

# **Annotationsschicht: Thematische Strukturen / Themenentwicklung**

Dokumentation

Projekt  
**Hypertextualisierung auf  
textgrammatischer Grundlage**  
([www.hytext.info](http://www.hytext.info))

Eva Anna Lenz  
Angelika Storrer

2003

Der Ausdruck „Thema“ bzw. „topic“ wird in der Linguistik und Textverstehensforschung sehr unterschiedlich definiert, entsprechend kann man unter einer als „thematische Struktur“ bezeichneten Annotationsschicht auch durchaus Verschiedenes verstehen. In der thematischen Annotationsschicht des HyTex-Projekts konzentrieren wir uns auf die thematische Strukturierung in der Tradition der Typologisierung thematischer Progressionsmuster und Thema-Rhema-Einheiten.<sup>1</sup> Das für HyTex entwickelte Annotationsschema für thematische Strukturen orientiert sich an den von Ludger Hoffmann in der „Grammatik der deutschen Sprache (GDS)“<sup>2</sup> beschriebenen Formen der Themenentwicklung, das an diese Tradition anknüpft und sie für die im Deutschen verfügbaren Strukturen und Sprachmittel ausführt. Unter einem Thema (topic) verstehen wir entsprechend „den kommunikativ konstituierten Gegenstand oder Sachverhalt, über den in einem Diskurs oder Text(-abschnitt) fortlaufend etwas gesagt wird“ (GDS, S. 509). Die Grundidee des Schemas besteht darin, Dokumente in thematische Segmente zu zerlegen, die wiederum zu thematischen Clustern zusammengefasst sind. Diese Cluster lassen sich entweder einfachen oder verschachtelten Formen der Themenentwicklung zuordnen.

Für die Hypertextualisierungsstrategien kann die thematische Annotationsschicht in zweierlei Weise zum Tragen kommen:

- Das Schema erlaubt es, jedes Thema aus der thematischen Annotation mit einem Verweis auf das entsprechende Konzept aus dem Wissensnetz TermNet zu versehen, soweit es dort modelliert ist (TermNet enthält bislang nur diejenigen Termini und Konzepte, die in den Dokumenten des Korpus auch definiert sind). Dies macht es potenziell möglich, vom Wissensnetz aus eine Liste mit allen Textsegmenten anzubieten, in denen das Thema abgehandelt wird. Umgekehrt kann in diesen Fällen zu jedem Thema ein Glossareintrag mit weiterführender Information aufgerufen werden.
- Die Annotation der thematischen Strukturebene wäre theoretisch ein sehr guter Ausgangspunkt für Strategien zur Modularisierung von Dokumenten. Allerdings scheint uns die thematische Annotation nur schwer automatisierbar zu sein. Da wir den zentralen Strategietyp „Modularisierung“ aber nicht von einer manuell vorzunehmenden Strukturierungsebene abhängig machen wollen, wird die thematische Ebene lediglich zur zusätzlichen Verfeinerung der Regeln herangezogen, die sich vornehmlich an der logisch-hierarchischen Dokumentstruktur orientieren.

Für die verschiedenen Formen der Themenentwicklung, wie sie in der Grammatik der deutschen Sprache beschrieben werden, wurde im HyTex-Projekt eine Dokumentgrammatik entwickelt, die es erlaubt, diese zu annotieren (XML-Schema „thematik.xsd“). Eine graphische Version findet sich im Anhang dieses Dokuments. Im Folgenden geben wir Hinweise für die Annotation, wobei die Beispielsätze ebenfalls aus der Grammatik der deutschen Sprache übernommen wurden. Zur besseren Übersichtlichkeit wurden die Bezüge, in denen übergreifende Themen und Teilthemen in den verschiedenen Themenentwicklungstypen zueinander stehen, graphisch dargestellt. Die gestrichelten Kästchen symbolisieren dabei Themencluster,

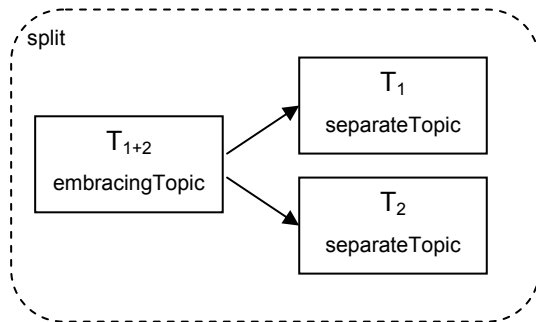
---

1 Z.B. Danes, Frantisek (1976): Zur semantischen und thematischen Struktur des Kommunikats. In: Danes, Frantisek/Viehweher, Dieter (Hgg.) (1976): Probleme der Textgrammatik. Berlin (Studia Grammatica XI), 29-40.

2 Zifonun, Gisela/Hoffmann, Ludger et al. (1997): Grammatik der deutschen Sprache. Berlin/New York; Bd. 1, Kap C6 „Thematische Organisation von Text und Diskurs, insb. die Seiten 535-591.

die Kästen mit durchgezogener Umrandung Themensegmente. Jedem Segment und Cluster kann fakultativ ein Element Thema (topic) zugeordnet werden, das im Hinblick auf die in HyTex verfolgten Strategien weiter in Thementypen subklassifiziert werden kann. Die aktuell verwendeten Thementypen werden abschließend erläutert.

## Themensplitting



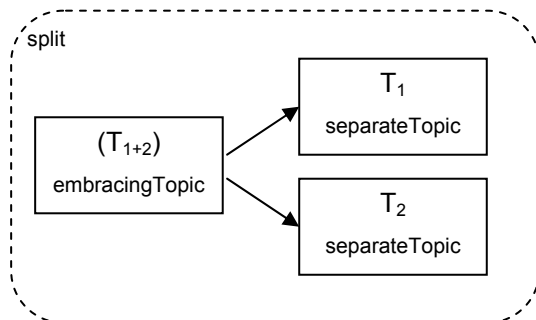
Themensplitting:

Übergang von einem mehrere Gegenstände umfassenden Thema  $T_{1+2}$  auf Einzelthemen  $T_1$  und  $T_2$ .

Hypothese:  $T_{1+2}$  steht zu  $T_1$  und  $T_2$  in der Klasse-Instanz Beziehung, der superclass-subclass-Beziehung (Hyperonymie) oder der Beziehung Klasse – Gruppe von Instanzen.

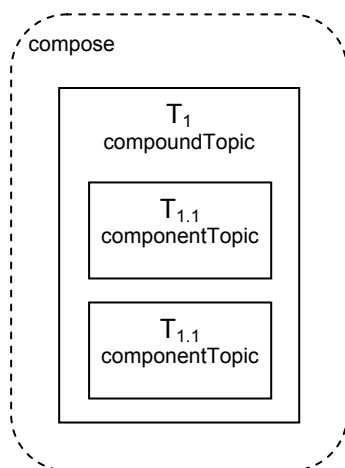
```
<tCluster type="split">
  <tSegment role="embracingTopic">...<topic>zwei Frauen</topic>...</tSegment>
  <tSegment role="separateTopic">...<topic>die eine</topic>...</tSegment>
  <tSegment role="separateTopic">...<topic>die andere</topic>...</tSegment>
</tCluster>
```

Es gibt genau ein Segment mit der Rolle „embracingTopic“, beliebig viele Segmente vom Typ „separateTopic“. Das Topic  $T_{1+2}$  im Segment mit der Rolle „embracingTopic“ muss nicht explizit genannt sein. Wenn es fehlt, dann wird statt `<topic>` `<noExplicitTopic>` benutzt. `<noExplicitTopic>` hat ein optionales Attribut `topicConceptName`, mit dem das Topic benannt werden kann:



```
<tCluster type="split">
  <tSegment role="embracingTopic"><noExplicitTopic topicConceptName="...">...</noExplicitTopic>...</tSegment>
  <tSegment role="separateTopic">...<topic>...</topic>...</tSegment>
  <tSegment role="separateTopic">...<topic>...</topic>...</tSegment>
</tCluster>
```

## Themenkomposition



Themenkomposition:

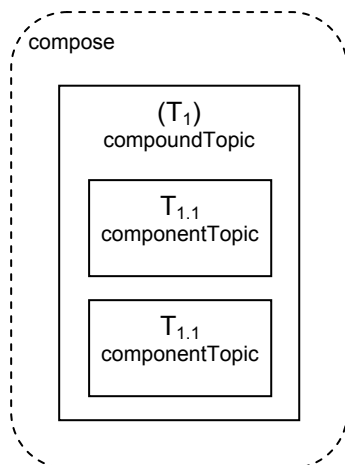
Ein Thema  $T_1$  wird durch die Behandlung der Subthemen  $T_{1.1}$  und  $T_{1.2}$ , die es konstituieren, bearbeitet.

Hypothese:  $T_1$  steht zu  $T_{1.1}$  und  $T_{1.2}$  in der Ganzes-Teil-Relation (Meronymie).

(Hinweis: Die Themen treten in der Reihenfolge  $T_1, T_{1.1}, T_{1.2}$  auf.)

```
<tCluster type="compose">
  <tSegment role="compoundTopic">...<topic>Narwal</topic>...</tSegment>
  <tSegment role="componentTopic">...
    <topic>sein aus dem Norwegischen stammender Name</topic>...</tSegment>
  <tSegment role="componentTopic">...<topic>beide Geschlechter</topic>...
</tSegment>
</tCluster>
```

Es gibt genau ein Segment mit der Rolle „compoundTopic“, beliebig viele Segmente vom Typ „componentTopic“. Das Topic  $T_1$  im Segment mit der Rolle „compoundTopic“ muss nicht explizit genannt sein. Wenn es fehlt, wird statt `<topic>` `<noExplicitTopic>` benutzt.



```
<tCluster type="compose">
  <tSegment role="compoundTopic"><noExplicitTopic topicConceptName="..." />
  ...</tSegment>
  <tSegment role="componentTopic">...<topic>...</topic>...</tSegment>
  <tSegment role="componentTopic">...<topic>...</topic>...</tSegment>
</tCluster>
```

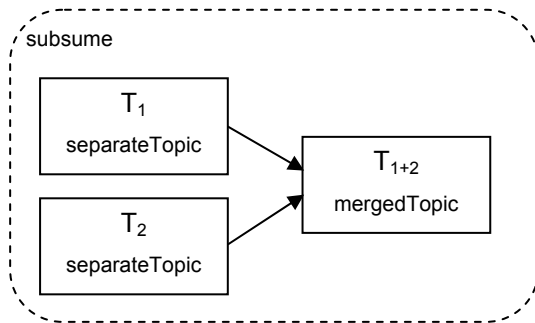
Da die Themenkomposition nicht immer klar vom Themensplitting unterschieden werden kann, lassen wir zu, dass in Zweifelsfällen bei der Annotation die (künstliche) Kategorie „splitcompose“ benutzt werden kann:

```

<tCluster type="splitcompose">
  <tSegment role="compoundTopic">...<topic>...</topic>...</tSegment>
  <tSegment role="componentTopic">...<topic>...</topic>...</tSegment>
  <tSegment role="componentTopic">...<topic>...</topic>...</tSegment>
</tCluster>

```

## Themensubsumtion



Themensubsumtion:  
Übergang von zwei oder mehr Themen  $T_1$  und  $T_2$  auf ein Thema  $T_{1+2}$ , in dem sie aufgehen.

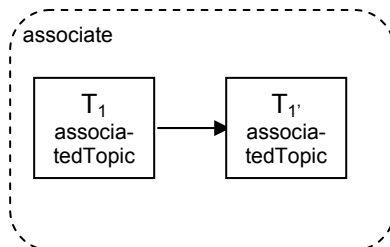
```

<tCluster type="subsume">
  <tSegment role="separateTopic">...<topic>Selma</topic>...</tSegment>
  <tSegment role="separateTopic">...<topic>Luise</topic>...</tSegment>
  <tSegment role="mergedTopic">...<topic>beide</topic>...</tSegment>
</tCluster>

```

Es gibt genau ein Segment mit der Rolle „mergedTopic“, beliebig viele Segmente vom Typ „separateTopic“.

## Themenassoziation



Themenassoziation:  
Von einem Thema  $T_1$  wird auf ein Thema  $T_1'$  übergeleitet.  
Zwischen beiden Themen besteht irgendeine Gemeinsamkeit.

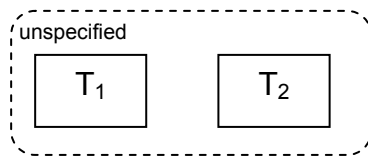
```

<tCluster type="associate">
  <tSegment role="associatedTopic">...<topic>Claudia</topic>...</tSegment>
  <tSegment role="associatedTopic">...<topic>ihre Handtasche</topic>...</tSegment>
</tCluster>

```

Es kann beliebig viele Segmente mit der Rolle „associatedTopic“ geben.

## Themenreihung und Themenfortführung



Themenreihung:  
Auf Thema T<sub>1</sub> folgt Thema T<sub>2</sub>. Die Themen haben hinsichtlich ihres Gegenstandes nichts gemein.

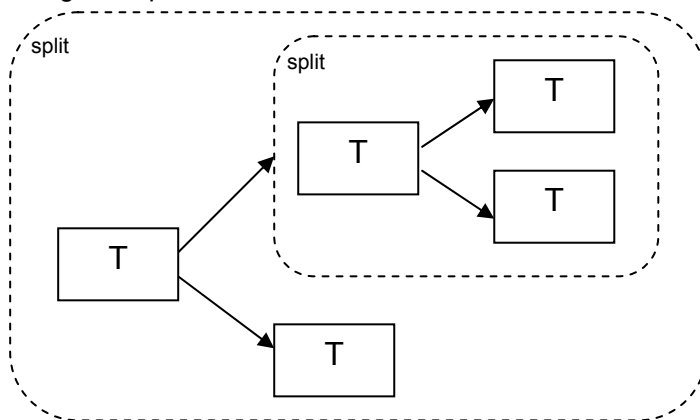
Themenreihung wird nicht explizit annotiert. (Die Annotation erfolgt implizit dadurch, dass zwei Segmente sich in demselben Cluster befinden, ohne dass ihre Beziehung zueinander durch Rollen ausgedrückt wird.)

Auch Themenfortführung wird nicht explizit annotiert. (Die Annotation erfolgt implizit dadurch, dass zwei Segmente an unterschiedlichen Stellen im Dokument mit demselben Thema ausgezeichnet werden.)

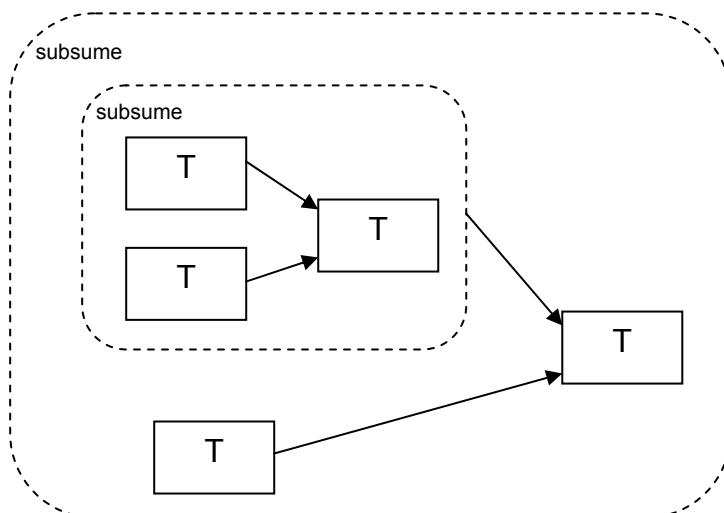
## Verkettungen verschiedener Formen der Themenentwicklung

Die Dokumentgrammatik lässt es zu, dass Cluster sowohl Segmente als auch andere Cluster enthalten können (auch gemischt). Damit wird es (syntaktisch) möglich, verschiedene Formen der Themenentwicklung rekursiv miteinander zu verknüpfen. Es sind jedoch nicht alle wünschenswerten Arten von Verknüpfungen möglich; umgekehrt gibt es syntaktisch mögliche Kombinationen, über deren Sinn noch diskutiert werden kann.

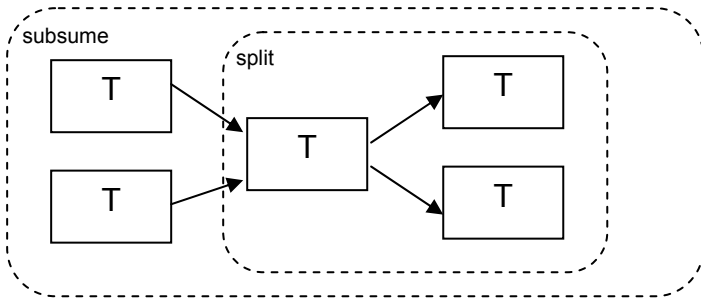
Hier einige Beispiele:



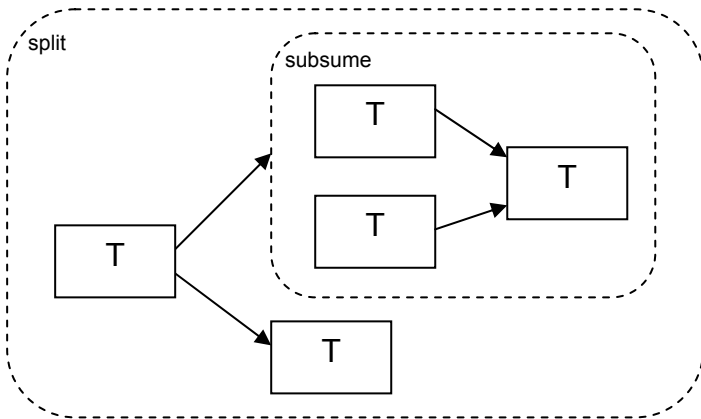
Möglich: Ein Thema wird immer feiner aufgesplittet.



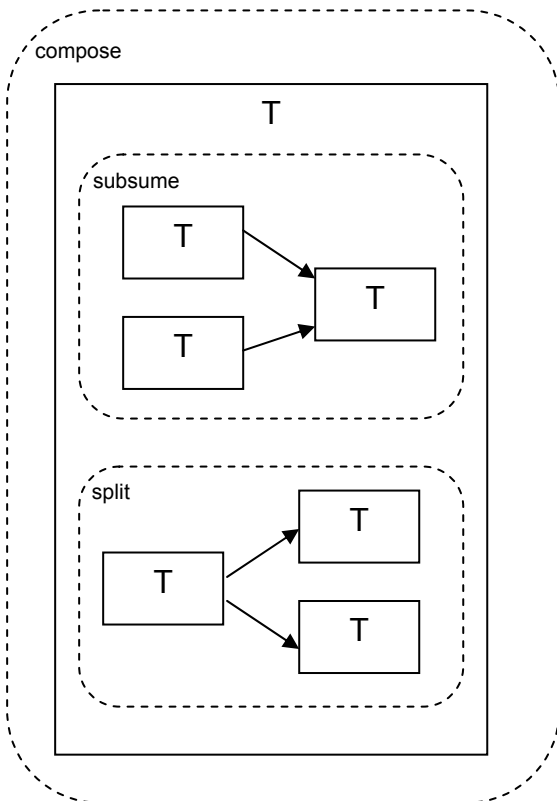
Möglich: Mehrere Themen werden sukzessive subsumiert.



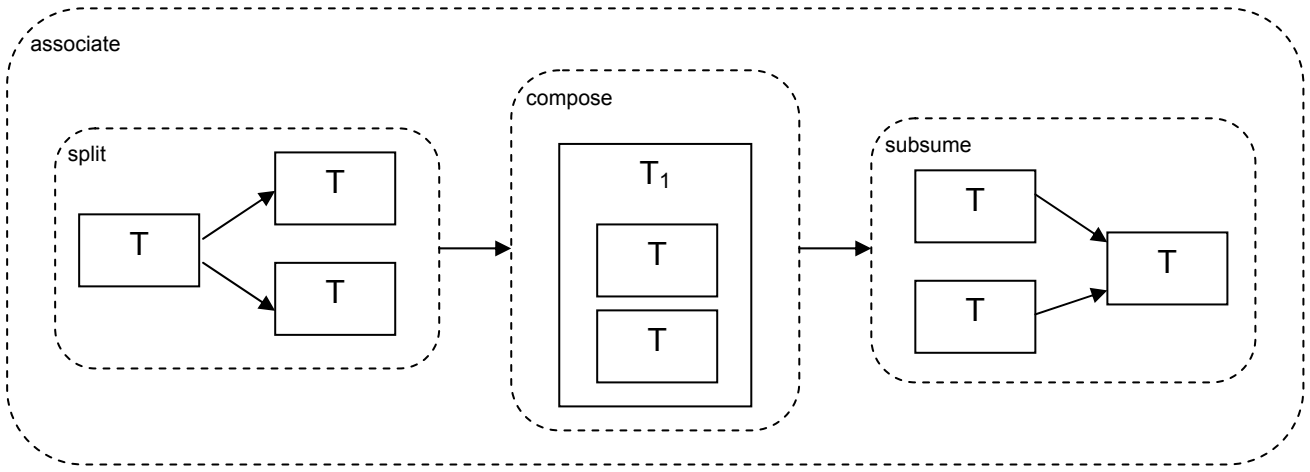
Möglich: Zwei Teilthemen werden in eines subsumiert, welches anschließend wieder aufgesplittet wird (in die ursprünglichen oder andere Teile).



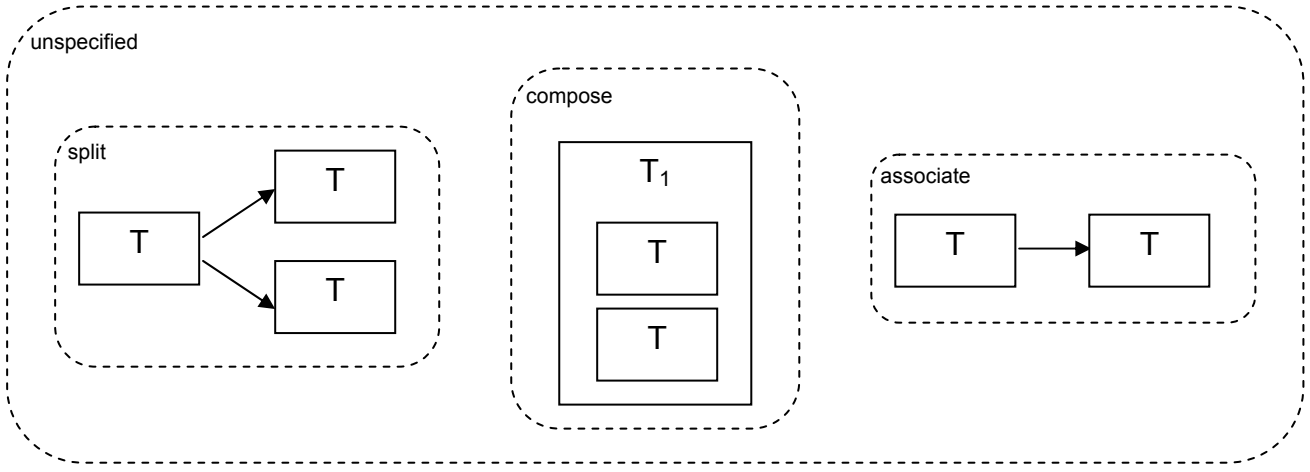
Syntaktisch möglich.



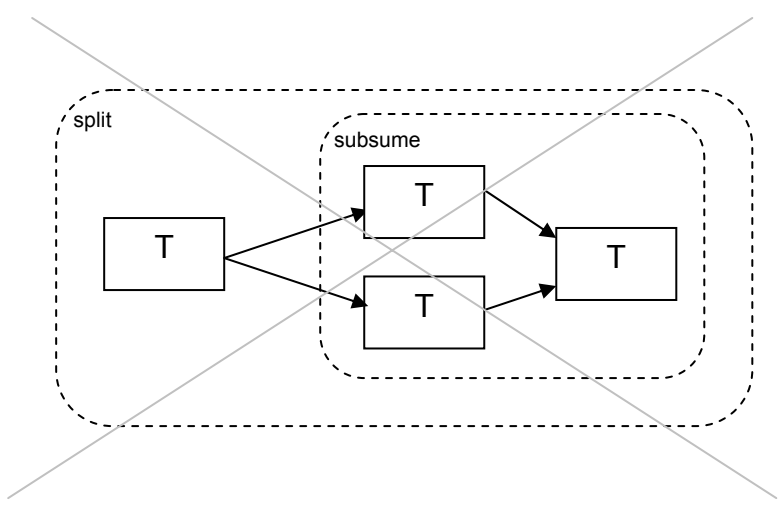
Möglich: Ein komplexes Thema wird durch zwei Teilthemen bearbeitet, wobei sich das erste ebenfalls aus zwei Teilthemen zusammensetzt und das zweite aufgesplittet wird.



Möglich: Mehrere komplex bearbeitete Themen werden assoziiert.



Möglich: Mehrere komplex bearbeitete Themen werden gereiht.



syntaktisch **nicht** möglich: Aufspaltung eines Themas, anschließend wieder Subsumtion.

## Annotation von Themen-Typen

Normalerweise hat jedes Segment (tSegment-Element) ein Topic-Element (topic) (Ausnahmen: siehe oben). Dabei lassen sich verschiedene Typen von Themen unterscheiden, die für die Weiterverarbeitung bzw. die Verknüpfung mit Wort- bzw. Konzept-Topiks im Wissensnetz TermNet relevant sind. Das Topic-Element (topic) hat deshalb ein optionales Attribut type mit den folgenden zulässigen Werten:

### Concept

Das concept-Attribut wird verwendet, wenn das Topic einen Gegenstand benennt:

#### Beispiel:

```
Die <topic type="concept">Spezifikation des Anker-Konzepts</topic> in (Tochtermann 1995) (s.u.) erweitert die einfache Unterscheidung von globalen und lokalen Ankern um eine gestufte Definition von Ankerbereichen auf den Präzisionsstufen Modul - Komponente - Bereich.
```

In diesem Fällen verweist das Thema auf ein Konzept-Topic im Wissensnetz (TermNet) (so weit dort aufgenommen).

### word

Wird verwendet, wenn ein sprachlicher Ausdruck selbst zum Thema wird, also z. B. in metakommunikativer Verwendung oder in definatorischen oder definitionsähnlichen Textsegmenten,

#### Beispiel.:

```
Unter der <topic type="word">Typisierung von Links (link typing)</topic> versteht man vielmehr die computerverarbeitbare Beschreibung der deklarativen und prozeduralen Linksemantik mit der Zielsetzung der Hypertextstrukturierung.
```

In diesem Fällen verweist das Thema auf ein Wort-Topic im Wissensnetz TermNet (soweit dort aufgenommen).

### problem

Wird verwendet, wenn es sich beim Thema um ein Problemfeld / eine Fragestellung handelt, die im Folgetext weiter behandelt wird: Die Kategorie ist zu verstehen im Sinne von Themenkonzepten, die ein Thema als Frage (quaestio) auffasst, die im Folgetext behandelt / beantwortet wird.<sup>3</sup> Sie ist nicht zu verwechseln mit der argumentativ-rhetorischen Funktion der Problematisierung z. B. als Textmusterelement in wissenschaftlichen Artikeln.

#### Beispiel:

```
<topic type="problem"> Zur Strukturierung der Objekte auf der konzeptionellen Modellierungsebene </topic> gibt es drei Strategietypen.
```

---

3 Z.B. Hellwig, Peter (1984): Grundzüge einer Theorie des Textzusammenhangs. In: Rothkegel, Annelly/Sandig, Barbara (Hgg.) (1984): Text - Textsorten - Semantik: linguistische Modelle und maschinelle Verfahren. Hamburg, 51-79; von Stutterheim, Christiane (1994): Quaestio und Textaufbau. In: Kornadt, Hans-Joachim/Grabowski, Joachim et al. (Hgg.) (1994): *Sprache und Kognition. Perspektiven moderner Sprachpsychologie*. Heidelberg, S. 251-272.

Themen von diesem Typ lassen sich mit dem Wort- und Konzept-orientierten Wissensnetz TermNet nicht sinnvoll verknüpfen.

**unspecified**

Verwenden wir, wenn sich ein Thema keiner der drei Kategorie eindeutig zuweisen lässt. Auch hier findet keine Verknüpfung zum TermNet statt.

## Anhang: Die verwendete Dokumentgrammatik thematik.xsd

Elements

[noExplicitTopic](#)

[tCluster](#)

[topic](#)

[tSegment](#)

### element **noExplicitTopic**

diagram

**noExplicitTopic**

Leeres Element, das anstelle des topic-Elements verwendet wird, wenn es kein explizit im Segment vorkommendes Topic gibt. Das optionale Attribut "topicConceptName" enthält einen Namen für das implizit vorhandene Topic (der mit einem Topic-Namen in der TopicMap übereinstimmen kann).

used by element [tSegment](#)

attributes	Name	Type	Use	Default	Fixed
	topicConceptName	xs:string	optional		

## element tCluster

diagram



children [tSegment](#) [tCluster](#)

used by element [tCluster](#)

attributes	Name	Type	Use	Default	Fixed
	type	xs:string	required		
	topicConceptName	xs:string	optional		
	role	xs:string	optional		

## element topic

diagram



Sprachlicher Ausdruck, mit denen ein Thema eingeführt wird.  
Das Attribut "topicConceptName" enthält den Namen des Konzepts, der mit einem Topic-Namen in der TopicMap übereinstimmen kann.  
Das optionale Attribut "type" gibt den Typ des Topics an. Erlaubte Werte: "concept", "word", "problem", "unspecified".

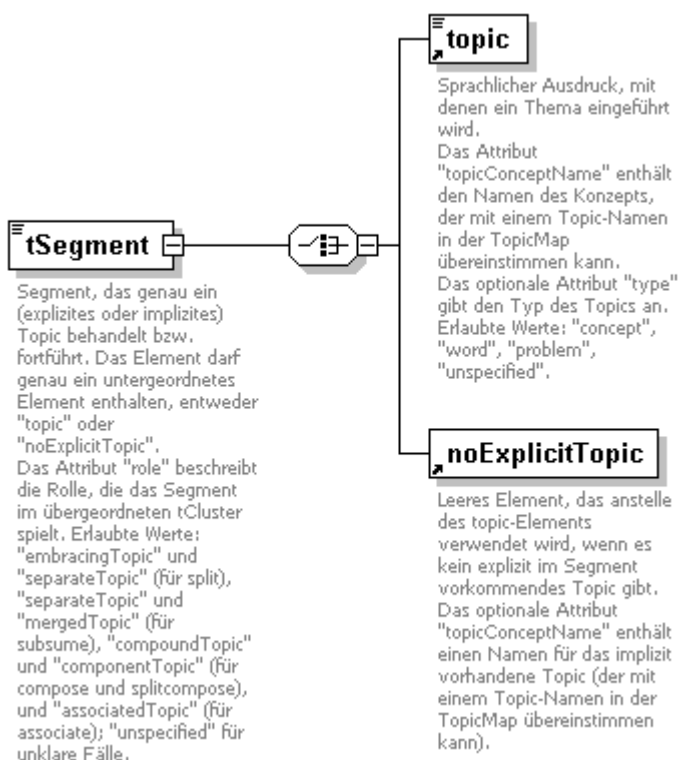
type extension of **xs:string**

used by element [tSegment](#)

attributes	Name	Type	Use	Default	Fixed
	topicConceptName	xs:string	optional		
	type	xs:string	optional		

## element tSegment

diagram



children [topic](#) [noExplicitTopic](#)

used by element [tCluster](#)

attributes	Name	Type	Use	Default	Fixed
	role	xs:string	required		