Michael Beißwenger, Angelika Storrer, Maren Runte

Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet

Abstract

Dieser Beitrag skizziert die Konzeption eines im Projekt „Hypertextualisierung auf textgrammtischer Grundlage“ (*HyTex*) modellierten terminologischen Wortnetzes (*TermNet*) zu den Fachtextdomänen Texttechnologie und Hypermedia. Schwerpunkt des Beitrags ist es zum einen, die Modellierung von *TermNet* in Hinblick auf fachsprachen- und domänenspezifische Merkmale vorzustellen, und zum anderen, die Anwendung von *TermNet* für die Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen zu erörtern.

1 Projektrahmen und Motivation des Ansatzes

Unter *Hypertextualisierung* versteht man die Aufbereitung von Dokumenten für die selektiven, interaktiven Nutzungsformen in einem Hypertextsystem, z.B. dem World Wide Web. Das Projekt *HyTex*[1] (*Hypertextualisierung auf textgrammatischer Grundlage*) sucht nach textgrammatisch geleiteten Verfahren für eine hypertextadäquate Aufbereitung selektiv organisierter Fachtexte (wissenschaftliche Artikel, technische Spezifikationen), d.h. eine Aufbereitung, die hypertexttypischen selektiven Rezeptionsformen optimal entgegenkommt. Auf der technischen Seite benötigt man für diese Aufgabe Konversionstools; auf der konzeptionellen Seite benötigt man Strategien und Verfahren für die folgenden beiden Teilaufgaben der Hypertextualisierung:

- *Segmentierung* (Zerlegung der Dokumente in Module).
- *Linking* (Verknüpfung der Module durch Hyperlinks).

Für die in *HyTex* entwickelte textgrammatisch geleitete Herangehensweise an diese Aufgaben gibt es zwei Leitlinien: (a) *Reversibilität* und (b) *Hypertextualisierung nach Kohärenzkriterien*.

Ad a): Reversibilität bedeutet, dass wir Hypertext-Sichten auf lineare Dokumente als zusätzliche Sichten generieren, die regelgeleitet aus textgrammatischem Markup und anderen Wissensquellen – z.B. aus dem in diesem Papier beschriebenen Terminologienetz – abgeleitet werden. Die sequentielle Struktur und der Originalwortlaut eines Dokuments bleiben dabei als eine mögliche Sicht auf das Dokument erhalten.[2] Damit geben wir dem Rezipienten die Möglichkeit, einen Text in der ursprünglichen linearen Form und Abfolge zu rezipieren, wenn er die Zeit dazu hat; die Hypertextsichten sind als zusätzliche Angebote für den eiligen Querleser gedacht.

Ad b): Das Ziel der Hypertextualisierung in unserem Ansatz ist es, Kohärenzbildungsprozesse beim selektiven Querlesen besser zu unterstützen als dies in Printmedien möglich ist und damit das Mehrwertpotenzial von Hypertexten auszureizen. Im Hinblick auf diese Zielsetzung spielen bei der Segmentierung und beim Linking *Kohärenzkriterien* eine zentrale Rolle. *Hypertextualisierung nach Kohärenzkriterien* ist eine Strategie, die Rainer Kuhlen (Kuhlen 1991) einer Strategie gegenüberstellt, die als „Hypertextualisierung nach formalen Texteigenschaften“ bezeichnet wird. Bei der Hypertextualisierung nach formalen Texteigenschaften erfolgt die Segmentation ausschließlich anhand der typographisch angezeigten Unterteilung in Kapitel, Unterkapitel und Abschnitte. Diese werden dann in Nachbildung der hierarchischen Dokumentenstruktur wieder durch Links verknüpft, d.h. die

Teil-Ganzes-Bezüge zwischen Kapiteln, Unterkapiteln und Abschnitten werden als Links nachgebildet und mit einem Inhaltsverzeichnis auf der Einstiegsseite verlinkt. Zusätzlich wird häufig ein Lesepfad gelegt, der in einer Tiefe-vor-Breite-Strategie auf genau demjenigen Weg durch den Hypertext führt, der der Abfolge im gedruckten Pendant entspricht. Der in *HyTex* verfolgte Ansatz hingegen legt den Schwerpunkt bei der Strategiebildung auf textgrammatisch geleitete, verfeinerte Segmentations- und Linkingtechniken, die die Kohärenzbildung des selektiv und quer lesenden Nutzers optimal unterstützen. Eine wichtiger Strategietyp ist dabei das sog. *Linking nach Wissensvoraussetzungen*, welches darauf abzielt, mit automatischen Verfahren Links zu genau denjenigen Textsegmenten zu generieren, deren Inhalte für das Verständnis des von einem selektiv zugreifenden Hypertextrezipienten aktuell rezipierten Moduls benötigt werden[3].

Hierbei orientieren wir uns an einem Szenario, das wir als *Hypertext-Rezeptionsumgebung für Fachtexte* bezeichnen. Das Szenario ist zugeschnitten auf Nutzungssituationen, in welchen sich ein Nutzer unter Zeitdruck und mit einer ganz speziellen Zielsetzung Wissen zu einem Fachgebiet erarbeiten muss, für welches er zwar bereits Wissensvoraussetzungen mitbringt, in dem er aber kein Experte ist. Beispiele, in denen solche Situationen auftreten, sind interdisziplinäre Projektarbeit, Wissenschaftsjournalismus, Fachlexikographie, sowie interdisziplinäres Arbeiten in Studium und Weiterbildung. Ganz unabhängig von WWW und Hypertext lesen Nutzer in solchen Situationen quer und partiell und rezipieren nur selektiv Teiltexte. Hypertextsichten kommen dieser Rezeptionsform nun prinzipiell entgegen, indem sie längere Dokumente bereits in modularisierter Form präsentieren und darin verschiedene Such- und Navigationsoptionen zur Verfügung stellen. Werden sequentiell organisierte Texte aber nur nach formalen Texteigenschaften hypertextualisiert, so besteht die Gefahr, dass dem selektiven Querleser wichtige Voraussetzungen für das korrekte Verständnis eines aktuell rezipierten Textausschnitts fehlen; schließlich sind die Teiltexte der einzelnen Module ja weiterhin auf die Ganzlektüre auf einem vorgebenen Leseweg hin formuliert. Dieses Kohärenzproblem bei der Hypertextrezeption[4] soll der bereits benannte Strategietyp eines „Linkings nach Wissensvoraussetzungen“ durch Generierung von Links und zusätzlichen Sichten (z.B. der Glossarsicht, vgl. Abschnitt 4) kompensieren. Die hierbei entwickelten Strategien verarbeiten Informationen aus drei verschiedenen Ebenen, die in Abb. 1 visualisiert sind und die sich folgendermaßen skizzieren lassen:

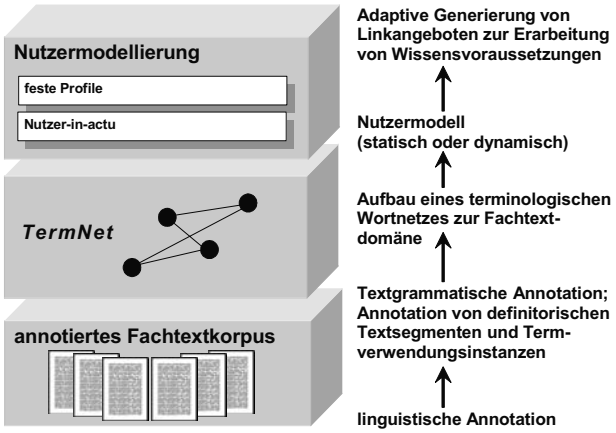

Abbildung 1: Der 3-Ebenen-Ansatz in HyTex

Auf der *Dokumentenebene* werden deutsche Fachtexte zum Thema „Texttechnologie“ und „Hypermedia“ textgrammatisch annotiert. Unser Korpus umfasst insgesamt 96 Dokumente, wobei neben Fachartikeln und normativ („Spezifikation“) oder didaktisch motivierten Dokumenttypen (z.B. „Tutorial“, „Einführung“, „Überblicksdarstellung“) auch diskursiv geprägte Textsorten wie FAQs, Mailinglist- oder Foren-Postings und Chat-Protokolle berücksichtigt sind. Das Korpus wird in Kooperation mit dem Tübinger *DEREKO*-Projekt automatisch linguistisch annotiert (Lemmatisierung, POS-Tagging, Chunk-Parsing). Auf dieser linguistischen Annotation aufsetzend erfolgt dann die textgrammati-

sche Annotation, d.h. die Auszeichnung von Koreferenzstrukturen, von rhetorischen und thematischen Strukturen und – für das im Beitrag fokussierte Thema besonders relevant – von Termverwendungsinstanzen und definitorischen Textsegmenten. Die Annotationen auf dieser Ebene erfassen also sowohl Aspekte der syntaktischen Kohärenz (z.B. Koreferenzbezüge zwischen Textelementen) als auch Aspekte der durch thematische und rhetorische Muster geleiteten globalen semantischen Textkohärenz.

Auf der *Ebene der Nutzermodellierung* werden in der laufenden Projektphase fixe Nutzerprofile verarbeitet; geplant ist im weiteren Verlauf aber auch die Verarbeitung von Nutzungsprotokollen, aus denen sich Anhaltspunkte über den aktuellen Wissensstand und den aktuellen Aufmerksamkeitsbereich des Nutzers ergeben können.

Im Mittelpunkt dieses Beitrags steht die *Ebene der Modellierung von terminologischem Wissen zu grundlegenden Konzepten und Lexemen der im Korpus repräsentierten Fachdomäne*[5]. Auf dieser Ebene nutzen wir eine erweiterte Form des in *WordNet* für semantisch-lexikalische Wortnetze entwickelten Beschreibungsmodells für die Modellierung einerseits zentraler Konzepte und andererseits der (terminologisch eingeführten) Lexeme, mit denen diese Konzepte in verschiedenen Publikationen und von verschiedenen Autoren versprachlicht werden, für das Linking und für die Generierung von mit den Dokumenten verlinkten Glossarsichten. Im folgenden Abschnitt werden wir zunächst auf einige verwandte Ansätze eingehen, um dann die theoretischen Grundlagen der Modellierung unseres terminologischen Netzes (*TermNet*) zu erläutern. Abschließend werden wir anhand eines Anwendungsbeispiels skizzieren, wie *TermNet* im Zusammenspiel mit den anderen der in Abb. 1 skizzierten Ebenen zur Hypertextualisierung eingesetzt wird.

2 Wortnetze und Ontologien für das Linking

Bislang wurden Wortnetze vor allem für statistisch basierte Strategien zur Hypertextualisierung eingesetzt. Der in Green (Green 1998) beschriebene Ansatz nutzt das englische *WordNet* (vgl. Fellbaum 1998) als lexikalische Ressource zur vollautomatischen Hypertext-Konversion. *WordNet* dient dabei vor allem dazu, aus den Ausgangstexten lexikalische Ketten („lexical chains"), d.h. Folgen von semantisch verwandten Textwörtern, zu extrahieren und dabei ggf. mehrdeutige Lesarten zu disambiguieren. Als Grundlage für das vollautomatische Linking dient die Berechnung der relativen Wichtigkeit der extrahierten Ketten für das jeweilige Textmodul (Paragraph) und ein Berechnungsmaß für die Ähnlichkeit zwischen Modulen, das auf der Ähnlichkeit der jeweiligen lexikalischen Ketten beruht. Überschreitet das Ähnlichkeitsmaß einen bestimmten Schwellenwert, so werden die entsprechenden Module verlinkt. Durch ein ähnliches Verfahren werden auch Links zwischen Modulen verschiedener Dokumente und zwischen ganzen Dokumenten erzeugt. Die Links, die auf statistischer Grundlage aufgrund von Ähnlichkeitsberechnungen ohne Berücksichtigung des Nutzungskontextes und der Wissensvoraussetzungen des Nutzers generiert werden, sind assoziative Links, die semantisch nicht weiter typisiert sind.

Im Gegensatz zu diesem Ansatz nutzen wir in *HyTex WordNet* nicht als lexikalische Ressource[6], sondern als Strukturierungskonzept für die Modellierung von Wissen über die Bedeutung und Verwendungsweise von zentralen Fachtermini der im Fachtextkorpus behandelten Domäne (also „Texttechnologie" und „Hypermedia"). Auf der Grundlage einer über dem Fachtextkorpus erzeugten Liste von Termini wird ein Terminologienetz im Stil lexikalisch-semantischer Wortnetze[7] erstellt und mit den Termverwendungsinstanzen in den Dokumenten verlinkt. Dieses Netz dient der Generierung semantisch

typisierter Links, anhand derer sich Nutzer genau diejenigen Wissensvoraussetzungen erarbeiten können, die für die korrekte Semantisierung der Verwendung von Termini in je konkreten Kontexten notwendig sind.

Insofern ist unsere Verwendung des Wortnetzes verwandt mit der Verwendung sog. „Ontologien" in *ontologiebasierten Hypertexten* (vgl. Miles-Board et al. 2001). In diesem Ansatz, der in verschiedenen Projekten zum Einsatz gebracht wurde[8], werden die Objekte sowie die Beziehungen zwischen ihnen in einer als Ontologie bezeichneten Wissensrepräsentation abgebildet, die dann die Strukturierung und die Verlinkung von Hypertextdokumenten motiviert. Die Ansätze verwenden Metadaten, um Dokumenteninhalte zu beschreiben, und generieren daraus automatisch Links zwischen den Dokumenten und der Ontologie. Im Vergleich zum *HyTex*-Ansatz, der zunächst nur auf dem geschlossenen Fachtextkorpus operiert, zeichnen sich diese Ansätze durch einen hohen Grad an Automatisierung und die Anwendbarkeit auf offene Korpora aus. Das zur Modellierung der Ontologien genutzte Inventar von Entitäten und Relationen ist allerdings im Vergleich zu dem Beschreibungsinventar für lexikalisch-semantische Wortnetze beschränkter, enthält aber dafür „ontologietypische" Entitäten (Instanzen) und entsprechende Relationen (z.B. *class-instance*), die für Inferenzen (d.h. die Gewinnung nicht explizit gespeicherten Wissens aus dem Modell durch logische Inferenzmechanismen) genutzt werden können.

Die im Folgenden dargestellten Modifikationen und Erweiterungen des Beschreibungsrahmens von *WordNet*, das in seiner ursprünglichen Form nicht ohne Weiteres für Inferenzen verwendet werden kann und auch nicht als Ontologie konzipiert wurde (vgl. Fischer 1998, Carr et al. 2001), versucht die Stärken des *WordNet*-Ansatzes bei der Modellierung von Sprachwissen mit den Vorteilen von Ontologien beim Durchführen von Inferenzen auf Wissensrepräsentationen zu verbinden und den „klassischen" Wortnetz-Ansatz auf die Erfordernisse des oben skizzierten Anwendungsszenarios hin zuzuschneiden.

3 Modellierung terminologischen Wissens in *TermNet*

Die in *TermNet* verfolgten Modellierungsprinzipien folgen im Grundsatz dem *WordNet*-Ansatz. Zwar ist der *WordNet*-Ansatz primär auf die Gemeinsprache ausgerichtet; doch obwohl sich für Fachsprachdomänen in verschiedenerlei Hinsicht (sowohl lexikalischer, konzeptueller wie auch funktionaler Art) „besondere" (d.h.: von der Gemeinsprache verschiedene) Regularitäten der Prägung, Etablierung und Verwendung lexikalischer Einheiten feststellen lassen, kann die prinzipielle Unterscheidung zwischen *words* (Lexemen) und *synsets* (Konzepten) auch für die Modellierung fachdomänenspezifischer Wortnetze vorteilhaft sein (insbesondere in Hinblick auf den in *HyTex* anvisierten Anwendungsbereich einer automatischen Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen beim selektiven Zugriff auf den Dokumentenpool). Allerdings waren bei der Konzeption von *TermNet* – eben in Hinblick auf den in Rede stehenden Anwendungsbereich – einige Modifikationen des in *WordNet* enthaltenen Relationeninventars notwendig. Diese (sowie die damit in Zusammenhang stehenden fachsprachdomänenspezifischen Besonderheiten) sollen im folgenden skizziert und anhand von Beispielen erläutert werden.

3.1 Behandlung der Synonymie in Fachtextdomänen

Die beiden grundlegenden Typen von Entitäten bei der Modellierung von Wortnetzen im Stile von *WordNet* werden auch in *TermNet* unterschieden: das *Lexem* („word" in der Terminologie von *WordNet*) und das *Synset*, welches eine (ein- oder mehrelementige) Menge von Lexe-

men darstellt, denen jeweils paarweise und restfrei ein Verbundensein über eine Synonymierelation zugesprochen werden kann. In *TermNet* umfasst ein Synset terminologische Ausdrücke, die in der Fachtextdomäne in ungefähr dasselbe Konzept lexikalisieren. So werden z.B. die Lexeme Hyperlink und Verweis demselben Synset zugeordnet, da sie von unterschiedlichen Autoren der Fachdomäne für identische oder zumindest *thematisch* identische Konzepte eingeführt und genutzt werden. Wie auch in *WordNet* wird der Ermittlung der Mitglieder eines Synsets ein weit gefasster Synonymiebegriff zu Grunde gelegt, welcher die Austauschbarkeit der Lexeme in mindestens einem Kontext postuliert. Hierzu ist natürlich anzumerken, dass ein Synonymiekonzept, wie es bei der Modellierung gemeinsprachlicher Wortnetze in durchaus begründbarer Weise angewendet werden kann, nicht ohne weiteres auch auf die Modellierung fachsprachlicher Wortnetze übertragbar ist. Während die Regeln der Verwendung gemeinsprachlicher Ausdrücke prinzipiell gebrauchs- und kontextabhängig sind und bedeutungsähnliche lexikalische Einheiten unter Absehung von je konkreten Verwendungskontexten in semantischer Hinsicht nicht trennscharf von einander geschieden werden können, ist es eine Besonderheit des Sprachgebrauchs in fachsprachlicher Kommunikation, dass jeder Autor prinzipiell als Herr seiner eigenen Benennungs- und Konzeptsysteme in Erscheinung treten kann. Typisch für Fachtexte ist, dass anhand definitorischen Sprachhandelns Konzepte explizit mit einer bestimmten Benennung versehen (und somit vom Autor terminologisiert) werden beziehungsweise Regeln für die Verwendung bestimmter Bezeichnungen fixiert werden (durch explizite Beschreibung der Konzepte, zu deren Benennung sie – zumindest im Textuniversum des betreffenden Autors – fungieren sollen). Unterschiedliche Autoren (und bisweilen sogar derselbe Autor in unterschiedlichen Abschnitten seiner wissenschaftlichen Biographie) können (a) dieselben Benennungen für unterschiedliche Konzepte oder (b) unterschiedliche Benennungen für gleiche oder ähnliche Konzepte oder aber auch (c) unterschiedliche Benennungen für unterschiedliche, aber thematisch identische Konzepte einführen und verwenden. Deshalb ist die konzeptuelle wie lexikalische Geordnetheit fachsprachlichen Konzept- und Bedeutungswissens somit immer textabhängig. Ein fachsprachliches Wort- und Konzeptnetz, welches den Anspruch erhebt, die *tatsächlichen* Verhältnisse abzubilden, müsste daher im Grunde für jeden einzelnen Fachtext (oder zumindest für die Fachsprache jedes einzelnen Autors der Domäne) individuell modelliert werden.

Unter fachsprachenlinguistischem Aspekt wäre die „autorsensitive“ Modellierung fachsprachlicher Wortnetze (und insbesondere ihr Vergleich) zweifelsohne von großem Interesse. In Hinblick auf den im *HyTex*-Projekt verfolgten Anwendungsrahmen (der auf die Unterstützung der selektiven Fachtextrezeption und nicht auf Untersuchungen zur lexikalischen und konzeptuellen Struktur von Fachsprachen abhebt; siehe Abschnitt 1) wäre eine solche „autorsensitive“ Modellierung allerdings weder praktikabel noch zielführend. Unser Anwendungsszenario geht davon aus, dass ein selektiv auf eine große Menge von (Fachtext-)Dokumenten zugreifender Nutzer mit Semi-Experten-Status sich möglichst schnell darüber informieren möchte, (a) welches Konzept einem in einem Textmodul verwendeten Terminus autorseitig unterliegt, und (b) in welcher Beziehung dieses Konzept zu den übrigen in der Fachtextdomäne relevanten Konzepten steht. Anforderung (a) versuchen wir dadurch nachzukommen, dass wir über ein Verfahren zur pragmatischen Gewichtung von Definitionen in Fachtexten dem Nutzer genau diejenige Definition im Vortext über ein Linkangebot zugänglich machen, von welcher wir annehmen, dass sie diejenige ist, an die sich der Autor des betreffenden Textes auch in seinem eigenen

Sprachhandeln hält[9]. Da Nutzer mit Semi-Experten-Status zum Verständnis der in der Fachtextdomäne behandelten Gegenstände zunächst einmal eines systematischen Überblicks über einzelne Konzepte in ihrer Beziehung zu anderen Konzepten der Domäne bedürfen, primär also an der Geordnetheit des in der Domäne behandelten Wissensauschnitts interessiert sind, versuchen wir Anforderung (b) dadurch nachzukommen, dass wir bei der Modellierung von *TermNet* ein ähnlich weitgefasstes Synonymiekonzept wie in *WordNet* ansetzen. Damit können wir den Nutzer (der in dem anvisierten Szenario primär an *Themen* und erst sekundär an deren ggf. variabler Konzeptualisierung und Benennung interessiert ist) von verschiedenen bedeutungs*ähnlichen* Termini, die von einzelnen Autoren zu Zwecken der Konzeptualisierung und Benennung desselben Themas geprägt wurden, zu ein- und demselben Synset unseres Wortnetzes führen. Von diesem Synset aus kann er sich dann über die Beziehung des damit repräsentierten Themas zu anderen Themen der Domäne orientieren. Dass Bedeutungs*ähnlichkeit* (im Sinne einer graduellen *Unähnlichkeit*) von Lexemen in Fachsprachen letztlich auch zugleich eine Unähnlichkeit der damit bezeichneten Konzepte bedeuten kann, wird dabei nicht geleugnet. Letztlich sollen sowohl *TermNet* als auch die auf seiner Basis generierten Linkangebote als Orientierungshilfen bei der selektiven Fachtextrezeption dienen. Für ein differenzierteres Eindringen in die Fachdomäne (zumindest, wenn sie – wie im *HyTex*-Korpus – durch linear, d.h. nicht speziell in Hinblick auf eine selektive Rezeption hin organisierte Texte repräsentiert ist) empfiehlt sich nach wie vor die Lektüre eines Textes von Anfang bis Ende (also gemäß dem vom Autor geplanten Rezeptionsverlauf). Da in vielen Situationen der Beschäftigung mit Fachtexten (Studium, Fachjournalismus, interdisziplinäres Arbeiten) die Bewältigung großer Dokumentenmengen unter Zeitdruck erfolgt und daher nur selektiv möglich ist, erscheint uns ein Wort- und Konzeptnetz, wie wir es mit *TermNet* aufbauen, dennoch als eine wertvolle Verständnishilfe (welche aber nicht beansprucht, die lineare und ausführliche Textrezeption qualitativ zu ersetzen).

3.2 Lexikalische Relationen

Als *lexikalische Relationen* fassen wir bei der Modellierung von *TermNet* solche Relationen, die zwischen Einheiten bestehen, die insgesamt ein Synset konstituieren. Aufgrund der oben beschriebenen fachsprachenspezifischen Abhängigkeit der strukturellen Gliederung des Konzeptbereichs von autorindividuellen Terminologisierungsoperationen sind die Synsets in *TermNet* so flexibel konzipiert, dass sie zwar im Idealfall Konzepte repräsentieren, im Falle konzeptueller Varianz aber auch für *thematische Ausschnitte* aus dem insgesamt in der Fachtextdomäne behandelten Gegenstandsbereich stehen können. Kriterium für die Zugehörigkeit zweier terminologischer Einheiten zu einund demselben Synset ist somit im Zweifelsfalle nicht nur die *konzeptuelle*, sondern auch die *thematische Identität*. Für die praktische Modellierungsarbeit bedeutet dies, dass zwei Termini, die (bei Betrachtung der von den jeweiligen Autoren qua Definition zugeordneten Konzepte) zwar partiell unterschiedliche Konzepte lexikalisieren, trotzdem demselben Synset zugewiesen werden können. So sind beispielsweise die Termini `Verknüpfung` (nach Kuhlen 1991) und Verweis (nach Tochtermann 1995) zwar konzeptuell unterschiedlich, aber thematisch identisch und in *TermNet* daher als Mitglieder desselben Synsets LINK[10] ausgewiesen.

Um Zusammenhänge zwischen den Termini, die demselben Synset angehören, in Hinblick auf die Besonderheiten des Sprachgebrauchs in der Fachtextdomäne differenzierter beschreiben zu können, haben wir in Erweiterung des *WordNet*-Modells (bei welchem die Mitglieder eines Synsets ganz allgemein über die Relation

der Bedeutungsähnlichkeit verbunden sind) einige weitere Relationen eingeführt, die als Spezifizierungen der Relation der Bedeutungsähnlichkeit aufgefasst werden können (siehe Abb. 2). So beschreibt die (symmetrische) Relation `ist_orthographische_Variante_zu` eine Synonymiebeziehung, die sich durch Varianz in der Schreibung eines Terminus ergibt (Beispiele: `referenzieller Link ist_orthographische_Variante_zu referentieller Link` und `Hyper-Link ist_orthographische_Variante_zu Hyperlink`). Die Relationen `ist_Akronym_zu` (mit der Konverse `ist_Vollform_zu`) und `ist_Abkürzung_zu` (mit der Konverse `ist_Expansionsform_zu`) beschreiben zwei terminologische Ausdrücke als synonym hinsichtlich der Tatsache, dass sich die Differenz ihrer Ausdrucksseiten auf wortbildungsmorphologische Prozesse zurückführen lässt (Beispiele: `HTML ist_Akronym_zu Hypertext Markup Language` und `Link ist_Abkürzung_zu Hyperlink`).

Eine weiteres Phänomen, dessen Berücksichtigung speziell für solche Domänen (wie die in unserem Korpus dokumentierten) relevant ist, deren Konzepte im Englischen entwickelt und terminologisiert und dann auf das Deutsche übertragen wurden, ist das der *sprachkontaktbedingten Lexemkonkurrenz*. Hierunter fassen wir solche Fälle der Synonymie, die sich daraus ergeben, dass ein englischer Ausdruck im Deutschen sowohl als Lehnwort als auch in Form einer oder mehrerer Lehnübersetzungen existiert, die zwar im Ausdruck verschieden sind, aber identisch verwendet werden. Da *TermNet* dem Nutzer die Möglichkeit bieten soll, zu einem gesuchten Konzept oder Thema sämtliche Ausdrücke zu ermitteln, die dieses Konzept oder Thema terminologisch lexikalisieren, werden Lehnwort- und Lehnübersetzungsbeziehungen explizit als solche modelliert. Zu diesem Zweck werden bei Bedarf

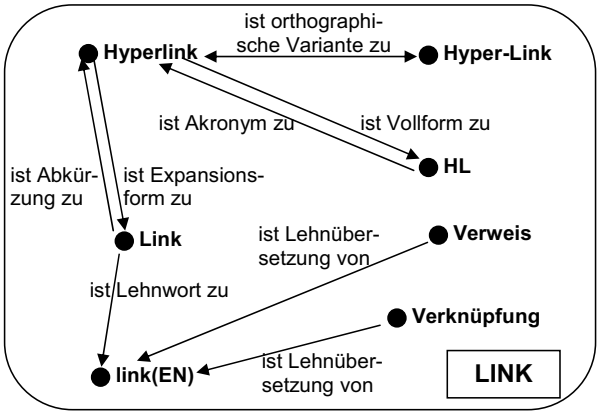

Abbildung 2: Lexikalische Relationen zwischen den Mitgliedern des Synsets LINK.

die entsprechenden englischen Lexeme in die betreffenden Synsets mitaufgenommen, um dann mit ihren deutschen Äquivalenten in folgender Weise verbunden werden zu können: `Link ist_Lehnwort_zu link(EN)` bzw. `Verweis ist_Lehnübersetzung_von link(EN)` bzw. `Verknüpfung ist_Lehnübersetzung_von link (EN)`. Für den Nutzer können die Termini `Link`, `Verweis` und `Verknüpfung` dann paarweise als *Lokalisierungsvarianten* ausgewiesen werden, die sich auf unterschiedliche Arten der lexikalischen Übernahme des mit dem englischen Ausdruck `link` verbundenen Konzepts in die deutsche Fachsprache zurückführen lassen. Durch die explizite Darstellung solcher durch Sprachkontakt verursachter Fälle von Synonymie kann *TermNet* zugleich eine „Brückenfunktion" erfüllen, insofern Lehnwörter und Lehnübersetzungen zu ihren jeweiligen englischen Ursprungswörtern in Beziehung gesetzt werden und sich somit auch die Möglichkeit eines Zugangs zur englischen Fachsprache eröffnet.

3.3 Konzeptuelle Relationen

Anhand der vorgestellten lexikalischen Relationen strukturieren wir Mengen von Termini als Mitglieder einzelner Synsets. Die Synsets wiederum repräsentieren – in unserer Modellierung wie auch in *WordNet – Konzepte* (mit der Einschränkung, dass in Fällen terminologiebil-

dungsbedingter konzeptueller Varianz in *TermNet* Termini auch unter dem weitergefassten Kriterium der thematischen Identität zu Synsets gruppiert werden können). Die Menge der vorhandenen Synsets wird über *konzeptuelle Relationen* strukturiert. Zentral ist hierbei die *Hyponymie*-Relation (`ist_hyponym_zu` mit der Konverse `ist_hyperonym_zu`), anhand derer der Konzeptbereich in hierarchischer Gliederung dargestellt werden kann und aus deren Anwendung sich paarweise Kohyponymie-Beziehungen für demselben Mutterknoten untergeordnete Geschwisterkonzepte ableiten lassen.

Zur vertikalen Strukturierung reicht die Hyponymie-Relation oft nicht aus, daher wird auch in *TermNet* die *Meronymie* als weitere, hierarchisierende Relation modelliert. Zur Beschreibung von Meronymiebeziehungen orientieren wir uns an den in *EuroWordNet* unterschiedenen Mernoymietypen (vgl. Vossen 1998), verwenden aber letztlich nur zwei davon, da für die zu modellierende Fachtextdomäne sowie in Hinblick auf das in *HyTex* verfolgte Anwendungsszenario lediglich Meronymiebeziehungen des Typs *Konstituenz* und *Gruppenzugehörigkeit* relevant sind. Die Bezeichnungen für die betreffenden Relationen (und ihre jeweiligen Konversen) wurden aus *EurWordNet* übernommen. So beschreiben wir die Konstituenzbeziehung zwischen `MODUL` und `HYPERTEXT` als `MODUL has_holo_part HYPERTEXT` und die Konverse als `HYPERTEXT has_mero_part MODUL`. Die Gruppenzugehörigkeit von `XHTML`, `XTM` und `NITF` zu `XML-SPRACHENFAMILIE` wird beschrieben als `XHTML, XTM, NITF has_holo_member XML-SPRACHENFAMILIE`, die Konverse als `XML-SPRACHENFAMILIE has_mero_member XHTML, XTM, NITF`.

Neben der Hyponymie-/Hyperonymie- sowie der Meronymie-/Holonymie-Relation umfasst unser Inventar an konzeptuellen Relationen auch noch die Relation der *Antonymie*. In Hinblick auf die anvisierte Anwendung (Linking nach Kohärenzkriterien) hat für die *TermNet*-Modellierung eine Opposition zwischen Konzepten ein stärkeres Gewicht als eine Opposition zwischen einzelnen Lexemen. Während es im Standard-*WordNet*-Konzept der Antonymie insbesondere um die Erfassung stilistischer Feinheiten geht (z.B.: `ascend` steht in lexikalischer Opposition zu `descend` und nicht zu `go down`), ist es für unser Anwendungsszenario wichtiger, dem Rezipienten zu vermitteln, dass zwei Konzepte – z.B. `WOHLGEFORMT` und `NICHT-WOHLGEFORMT` aus der Domäne „Texttechnologie“ – sich wechselseitig ausschließen, als ihm zu vermitteln, ob der Terminus `nicht-wohlgeformt` eher mit dem Terminus `wohlgeformt` oder dem synonymen Lehnwort `well-formed` kontrastiert.

In Hinblick auf die Modellierung fachsprachlicher Domänen reicht die aus der Hyponymie-/Hyperonymie-Beziehung abgeleitete Kohyponymie-Beziehung in einigen Fällen nicht aus, um spezielle Ausschlussbeziehungen zwischen Paaren oder Gruppen von Kohyponymen, die aus typologisch motivierten terminologischen Konzeptualisierungsprozessen resultieren, zu beschreiben. So stehen beispielsweise bestimmte Hyponyme zu `LINK` nicht gleichrangig nebeneinander, sondern schließen sich bezüglich der Möglichkeit ihres Zutreffens auf einen konkreten Vertreter des Typs `LINK` gruppenweise wechselseitig aus: `INTRAHYPERTEXTUELLER LINK`, `INTERHYPERTEXTUALLER LINK` und `EXTRAHYPERTEXTUELLER LINK` schließen sich wechselseitig aus, während dies z.B. für `INTERHYPERTEXTUELLER LINK` und `UNIDIREKTIONALER LINK` nicht der Fall ist. Die dem Konzept `LINK` untergeordneten Hyponyme bilden Gruppen, die in sich jeweils durch ein bestimmtes (durch typologische Differenzierung bei der terminologischen Konzeptualisierung motiviertes) klassifikatorisches Merkmal bestimmt sind. So kann man

die Klasse der Links (als Klasse von Objekten in der Welt) nach dem Merkmal der Direktionalität z.B. in UNIDIREKTIONALE LINKS und BIDIREKTIONALE LINKS unterteilen, während man nach dem Verhältnis von Ausgangsanker und Zielanker zwischen INTRAHYPERTEXTUELLEN LINKS, INTERHYPERTEXTUELLEN LINKS und EXTRAHYPERTEXTUELLEN LINKS unterscheiden kann. Wenn man nur die Hyponymie-Relation kodiert, so wird ein für das Folgern wichtiger Aspekt verdeckt, nämlich derjenige, dass ein individueller Link zwar zugleich eine Instanz der Konzepte INTERTEXTUELLER LINK und UNIDIREKTIONALER LINK sein kann, aber nicht zugleich Instanz der Konzepte UNIDIREKTIONALER LINK und BIDIREKTIONALER LINK. Dieses Phänomen tritt natürlich nicht nur in Fachdomänen auf, sondern auch in der Allgemeinsprache: Ein individuelles Pferd kann zwar gleichzeitig zu den Klassen HENGST und RAPPE gehören; es kann aber nicht gleichzeitig RAPPE und SCHIMMEL sein, zumindest nicht in der uns vertrauten nicht-fiktionalen Tierwelt. Die Kohyponyme auf derselben Hierachiestufe können also durch klassifikatorische Merkmale weiter struktuiert sein, was dazu führt, dass den Kohyponymen mit demselben Klassifikationsmerkmal extensional disjunkte Teilmengen von Instanzen entsprechen.

Das Standard-Modell von *WordNet* bietet bislang keine Lösung, um diesen Sachverhalt zu erfassen. Aus diesem Grund haben wir das Modell erweitert, indem wir Attribute eingeführt haben, über die wir inferieren können, welche Gruppen (Mengen) von Kohyponymen in einer Relation wechselseitiger *Disjunktivität* stehen (siehe Abb. 3). Durch die Zuweisung von Attributen, anhand derer sich einzelne Mengen von alternativen Konzepten über einen einmalig zu vergebenden Attributwert (z.B. „Direktionalität“) definieren lassen, kann Disjunktivität zwischen Kohyponymenmengen modelliert werden, *ohne* dass dabei auf die Einführung von künstlichen Konzepten (Pseudokonzepten in der hierarchischen Modellierung) zurückgegriffen werden muss, die anschließend auf der Präsentationsebene wieder herausgefiltert werden müssen.

4 Nutzung von TermNet für das automatische Linking: Ein Anwendungsbeispiel

Das richtige Verständnis von Fachtexten hängt zu einem nicht unerheblichen Anteil davon ab, dass die in der Domäne eingespielten terminologischen Einheiten mit den entsprechenden Konzepten verbunden und Unterschiede und Ähnlichkeiten zwischen verschiedenen Benennungen für dasselbe Konzept erkannt werden. Im Rahmen des im ersten Abschnitt skizzierten Strategietyps „Linking nach Wissensvoraussetzungen“ versuchen wir deshalb, Wissen über die Verwendungsregeln von Termini und über Bezüge zwischen den Konzepten unserer Fachtextdomäne über Linkangebote rekonstruierbar zu machen.[11] Die Generierung dieser Linkangebote erfolgt auf der Grundlage einer teilautomatischen Annotation sowohl von definitorischen Textsegmenten (also Textsegmenten, in denen Verwendungsregeln für Termini spezifiziert werden), als auch von Verwendungen der entsprechenden Termini in den Dokumenten (den Termverwendungsinstanzen). Um den Nutzern die für sie relevanten Informationen geben zu können, werden Termverwendungsinstanzen verlinkt mit automatisch generierten Glossarsichten, die Aufschluss über die spezifische Verwendung eines Terminus in der Fachtextdomäne geben.

Das in Abb. 4 gezeigte Beispiel für die nach diesem Prinzip generierten Sichten soll illustrieren, wie die Linkingstrategien in einem konkreten Nutzungskontext zusammenspielen. Ein Nutzer, der im Zuge der selektiven Textlektüre im oben rechts angezeigten Modul einsteigt, trifft auf eine Verwendungsinstanz des Terminus Link, die in der Hypertextsicht als Linkanzeiger

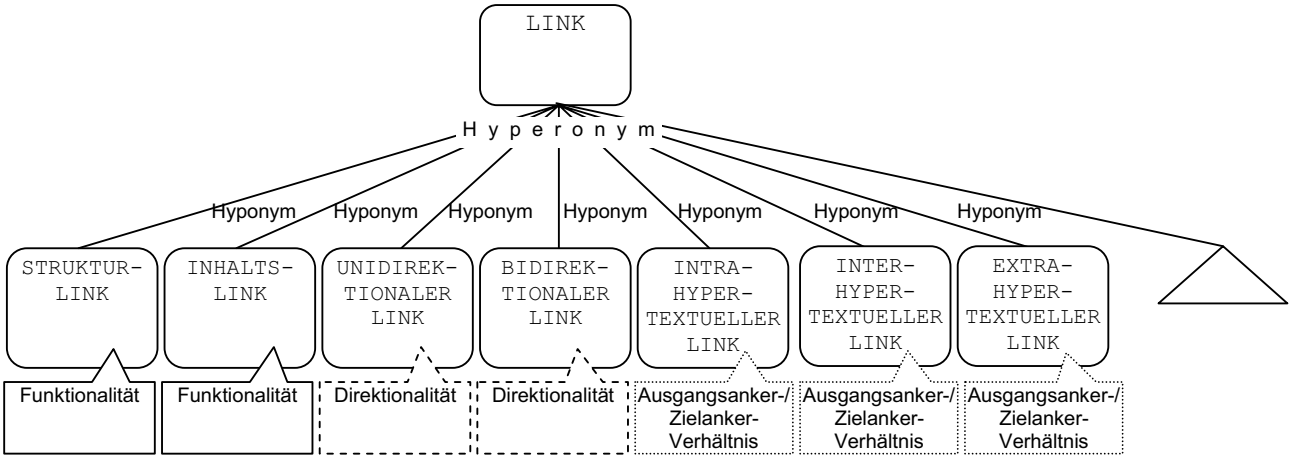

Abbildung 3: Veranschaulichung zur attributbasierten Modellierung der Disjunktivität in *TermNet*

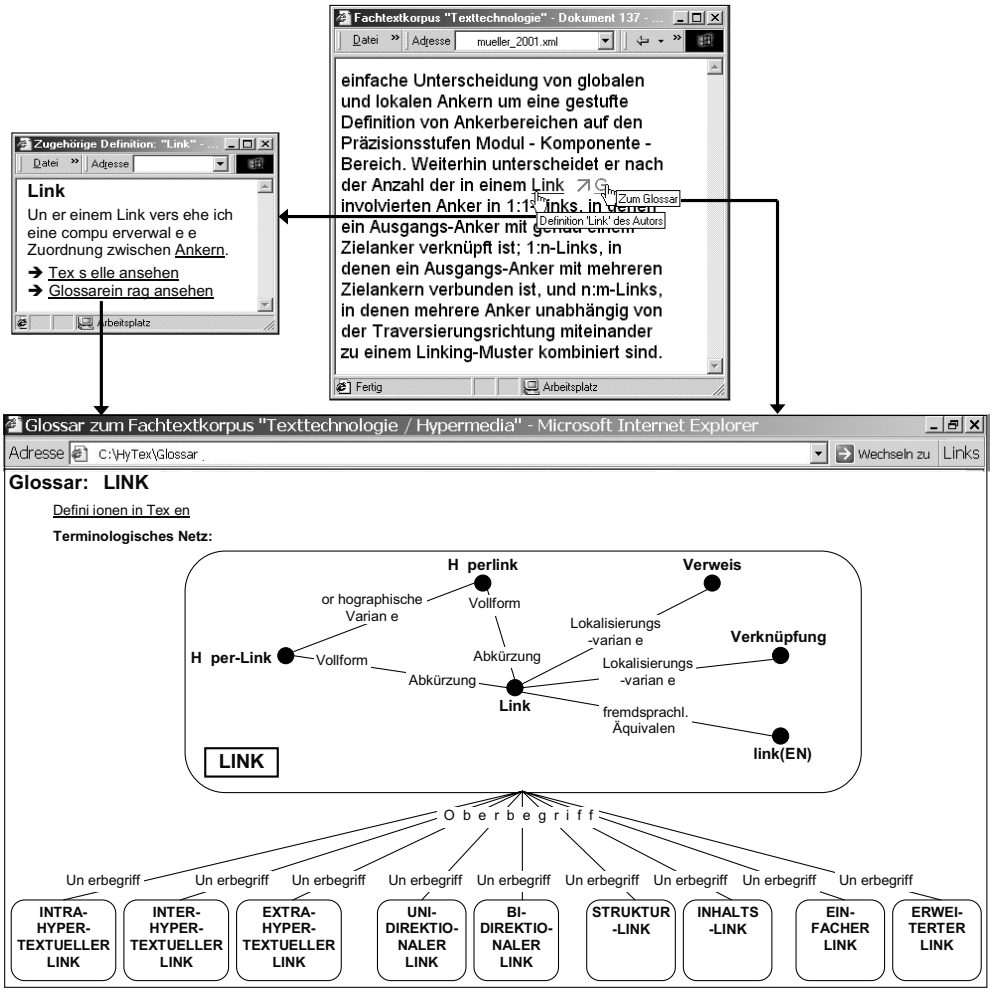

Abbildung 4: Veranschaulichung zur Vernetzung einer Termverwendungsinstanz (hier: „Link“) in einem Dokumentenmodul mit der zugehörigen Definition des Autors sowie mit einer automatisch generierten Glossarsicht, welche Ausschnitte aus dem terminologischen Netz (hier: das Synset LINK mit den zugehörigen lexikalischen Einheiten und seiner konzeptuellen Umgebung) in grafisch visualisierter Form präsentiert.

repräsentiert ist. Will sich der Nutzer über die Bedeutung des Terminus informieren, so erhält er durch Aktivierung des Linkanzeigers zunächst das oben links angezeigte Definitionsfenster, das die im Vortext vom Autor des Dokuments eingeführte Definition präsentiert. Diese enthält wiederum den Terminus `Anker`, der ebenfalls mit einer entsprechenden Definition und einem Glossareintrag verknüpft ist. Durch Aktivieren des Linkanzeigers „Textstelle ansehen" kann sich der Nutzer bei Bedarf das textuelle Umfeld der Definition anzeigen lassen; ein Mausklick auf den Linkanzeiger „Glossareintrag ansehen" erzeugt den im Fenster unten dargestellten Glossareintrag. Dieser Glossareintrag ist für Nutzer gedacht, die in der Lektüre des ursprünglichen Textes pausieren und sich zunächst vertiefend mit dem hinter dem Terminus `Link` stehenden Konzept vertraut machen möchten. Der Glossareintrag bietet als weitere Konzeptualisierungshilfen Verknüpfungen zu Definitionen des Terminus `Link` in verschiedenen Fachtextdokumenten an (diese Verknüpfungen operieren über dem Lexem `Link`). Des weiteren präsentiert der Glossareintrag einen grafisch visualisierten Ausschnitt aus dem terminologischen Netz. Dieser zeigt die lexikalische und konzeptuelle Umgebung des Terminus, wobei die Kanten durch Relationennamen gelabelt und die beiden Knotentypen (Konzept und Lexem) in der Darstellung unterschieden sind. Die Knoten des Netzes sind ebenfalls als Linkanzeiger gestaltet, bei deren Aktivierung ein entsprechender neuer Glossareintrag generiert wird. Auf diese Weise kann sich der Nutzer mit den jeweiligen Konzepten in derjenigen Detailtiefe auseinandersetzen, die seinem aktuellen Informationsbedarf und seinem Vorwissensstand am besten entspricht.

Unser Strategieansatz bietet dem Nutzer bei der selektiven Textlektüre somit zwei Arten von terminologiebezogenen Informationsangeboten:

- *Informationen über den Terminus selbst*. Dazu gehört sowohl die Angabe der Definition des Terminus durch den Autor des Primärtextes selbst als auch die Möglichkeit, andere mit demselben terminologischen Ausdruck verbundene Konzeptualisierungen in den Dokumenten des Fachtextkorpus einzusehen.
- *Informationen über die Beziehungen des Terminus zu anderen Termini*. Im Glossar werden zu jedem Terminus dessen lexikalische und konzeptuelle Relationen zu anderen Termini dargestellt. Hierbei ist es auch möglich, zu den jeweils „verwandten" Termini zu gelangen und von dort aus Informationen über diese zu erhalten.

TermNet spielt dabei sowohl bei der Verknüpfung der Termini mit den Termverwendungsinstanzen und den dafür relevanten Definitionen eine Rolle, als auch bei der Generierung der Glossarsichten und der Einbettung eines Konzepts in sein jeweiliges konzeptuelles und lexikalisches Umfeld.

Anmerkungen

[1] HyTex wird seit April 2002 an der Universität Dortmund durchgeführt (*http://www.hytex.info*) und ist ein Teilprojekt der Forschergruppe „Texttechnologische Informationsmodellierung" (*http://www.text-technology.de*), die sich mit den theoretischen Grundlagen und Methoden der Modellierung von Sprachdaten mit Markup-Sprachen (insbesondere XML und Tochterstandards) beschäftigt.

[2] Streng genommen handelt es sich nicht um Reversibilität im Sinne eines Umkehrprozesses, sondern die Hypertextualisierung erfolgt „on the fly" auf der Basis der textgrammatischen Annotationen der weiterhin in der ursprünglichen Form verfügbaren Ausgangstexte. Wir verwenden den Ausdruck „reversibel", um uns von Ansätzen zur Hypertextkonversion abzugrenzen, in denen der ursprüngliche Text irreversibel in Form und Struktur umgestaltet wird.

[3] Vgl. Beisswenger et al. 2002 und Lenz et al. 2002.

[4] Vgl. Storrer 2002.

[5] Da Fachdomänen, wenn man sie auf der Grundlage von Textkorpora (und nicht etwa wissenschaftssoziologisch oder in Hinblick auf die mündliche Fachkommunikation) betrachtet, stets nur über die in den jeweiligen Korpora (die grundsätzlich immer nur einen Ausschnitt der für die Fachdomäne relevanten bzw. konstitutiven Texte umfassen) enthaltenen Textdokumente repräsentiert werden, sprechen wir im Folgenden in bezug auf den Bezugsbereich unserer Modellierung von einer Fachtextdomäne. Dies ist zu lesen als „Die Fachdomäne, so wie sie sich in den zugrunde gelegten Texten zeigt". Den Ausdruck Fachdomäne verwenden wir nur dann, wenn wir uns ganz allgemein auf das Thema unserer Modellierung, nicht aber auf deren – mit dem Zuschnitt unseres Korpus gegebenen – Bezugsbereich beziehen.

[6] Für unser deutsches Korpus käme ein Nachbau des Ansatzes von Green ohnehin nur mit Hilfe des deutschen GermaNet (vgl. z.B. Kunze & Wagner 2001) in Frage.

[7] Der Ausdruck lexikalisch-semantische Wortnetze bezeichnet nach Kunze 2001 generell einen Modellierungsansatz für Sprach- und Konzeptwissen im Stile des WordNet-Projekts, der zunächst an der Universität Princeton für das Englische entwickelt und später für viele andere Sprachen ausgebaut wurde.

[8] Z.B. COHSE (Carr et al. 2001), OntoPortal und ESKIMO (Miles-Board et al. 2001), On2broker (Decker et al. 1998).

[9] Siehe hierzu Beisswenger et al. 2002.

[10] Um die unterschiedlichen Modellierungseinheiten in TermNet, wenn sie im Rahmen von Beispielen oder Erwähnungen benannt werden, typographisch von einander zu unterscheiden, geben wir hier und im Folgenden Lexeme (Termini) in Normalschrift, die Namen von Konzepten hingegen in Versalien an.

[11] Die verschiedenen Teilaspekte der Strategien sind in Beisswenger et al. 2002; Lenz et al. 2002 und Lenz et al. 2004 beschrieben.

Literatur

Beisswenger, M.; Lenz, E. A.; Storrer, A. (2002). „Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen." In: Busemann, S. (Hrsg.) (2002). Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002), Saarbrücken, September / Oktober 2002. Saarbrücken: Deutsches Institut für Künstliche Intellgenz (DFKI) [DFKI Document D-02-01], 187-191.

Carr, L.; Bechhofer, S.; Goble, C.; Hall, W. (2001). "Conceptual Linking: Ontology-based Open Hypermedia." In: Proceedings of the 10th International World Wide Web Conference, Hong Kong, May 2001, *http://www.ecs.soton.ac.uk/~lac/WWW10/ConceptualLinking.html* [accessed April 2004].

Decker, S. et al. (1998). "Ontobroker in a Nutshell." In: Constantine, N.; Stephanidis, C. (eds.) (1998). Research and Advanced Technologies for Digital Libraries. Proceedings of the 2nd European Conference on Research and Advanced Technology For Digital Libraries (ECDL'98). Berlin et al.: Springer [= LNCS 1513], 540-651, *http://citeseer.ist.psu.edu/89427.html* [accessed April 2004].

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Green, S. J. (1998). "Automated Link Generation: Can We Do Better than Term Repetition?" In: Proceedings of the 7th International World Wide Web Conference. Computer Networks and ISDN Systems, 30 (1-7) (1998), 87-84, *http://citeseer.nj.nec.com/green98automated.html* [accessed April 2004].

Kuhlen, R. (1991). Hypertext: ein nicht-lineares Medium zwischen Buch und Wissensbank. Berlin et al.: Springer [Edition SEL-Stiftung].

Kunze, C. (2001). „Lexikalisch-Semantische Wortnetze." In: Carstensen, K.-U. (Hrsg.) (2001). Computerlinguistik und Sprachtechnologie: eine Einführung. Heidelberg, Berlin: Spektrum Akademischer Verlag, 386-393.

Kunze, C.; Wagner, A. (2001). „Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche." In: Lemberg, I.; Schröder, B.; Storrer, A. (eds.) (2001). Chancen und Perspektiven computergestützter Lexikographie. Tübingen: Niemeyer [= Lexicographica Series Maior Vol. 107], 229-246.

Lenz, E. A.; Beisswenger, M.; Storrer, A. (2002). „Hypertextualisierung mit Topic Maps – ein Ansatz zur Untersützung des Textverständnisses bei der selektiven Rezeption von Fachtexten." In: Tolksdorf, R.; Eckstein, R. (Hrsg.) (2002). Proceedings Workshop XML-Technologien für das SemanticWeb (XSW 2002), Berlin, Juni 2003. Bonn: Köllen Verlag [= GI-Edition - Lecture Notes in Informatics (LNI), P-14], 151-159.

Lenz, E. A.; Birkenhake, B.; Maas, J. F. (2004). Von der Erstellung bis zur Nutzung: Wortnetze als XML Topic Maps. In diesem Band, 127-136.

Miles-Board, T.; Kampa, S.; Carr, L.; Hall, W. (2001). "Hypertext in the Semantic Web." In: Proceedings of the 12th ACM Conference on Hypertext and Hypermedia (HT '01), 237-238.

Storrer, A. (2002). "Coherence in Text and Hypertext." In: Document Design 3(2) (2002), 156-168.

Vossen, P. (ed.) (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.