Richtlinien für die Annotation der Makrostruktur in Pop-Scie

Stand: 27.05.2011
Autor: Karin Maksymski

Die vorliegenden Richtlinien gelten für den ersten Schritt bei der Annotation der Texte im Projekt Pop-Scie. Ausganspunkt der Annotation sind die "rohen" Textdateien der Korpustexte (.txt-Dateien). Bevor diese nun im zweiten Schritt in die Annotationspipeline eingespeist werden (zur Annotation der Wortarten, der Dependenzen und der syntaktischen Funktionen), soll mithilfe einer (manuellen) *Makrostruktur-Annotation* markiert werden, zu welchem Teil des Artikels der Text gehört bzw. welche zusätzlichen Elemente im Artikel vorhanden sind.

Diese Richtlinien geben einen Überblick über die dabei verwendeten *Bezeichnungen* und die *Regeln*, die bei der Annotation befolgt wurden.

Art der Annotation

Tags

Die Annotation wird in Form von Abkürzungen in spitzen Klammern (Tags) vorgenommen. Alle Elemente, die keinen Text enthalten oder deren Text nicht Teil des Artikels ist (z.B. Bilder ohne Bildunterschrift), werden nicht markiert bzw. aus dem Text entfernt. Zwischentitel etc. werden zunächst nicht nummeriert, da dies später automatisch erfolgt.

Formate

Am Anfang jedes Dokuments steht die Info zur XML-Version:

<?xml version="1.0" encoding="UTF-8"?>

Die fertig annotierten Dateien werden als .txt-Datei mit der Dateiendung *_ms.txt* gespeichert.

Textblöcke

Die Abkürzungen gliedern sich in verschiedene Gruppen. Auf der obersten (sehr allgemeinen) Ebene wird zwischen verschiedenen Formen der Textdarstellung unterschieden, und zwar den Textblöcken:

M	main = Haupttext
I	Info = Infokasten (auch Informationen, die klar vom Text abgetrennt, aber nicht in einem gezeichneten Kasten stehen, werden so markiert)
B	Bild = Beschriftung einer Grafik, eines Fotos o.ä.
Z	Zusatz = einzelne Zeilen oder Kurztexte mit Metainformationen oder anderem Bezug zum Haupttext

Die Blöcke M und I enthalten im Gegensatz zu Z immer mehrere Elemente. Eine vom Haupttext abgesetzte Zeile mit einem Literaturhinweis gilt demnach nicht als Infokasten, sondern als Zusatz.

Block B kann mehrere Elemente enthalten (z.B. Fließtext mit Titel), muss aber nicht (nur Fließtext, also einfache Bildbeschriftung).

Die Blöcke B und Z können auch innerhalb der anderen Textblöcke vorkommen, z.B.:

- ein Bild mit entsprechender Bildbeschriftung, das in einem Infokasten neben einem längeren Textabschnitt steht (als B in Kästen zählen Texte, die explizit auf ein Bild Bezug nehmen)
- Metainformationen zum Autor, die am Schluss des Haupttextes gegeben werden

Textteile (Gliederung)

Eine Ebene tiefer gibt es für die Textblöcke M, B und I die folgenden Textteile:

T	Titel
OT	kleine gedruckte Überschrift über dem eigentlichen Titel
UT	kleine gedruckte Überschrift unter dem eigentlichen Titel
VS	Vorspann / Lead, oft fett oder kursiv gedruckt (Text wird als VS und nicht als UT gekennzeichnet, wenn er aus mehreren Sätzen besteht oder aber nur aus einem Satz, der mit einem Punkt abgeschlossen wird)
ZT	Zwischenüberschrift im Text
B	nur für I, s.o.
Z	nur für M und I, s.o.
(FT)	(Fließtext)

Der Fließtext wird nicht eigens gekennzeichnet. Entsprechend ist alles Nicht-Markierte Fließtext. Die Elemente in dieser Liste sind (bis auf B und Z) nur auf dieser Ebene zu finden. Überschriften werden nur in den genannten Textblöcken annotiert. In Z tauchen sie bisweilen zwar auch auf (z.B. in Fußnoten oder Literaturhinweisen bei ZEIT Wissen), werden dort aufgrund der insgesamt geringen Textlänge jedoch ignoriert (so ist auch gesichert, dass Z nicht, wie bei den Textblöcken M und I der Fall, mehrere Elemente enthält).

Zum Unterschied zwischen B und I: Generell kann man sagen, dass I komplexer ist als B. B kann zwar, wie oben erwähnt, auch mehrere Elemente haben; das trifft jedoch tatsächlich nur auf die Kombination Titel mit Fließtext zu. Wenn dagegen z.B. eine Grafik mit Titel und mehreren Beschriftungen vorliegt, gilt der Gesamtkomplex als Infokasten.

Textteile (inhaltliche Funktion)

Eine dritte Ebene gibt es nur noch für Z bzw. in einem Fall auch für B. Da Z eine eher schwammige Kategorie ist, müssen in jedem Fall genauere Informationen bezüglich des jeweiligen Inhalts gegeben werden. Diese werden über den Zusatz "inhalt" in das Tag <Z> mit aufgenommen (Beispiele siehe unten). Hierzu zählen Metainformationen wie:

AA	Autorenangabe (im Prinzip nur Name des Autors)
AI	Autoreninfo = weitere Infos zum Autor

Außerdem fallen auch Hinweise auf etwas außerhalb des Textes sowie Zusätze mit Bezug zum Textinhalt in diese Kategorie:

LH	Literaturhinweis
V	Verweis = Metatext außerhalb des Haupttextes, der meist als einzelne Zeile und z.B. kursiv geschrieben, auf Grafiken, Audioversion des Artikels im Internet etc. verweist
FN	Fußnote
TZ	Textzitat = Satz oder Aussage aus dem Text (Zitat oder Quasi-Zitat), der außerhalb des Haupttextes als Blickfänger steht (wird als TZ und nicht als B gekennzeichnet, auch wenn daneben ein Bild o.ä. ist – vorausgesetzt, das Bild illustriert den Text und nicht umgekehrt; siehe z.B. *ZEIT Wissen*)

Beispiele:

<Z inhalt="AA"> => Zusatzzeile mit Name der Autoren

<Z inhalt="FN"> => Fußnote außerhalb des, aber mit Bezug zum Haupttext

Da die Autoreninfo häufig neben einem Foto der Autorin platziert ist und folglich als Bildbeschriftung zählt, ist in diesem Fall auch eine nähere Definition des Textblocks B möglich:

<B inhalt="AI">NameAutorin</B>

[Gäbe es kein Foto dazu, sähe die Annotation so aus:

<Z inhalt="AI">NameAutorin</Z>]

Es gibt einen Fall, wo nicht ein Textblock oder ein Teil davon, sondern einzelne Wörter oder Wortgruppen im Fließtext annotiert werden. Hierbei handelt es sich um Web-Links im Textblock M, die nicht durch die Länderabkürzung oder durch "www" oder "http", sondern nur durch "unterstrichen" markiert sind. Diese Art der Verlinkung taucht nur in einigen ZEIT-Artikeln auf und wird mit den Tags <V> und </V> annotiert.

In diesem Fall steht V also nicht als Inhaltsattribut in einem "Z"-Tag, sondern als eigenes Tag.

beachten: keine Leerzeichen vor und nach den Tags