



Segmentierungsregeln zur Erzeugung von Modulsichten

Dokumentation

Projekt
**Hypertextualisierung auf
textgrammatischer Grundlage**
(www.hytex.info)

Eva Anna Lenz

2004

Bei der Hypertextualisierung sind zwei Teilaufgaben zu bewältigen: Die Segmentierung (Modularisierung), d.h. die Zerlegung eines linearen Textes in einzelne Module (oft auch Hypertext-Knoten genannt), und das Linking, d.h. die (Re-)Konnexion dieser Teile. Bei der Modularisierung nehmen wir jedoch keine endgültige „Zerteilung“ des Textes vor, sondern erzeugen für den Benutzer Modul-Sichten, die es ihm ermöglichen, auch vorangehende und nachfolgende Module – und damit Kontext – mit anzuzeigen („Sichtfelderweiterung“¹).

Die einfache Segmentierungsstrategie „ein Paragraph wird ein Modul“ wurde im HyTex-Projekt auf zweierlei Weise verfeinert: zum einen wurden kohäsive Bezüge zwischen Modulen durch Links und Sichtfelderweiterungen aufgelöst; zum anderen wurden andere Textstrukturelemente als Paragraphen (aus der Annotationsebene „logische Textstruktur“²) ebenfalls zur Modularisierung herangezogen. Beispielsweise erschien es uns sinnvoll, einen Abschnitt, der Literaturangaben enthält, als Modul anzubieten. Es gab jedoch viele Elemente – auch Paragraphen – bei denen die Segmentierungsstrategie vom Kontext abhängt: Ein Paragraph (`<doc:para>`), der direktes Kind eines Abschnitts (`<doc:section>`) ist, sollte ein Modul werden, aber ein Paragraph, der einen Listenelement umfasst (unterhalb von `<doc:itemizedlist>` oder `<doc:orderedlist>`), soll kein eigenes Modul erzeugen. Eine Liste wiederum soll ebenfalls zum Modul werden, wenn sie direkt unterhalb eines Abschnitts liegt, aber nicht, wenn sie in einen Paragraphen eingebettet ist.

Um solche kontextabhängigen Segmentierungsregeln schnell ändern und testen zu können, wurden sie nicht zusammen mit den Linking-Regeln direkt in XSLT implementiert, sondern deklarativ in einer externen Datei beschrieben. Wir entschieden uns dabei für ein textbasiertes Format, das XPath-Prädikate³ enthält. Eine Segmentierungsregel besteht dabei aus zwei Teilen:

- 1) der Angabe eines XML-Elements (inklusive Namensraum), welches ein Modul erzeugen soll, z.B. `doc:para` oder `doc:bibliodiv`
- 2) eine kontextuelle Bedingung des XML-Elements, ausgedrückt durch ein XPath-Prädikat. Diese Bedingung dient als Filter, d.h. es werden nur diejenigen XML-Elemente zur Modularisierung herangezogen, die der Bedingung genügen. Mit XPath-Prädikaten können z.B. Eltern- oder Kind-Elemente beschrieben werden. Wie üblich werden sie in eckigen Klammern angegeben, z.B.
`[parent::doc:section or parent::doc:article]`.

Die im HyTex-Projekt verwendeten Modularisierungsregeln sind im Anhang dieses Dokuments aufgeführt.

Zur Auswertung der Regeln wird die Textdatei zunächst geparkt und in ein XML-basiertes Format übersetzt, welches mittels einer DTD validiert werden kann. Die XML-Datei enthält nun die Modularisierungsregeln in einer Form, die sich leicht durch XSLT weiterverarbeiten lässt. Der Parser für die Textdatei wurde mit dem Parser-Generator ANTLR (<http://www.antlr.org>) erzeugt. Das XSLT-Skript `AllXML2HTML.xsl` liest die

-
- 1 Siehe hierzu die Dokumentation: *Strategien zur Herstellung kohäsiv autonomer Modulsichten* (Eva Anna Lenz und Angelika Storrer; <http://www.hrz.uni-dortmund.de/~hytex/hytex/Publikationen/strategien-kohaesive-autonomie.pdf>).
 - 2 Siehe *Dokumentation zur Annotationsschicht: Dokumentenstruktur* (Eva Anna Lenz und Harald Lungen; <http://www.hrz.uni-dortmund.de/~hytex/hytex/Publikationen/fgtt-docbook-docu.pdf>).
 - 3 Beschrieben z.B. bei Kay, Michael: *XSLT Programmer's Reference*. Wrox Press, 2nd edition, 2001.

erzeugte XML-Datei mit den Modularisierungsregeln aus und interpretiert diese, d.h. es erzeugt entsprechende Modulsichten. Dieses Skript übernimmt neben der Teilaufgabe der Modularisierung auch die der Verlinkung, indem es alle drei Annotationsebenen (logische Textstruktur, Definitionen und Termverwendungsinstanzen, Koreferenz und Konnektive) auswertet und entsprechende Links generiert⁴.

4 Siehe die Dokumentation *Erzeugung modularisierter und verlinkter Hypertextsichten in der HyTex-Pilotversion (Stylesheet AllXML2HTML.xsl)* (Benjamin Birkenhake; <http://www.hrz.uni-dortmund.de/~hytex/hytex/Publikationen/html-sichten-aus-xml-annotationen.pdf>).

Anhang: Die verwendeten Segmentierungsregeln zur Erzeugung von Modulsichten

Datei rules.txt:

```
<!-- Segmentierungsregeln -->

segRule:
para[parent::doc:section or parent::doc:sect1 or
parent::doc:sect2 or parent::doc:sect3 or parent::doc:sect4 or parent::doc:sect5
or parent::doc:abstract or parent::doc:appendix or parent::doc:article or
parent::doc:bibliography]

segRule:
log:table[parent::doc:section or parent::doc:sect1 or
parent::doc:sect2 or parent::doc:sect3 or parent::doc:sect4 or parent::doc:sect5
or parent::doc:appendix or parent::doc:bibliography]

segRule:
log:figure[parent::doc:section or parent::doc:sect1 or
parent::doc:sect2 or parent::doc:sect3 or parent::doc:sect4 or parent::doc:sect5
or parent::doc:appendix or parent::doc:bibliography]

segRule:
doc:programlisting[parent::doc:section or parent::doc:sect1 or
parent::doc:sect2 or parent::doc:sect3 or parent::doc:sect4 or parent::doc:sect5
or parent::doc:appendix or parent::doc:bibliography]

segRule:
doc:itemizedlist[parent::doc:section or parent::doc:sect1 or
parent::doc:sect2 or parent::doc:sect3 or parent::doc:sect4 or parent::doc:sect5
or parent::doc:appendix or parent::doc:bibliography]

segRule:
doc:orderedlist[parent::doc:section or parent::doc:sect1 or
parent::doc:sect2 or parent::doc:sect3 or parent::doc:sect4 or parent::doc:sect5
or parent::doc:appendix or parent::doc:bibliography]

segRule:
doc:glosslist[parent::doc:section or parent::doc:sect1 or
parent::doc:sect2 or parent::doc:sect3 or parent::doc:sect4 or parent::doc:sect5
or parent::doc:appendix or parent::doc:bibliography]

segRule:
log:footnoteSect

segRule:
doc:bibliography

segRule:
doc:bibliodiv

segRule:
log:blockemphasis[parent::doc:section or parent::doc:sect1 or parent::doc:sect2 or
parent::doc:sect3 or parent::doc:sect4 or parent::doc:sect5 or
parent::doc:appendix or parent::doc:bibliography]

segRule:
doc:blockquote[parent::doc:section or parent::doc:sect1 or parent::doc:sect2 or
parent::doc:sect3 or parent::doc:sect4 or parent::doc:sect5 or
parent::doc:appendix or parent::doc:bibliography]
```