geworden.

Bei den mit MAH beginnenden Zeilen ist an der Zahl der Punkte vor der geöffneten spitzen Klammer abzulesen, wie viele übergeordnete Elemente zu dem Element mit dem betreffenden Start-Tag gehören. Das jeweils nächste übergeordnete Element ist dasjenige, vor dessen Start-Tag in den vorher aufgeführten Zeilen genau ein Punkt weniger steht.

Die hier in eckige Klammern eingeschlossenen Zahlen stehen in der Datei als Kommentar in eigenen Zeilen. Diese Zahlen finden sich auch in den übrigen von *TAGS erzeugten Listen an den Stellen, an denen das entsprechende Tag vorkommt.

2. Liste der Tags (Datei hound.tag)

In dieser Liste sind alle Tags einschließlich der jeweils übergeordneten (ausgeschriebenen) Start-Tags enthalten. Die Nummer in eckigen Klammern gibt an, als wievieltes ein Tag zum ersten Mal im Text vorkommt.

```
<text > [1]
<text ><front> [2]
<text ><front><head> [3]
<text ><front><head></head> [4]
<text ><front></front> [5]
<text ><div > [6]
<text ><div ><head> [7]
<text ><div ><head></head> [8]
```

```
<text ><div ><p> [9]
<text ><div ><p><s> [10]
<text ><div ><p><s></s> [11]
<text ><div ><p></p> [12]
<text ><div ><quote> [13]
<text ><div ><quote><p> [14]
<text ><div ><quote><p><s> [15]
<text ><div ><quote><p><s></s> [16]
<text ><div ><quote><p></p> [17]
<text ><div ><quote></quote> [18]
<text ><div ><letter> [20]
<text ><div ><letter><p> [21]
<text ><div ><letter><p><s> [22]
<text ><div ><letter><p><s></s> [23]
<text ><div ><letter><p></p> [24]
<text ><div ><letter><signed> [25]
<text ><div ><letter><signed><s> [26]
<text ><div ><letter><signed><s></s> [27]
<text ><div ><letter><signed></signed> [28]
<text ><div ><letter></letter> [29]
<text ><div ></div> [19]
<text ></text> [30]
```

3. Tags mit Stellenangaben (Datei hound.prt)

Von diesem Protokoll ist nur der Anfang wiedergegeben. Hinter jedem Tag steht wie in den übrigen Listen in eckigen Klammern die Nummer, die angibt, als wievieltes das entsprechende Tag zum ersten Mal im Text vorkommt. Dahinter sind die Textstellen (Seiten- und Zeilennummer) aufgeführt, an denen das Tag im Text vorkommt. Die Reihenfolge der Tags richtet sich nach ihrem ersten Vorkommen im Text.

```
<text > [1] 1.1
<text ><front> [2] 1.1
<text ><front><head> [3] 1.2
<text ><front><head></head> [4] 1.2
<text ><front></front> [5] 1.2
<text ><div > [6] 1.3 254 636 978 2.432 825
      3.174 688 942 4.567 935 5.400 854 6.296
      712
<text ><div ><head> [7] 1.4 255 637 979
      2.433 826 3.175 689 943 4.568 936 5.401
      855 6.297 713
<text ><div ><head></head> [8] 1.4 255 637
      979 2.433 826 3.175 689 943 4.568 936
      5.401 855 6.297 713
<text ><div ><p> [9] 1.5 17 18 20 22 28 32 33
      36 41 46 55 65 69 72 78 80 86 90 93 95 97
      107 120 123 141 151 158 161
```

4. Tags in der Textreihenfolge (Datei hound.tat)

Von dieser Liste, die für den Beispieltext insgesamt 10660 Zeilen umfaßt, sind nur der Anfang und das Ende wiedergegeben.

```
<text > [1] 1.1
<text ><front> [2] 1.1
<text ><front><head> [3] 1.2
<text ><front><head></head> [4] 1.2
<text ><front></front> [5] 1.2
<text ><div > [6] 1.3
<text ><div ><head> [7] 1.4
<text ><div ><head></head> [8] 1.4
<text ><div ><p> [9] 1.5
<text ><div ><p><s> [10] 1.5
<text ><div ><p><s></s> [11] 1.7
<text ><div ><p><s> [10] 1.7
<text ><div ><p><s></s> [11] 1.9
<text ><div ><p><s> [10] 1.9
<text ><div ><p><s></s> [11] 1.11
. . . . . .
<text ><div ><p><s> [10] 7.79
<text ><div ><p><s></s> [11] 7.81
<text ><div ><p></p> [12] 7.81
<text ><div ></div> [19] 7.82
<text ></text> [30] 7.82
```

In der Datei hound.mac (siehe oben Liste 1) müssen hinter den Tags die Steueranweisungen für den Satz ergänzt werden. Für den Beispieltext empfehlen sich folgende Parameter (die Kommentarzeilen sind hier weggelassen):

```
MAC     <body> \
MAH     <text >
MAH     . <front>
MAH     . . <head> &&#f+
MAH     . . </head>#f–&&{
MAH     . </front>
MAH     . <div >
MAH     . . <head> &&
MAH     . . </head>&&{
MAH     . . <p> $
MAH     . . . <s>
MAH     . . . </s>
MAH     . . </p>
MAH     . . <quote> $$
MAH     . . . <p> $
MAH     . . . . <s>
MAH     . . . . </s>
MAH     . . . </p>
MAH     . . </quote>$${
MAH     . . <letter> $$#/+
MAH     . . . <p> $$$0$$$#/+
MAH     . . . . <s>
MAH     . . . . </s>
MAH     . . . </p>#/–
MAH     . . . <signed> $$$@–0#/+
MAH     . . . . <s>
MAH     . . . . </s>
MAH     . . . </signed>#/–
```

```
MAH     . . </letter>$${
MAH     . </div>
MAH     </text> $$$=$$$@–zTHE END
```

Einige Makros kommen in diesem Beispiel mehrfach in jeweils unterschiedlichem Kontext vor. Von diesen werden je nach Kontext die Makros <head> und </head> sowie und typographisch jeweils unterschiedlich behandelt: die zwischen <front> und </front> stehende Überschrift wird fett gedruckt, die im Text stehenden Überschriften normal. Ebenso wird der Text der einzelnen Abschnitte zwischen <letter> und </letter> im Unterschied zu dem der übrigen Abschnitte kursiv gedruckt.

Noch nicht ausreichend berücksichtigt sind in diesem Beispiel Tags, die Attribute enthalten. Sie sind in der Makro-Liste daran erkennbar, daß als letztes Zeichen vor der geschlossenen spitzen Klammer ein Blank steht (dies ist bei den Makros <text > und der Fall). Diese Makros gelten für alle gleich beginnenden Tags, unabhängig davon, welche Attribute sie haben (eine Liste der tatsächlich vorkommenden Attribute wird vom Makro *TAGS mit ausgegeben, siehe oben in Liste 1).

Tags mit Attributen können zwar auch jetzt schon mit ihrem vollen Wortlaut (d. h. mit allen Attributen) berücksichtigt werden; sie müssen dazu lediglich vollständig (statt nur bis zum ersten Blank) mit dem Parameter MAH angeführt werden. Wenn es sich dabei nur um einige wenige unterschiedliche Attribute handelt (z. B. wenn bei Hervorhebungen die Art der Hervorhebung über ein Attribut angegeben ist), ist dies auch leicht durchführbar. Dieses Vorgehen verbietet sich aber, wenn als Attribute z. B. Verweisziele oder laufende Nummern angegeben sind. Wollte man dies komplett ins Makro aufnehmen, so wären ggf. mehrere tausend verschiedene Makros zu definieren.

Da dies nicht praktikabel ist, ist geplant, für solche Fälle in einer späteren Fassung von TUSTEP eine andere Lösung vorzusehen.

Der Anfang des nach TEI-Lite ausgezeichneten Textes aus dem Oxford Text Archive:

```
<text id=HOUND n='Hound of the Baskervilles'><front>
<head>The Hound of the Baskervilles.</head></front><body>
<div type=chap n="hound.01">
<head> Mr. Sherlock Holmes</head>
<p><s>Mr. Sherlock Holmes, who was usually very late in the mornings,
save upon those not infrequent occasions when he was up all night,
was seated at the breakfast table. </s><s>I stood upon the hearth-rug
and picked up the stick which our visitor had left behind him the
night before. </s><s>It was a fine, thick piece of wood,
bulbous-headed, of the sort which is known as a &odq;Penang
lawyer.&cdq; </s><s>Just under the head was a broad silver band
nearly an inch across. </s><s>&odq;To James Mortimer, M.R.C.S., from
his friends of the C.C.H.,&cdq; was engraved upon it, with the date
&odq;1884.&cdq; </s><s>It was just such a stick as the old-fashioned
family practitioner used to carry &mdash; dignified, solid, and
reassuring. </s></p>
<p><s>&odq;Well, Watson, what do you make of it?&cdq;
</s></P><P><s>Holmes was sitting with his back to me, and I had given
him no sign of my occupation. </s></p>
```

Wilhelm Ott
ott@zdv.uni-tuebingen.de
