Richtlinien zur Annotation *von Themenbezeichnern in Fachtexten*

Tobias Claas, Alexander Kurek und Irene Cramer

Projekt HyTex (www.hytex.info)
Institut für deutsche Sprache und Literatur,
Technische Universität Dortmund
Stand: 2008-02-24

1. Aufgabenstellung

Im Rahmen des Projekts HyTex untersuchen wir Verfahren, die es einem Nutzer mit Grundkenntnissen in einem bestimmten Fach ermöglichen sollen, sich schnell und unkompliziert in einen neuen (aber seinem Fach verwandten) Themenbereich einzuarbeiten. Unsere bisherige Arbeit (siehe www.hytex.info) stellt dafür bereits einige Funktionalitäten zur Verfügung, diese sollen nun um verschiedene Aspekte der Verlinkung nach thematischen Gesichtspunkten erweitert werden.

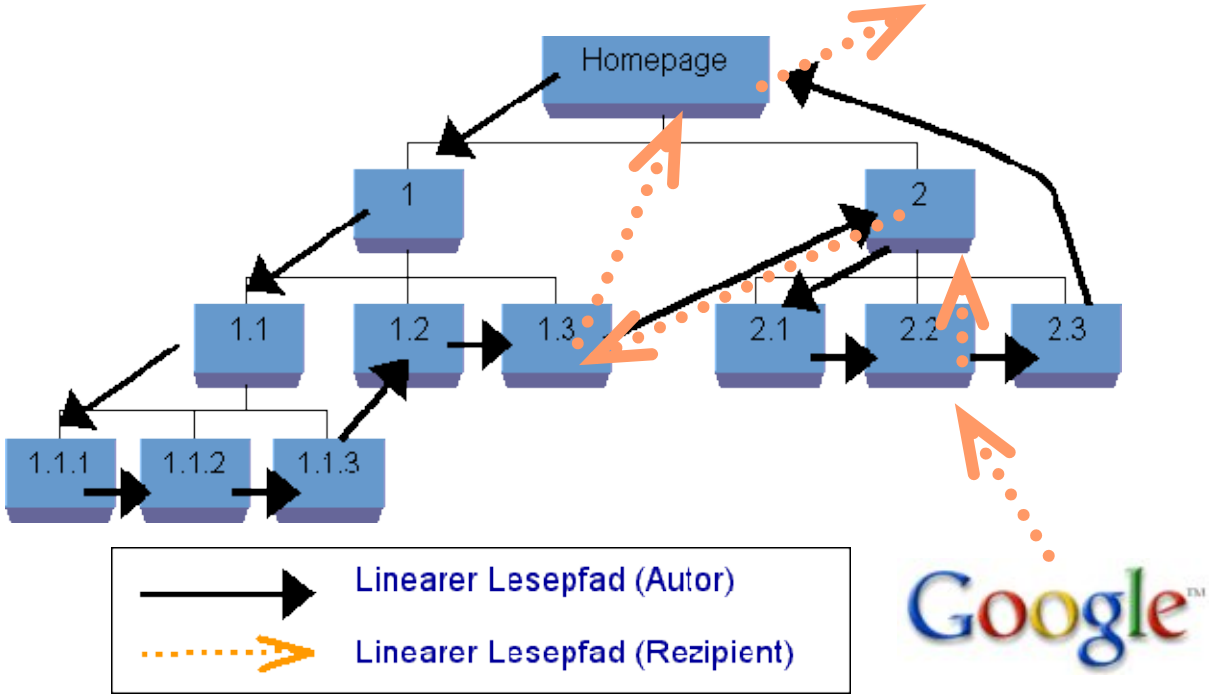
Abbildung 1 zeigt das grundlegende Szenario, das wahrscheinlich jeder Internetnutzer bereits kennen gelernt hat. Angenommen ein Nutzer sucht mit Hilfe von Google nach Informationen zu einem bestimmten Begriff oder Konzept, so kann es passieren, dass er über die Suchergebnisse von Google mitten in ein längeres Dokument hineinspringt.

An der aktuellen Stelle (in Abbildung 1 Kapitel 2.2) fehlen dem Nutzer daher möglicherweise einige Informationen, die in den Kapiteln weiter vorne (in Abbildung 1, Kapitel 1 bis 2.1) bereits angesprochen wurden.

Um den Text an der aktuellen Position am besten zu verstehen, müsste der Nutzer natürlich nun alles von Anfang an lesen. Diese Vorgehensweise ist allerdings sehr zeitraubend und in der Praxis oft nicht realistisch – zumal den Nutzer ja evtl. wirklich nur ein ganz bestimmter Aspekt interessiert.

Um in genau diesem Szenario eine Vereinfachung für den Nutzer anbieten zu können, wollen wir den thematischen Verlauf des gesamten Dokuments in einigen Schlüsselwörtern, wir nennen diese **Themenbezeichner**, in einer Übersicht anbieten. Ein Beispiel hierfür ist in Abbildung 2 dargestellt.

Abbildung 1: Anwendungsszenario



Wie man sieht, sind ergänzend zu den Überschriften, die selbst wieder mit den eigentlichen Textabschnitten verlinkt sind, pro Abschnitt (bzw. pro Überschrift) ein bis drei Themenbezeichner (thematische Schlüsselwörter) aufgelistet.

Kommen wir zurück auf das oben skizzierte Anwendungsszenario: Angenommen der Nutzer bemerkt, dass ihm für das Verständnis des Textes an seiner aktuellen Position offenbar der Kontext fehlt, er also irgendeinen der Bereiche weiter vorne im Dokument hätte lesen müssen. In diesem Fall kann er die Themenbezeichner (vgl. Abbildung 2) als Hinweis dafür nutzen, wo er die ihm fehlende Information im Dokument vermutlich finden wird.

Themenbezeichner geben also darüber Aufschluss, welche Informationen in bestimmten Bereichen eines Dokuments thematisiert werden; allerdings haben wir es mit einem ganz bestimmten Anwendungszenario zu tun – wir interessieren uns also besonders für solche Ausdrücke, die nicht unbedingt häufig genannt werden, sondern die das Thema eines bestimmten Abschnitts sind, die also in einem Abschnitt (evtl. umfassend/detailliert) erläutert werden.

Abbildung 2: Beispiel Themenkarte als Ergebnis der Annotation von Themenbezeichnern

Struktur der vorliegenden Annotationsrichtlinien: In Abschnitt 2 werden zunächst die grundlegenden Merkmale von Themenbezeichnern zusammengefasst. In Abschnitt 2.1 wird erläutert, wie man Themenbezeichner in einem Fachtextkorpus ermittelt und annotiert. Dieser Abschnitt geht auch auf die Aspekte ein, die sich in der bisherigen Annotationsarbeit als besonders kritisch erwiesen haben (vgl. Abschnitt 2.2 bis 2.7). In Abschnitt 3 erklären wir dann, wie die Annotation technisch umgesetzt werden soll.

Die entsprechend der vorliegenden Annotationsrichtlinien aufbereiteten Daten sollen im Rahmen des Projekts HyTex verwendet werden, um automatische Verfahren für

die Extraktion von Themenbezeichnern zu entwickeln. Aus diesem Grund ist es besonders wichtig, dass die Annotation systematisch und gewissenhaft durchgeführt wird.

Vorstudien haben gezeigt, dass die Annotation von Themenbezeichnern eine komplexe Aufgabenstellung ist. Wir haben daher die vorliegenden Annotationsrichtlinien so konzipiert, dass die Annotatoren möglichst schnell und sicher entscheiden können, welche Wörter (Wortgruppen) die besten Themenbezeichner für ein Textsegment (auch Modul genannt) sind.

Im Hinblick auf unser Anwendungsszenario haben wir daher in Kauf genommen, dass nicht in jedem einzelnen Fall die möglicherweise optimalen Themenbezeichner annotiert werden. Unsere Experimente zeigen aber, dass man über "optimale" Themenbezeichner sehr lange diskutieren kann – was wissenschaftlich gesehen interessant sein mag, für die konkrete Annotationsarbeit aber leider nicht praktikabel ist.

Wie das auch in anderen Bereichen der Korpusannotation üblich ist, gehen wir davon aus, dass eine Einarbeitungszeit von ca. einer Stunde notwendig ist, um eine grundlegende Sicherheit im Verständnis der in diesen Annotationsrichtlinien vorgestellten Konzepte zu erlangen. Weitere 30 Minuten sollte es dauern, bis die technische Umsetzung der Annotation klar geworden ist. Erst danach kann mit der eigentlichen Annotationsarbeit begonnen werden.

Möglicherweise treten auch nach der Einarbeitung noch Unsicherheiten auf. Bestimmte Textsegmente sind besonders schwer zu annotieren, das sollte aber nicht zu Unsicherheit im Bezug auf die Annotationsaufgabe führen. Durch den Vergleich der Annotation verschiedener Annotatoren, können wir problematische Textsegmente leicht identifizieren; diese Bereiche werden dann in einer speziellen Sitzung von besonders erfahrenen Annotatoren gemeinsam überarbeitet.

2. Annotation von Themenbezeichnern – grundlegende Aspekte

Themenbezeichner können sein:

- Substantive oder
- einfache Nominalphrasen.

Einfache Nominalphrasen im Sinn der vorliegenden Annotationsrichtlinien weisen eine der folgenden Strukturen auf:

Adj + N	Bspl.: syntaktische Links
NE + NE	Bspl.: Vannevar Bush
Adj + V (nominalisiert)	Bspl.: 'Weltweit Warten'
Partikel + N	Bspl.: Nicht-Linearität
Adj + N	Bspl.: nicht-lineare Schreibtechnologie, lineares Schreiben
V + N	Bspl.: gesprochene Sprache, geschriebene Sprache
NE + N	Bspl.: Züricher Textqulitätenmodell

Da es sich um ein Fachtextkorpus handelt, können außerdem bestimmte Fachtermini als Themenbezeichner fungieren, selbst wenn sie keine Substantive sind und auch keine der oben dargestellten Strukturen aufweisen. Alle Fachtermini, die bei dieser Annotationsarbeit berücksichtigt werden dürfen, sind im Anhang dargestellt.

Wörter (Wortgruppen),
- die **nicht in der Liste der Fachtermini** im Anhang auftauchen und
- **keine Substantive** sind und
- **auch keine der** oben aufgelisteten **Strukturen** aufweisen,

dürfen **nicht als Themenbezeichner annotiert** werden**.**

Die Themenbezeichner werden pro Textabschnitt ermittelt und später ergänzend zur Überschrift des Abschnitts dem Nutzer des Systems angeboten, wie in Abbildung 2 dargestellt. Die Textabschnitte (im Folgenden auch Module genannt) sind für die Annotation vorgegeben und müssen nicht selbst ermittelt werden. Genauere Informationen dazu sind in Abschnitt 3 zu finden.

Pro Modul sollen ein bis drei Themenbezeichner ermittelt werden. Bestimmte Module werden allerdings aufgrund ihrer Struktur bzw. ihres Inhalts bei der Annotation nicht berücksichtigt.

Bei der **Annotation nicht berücksichtigt** werden:
- das Inhaltsverzeichnis,
- Literatur- und Autorenangaben,
- Inhalte von Tabellen,
- Textsegmente, die Quellcode darstellen,
- der Anhang und
- Fußnoten.

Ebenfalls ausgeschlossen sind Texte innerhalb von Bildern. Grundsätzlich zulässig sind allerdings Themenbezeichner innerhalb von Bild- oder Tabellenüberschriften/-unterschriften.

2.1 Wie findet man Themenbezeichner?

Der erste Schritt bei der Annotation von Themenbezeichnern ist die Suche nach allen Kandidaten, die als Themenbezeichner in Frage kommen. Dazu werden also zunächst alle Substantive, einfachen Nominalphrasen und Fachtermini im Text herausgesucht (z.B. mit Textmarker hervorgehoben). In diesem Schritt werden neben dem eigentlichen Modultext auch die Überschrift sowie Bildüber- und -unterschriften mit berücksichtigt.

Wörter, Nominalphrasen oder Fachtermini, die mehr als einmal im Modul (evtl. in unterschiedlichen Kasus, Plural etc.) auftauchen, werden nur einmal in der Liste der Kandidaten berücksichtigt.

Abbildung 3: Beispiel Markierungen der Kandiadaten


[ ] [ ] [Storrer_2000] [Modul Nr. 43] [<42] [43 :: 55] [44>]

4. Überlegungen zur "Nicht-Linearität" von Hypertexten

Es war Ted Nelson, der Hypertext als nicht-lineare Schreibtechnologie ("non-sequential writing") dem "herkömmlichen" linearen Schreiben gegenüberstellte. Seine Ansicht, Hypertext befreie den Schreibenden von der Bürde der Sequenzierung, ruft jedoch bei vielen Textwissenschaftlern Skepsis und Befremden hervor, und zwar aus zwei Gründen:

- Gesprochene Sprache wird notwendigerweise in zeitlichem Nacheinander übermittelt. Die geschriebene Sprache simuliert dieses Nacheinander durch die räumliche Anordnung von Schriftzeichen entlang einer konventionell festgelegten Schriftrichtung. Geschriebene Sprache hat also - im Vergleich etwa mit Bildern und Diagrammen - zumindest im Mikrobereich, also innerhalb von schriftlich kodierten Hypertext-Modulen, eine lineare Ausrichtung. 23

- Aber auch im Makrobereich, d.h. bei der Frage der Anordnung der Module untereinander, fällt es bei vielen Textsorten schwer, sich einen Verzicht auf einen vom Autor vorgegebenen Leseweg vorzustellen. Um zwei drastische Beispiel zu wählen: Wie soll es einem nicht-sequenzierten Krimi gelingen, einen Spannungsbogen aufzubauen? Wie kann ein Witz funktionieren, bei dem es dem Rezipienten freisteht, die Pointe zuerst, mittendrin oder zuletzt zu lesen? Ein längerer Gedankengang bedarf der Entfaltung; die Kunst der Sequenzierung in argumentativen Texten, in denen es darum geht, andere von einer Position zu überzeugen, wird schon in der klassischen Rhetorik gelehrt. Die Metapher von der Textplanung als Wegplanung reicht von der Rhetorik Quintilians24 bis hin zum Züricher Textqualitätenmodell25. Untersuchungen und Modelle der Lern- und Verstehensforschung zeigen, dass die Sequenzierung von Inhalten für die Kohärenzbildung eine zentrale Rolle spielt26.

[Anfang] [<42] [43 :: 55] [44>]

Abbildung 3 zeigt ein Beispiel, in dem alle Themenbezeichner-Kandidaten eines Moduls markiert sind. Für diese Liste wird dann ermittelt, welche der Kandidaten aufgrund ihrer Relevanz für den Text als Themenbezeichner annotiert werden sollen:

Dazu werden zunächst diejenigen Kandidaten aussortiert, die nur einmal im Text erwähnt und auch nicht später wieder aufgegriffen werden – also diejenigen Kandidaten, die offensichtlich nicht den thematischen Fokus des Abschnitts darstellen können.

Auf der Grundlage dieser (evtl. bereits etwas reduzierten) Liste werden dann mit Hilfe bestimmter Heuristiken (~ Regeln) die Themenbezeichner eines Moduls ermittelt. Dabei geht man wie folgt vor:

- für das Auftauchen in einer **Überschrift** bekommt ein Kandidat **2 Punkte** auf seinem Relevanz-Kontostand gutgeschrieben;
- für **jedes (evtl. auch flektierte) Auftauchen** im normalen Modultext bekommt ein Kandidat **1 Punkt** auf seinem Kontostand gutgeschrieben;
- für jede Nennung von **Synonymen** bekommt ein Kandidat **1 Punkt** gutgeschrieben;
- für jeden **anaphorischen Verweis** bekommt ein Kandidat **0,5 Punkte** gutgeschrieben;
- für jede **lexikalisierte Referenz** bekommt ein Kandidat **0,5 Punkte** gutgeschrieben.

Auf der Grundlage dieser Punktebewertung aller Kandidaten kann dann entschieden werden, welche ein bis drei Wörter/Wortgruppen als Themenbezeichner des Moduls annotiert werden sollen.

In der Regel sind das die drei Kandidaten mit den meisten Punkten.

Allerdings müssen Themenbezeichner mindestens eine Punktzahl von 2,5 aufweisen.

Weist keiner der Kandidaten eine ausreichende Punktzahl auf, so wird für das Modul kein Themenbezeichner annotiert.

Weisen mehr als drei Wörter/Wortgruppen mehr als 2,5 Punkte auf und sind mehr als drei Wörter/Wortgruppen in der Rangliste ganz oben mit derselben Punktzahl vertreten (Beispiel: Kandidat 1 hat 4 Punkte, die Kandidaten 2-5 jeweils 2,5 à Themenbezeichner des Moduls werden die Kandidaten 1-5), so müssen für die endgültige Annotationsentscheidung zunächst folgende Fragen beantwortet werden:

- Können Kandidaten aus der Liste der rangbesten, z.B. weil sie synonym oder teilsynonym sind, zusammengefasst werden? Falls ja, werden die (teil-) synoymen Kandidaten zusammengefasst und die nun drei rangbesten Kandidaten als Themenbezeichner des Moduls annotiert.
- Sind alle rangbesten Kandidaten im Modul gleichmäßig verteilt? Gleichmäßig verteilt bedeutet, es liegen zwischen den Nennungen (auch Synonym, anaphorische Strukturen etc.) eines Kandidaten immer weniger als drei Sätze, wobei als Satzgrenzenmarkierung "." und ";" gelten. Falls ein Kandidat nicht gleichmäßig verteilt im Modul auftaucht, kann er nicht als Themenbezeichner des Moduls gelten.

Erst wenn beide Fragen beantwortet und die Rangliste entsprechend der Antworten überarbeitet wurde, kann die endgültige Annotationsentscheidung getroffen werden. Sind nun immer noch mehr als drei Kandidaten bzgl. ihrer Punktzahl gleichauf, so werden alle rangbesten Kandidaten als Themenbezeichner des Moduls bewertet – in jedem anderen Fall nur die drei besten.

Abbildung 4: Beispiel für das Zusammenfassen von Kandidaten

In einem Textmodul bildeten sich, nach oben genannter Zählung, folgende vier Kandidaten heraus:

1. Hypertext (5 Punkte)
2. lineares Schreiben (3 Punkte)
3. nicht-lineares Schreiben (3 Punkte)
4. Kohärenzen (3 Punkte)

Somit liegen in diesem Beispiel vier Kandidaten vor – das ist einer zuviel. Die Kandidaten 2 und 3 lassen sich zum Oberbegriff *Schreiben* zusammenfassen. Somit liegt eine neue Punkteverteilung vor:

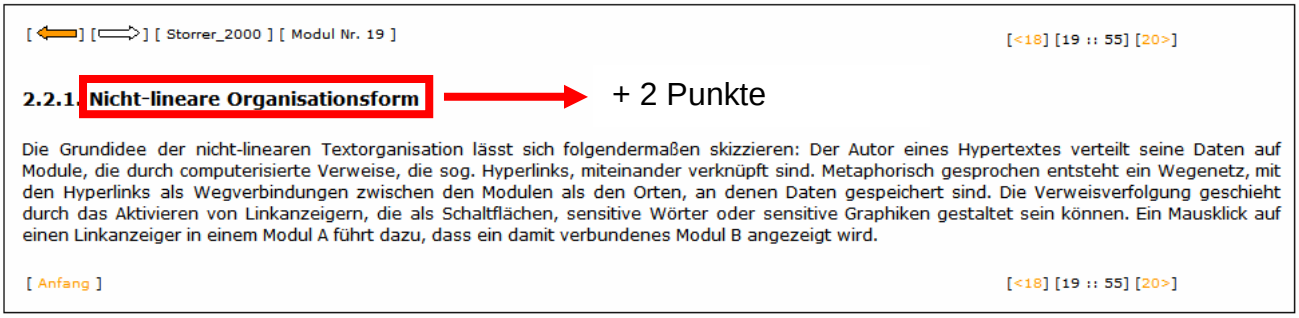
1. Schreiben (6 Punkte)
2. Hypertext (5 Punkte)
3. Kohärenzen (3 Punkte)

2.2 Der Umgang mit Überschriften

Substantive, Nominalphrasen und Fachtermini aus Modulüberschriften können als Themenbezeichner fungieren. In unseren bisherigen Experimenten bisher haben wir die Erfahrung gemacht, dass Wörter und Wortgruppen in einer Überschrift häufig besonders relevant sind. Daher haben wir uns entschieden, diese besonders stark zu gewichten. Substantive, Nominalphrasen und Fachtermini, die als Themenbezeichner-Kandidaten identifiziert wurden, werden mit 2 Punkten gewichtet.

Wörter und Wortgruppen aus einer Überschrift dürfen aber nur dann als Themenbezeichner annotiert werden, wenn sie mindest ein weiteres Mal im eigentlichen Modultext genannt wurden (entweder wörtlich, mit einem synonymen Ausdruck oder als lexikalisierte Referenz).

Abbildung 5: Beispiel Kandidat in Überschrift

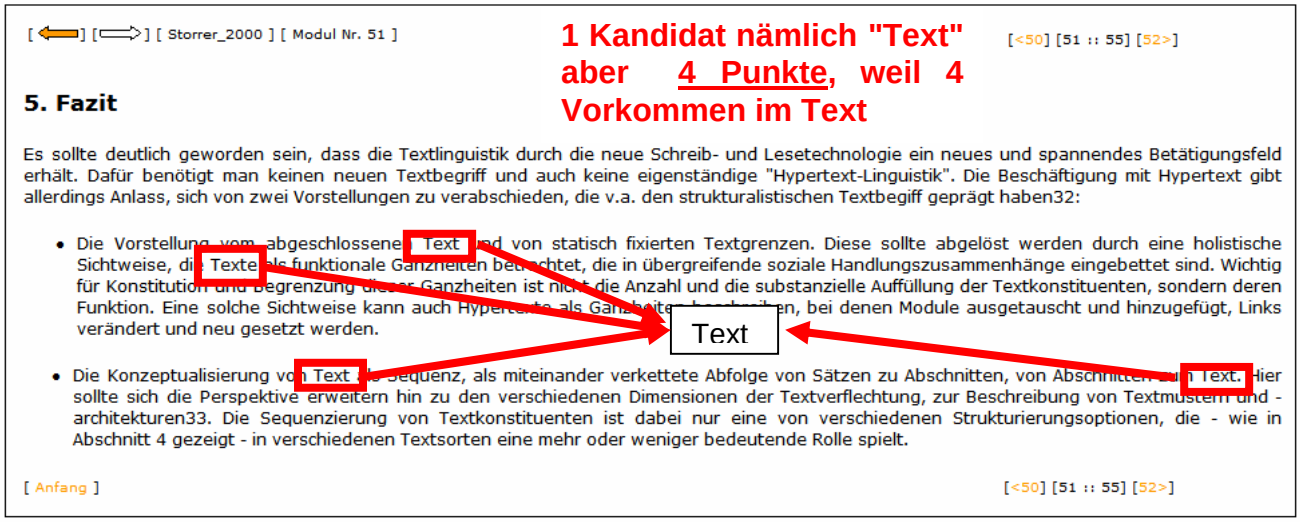


2.3 Der Umgang mit flektierten Formen

Substantive, Nominalphrasen und Fachtermini können flektiert (Kasus, Numerus) in den Modultexten vorkommen. Flektierte Formen werden jedoch nicht als unterschiedliche Themenbezeichner-Kandidaten behandelt. Alle verschiedenen

Formen – der Kandidat in jedem Kasus, Numerus – werden als zu einem Kandidat gehörend bewertet. Für jede Nennung wird daher dem Kandidaten 1 Punkt auf seinem Konto gutgeschrieben.

Abbildung 6: Beispiel flektierte Form



2.4 Der Umgang mit synonymen Formen

Häufig wird – besonders aus stilistischen Gründen – in einem Text die Formulierung bzw. die Benennung für Konzepte variiert. Dazu werden insbesondere synonyme Formen, Abkürzungen bzw. expandierte Abkürzungen oder in Fachtexten auch Übersetzungen (à "fremdsprachige Synonyme") etc. verwendet. Diese Formen erhöhen die Bewertung eines Kandidaten um jeweils 1 Punkt. Einige Beispiele hierfür sind in Abbildung 7, Abbildung 8 und Abbildung 9 dargestellt.

Abbildung 7: Beispiel synonyme Form

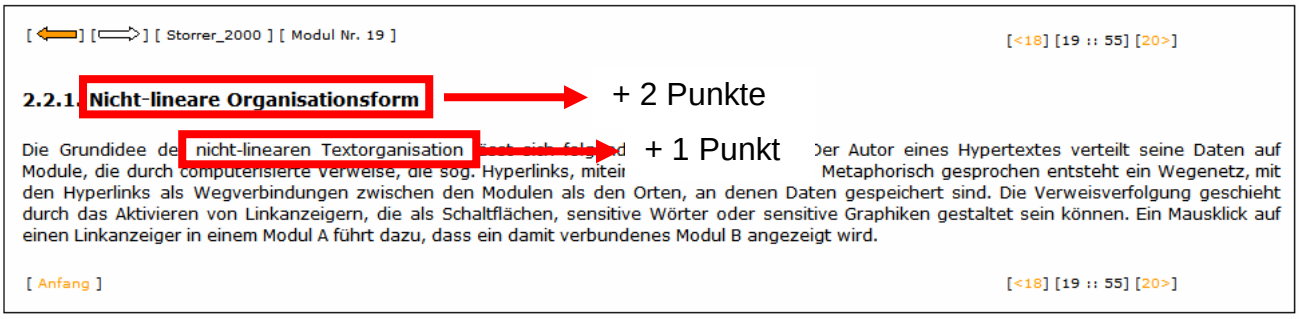


Abbildung 8: Beispiel Abkürzung

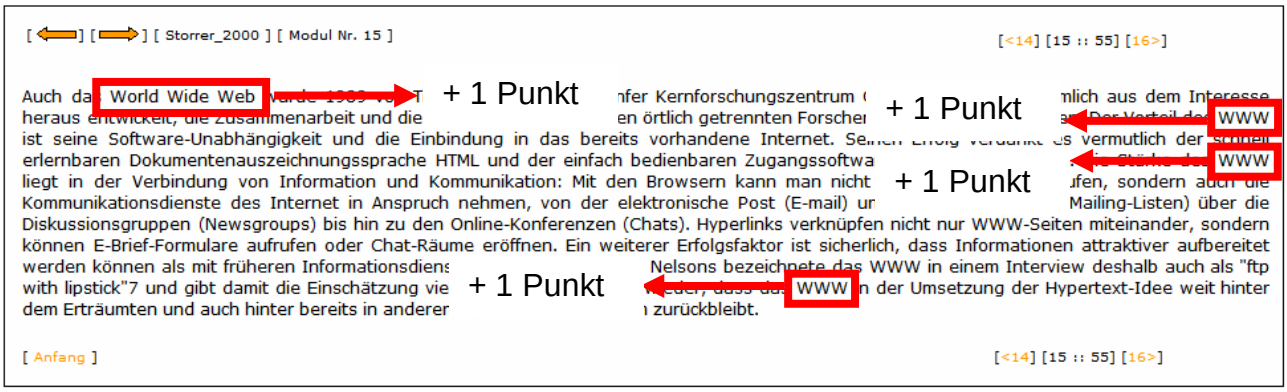
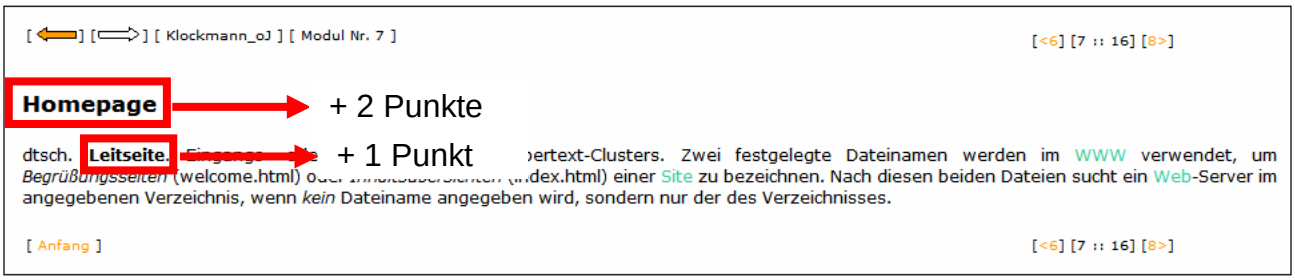


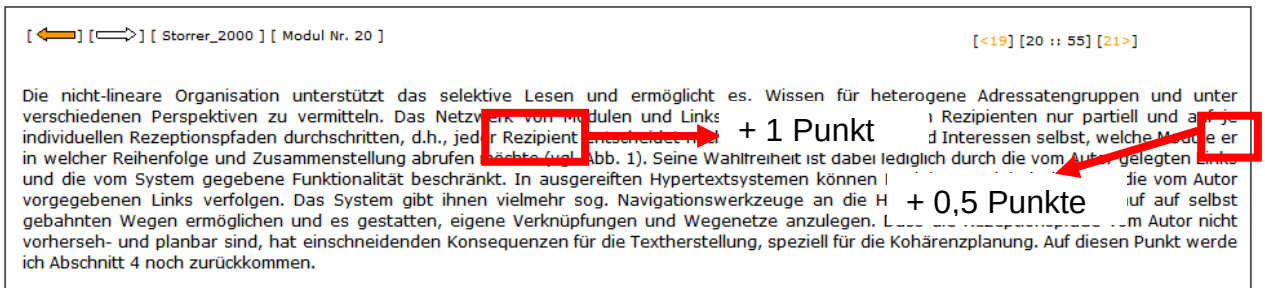
Abbildung 9: Beispiel Übersetzung



2.5 Der Umgang mit anaphorischen Verweisen

Aus den verschiedenen Varianten für die Realisierung von anaphorischen Verweisen haben wir nur eine kleine Menge ausgewählt, die in den vorliegenden Annotationsrichtlinien berücksichtigt werden sollen. Generell liegt ein anaphorischer Verweis vor, wenn ein Satzteil auf einen anderen, vor ihm im Text stehenden Satzteil verweist. In der Annotation von Themenbezeichnern werden nur solche anaphorischen Verweise berücksichtigt, die über Pronomen/Proformen realisiert werden. Anaphorische Verweise auf Themenbezeichner-Kandidaten werden, wie im Beispiel in Abbildung 10 dargestellt, mit 0,5 Punkten bewertet.

Abbildung 10: Beispiel anaphorische Struktur



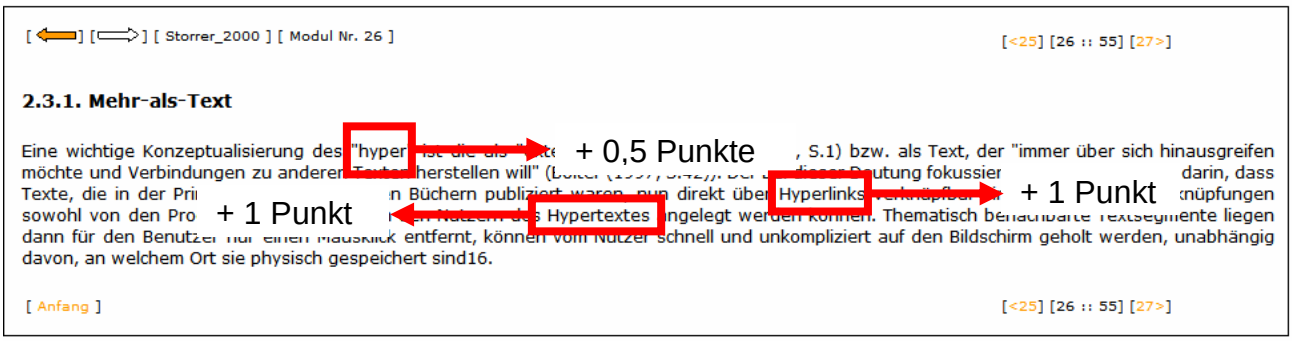
2.6 Der Umgang mit lexikalisierter Referenz

Neben den anaphorischen sollen auch solche Verweise in der Annotation von Themenbezeichnern berücksichtigt werden, die sich lexikalisch manifestieren. Hierbei gibt es zwei grundsätzliche Möglichkeiten:

- Ausgehend von einer Nominalphrase als Themenbezeichner-Kandidat liegt ein lexikalischer Verweis dann vor, wenn im Text später ein Teil der Nominalphrase – etwa ein Adjektiv – wieder aufgegriffen wird.
- Ausgehend von einem Kompositum als Themenbezeichner-Kandidat liegt ein lexikalischer Verweis dann vor, wenn im Text später ein Teil des Kompositums – etwa Kopf des Kompositums – wieder aufgegriffen wird.

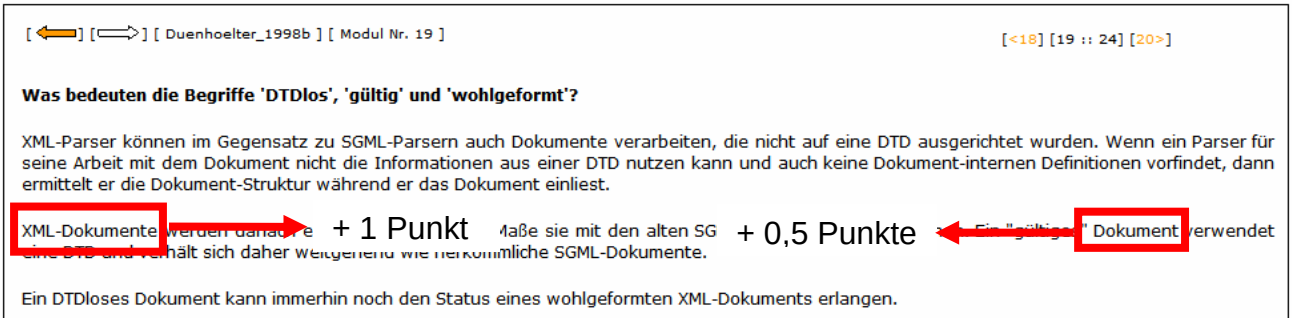
Für jede lexikalisierte Referenz werden dem Kandidaten 0,5 Punkte gutgeschrieben. Dabei spielt es keine Rolle, welcher der beiden oben angesprochenen Fälle vorliegt. Beispiele hierfür sind in Abbildung 11 und Abbildung 12 dargestellt.

Abbildung 11: Beispiel lexikalische Referenz – Adjektiv



Das Beispiel zeigt, dass "hyper" sowohl den Punktestand von "Hyperlinks" als auch den von "Hypertextes" um 0,5 Punkte erhöht.

Abbildung 12: Beispiel lexikalische Referenz – Kompositum



2.7 Sonderfall: Verweise auf Bücher und Kapitel

Manchmal wird in den Modulen thematisch auf andere Module oder sogar Texte außerhalb der hier bearbeiteten Textsammlung verwiesen. Kandidaten, für die lediglich auf andere Textsegmente verwiesen wird, können natürlich nicht als Themenbezeichner des aktuellen Moduls bewertet werden. Einige Beispiele für Formulierungen sind in Abbildung 13 dargestellt.

Abbildung 13: Beispielformulierungen für Verweise auf Bücher und Kapitel

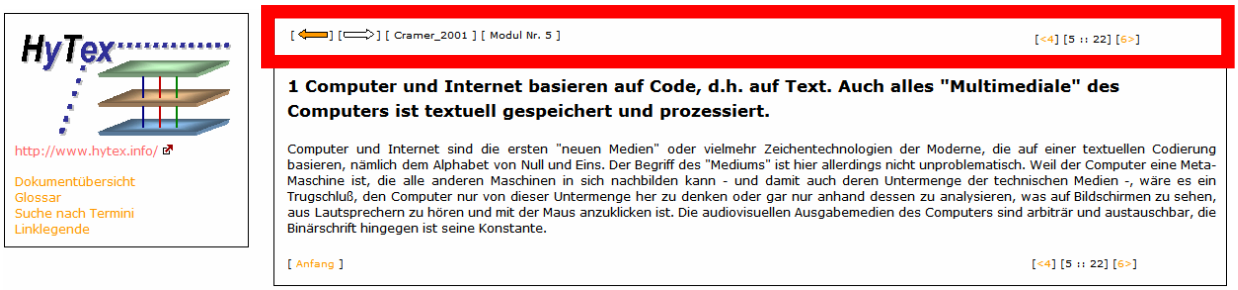
Kapitel ... erklärt/erläutert/beschreibt … in Kapitel … wird … erklärt/erläutert im folgenden Abschnitt werde ich auf … näher eingehen … werde ich in … erläutern/erklären/darlegen

In unseren Experimenten hat sich gezeigt, dass bestimmte Formulierungen einen klaren Hinweis darauf geben, dass ein vorliegendes Substantiv, eine Nominalphrase oder ein Fachterminus lediglich genannt, nicht aber an der aktuellen Textstelle wirklich erläutert werden soll. Im Zusammenhang mit den in Abbildung 13 abgebildeten und ähnlichen Formulierungen kommen in der Regel keine Themenbezeichner vor.

3. Technische Umsetzung der Annotation

Nachdem die Themenbezeichner eines Moduls wie oben beschrieben ermittelt wurden, müssen sie für die weitere, z. T. automatische Verarbeitung in ein entsprechendes Dateiformat überführt werden. Dazu werden alle Themenbezeichner eines Textes (also alle Themenbezeichner aller Module des gesamten Textes) in eine baumartige Struktur in einer txt-Datei (einfache Textdatei)[1] abgelegt werden.

Abbildung 14: Screenshot – Modulseite
(Markierung zeigt Textkennung und Navigationsleiste des HyTex-Prototyps)



Für die Annotation wird der HyTex-Demo-Prototyp verwendet. Diesen findet man unter

http://www.hytex.uni-dortmund.de/HyTex-Demo/output/Texte/index-htversioneval.externalLinksfalse.html

Abbildung 14 zeigt einen Screenshot des Prototyps. Module im Sinn der hier vorliegenden Annotationsrichtlinien sind die Module des Prototyps, ein solches Modul ist in Abbildung 14 dargestellt. Der Bereich, der anzeigt, in welchem Modul man sich aktuell befindet, ist in Abbildung 14 rot markiert. Diese Information muss zusätzlich zu den ein bis drei Themenbezeichnern pro Modul angegeben werden (vgl. Annotationsstruktur weiter unten).

[1] Eine einfache Textdatei kann z.B. mit WordPad, emacs oder jedem anderen Texteditor bearbeitet werden. Ungünstig aber zulässig ist auch das doc-Format.

Zur eindeutigen Identifikation der Annotation muss jede txt-Datei nach dem folgenden Schema benannt werden.

```
Textkennung_JJ-MM-TT.txt
```

Dabei bezeichnet:

- Textkennung die Textkennung im HyTex-Korpus,
- JJ das Jahr,
- MM den Monat und
- TT den Tag der Annotation.

Ein Beispiel für einen derartigen Dateinamen (vgl. auch Abbildung 16) ist

```
Arnold_2001_07-12-11.txt
```

Die Themenbezeichner für einen kompletten Text (also alle Themenbezeichner für alle Module) werden in der so benannten Datei nach einer bestimmten Struktur notiert, diese ist in Abbildung 15 dargestellt.

Abbildung 15: Schema der Annotation

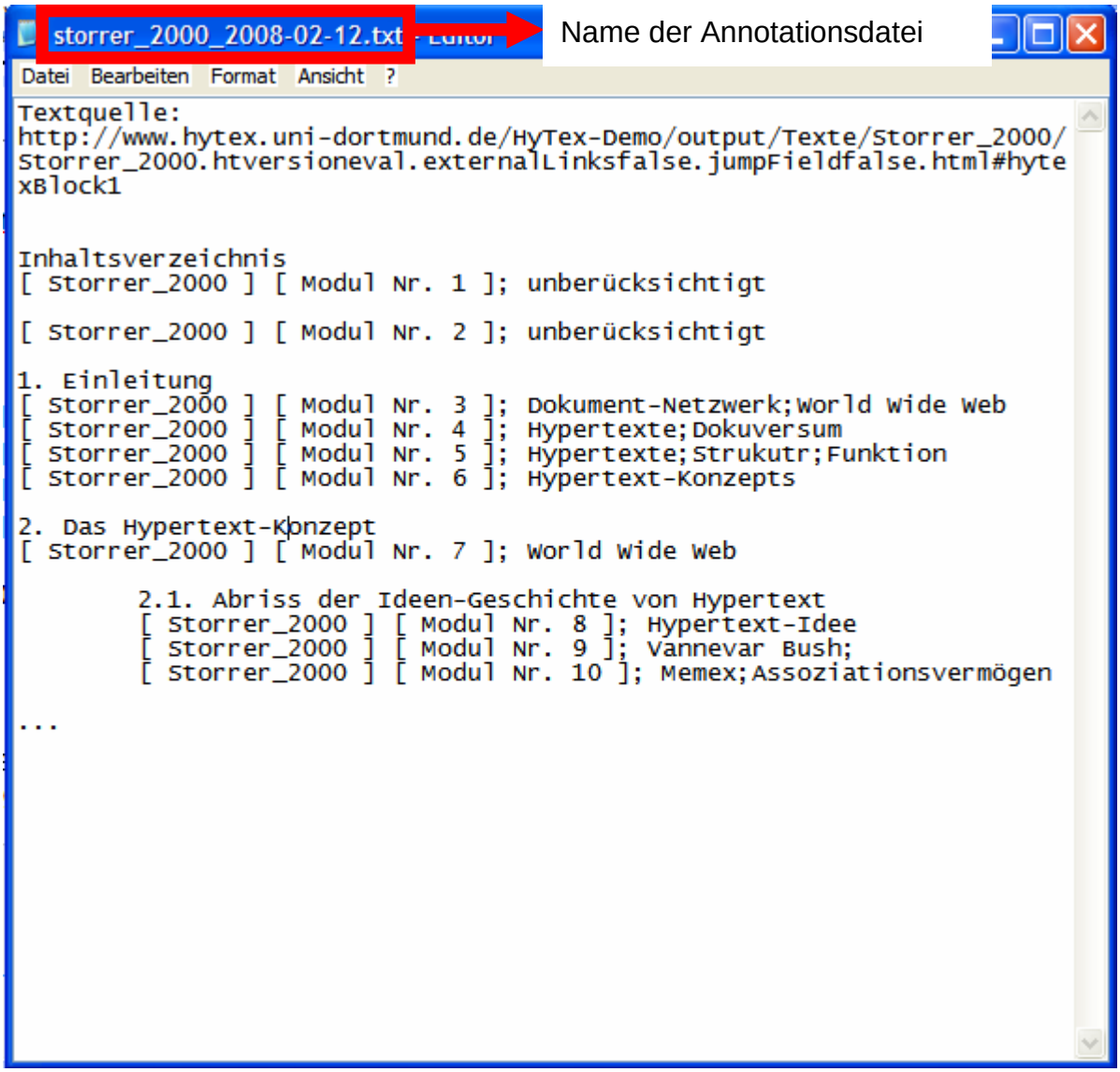
```
Textquelle:¶
Zieladresse¶
¶
¶
Modulüberschrift1¶
[ Textkennung ][ Modulnummer ]; TB1;TB2; ... TBn¶
[ Textkennung ][ Modulnummer ]; TB1;TB2; ... TBn¶
¶
->|     Modulüberschrift1.1¶
->|     [ Textkennung ][ Modulnummer ]; TB1;TB2; ... TBn¶
->|     [ Textkennung ][ Modulnummer ]; TB1;TB2; ... TBn¶
->|     [ Textkennung ][ Modulnummer ]; TB1;TB2; ... TBn¶
->|     [ Textkennung ][ Modulnummer ]; TB1;TB2; ... TBn¶
¶
->|     ->|    Modulüberschrift1.1.1¶
->|     ->|     [ Textkennung ][ Modulnummer ]; TB1;TB2; ... TBn¶
->|     ->|     [ Textkennung ][ Modulnummer ]; TB1;TB2; ... TBn¶
¶
->|     Modulüberschrift2¶
->|     [ Textkennung ][ Modulnummer ]; TB1;TB2; ... TBn¶
->|     [ Textkennung ][ Modulnummer ]; TB1;TB2; ... TBn¶

(...)
```

Dabei bezeichnet:

- Zieladresse die Zieladresse des ersten Moduls des zu annotierenden Textes,
- ¶ einen Zeilenumbruch,
- Modulüberschrift die Überschrift eines Moduls,
- Textkennung die Textkennung im HyTex-Korpus,
- Modulnummer die Modulnummer des annotierten Moduls,
- $TB_1;TB_2;$ ... TB_n die einzelnen Themenbezeichner, die durch Semikola (ohne Leerzeichen) getrennt werden und
- ->| einen Tabulatorvorschub.

Ist für ein Modul kein Themenbezeichner zu annotieren, beispielsweise für das Inhaltsverzeichnis, so wird an Stelle eines Themenbezeichners "unberücksichtigt" vermerkt. Eine Datei könnte somit aussehen wie in Abbildung 16 dargestellt.

Abbildung 16: Annotationsbeispiel



4. FAQ

Was ist mit "gleichmäßiger Verteilung" gemeint? (Vgl. 2.1)

Tritt ein Kandidat zu Anfang eines Moduls in Erscheinung und wird gegen Ende eines Moduls noch ein-/zweimal genannt, so kann man sagen, dass dieser unregelmäßig im Modul verteilt ist. Sind die weiteren Kandidaten nicht so stark verteilt, so ist das ein Indiz dafür, dass es sich bei diesem

ungleichmäßig vorkommenden Kandidaten nicht um einen Themenbezeichner handelt.

Ich habe mehr als drei Kandidaten in einem bestimmten Modul gefunden, die sich nicht reduzieren lassen. Was ist zu tun?

Dann müssen alle Kandidaten annotiert werden.

Muss ich die Textquellenangabe in der Annotationsdatei abschreiben oder gibt es einen einfacheren Weg diese zu übernehemen?

Ja, den gibt es: Man muss mit den Mauscursor auf den Link des zu annotierenden Dokuments plazieren und dann mit der rechten Maustaste auf diesen klicken. Nun wählt man im erschienenen Menü "Verknüpfung kopieren" (IE) oder "Link-Adresse kopieren" (Opera). Bei Firefox reicht der Rechtsklick. Nun kann man die Adresse in der Textdatei an der Stelle einfügen (Über das Menü Bearbeiten – Einfügen).

Anhang

Liste der Fachtermini aus TermNet

1-Anker-Link	one-to-one link	1:1-Beziehung	1:1-Verknüpfung	1:1-Verweis
1:1-Relation	1:1-Link	1:n-Link	1:n-Verweis	1:n-Beziehung
1:n-Verknüpfung	one-to-many link	1:n-Relation	Attribut vom Typ actuate	Attribut actuate
actuate-Attribut	Adressangabe-Attribut	Anker	Annotierung	gloss
Anmerkung	Annotation	Glossierung	annotation	SGML-Architekturen-Definition
	ttribut arcrole	arcrole-Attribut	Attribut vom Typ arcrole	Beteiligter einer Assoziation
association member	Mitspieler einer Assoziation	association role player	Assoziationsrolle	association role
Beteiligtenrolle	Mitspielerrolle	semantische Rolle	association role type	Assoziationsklasse
association type	Assoziationstyp	association	Assoziation	Attribut zur Typfestlegung
Ausgangsanker	Start-Ressource	Verknüpfungsstartpunkt	ausgehende Kante	Markup
Auszeichnung	markup	authoring tool	Autorenwerkzeug	Basisname
Grundform des Topic-Namen	base name	bidirektionaler Verweis	bidirektionaler Link	bidirektionale Relation
bidirektionale Verknüpfung	Brotkrumen	bread crumb	Browser	Internet-Browser
Web-Browser	Webbrowser	Internetbrowser	CSS	Cascading Style Sheet
CSS-Standard	history	Dialoghistorie	Chronik	Cross-Media-Publishing
Multiple-Media-Publishing	Datenfilter	datenorientierter Filter	deduktiver Link	deduktive Relation

deduktive Verknüpfung	deduktiver Verweis	Document Object Identifier	DOI	Dokument
Dokumentinstanz	Dokument-Instanz	Dokumenttyp-Definition	DTD	dynamischer Pfad
dynamischer Hypertrail	E-Text	einfacher Link	simple link	simple Link
einfaches Dokumenten-Modul	eingebetter Link	embedded link	eingehende Kante	Tag
tag	Element	Element vom Typ arc	arc-Element	extended-Element
Element vom Typ extended	locator-Element	Element vom Typ locator	resource-Element	Element vom Typ resource
Element vom Typ simple	simple-Element	Element vom Typ title	title-Element	entfernte Ressource
extended link	erweiterter Link	XML-Standard	XSL-Standard	extensional definierter Verweis
extensional definierte Verknüpfung	extensional definierter Link	externer Verweis	extrahypertextuelle Relation	extrahypertextueller Verweis
externer Link	externe Verknüpfung	extrahypertextueller Link	extrahypertextuelle Verknüpfung	externe Relation
Filter	Filterung	Filterwerkzeug	Forced March	Attribut from
Attribut vom Typ from	from-Attribut	geführte Tour	guided tour	globaler Anker
Attribut href	href-Attribut	Attribut vom Typ href	HTML	HyperText Markup Language
Hypertext Markup Language	HTML-Standard	HTML-Dokument	HTML-Element	Hyperdokument
Hypertext	Hypertextbasis	Hypertextdokument	Hypertextsystem	individueller locator
Informationsmodellierung	inhaltliche Kontextualisierungshilfe	Inhalts-Link	Inhaltslink	intensional definierte Verknüpfung
intensional definierter Link	intensional definierter Verweis	interhypertextuelle Verknüpfung	interhypertextueller Verweis	interhypertextuelle Relation
interhypertextueller Link	interne Verknüpfung	intrahypertextueller Link	intrahypertextuelle Relation	interner Verweis
intrahypertextuelle Verknüpfung	interne Relation	intrahypertextueller Verweis	interner Link	unsichtbarer Link
verborgener Link	invisible Link	invisible link	SGML-Standard	ISO-Standard

	opic Maps	HyTime-Standard	Kante	arc
Index	Katalog	Komponente	konforme XLink-Anwendung	XLink-konforme Anwendung
Kontextualisierungshilfe	konzeptionelle Ebene	Attribut vom Typ label	Attribut label	label-Attribut
leeres Element	leeres XML-Element	Favorit	bookmark	Bookmark
Lesezeichen	Beziehung	Hyperlink	Relation	Verknüpfung
link	Link	Verweis	hyperlink	Hyper-Link
Verweis mit eingebetteter Anzeige	Relation mit eingebetteter Anzeige	Verknüpfung mit eingebetteter Anzeige	Link mit eingebetteter Anzeige	Verknüpfung mit ersetzender Anzeige
Relation mit ersetzender Anzeige	Verweis mit ersetzender Anzeige	Link mit ersetzender Anzeige	Verknüpfung mit paralleler Anzeige	Verweis mit paralleler Anzeige
Relation mit paralleler Anzeige	Link mit paralleler Anzeige	Link-Datenbank	Linkdatenbank	Linkbasis
link base	Link-Bank	Linkbase	Link-Etikett	Linktitel
link title	Linketikett	Link-Titel	Link-Explikation	Link-Kennzeichnung
Linkanzeiger	Verweisanzeiger	Link-Button	Button	Verknüpfungsanzeiger
Linking-Muster	lokale Ressource	lokaler Anker	Medienobjekt	Meta-DTD
Meta-Sprache	Metadaten	metadatenorientierter Filter	Metafilter	Metasuchmaschine
Knoten	Hypertextmodul	informationelle Einheit	Seite	Informationseinheit
Hypertext-Modul	Modul	multidirektionaler Link	n:m-Relation	n:m-Link
many-to-many link	n:m-Beziehung	n:m-Verweis	n:m-Verknüpfung	Namensraum
namespace	Namespace	Navigationshilfe	navigatorische Kontextualisierungshilfe	nicht traversierbarer Link
nicht valides XML-Dokument	nicht valides Dokument	nicht-valide Instanz	nicht valide Dokumentinstanz	nicht-valides Dokument
nicht-valides XML-Dokument	nicht valide Instanz	nicht-valide Dokumentinstanz	objektiver Link	Topik-Anker
occurrence	topic occurrence	Vorkommensangabe	Topic-Anker	Vorkommensangaben-Typ
Topic-Anker-Typ	occurrence type	Orientierungshilfe	OWL	Web Ontology Language

OWL-Standard	Pfad	trail	Trail	Hypertrail
hypertrail	Portal	Webportal	Web-Portal	pragmatischer Link
Präsentations- und Interaktionsebene	PSI	public subject descriptor	published subject indicator	Ranking
ranking	RDF-Standard	referentielle Verknüpfung	referential link	referentieller Verweis
referentieller Link	referenzieller Link	referenzielle Verknüpfung	referenzieller Verweis	unspezifizierter Link
Resource Description Framework	RDF	Ressource	resource	retrospektive Hilfe zur Navigation
retrospektive Hilfe	role-Attribut	Attribut vom Typ role	Attribut role	backtracking
Backtracking	Rücksprung	Gültigkeitsbereich	Skopus	scope
semantic attribute	semantisches Attribut	semantic web	semantisches Web	separat angeordneter Link
Standard Generalized Markup Language	SGML	SGML-Anwendung	SGML-Deklaration	sgml declaration
SGML-Dokument	SGML-Element	Attribut vom Typ show	show-Attribut	Attribut show
Sicht	Sichtenmodul	Sichten-Modul	Sichtenknoten	Sichten-Knoten
site	Site	Speicherebene	statischer Hypertrail	statischer Pfad
Strukturlink	Struktur-Link	subjektiver Link	Template	texttechnologischer Standard
Standard	Themenbereich	Thematik	theme	third party arc
Third-Party-Kante	title-Attribut	Attribut title	Attribut vom Typ title	Attribut to
Attribut vom Typ to	to-Attribut	topic	Topik	Topic
Topic-Charakteristik	Topic-Merkmal	topic characteristic	Charakteristik	Topik-Merkmal
Topik-Charakteristik	Topic Map	topic map	Topic-Map	Topic-Name
Topik-Name	topic name	Topik-Typ	Topic-Typ	topic type
Übergang	Traversierung	Traversierungsattribut	Attribut type	type-Attribut
Attribut vom Typ type	getypter Verweis	typisierter Link	getypter Link	typisierte Verknüpfung
typisierter Verweis	getypte Verknüpfung	typed link	Linktypisierung	link typing
Typisierung von Links	Typisierung	Link-Typisierung	Überblickshilfe	unidirektionale

				Verknüpfung
unidirektionale Relation	unidirektionaler Link	unidirektionaler Verweis	URI reference	URI-Referenz
Adressangabe	Uniform Resource Identifier Reference	valides XML-Dokument	valide Instanz	valide Dokumentinstanz
valides Dokument	Nebenform des Topic-Namen	Variantenname	variant name	Verhaltensattribut
W3C-Standard	Waise	orphan page	Waisen-Seite	Waisenseite
wohlgeformte Dokument-Instanz	wohlgeformtes XML-Dokument	wohlgeformtes Dokument	WWW	Web
World Wide Web	XHTML	XHTML-Standard	XML Linking Language	XLink
XLink-Anwendung	XLink-Attribut	XLink-Elementtyp	XLink-Element	XLink-konformes XML-Element
XLink-Link	XLink-Link-Element	Extensible Markup Language	eXtensible Markup Language	XML
XLink-Standard	XML Linking Languages	XPath-Standard	XPointer-Standard	XML Topic Maps
XTM	Topic-Map-Standard	XML-Anwendung	XML-Attribut	XML-Dokument
XML-Element	XPath	XML Path Language	XPointer	XML Pointer Language
Extensible Stylesheet Language	eXtensible Stylesheet Language	XSL	Verknüpfungszielpunkt	Zielanker
End-Ressource	zusammengesetztes Modul	zusammengesetzter Knoten		