

Modell eines automatisierbaren syntaktischen Metathesaurus und seine Eignung für parlamentarische Thesauri im Internet

MARTIN FENSKE

Es werden Konzepte eines syntaktischen Metathesaurus, der identische Benennungen und deren Relationen aufeinander abbildet, vorgestellt und von semantischen Metathesauri abgegrenzt. Dieses aus mehreren Konzepten bestehende Modell bietet sich für die automatische Zusammenführung weitgehend übereinstimmender Thesauri zu einem virtuellen Metathesaurus an, den man in Internetportale und Suchmaschinen integrieren kann. Besondere Vorteile sind hierbei das günstige Kosten-Nutzen-Verhältnis und die geringen technischen Anforderungen eines syntaktischen Metathesaurus. Es werden die Inkonsistenzen bei der Zusammenführung inhomogener Thesauri ausführlich beschrieben und Möglichkeiten zur Konsistenzverbesserung angeboten. Ein syntaktisches Thesauruskonzept eignet sich für den Einsatz bei der Websuche in Parlamentsinformationssystemen, wie z. B. dem Parlamentsspiegel, einer Datenbank zum Nachweis der deutschen Parlamentsmaterialien.

Websuchen mit Makro- und Metathesauri

Datenbanken und Kataloge werden im Internet über Suchmaschinen, Portale oder virtuelle Informationssysteme zusammengeführt, um das Retrieval zu erleichtern. Aufgrund wirtschaftlicher und technischer Zwänge gehen dabei oft wesentliche Retrievalfunktionen und Zusatzinformationen der einzelnen Datenbanken und Kataloge verloren.¹ Dies gilt vor allem für die Bereitstellung von Thesaurusinformationen: Thesaurusrelationen stehen meist nicht oder nur sehr umständlich zur Verfügung; Widersprüche zwischen den einzelnen Thesauri werden weder aufgelöst noch überschaubar dargestellt. Durchaus typisch ist hier der Karlsruher Virtuelle Katalog², der wie viele der zugrunde liegenden Bibliothekskataloge weder über Schlagwortindices noch über

¹ Bei übergreifenden Internetsuchen in heterogenen Datenbeständen treten im Vergleich zu direkten Suchen in einzelnen Datenbanken zwischen 5 % und 70 % Informationsverluste auf [12, S. 116].

² URL: <http://www.ubka.uni-karlsruhe.de/kvk.html>.

Thesaurusrelationen verfügt. Das Umweltportal Deutschland (PortalU)³ bietet hingegen getrennt von der eigentlichen Dokumentsuche eine komplexe Suchoberfläche für die beteiligten Thesauri an (sog. Rechercheassistent), obwohl über die Ergebnisanzeige der am PortalU beteiligten Systeme, wie z. B. der Umweltliteratur-Datenbank des Umweltbundesamts, eine Navigation mit Thesaurusrelationen und ein schneller Wechsel zu den dazugehörigen Dokumenten möglich ist. Dadurch werden Benutzer gezwungen, ständig zwischen dem Portal und den beteiligten Subsystemen zu wechseln. Im Gegensatz dazu kann durch den Einsatz von Makro- und Metathesauri die Thesaurusintegration bei der Websuche und damit das Kosten-Nutzen-Verhältnis des Thesauruseinsatzes insgesamt verbessert werden.

Wenn es gelingt, system- und institutionenübergreifend einen gemeinsamen Thesaurus zu erarbeiten, steht ein konsistenter Thesaurus für Indexierung und datenbankübergreifende Suche zur Verfügung. Da ein Thesaurus die Fragen der Nutzer an das jeweilige Informationssystem widerspiegelt, vgl. [5, S. 106], eignet sich ein Thesaurusverbund vor allem dann, wenn sich die Fragen an die beteiligten Systeme überlappen oder ergänzen. Je widersprüchlicher die Anforderungen sind, desto schwieriger wird es, einen gemeinsamen Thesaurus zu entwickeln.

Metathesauri ersetzen nicht die einzelnen Thesauri, sondern führen diese nur für das Retrieval im Nachhinein zusammen. Von den semantischen Metathesauri ist das Unified Medical Language System (UMLS) am bekanntesten. „Es stellt kein vereinheitlichtes Begriffssystem her, wie es Macrothesauri wie etwa der Standard-Thesaurus Wirtschaft leisten, vgl. [17], sondern bildet verschiedene Begriffssysteme bzw. -welten aufeinander ab. Diese Abbildung geht von den Bedeutungen der verwendeten Benennungen aus und wird über semantische Netze realisiert.“ [14, S. 202]. Beim UMLS erfolgt die Zuordnung der Begriffssysteme zu den semantischen Netzen in einem mehrstufigen Verfahren durch Bearbeiter [8]. Semantische Metathesauri sind in der Regel mit erheblichen Personalkosten verbunden. Unter bestimmten Voraussetzungen können sie aber auch automatisch geführt werden, z. B. wenn in den beteiligten Systemen überwiegend die gleichen Dokumente enthalten sind. Bei der Begriffsintegration kann zwischen dokumentenbestands-, thesaurus- und anfragebasierten Ansätzen unterschieden werden, [10, S. 36-38]. Ein automatisches Verfahren basiert meist auf einer Verbindung mehrerer Ansätze. Bei Agogino [1] werden Korrelationen zwischen indexierten Benennungen mit einer Analyse von Nutzeranfragen und einer Stichwortextraktion kombiniert. Derartige Verfahren sind nicht nur technisch aufwendig, meist reicht auch die

³ URL: <http://www.portalu.de>.

Qualität der ermittelten Korrelationen für eine vollautomatische Bearbeitung nicht aus.⁴

Im Gegensatz zu semantischen Metathesauri baut der im Folgenden dargestellte syntaktische Metathesaurus ausschließlich auf den vorhandenen Thesaurusstrukturen auf, da er nur identische Benennungen (also keine Begriffe)⁵ und deren Relationen⁶ aufeinander abbildet. Er eignet sich für die Zusammenführung von Thesauri, die zu einem wesentlichen Teil identische Benennungen und Relationen nutzen, gleichzeitig aber auch abweichende Benennungen und Relationen zulassen. Dies ist typisch für Datenbanken mit einem gemeinsamen Thesaurus, der jedoch aus pragmatischen Gründen auf die Bedürfnisse des jeweiligen Systems angepasst oder um zusätzliche Benennungen und Relationen erweitert wird. Da das einfachste Modell eines syntaktischen Metathesaurus auf einer reinen Kumulation unterschiedlicher Benennungen und Thesaurusrelationen basiert, eignet es sich für die Bereitstellung eines virtuellen Metathesaurus. Durch automatische Verfahren kann die Konsistenz des syntaktischen Metathesaurus erheblich verbessert werden.

Thesaurusrelationen

Die Beschreibung des Modells eines syntaktischen Metathesaurus konzentriert sich auf die zentralen Thesaurusstrukturen entsprechend der DIN 1463-1 [3]: Deskriptoren, Nichtdeskriptoren und ihre Äquivalenz-, Hierarchie- sowie Assoziationsrelationen. Thesaurusrelationen bestehen sowohl zwischen zwei Deskriptoren (D) als auch zwischen einem Nichtdeskriptor (ND) und seinem Deskriptor (Vorzugsbenennung⁷). Sie sollen zur besseren Übersicht entsprechend dem Entity-Relationship-Modell als 1-n-Relation oder als n-n-Relationen beschrieben werden. Bei einer 1-n-Relation kann das erste

⁴ Für die automatische Zuordnung von Nicht-MeSH- zu MeSH-Begriffen des UMLS gibt Bodenreider [2] eine Relevanzrate von 61 Prozent an.

⁵ Da die weit überwiegende Zahl der eingesetzten Thesauri natürlichsprachig ist, soll abweichend von der DIN 1463-1 [3] im Folgenden auch dann „Benennung“ verwendet werden, wenn Aussagen für alle Arten von „Bezeichnungen“ gelten.

⁶ Auf eine Ergänzung von Thesaurusstrukturen, z. B. von fehlenden Hierarchieebenen, wird im Gegensatz zu den semantischen Modellen von Nikolai [10] und Doerr [4] verzichtet, da hierzu komplexe Verfahren und ein erhöhter Personalaufwand erforderlich sind.

⁷ Im Folgenden sollen Deskriptoren immer dann als Vorzugsbenennungen bezeichnet werden, wenn ihre Relationen zu Nichtdeskriptoren von Bedeutung sind.

Element mit beliebig vielen (also n) weiteren Elementen in Relation stehen, die alle jeweils nur mit diesem einen (also 1) Element eine Relation haben. Bei einer n-n-Relation hingegen können alle beteiligten Elemente beliebig viele Thesaurusrelationen mit anderen Elementen haben. In einer Tabelle lassen sich die wichtigsten Relationsarten wie folgt darstellen:⁸

Tabelle 1: Thesaurusrelationen

	1-n-Thesaurusrelation	N-n-Thesaurusrelation
Deskriptor (D)	1D-nD (Hierarchierelation einer Monohierarchie)	nD-nD (Hierarchierelation einer Polyhierarchie bzw. Assoziationsrelation)
Nichtdeskriptor (ND)	1D-nND (Äquivalenzrelation)	-

Auch hiervon abweichende Relationsarten, vgl. [3; 22], können in der o. g. Art beschrieben und meist analog behandelt werden:

Wenn im Fall von Thesauri ohne Vorzugsbenennungen alle Benennungen, also auch Synonyme, zur Indexierung zugelassen werden, so handelt es sich um 1D-nD-Äquivalenzrelationen, bei denen über entsprechende Retrievalfunktionen die Äquivalenzklassen zusammengeführt werden. Diese Relationen können analog zu 1D-nND-Äquivalenzrelationen oder teilweise auch wie 1D-nD-Hierarchierelationen bearbeitet werden. Hierzu müsste das beschriebene Modell jedoch noch in Details angepasst werden.

Wenn Unterbegriffe als Nichtdeskriptoren verwendet werden, so handelt es sich zwar um 1D-nND-Hierarchierelationen, die aber wie Äquivalenzrelationen mit Quasisynonymen verwaltet und daher auch im Metathesaurus so behandelt werden sollten.

Wenn statt eines Nichtdeskriptors mehrere speziellere Vorzugsbenennungen zur Auswahl stehen oder bei einer auf morphologischer Zerlegung basierenden Dokumentationssprache statt mit einem Nichtdeskriptor (präkombinierte Benennung) mit mehreren Deskriptoren zugleich indexiert werden soll (für ein postkoordiniertes Retrieval), ist eine nD-nND-Hierarchierelation oder eine nD-1ND-Äquivalenzrelation gegeben. Da der Metathesaurus nur dem Retrieval

⁸ Wenn 1 oder n einem D oder ND vorangestellt werden, so soll dies ausschließlich das Verhältnis zwischen Benennungen nach dem Entity-Relationship-Modell darstellen (z. B. 1D-nND), während nachgestellte Ziffern und deren Platzhalter bei den weiter unten beschriebenen Regeln zur Unterscheidung von Mengen und deren Elementen verwendet werden (z. B. ND_n als Menge der Nichtdeskriptoren 1 bis n).

dient, lassen sich die nD-nND-Hierarchierelationen ohne wesentliche Nachteile wie eine nD-nD-Relation bearbeiten. Bei nD-1ND-Äquivalenzrelationen wäre eine erhebliche Erweiterung des beschriebenen Thesaurusmodells erforderlich.

Aufgrund der Unterschiede zwischen den beteiligten Thesauri ist der Metathesaurus stets komplexer als die jeweiligen Einzelthesauri. Auf eine Differenzierung von Strukturunterschieden wird daher zugunsten einer besseren Übersichtlichkeit so weit wie möglich verzichtet. Es sollen nicht mehr als die in Tabelle 1 genannten Thesaurusrelationen dargestellt werden. Da bei nicht primär hierarchisch gegliederten Thesauri auf eine Differenzierung zwischen Hierarchie- und Assoziationsrelationen verzichtet werden kann, wird beim Modell eines kumulierenden syntaktischen Metathesaurus nur zwischen Äquivalenz- und Nichtäquivalenzrelationen unterschieden.

Kumulierender syntaktischer Metathesaurus

Wortfelder

Der syntaktische Metathesaurus fasst identische Benennungen unterschiedlicher Thesauri zu einer Benennung zusammen. Die Thesaurusrelationen werden auf die entsprechende Benennung des Metathesaurus übertragen. Dadurch werden die Wortfelder der einzelnen Thesauri aufeinander zu einem einzigen Wortfeld des Metathesaurus abgebildet (kumuliert): Der Metathesaurus vereinigt die unterschiedlichen Relationen der beteiligten Thesauri, indem Thesaurusrelationen, bei denen die Art der Relation und die beteiligten Benennungen identisch sind, zu einer einzigen Thesaurusrelation zusammengefasst werden.

Wenn wir uns die Benennungen eines Thesaurus und seine Relationen als eine Abbildung auf einer zweidimensionalen Ebene vorstellen (s. Abb. 1), so sind die Benennungen b_1 bis b_n Kreise (Deskriptoren) bzw. Quadrate (Nichtdeskriptoren) auf dieser Ebene und die Relationen Pfeile zwischen diesen Objekten. Die Pfeilspitze stellt die Richtung einer Relation dar. Hierbei wird deutlich, dass „ein Wortfeld nichts anderes ist, als eine auf eine Ebene projizierte Thesaurushierarchie“ [18, S. 469].

Für jeden Thesaurus können wir uns eine solche zweidimensionale Ebene denken, wobei alle Ebenen parallel zueinander verschoben sind. Identische Benennungen und Relationen haben auf diesen Ebenen immer die gleichen Koordinaten. Ein Metathesaurus stellt nun eine Projektion aller Ebenen (Thesauri) auf eine einzige zweidimensionale Ebene dar. Die Projektionsrichtung wird in Abbildung 1 durch gestrichelte Pfeile dargestellt.

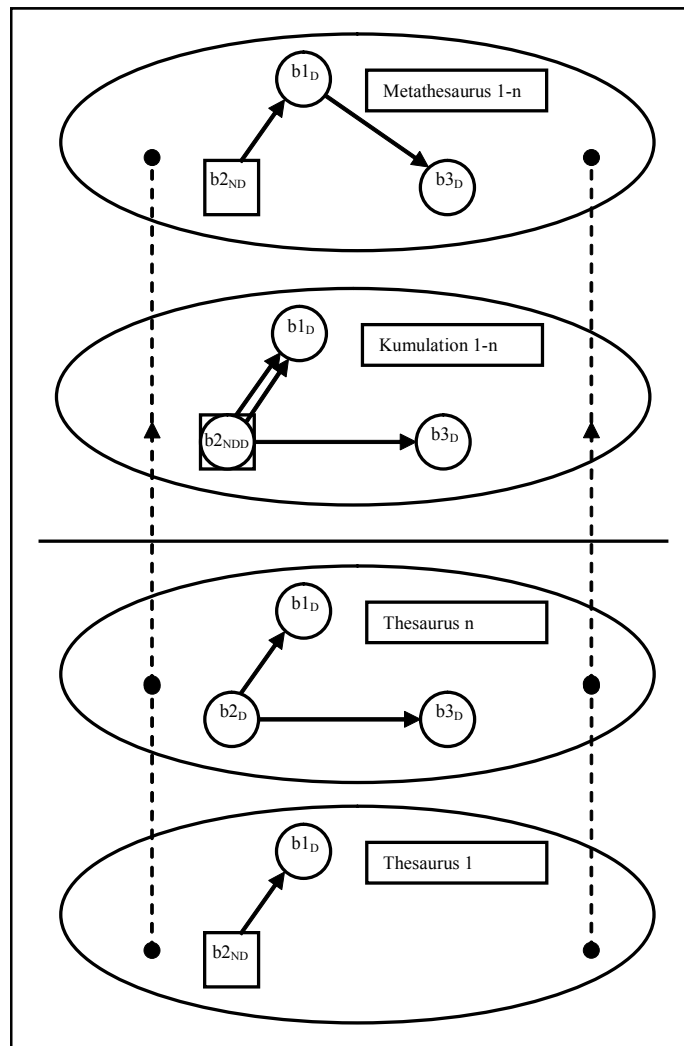


Abb. 1: Thesaurusebenen im Metathesaurus

Betrachtet man nur ein einziges kumuliertes Wortfeld des Metathesaurus, so sind zwei Arten inkonsistenter Syntax erkennbar, und zwar:

1. wenn eine Benennung, die in einem Thesaurus Deskriptor ist, in einem anderen Thesaurus Nichtdeskriptor ist (Quadrat und Kreis überlagern sich) und
2. wenn zwischen zwei Benennungen verschiedenartige Relationen vorhanden sind (mehrere Pfeile zwischen zwei Objekten).

Retrieval mit einem inkonsistenten Metathesaurus

Mit den zuvor beschriebenen Bearbeitungsschritten liegt ein ausschließlich kumulierender Metathesaurus vor, den man durch einen kleinen Trick trotz inkonsistenter Syntax sinnvoll für ein Retrieval bereitstellen kann: Dargestellt wird immer nur ein Wortfeld zum jeweils aktuellen Suchbegriff, dessen Relationen soweit wie möglich reduziert werden. Dieser Metathesaurus stellt eine unveränderte Projektion aller vereinheitlichten Thesaurusstrukturen der beteiligten Thesauri dar.

Um die Komplexität und die Zahl der Relationen zu reduzieren, soll bei der Darstellung von Thesaurusrelationen nur zwischen Äquivalenzrelationen zu Deskriptoren und zu Nichtdeskriptoren sowie Nichtäquivalenzrelationen differenziert werden. Damit können alle Nichtäquivalenzrelationen zwischen identischen Benennungen als Dubletten angesehen und zu jeweils einer einzigen Relation zusammengefasst werden. Hierarchierelationen werden also wie Assoziationsrelationen dargestellt.

Zur besseren Verdeutlichung wird exemplarisch beschrieben, wie ein kumuliertes Wortfeld in einem Internetbrowser dargestellt werden könnte:

Ein Wortfeld bezieht sich immer auf die bei der Suche im Thesaurus gefundene Benennung (Suchbegriff), unabhängig davon, ob sie ein Deskriptor oder ein Nichtdeskriptor ist. Analog zur Dokumentsuche wird also vom Suchbegriff ausgegangen, um dessen verschiedene Rollen (Vorzugsbenennung oder Nichtdeskriptor) und deren Relationen zu anderen Benennungen wiederzugeben.

Jede Benennung kommt im Wortfeld nur einmal vor. Relationen zu Vorzugsbenennungen und zu Nichtdeskriptoren haben Vorrang vor Assoziations- und Hierarchierelationen. Durch die Benennung der jeweiligen Relationsart wird erkennbar, wenn mehrere widersprüchliche Relationen vorliegen. Benennungen werden mit ihren Wortfeldern verlinkt, wenn sie in mindestens einem Thesaurus Deskriptor sind, da es sich ansonsten ausschließlich um Äquivalenzrelationen zum aktuellen Suchbegriff handelt. Ein zusätzlicher Link oder Button ermöglicht die Suche im Dokumentenbestand mit dem aktuellen Suchbegriff.

Vorzugsbenennungen zum Suchbegriff (Äquivalenzrelation) werden hinter der Floskel „Benutzt als:“, Nichtdeskriptoren zum Suchbegriff (Äquivalenz-

relation) hinter der Floskel „Benutzt für:“ ausgegeben. Benennungen, die ausschließlich Nichtäquivalenzrelationen zum Suchbegriff haben, werden hinter der Floskel „Siehe auch“ dargestellt. Von der weiteren Benennung, in diesem Fall dem aktuellen Suchbegriff, wird auf die engere Benennung verwiesen.

Gibt es zu Benennungen des Wortfelds zugleich Äquivalenz- und Nichtäquivalenzrelationen, so ist es erforderlich, dem Nutzer deutlich zu machen, dass diese Benennungen nur teilweise mit der verlinkten Vorzugsbenennung bzw. dem verlinkten Nichtdeskriptor übereinstimmen. Die Floskel „Benutzt als:“ wird daher durch „Teilweise benutzt als:“ bzw. die Floskel „Benutzt für:“ durch „Teilweise benutzt für:“ ersetzt. Sind die Nichtdeskriptoren des Wortfelds zugleich auch Vorzugsbenennung des aktuellen Suchbegriffs, so werden sie nur einmal, und zwar hinter der Floskel „Teilweise benutzt als und zugleich benutzt für:“ ausgegeben.

An einer einfachen Suche nach „Forelle“ im Metathesaurus soll, abhängig von den Strukturen der zugehörigen Thesauri, gezeigt werden, wie sich die Wortfelder verändern:

1. Wenn ein Thesaurus die Vorzugsbenennung „Forelle“ mit dem Nichtdeskriptor „Salmo trutta“ (lateinisch für Forelle) und ein anderer Thesaurus den Deskriptor „Forelle“ mit Assoziationsrelationen zu „Regenbogenforelle“ und „Meerforelle“ enthält, so zeigt der Metathesaurus je nach gewähltem Suchbegriff folgende Wortfelder:

a.) Suchbegriff: Forelle

Benutzt für: Salmo trutta

Siehe auch: Meerforelle, Regenbogenforelle

b.) Suchbegriff: Salmo trutta

Benutzt als: Forelle

2. Wenn nun zusätzlich ein Thesaurus auch noch einen Nichtdeskriptor „Regenbogenforelle“ mit der Vorzugsbenennung „Forelle“ und ein anderer Thesaurus einen Nichtdeskriptor „Forelle“ mit der Vorzugsbenennung „Fisch“ liefert, so sieht dies im Metathesaurus wie folgt aus:

Suchbegriff: Forelle

Teilweise benutzt als: Fisch

Teilweise benutzt für: Regenbogenforelle, Salmo trutta

Siehe auch: Meerforelle

3. Stellt nun ein weiterer Thesaurus entgegengesetzt zum ersten Thesaurus eine Vorzugsbenennung „Salmo trutta“ mit dem Nichtdeskriptor „Forelle“ bereit, so verändert sich das Wortfeld erneut:

Suchbegriff: Forelle

Teilweise benutzt als und zugleich benutzt für: Salmo trutta

Teilweise benutzt als: Fisch

Teilweise benutzt für: Regenbogenforelle

Siehe auch: Meerforelle

Ohne Zweifel ist diese Darstellung eines Wortfelds nicht völlig selbst-erklärend. Dies gilt insbesondere für Extremfälle mit besonders widersprüchlichen Relationen. Da nicht mehr als die im letzten Beispiel dargestellten vier Relationsgruppen unterschieden werden, sollte sich die Komplexität des Metathesaurus jedoch für Endnutzer noch bewältigen lassen.

Einsatz als virtueller Metathesaurus

Bei Portalen und Suchmaschinen, die keinen zentralen Datenbestand führen und nur über Schnittstellen Daten bei der jeweiligen Suche „virtuell“ zusammenführen, werden Thesauri nur dann bereitgestellt, wenn sie über einfache Schnittstellen und ohne aufwendige Datenaufbereitungen genutzt werden können. Werden diese Thesauri beim Retrieval zu einem gemeinsamen Thesaurus zusammengeführt, handelt es sich um einen virtuellen Metathesaurus.

Der ausschließlich kumulierende syntaktische Metathesaurus ist gut für den Einsatz als virtueller Metathesaurus geeignet, da zur Darstellung der Wortfelder nur wenige Informationen der beteiligten Thesauri erforderlich sind und daher performant aufbereitet werden können:

- Pro Thesaurus wird nur einmal mit dem Suchbegriff nach seiner Vorzugsbenennung und ihren Relationen zu Deskriptoren und Nichtdeskriptoren gesucht.
- Für alle Nichtdeskriptoren zum Suchbegriff wird geprüft, ob sie zugleich in mindestens einem Thesaurus Deskriptor sind.

Automatische Optimierung eines syntaktischen Metathesaurus

Beim syntaktischen Metathesaurus mit ausschließlich kumulierenden Wortfeldern treten noch widersprüchliche Strukturen auf, die bereits ausschließlich über thesaurusbasierte Verfahren automatisch bereinigt werden können. Als Ergebnis steht dann ein der Form nach konsistenter Metathesaurus zur Verfügung.

Konflikte zwischen den beteiligten Thesauri führen dazu, dass der Meta-thesaurus gegen folgende Anforderungen an einen Thesaurus verstößt:⁹

1. Eine Benennung darf entweder nur Deskriptor oder Nichtdeskriptor sein.
2. Ein Nichtdeskriptor hat nur eine Vorzugsbenennung.
3. Die Vorzugsbenennung eines Nichtdeskriptors muss ein Deskriptor sein.
4. a) Ein Deskriptor darf direkt oder indirekt nicht zugleich sein eigener Nichtdeskriptor sein.
b) Dies gilt analog für das Verhältnis zwischen Oberbegriff und Unterbegriff.
5. In einem Thesaurus sollte es keine Hierarchien geben, die nur eine Teilmenge einer anderen Hierarchie darstellen und daher redundant sind.
6. Eine Umwandlung einer Monohierarchie in eine Polyhierarchie ist nicht zulässig.
7. Zwischen zwei Benennungen darf es nur eine Relation geben (paarweise disjunkte Relation).
8. Selbstverweise sind nicht zulässig.

Nur die Umwandlung einer Monohierarchie in eine Polyhierarchie leitet sich nicht aus der DIN 1463-1 [3] ab. Da derartige Veränderungen einen erheblichen Eingriff in die Thesaurusstruktur darstellen und unter Umständen verhindert werden sollen, wurde beim syntaktischen Modell eine Möglichkeit zur Unterdrückung von Polyhierarchien berücksichtigt. Bei Thesauri, die im Gegensatz zu Klassifikationen meist polyhierarchisch sind, dürfte dieses Problem jedoch nicht von grundsätzlicher Bedeutung sein.

Aufbauend auf dem Prinzip des kleinsten gemeinsamen Nenners können Verstöße gegen Anforderungen des Metathesaurus durch verschiedene automatische Bearbeitungsschritte korrigiert werden. Bei der Zusammenführung von Benennungen werden deren Relationen ebenfalls mitgeführt (vererbt) und um Dubletten bereinigt. Zur besseren Verdeutlichung werden die einzelnen Bearbeitungsschritte im Folgenden jeweils durch ein einfaches Beispiel und eine formelartige Regel¹⁰ ergänzt.

⁹ Bei Nikolai [10] werden die dargestellten Anforderungen mit Ausnahme von Nummer 1 und 4 a) auf der Grundlage von Sintichakis [15, S. 132], und Viegener [21, S. 68-69], als Invarianten eines Thesaurus behandelt und Algorithmen zur Bearbeitung von Konflikten beschrieben. Mili [9] beschäftigt sich mit Verfahren zur Bearbeitung von Verstößen gegen die Anforderungen 4 b), 5 und 7.

¹⁰ Zur Beschreibung der automatischen Verfahrensregeln seien folgende Mengen definiert: Die Menge der Benennungen ($B_i = \{b_1, b_2, \dots, b_i\}$). Die Mengen

Nichtdeskriptordominanz

Ein Deskriptor kann nachträglich problemlos in einen Nichtdeskriptor umgewandelt werden, da er direkt mit den indexierten Dokumenten verbunden ist. Umgekehrt ist eine entsprechende Umwandlung eines Nichtdeskriptors, dessen Dokumente mit der Vorzugsbenennung indexiert wurden, ohne Kenntnis aller unter seiner Vorzugsbenennung indexierten Dokumente nicht möglich. Bei einem thesaurusbasierten Verfahren ist daher die Festlegung einer Benennung als Nichtdeskriptor nicht umkehrbar, so dass man von einer Dominanz des Nichtdeskriptors sprechen kann. Wenn eine Benennung in den beteiligten Thesauri zugleich Deskriptor und Nichtdeskriptor ist (Verstoß gegen Anforderung 1), wird sie im Metathesaurus zum Nichtdeskriptor.

Regel 1 (s. Abb. 2):

$$(b_{1D}, b_{2ND}) + b_{2D} \Rightarrow (b_{1D}, b_{2ND})$$

Beispiel: Ein Nichtdeskriptor „Regenbogenforelle“ zu einer Vorzugsbenennung „Forelle“, der zugleich in einem anderen Thesaurus Deskriptor ist, führt dazu, dass im Metathesaurus „Regenbogenforelle“ Nichtdeskriptor mit der Vorzugsbenennung „Forelle“ wird.

Mehrdeutige Nichtdeskriptor-Deskriptor-Relation

Indem ein Nichtdeskriptor im beteiligten Thesaurus immer nur mit einer Vorzugsbenennung in Relation steht, ist sichergestellt, dass alle Dokumente zu dieser Benennung unter einem einzigen Deskriptor stehen. Wenn nun identische Nichtdeskriptoren in den beteiligten Thesauri mit verschiedenen Vorzugsbenennungen in Relation stehen (Verstoß gegen Anforderung 2), so wird aus mehreren eindeutigen Relationen (1D-nND-Relation) eine mehrdeutige Nichtdeskriptor-Deskriptor-Relation des Metathesaurus

der Deskriptoren ($D_j = \{b_{1D}, b_{2D}, \dots, b_{jD}\}$), der Nichtdeskriptoren ($ND_k = \{b_{1ND}, b_{2ND}, \dots, b_{kND}\}$), der Unterbegriffe ($DU_i = \{b_{1DU}, b_{2DU}, \dots, b_{iDU}\}$) und der Oberbegriffe ($DO_o = \{b_{1DO}, b_{2DO}, \dots, b_{oDO}\}$). Eine Deskriptorgruppe DG_p ist die Menge der Deskriptoren 1 bis p, die im Metathesaurus zusammengefasst und wie ein einziger Deskriptor behandelt werden ($DG_p = \{b_{1D}, b_{2D}, \dots, b_{pD}\}$). Eine Thesaurusrelation wird als Wertepaar zweier Benennungen dargestellt, z. B. eine Äquivalenzrelation zwischen dem Nichtdeskriptor b2 und einer beliebigen Vorzugsbenennung b1 als (b_{1D}, b_{2ND}) . Um die Lesbarkeit zu vereinfachen, steht bei Regeln mit n+1 Benennungen b_x stellvertretend für b_{n+1} .

(nD-nND-Relation), so dass ein Nichtdeskriptor n Vorzugsbenennungen hat (1ND-nD-Stern). Durch eine Zusammenführung der Vorzugsbenennungen zu einem „Deskriptor“ kann die Eindeutigkeit dieser Relation wiederhergestellt werden. Statt einer einzigen Benennung besitzt dieser „Deskriptor“ eine Gruppe gleichrangiger Benennungen, wenn aufgrund der Thesaurusrelationen diese Benennungen ähnlich wie Synonyme bzw. Quasisynonyme behandelt werden sollen. Er wird daher als Deskriptorgruppe bezeichnet.

Regel 2 (s. Abb. 2):

$$(b_{1D}, b_{X_{ND}}) + (b_{2D}, b_{X_{ND}}) + \dots + (b_{nD}, b_{X_{ND}}) \Rightarrow (DG_n, b_{X_{ND}})$$

Beispiel: Ist „Regenbogenforelle“ Nichtdeskriptor sowohl zur Vorzugsbenennung „Forelle“ als auch zu „Salmo trutta“, so wird daraus eine Äquivalenzrelation zur Deskriptorgruppe „Forelle/ Salmo trutta“.

Benennungsketten

Benennungen, die über mehrstufige Hierarchie- oder Äquivalenzrelationen verbunden sind, werden im Folgenden als Benennungsketten bezeichnet. Eine Benennungskette (kurz: Kette) hat mindestens zwei gleichartige Relationen. Im Metathesaurus können nach der Zusammenführung übereinstimmender Benennungen nicht nur mehr, sondern auch längere Benennungsketten als in den beteiligten Thesauri vorkommen.

Nichtdeskriptorkette

Die oben beschriebene Nichtdeskriptordominanz führt dazu, dass auch solche Deskriptoren in Nichtdeskriptoren umgewandelt werden, die selbst Vorzugsbenennung für weitere Nichtdeskriptoren sind. In diesem Fall wären Nichtdeskriptoren des Metathesaurus nur noch indirekt, also über einen anderen Nichtdeskriptor, mit einem Deskriptor verbunden (Verstoß gegen Anforderung 3 durch eine Kette 1D-nND-nND...). Beim Navigieren zwischen dem Nichtdeskriptor und seinem Deskriptor (Vorzugsbenennung) können unnötige Zwischenschritte vermieden werden, indem alle beteiligten Nichtdeskriptoren direkt in Relation zum Deskriptor der Benennungskette gesetzt werden. Eine 1D-nND-Kette wird also zu einem 1D-nND-Stern.

Regel 3 (s. Abb. 2):

$$\begin{aligned} & (b_{X_D}, b_{1_{ND}}) + (b_{1_D}, b_{2_{ND}}) + \dots + (b_{n-1_D}, b_{n_{ND}}) \Rightarrow \\ & (b_{X_D}, b_{1_{ND}}) + (b_{1_{ND}}, b_{2_{ND}}) + \dots + (b_{n-1_{ND}}, b_{n_{ND}}) = b_{X_D}\text{-ND}_n\text{-Kette} \Rightarrow \\ & (b_{X_D}, b_{1_{ND}}) + (b_{X_D}, b_{2_{ND}}) + \dots + (b_{X_D}, b_{n_{ND}}) = b_{X_D}\text{-ND}_n\text{-Stern} \end{aligned}$$

Beispiel: Wenn „Regenbogenforelle“ in einem Thesaurus Nichtdeskriptor zur Vorzugsbenennung „Forelle“, „Forelle“ jedoch in einem anderen Thesaurus Nichtdeskriptor zur Vorzugsbenennung „Fisch“ ist, so werden im Meta-thesaurus sowohl „Regenbogenforelle“ als auch „Forelle“ Nichtdeskriptoren zum Deskriptor „Fisch“.

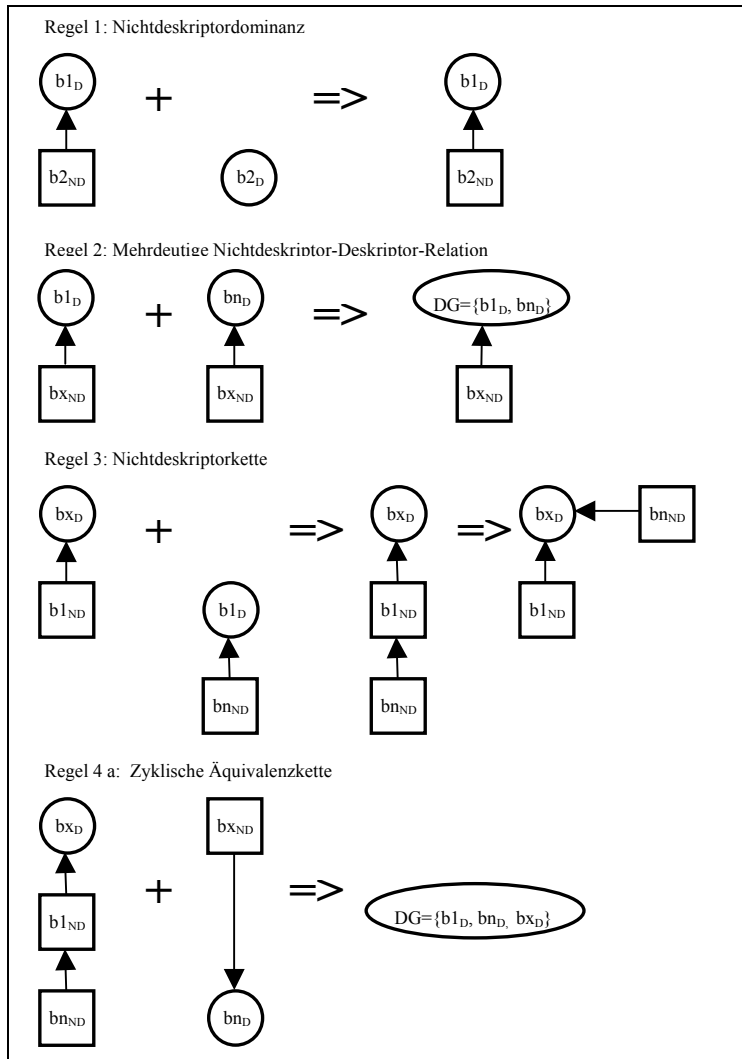


Abb. 2: Konfliktbeseitigung bei Äquivalenzrelationen

Zyklische Hierarchie- oder Äquivalenzkette

Es ist offensichtlich, dass ein Deskriptor nicht zugleich direkt oder indirekt sein eigener Nichtdeskriptor und ein Oberbegriff nicht zugleich sein eigener Unterbegriff sein darf. Im Gegensatz zu Assoziationsrelationen sind bei Hierarchie- und Äquivalenzrelationen rekursive Bezüge durch direkte und indirekte zyklische Ketten nicht zulässig, da diese Relationen von eindeutigen Vorzugsbenennungen bzw. Über- und Unterordnungen abhängen (geschlossene Nichtdeskriptor- oder Deskriptorkette bzw. auch eine Mischung aus beiden Arten von Benennungsketten). Wenn im Metathesaurus aufgrund von Widersprüchen zwischen den beteiligten Thesauri zyklische Hierarchie- oder Äquivalenzketten auftreten (Verstoß gegen Anforderung 4), ist es nicht möglich, automatisch zu entscheiden, welche dieser alternativen Relationen „falsch“ sind und wie sie für den Metathesaurus bearbeitet werden müssten. Es sollen daher alle an der zyklischen Kette beteiligten Benennungen zu einer Deskriptorgruppe zusammengefasst werden. Eine Beschränkung von Anforderung 4 auf transitive Relationen¹¹[10, S. 63] ist mathematisch gesehen nahe liegend, jedoch aus dokumentarischer Sicht nicht sinnvoll. Widersprüche treten auch bei einer zyklischen Kette von nicht transitiven Bestandsrelationen auf, z. B. dadurch dass ein enges Schlagwort zum Oberbegriff eines weiten Schlagworts werden kann.

Regel 4 a (Zyklische Äquivalenzkette s. Abb. 2):

$$b_{x_D}\text{-ND}_n\text{-Kette} + (b_{n_D}, b_{x_{ND}}) = \text{geschlossene ND}_{n+x}\text{-Kette} \Rightarrow \text{DG}_{n+x}$$

Regel 4 b (Zyklische Hierarchiekette s. Abb. 3):

$$(b_{1_{DO}}, b_{2_{DU}}) + (b_{2_{DO}}, b_{3_{DU}}) + \dots + (b_{(n-1)_{DO}}, b_{n_{DU}}) + (b_{1_{DU}}, b_{n_{DO}}) \\ = \text{D}_n\text{-Hierarchiekette} + (b_{1_{DU}}, b_{n_{DO}}) = \text{geschlossene D}_n\text{-Hierarchiekette} \Rightarrow \\ \text{DG}_n$$

Beispiel: Wenn es in zwei beteiligten Thesauri Äquivalenzrelationen zwischen „Forelle“ und „Salmo trutta“ gibt, aber in einem Fall die eine und im anderen Fall die andere Benennung Nichtdeskriptor ist, dann wird aus beiden Benennungen eine Deskriptorgruppe „Forelle/ Salmo trutta“.

¹¹ Für transitive Relationen gilt: $b_{x_D} - b_{y_D} + b_{y_D} - b_{z_D} \Rightarrow b_{x_D} - b_{z_D}$. Relationen mit Nichtdeskriptoren entziehen sich jedoch dieser mathematischen Betrachtungsweise [13].

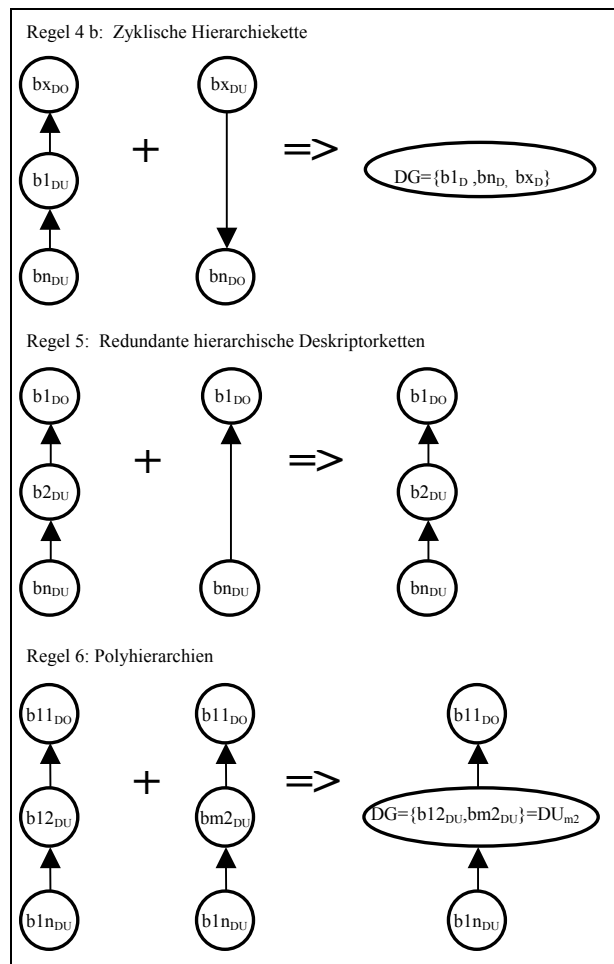


Abb. 3: Konfliktbeseitigung bei Hierarchierelationen

Redundante hierarchische Deskriptorketten

Eine Hierarchie, die nur eine Teilmenge einer anderen Hierarchie darstellt (Verstoß gegen Anforderung 5), ist redundant und sollte daher entfallen. Dies ist auch der Fall, wenn ein Deskriptor in einem beteiligten Thesaurus eine oder mehrere hierarchische Ebenen tiefer angeordnet wird als in einem anderen. Der kleinste gemeinsame Nenner ist in diesem Fall die Hierarchie mit den meisten Ebenen, da sie alle Elemente der anderen Hierarchie enthält [9, S. 214]. Etwas

abstrakter formuliert: Hat nur eine von zwei hierarchischen Deskriptorketten zwischen zwei identischen Deskriptoren mindestens einen zusätzlichen Deskriptor, so entfällt an dieser Stelle die direkte Relation zwischen den identischen Deskriptoren zugunsten der längeren Kette, da sie nur eine Teilmenge der längeren Deskriptorkette darstellt (s. Abb. 3). Wenn die Hierarchieketten unterschiedliche Reihenfolgen haben, so könnte eine Rangfolge der Thesauri darüber entscheiden, welche Reihenfolge in den Metathesaurus übernommen wird.

Einerseits ist eine Eliminierung impliziter Relationen, die sich vor allem bei Abstraktionsrelationen durch die Transitivität von Ketten ergeben können, sinnvoll, [10, S. 70]; [21, S. 64]. Da andererseits viele Thesauri keine eindeutigen Unterscheidungsmerkmale zwischen Abstraktions- und Bestandsrelationen liefern können, wurde beim hier dargestellten Modell darauf verzichtet.

Regel 5 (s. Abb. 3):

$$(b1_{DO}, b2_{DU}) + (b2_{DO}, b3_{DU}) + \dots + (b(n-1)_{DO}, bn_{DU}) + (b1_{DO}, bn_{DU}) \\ = D_n\text{-Hierarchiekette} + (b1_{DO}, bn_{DU}) \Rightarrow D_n\text{-Hierarchiekette}$$

Beispiel: In einem beteiligten Thesaurus wird folgende hierarchische Kette dargestellt: Oberbegriff „Fisch“ mit dem Unterbegriff „Forelle“ und dieser mit dem Unterbegriff „Regenbogenforelle“. Ein anderer Thesaurus kennt nur den Oberbegriff „Fisch“ mit dem Unterbegriff „Regenbogenforelle“. Da die Hierarchie des zweiten Thesaurus im ersten mit einer zusätzlichen Zwischenebene enthalten ist, kann die direkte Hierarchierelation zwischen „Regenbogenforelle“ und „Fisch“ entfallen.

Entstehung von Polyhierarchien

Haben mehrere hierarchische Deskriptorketten zwischen identischen Deskriptoren abweichende Deskriptoren, so entstehen zusätzliche polyhierarchische Strukturen. Da Thesauri meist polyhierarchisch strukturiert sind, spricht einiges dafür, diese Strukturveränderungen zuzulassen. Sollen nachträglich entstandene Polyhierarchien dennoch stören (Verstoß gegen Anforderung 6), z. B. weil grundsätzlich eine Mono-Hierarchie erhalten bleiben soll, kann durch die Bildung einer Deskriptorgruppe der alternativen Deskriptoren eine monohierarchische Struktur wiederhergestellt werden. Dies ist jedoch problematisch, da diese Deskriptoren unterschiedlichsten Hierarchieebenen entsprechen und daher als Deskriptorgruppe inhomogen sein können. Außerdem lässt sich die hier angegebene Regel nur auf gleichlange Ketten anwenden.

$$\begin{aligned} & \text{Regel 6 (s. Abb. 3):} \\ & (b11_{DO}, b12_{DU}) + (b12_{DO}, b13_{DU}) + \dots + (b1(n-1)_{DO}, b1n_{DU}) + \\ & \begin{matrix} \vdots \\ (b11_{DO}, b12_{DU}) + (b12_{DO}, b13_{DU}) + \dots + (b1(n-1)_{DO}, b1n_{DU}) \Rightarrow \\ \vdots \\ (b11_{DO}, DU_{m2}) + (DO_{m2}, DU_{m3}) + \dots + (DO_{m(n-1)}, b1n_{DU}) \end{matrix} \end{aligned}$$

Beispiel: Eine Kombination der hierarchischen Kette Oberbegriff „Fisch“ - Unterbegriff „Meeresfisch“ - Unterbegriff „Meerforelle“ mit einer hierarchischen Kette Oberbegriff „Fisch“ - Unterbegriff „Forelle“ - Unterbegriff „Meerforelle“ würde bei Einhaltung der Monohierarchie eine inhomogene Deskriptorgruppe „Meeresfisch/Forelle“ erzeugen.

Paarweise disjunkte Relationen und Selbstverweise

Die Zusammenführung der Benennungen beteiligter Thesauri und die Bildung von Deskriptorgruppen können dazu führen, dass zwischen zwei Benennungen mehr als eine Relation und auch Relationen von einer Benennung zu sich selbst (Selbstverweise) entstehen. Der erste Fall verstößt gegen Anforderung 7, so dass die Relationen nicht mehr paarweise disjunkt sind. Dies ist jedoch für eine eindeutige terminologische Kontrolle erforderlich. Durch eine Rangfolge der Relationsarten, ggf. kombiniert mit einer Rangfolge der Thesauri, können Relationen mit geringerer Priorität im Metathesaurus unterdrückt werden, vgl. [9; 10]. Bei den Relationsarten sollten aufgrund der Nichtdeskriptordominanz Relationen mit Nichtdeskriptoren Vorrang haben (Regel 7). Ansonsten hätte die jeweils schwächste Relation Vorrang (in der Regel Assoziationsrelationen). Im zweiten Fall, der vor allem bei der Zusammenführung von Nichtdeskriptoren und Deskriptoren sowie bei der Vererbung von Relationen eines Deskriptors auf seine Deskriptorgruppe von Bedeutung sein dürfte, liegt ein Verstoß gegen Anforderung 8 vor. Durch die Eliminierung von Selbstverweisen kann dieser Konflikt beseitigt werden (Regel 8).

Rahmenbedingungen für den Einsatz syntaktischer Metathesauri

Ausgehend von den hinter den Benennungen stehenden Begriffen ist auch bei identischen Benennungen Inhomogenität zwischen den Thesauri erkennbar. Folgende Ausgangssituationen können dazu führen, dass zu einer Benennung eines Thesaurus keine bzw. eine Benennung in einem anderen Thesaurus gefunden wird:

Eine Benennung für einen Begriff

Im idealtypischen Fall eines syntaktischen Metathesaurus stehen identische Benennungen beteiligter Thesauri auch für identische Begriffe. Meist bedeutet dies, dass eine Benennung für einen Begriff des Thesaurus steht. Wird keine identische Benennung gefunden, so bedeutet dies im Umkehrschluss, dass der jeweilige Begriff nicht vorhanden ist. Benennungen und ihre Relationen können daher weitgehend problemlos kumuliert werden. Dieser Fall wird jedoch nicht immer gegeben sein.

Eine Benennung für verschiedene Begriffe

Es kommt auch vor, dass bereits in einem einzigen Thesaurus eine Benennung für mehrere Begriffe steht. Dies gilt vor allem bei Polysemen und Vorzugsbenennungen von Quasisynonymen. In der Regel sollte von den Bearbeitern der beteiligten Thesauri gewährleistet werden, dass hierbei sich widersprechende und irreführende Relationen ausgeschlossen sind. Solche Fälle können jedoch auch dazu führen, dass eine Benennung in einem Thesaurus mehreren Benennungen in anderen Thesauri entspricht (z. B. bei unterschiedlich benannten Homonymen). Dadurch ist es möglich, dass teilidentische Benennungen nicht erkannt oder aber zu einer Deskriptorgruppe mit sehr ähnlichen Benennungen zusammengeführt werden.

Wenn nun aber die Benennung eines Homonyms in einem beteiligten Thesaurus für einen Begriff und in einem anderen Thesaurus für einen anderen Begriff steht, wäre in den jeweiligen Thesauri ein erläuternder Homonymzusatz zu den Benennungen nicht erforderlich (z. B. „Zensur“ als Schulnote in einem beteiligten Thesaurus und „Zensur“ als Kontrolle der Presse in einem anderen). Werden aber beide Thesauri kumuliert, dann führt dies zu einer unkontrollierten Mischung zweier Begriffe. Bei der Kumulation von Thesauri zu sehr unterschiedlichen Sachgebieten kann dies zu völlig abwegigen Homonymen und Wortfeldern führen. Eine nachträgliche Kontrolle der Polyseme ist nur möglich, wenn eine semantische Analyse der Thesaurusstruktur und der indexierten Dokumente erfolgt.

Verschiedene Benennungen für einen Begriff

Bei weitgehend übereinstimmenden Thesauri können unterschiedliche Benennungen für denselben Begriff auftreten. Die Ursachen dafür können kleine Ansatzungsunterschiede (Singular oder Plural, Eingabefehler, terminologische Fehler etc.) oder unterschiedlich angesetzte Polyseme sein. Beim der Form nach konsistenten syntaktischen Metathesaurus, dessen Konflikte durch die

o. g. Automatisierungsregeln bereinigt wurden, bleiben entsprechende Benennungen getrennt, obwohl ein Endnutzer sie leicht als vergleichbar erkennen kann. Hier würde eine halbautomatische Korrektur des syntaktischen Thesaurus ausreichen, um inhaltliche Inkonsistenzen zu beseitigen.

Gehören die beteiligten Thesauri zu unterschiedlichen natürlichen Sprachen (Mehrsprachigkeit) oder unterschiedlichen Dokumentationssprachen (präkombiniert oder postkoordiniert, natürlichsprachig oder Kunstsprache etc.), dann stehen regelmäßig unterschiedliche Benennungen für denselben Begriff. Dieses Problem kann auch durch eine abweichende Indexierungspraxis verursacht werden (insbesondere bei unterschiedlicher Indexierungsspezifität, d. h. Verwendung eines engen oder weiten Schlagworts). Im Gegensatz zu einem semantischen Metathesaurus kann ein syntaktischer Metathesaurus diese Thesauri nicht zufrieden stellend zusammenführen.

Eine unterschiedliche Indexierungspraxis führt auch zu einem grundsätzlichen Problem von Metathesauri: Wenn ein Endnutzer ein enges Schlagwort aufgrund eines beteiligten Thesaurus gefunden hat, nimmt er an, die wichtigsten Dokumente zu dem entsprechenden Begriff seien unter diesem Deskriptor indexiert. In einem anderen beteiligten Thesaurus, der das enge Schlagwort nicht enthält, könnten entsprechende Dokumente jedoch unter einem weiteren Schlagwort indexiert worden sein. Die Erwartung des Endnutzers an den Metathesaurus richtet sich daher nach dem Thesaurus mit der höchsten Indexierungsspezifität.

Beispiele:

- Geringe Ansetzungsunterschiede können zu „Wahlbeamter“, „Wahlbeamter/ Wahlbeamtin“ und „Wahlbeamte“ führen.
- Postkoordiniert würde man mit „Verwaltung“ und „Land (Gebietskörperschaft)“, präkombiniert mit „Landesverwaltung“ indexieren.
- Bei geringerer Indexierungsspezifität würde man mit „Verfassung“, bei höherer Indexierungsspezifität mit „Landesverfassung“ oder „Grundgesetz“ oder „Kommunalverfassung“ indexieren.

Man kann zwar versuchen, den Empfehlungen von Doerr [4] zu folgen, indem man zur Vermeidung semantischer Probleme Thesauri besser dokumentiert und ihre Relationen nur noch zwischen Begriffen und nicht zwischen Benennungen anlegt. Dennoch wird dies in der Praxis genauso wenig zur umfassenden Beseitigung dieser Probleme beitragen wie seine Annahme, dass das Erkennen von Eingabe- und Indexierungsfehlern auch tatsächlich zur Fehlerbehebung führt. Oftmals ist eine zeitnahe Anpassung beteiligter Thesauri nicht möglich oder wird aus Sicht der beteiligten Systeme grundsätzlich abgelehnt. In der Praxis wird die Inhomogenität zwischen den Thesauri daher

selbst bei semantischen Metathesauri nur gemildert, nicht jedoch vollständig beseitigt werden.

Eine automatische Normierung der Ansetzung von Benennungen [10, S. 146-148] verhindert, dass schon geringe Ansetzungsunterschiede der Benennungen Deskriptorgruppen verursachen oder die Zusammenführung identischer Benennungen stören. Da eine normierte Benennung nicht zur Darstellung beim Retrieval geeignet sein dürfte, sollte über eine Rangfolge der beteiligten Thesauri die im Metathesaurus stattdessen darzustellende Benennung festgelegt werden.

Die für den syntaktischen Metathesaurus vorgeschlagene Bildung von Deskriptorgruppen stellt eine gut handhabbare, einfache technische Lösung dar, die abhängig von den beteiligten Thesauri jedoch zu erheblichen inhaltlichen Inkonsistenzen führen kann:

- Schon eine einzige Relation eines beteiligten Thesaurus, also auch ein Eingabefehler, kann eine Benennungskette in eine Deskriptorgruppe verwandeln. Eine Beseitigung oder Korrektur der fehlerhaften Relation wäre hier sinnvoller.
- Die zuvor beschriebenen Ansetzungsunterschiede bei Benennungen können zur Bildung von Deskriptorgruppen mit nahezu identischen Benennungen führen, die man lieber durch eine Benennung ersetzen würde.
- Der Wunsch nach Vermeidung von Polyhierarchien führt zu Deskriptorgruppen, bei denen die Benennungen ganz unterschiedlichen Hierarchieebenen entstammen, so dass sie nicht als homogene Gruppen empfunden werden (inhomogene Deskriptorgruppen).

Semantische Thesauruskonzepte behandeln Konsistenzprobleme, die im syntaktischen Modell zur Bildung von Deskriptorgruppen führen, indem sie z. B. eine zyklische Hierarchiekette durch Eliminierung oder Umwandlung mindestens einer beteiligten Relation unterbrechen, vgl. [10, S. 89-90; 9, S. 214]. In der Regel kann erst durch den Einsatz von Bearbeitern vermieden werden, dass alle an der Kette beteiligten Relationen eliminiert oder in Assoziationsrelationen umgewandelt werden. Für einen automatisch geführten Metathesaurus stellt daher nur der Ersatz aller an zyklischen Ketten beteiligten Hierarchierelationen durch Assoziationsrelationen eine sinnvolle Alternative zur Bildung von Deskriptorgruppen dar.¹²

¹² Analog könnten auch zyklische Äquivalenzketten auf der Basis von 1D-nD-Äquivalenzrelationen in Assoziationsrelationen umgewandelt werden.

Nachgelagerte halbautomatische Verfahren

Wie bereits festgestellt wurde, lassen sich durch die oben beschriebenen vollautomatischen Verfahren nicht alle möglichen Inkonsistenzen beseitigen. Ein einfaches, manuell geführtes Regelwerk für automatisch durchzuführende Korrekturmechanismen kann insbesondere zur Beseitigung ungewollter Deskriptorgruppen und zum gezielten Zusammenführen von Benennungen genutzt werden.

Dies ist über folgende Regeln möglich:

1. Es wird zwischen zwei Benennungen eine neue Relation angelegt.
2. Eine Relation zwischen zwei Benennungen wird entfernt oder durch eine Relation einer anderen Relationsart ersetzt.
3. Eine Benennung wird zu Gunsten einer neuen Benennung entfernt.

Problematische Deskriptorgruppen können durch die erste Regel erzeugt, durch die zweite Regel vermieden und durch die dritte Regel umbenannt werden. Wenn die Benennungen einer Regel auch über unscharfe Suchbegriffe definiert werden (Endmaskierung, Kombination von Suchbegriffen über Bool'sche Operatoren etc.), kann der Aufwand zur Pflege des Regelwerks erheblich begrenzt werden.

Ergänzend zu diesem Regelwerk kann auch noch ein im Zweifelsfall maßgeblicher Leitthesaurus oder eine Rangfolge der beteiligten Thesauri festgelegt werden. Reichen einfache halbautomatische Verfahren nicht aus, um eine ausreichende Konsistenz herzustellen, wäre neben der Wahl eines anderen Thesauruskonzepts (semantischer Metathesaurus oder Makrothesaurus) auch eine Kombination des syntaktischen Metathesaurus mit semantischen Verfahren möglich.

Indexierungsspezifität, Inkohärenz und Polyhierarchien

Die höchste Indexierungsspezifität, die ein Metathesaurus bieten kann, ist die Indexierungsspezifität der jeweils beteiligten Einzelthesauri. Der rein kumulative syntaktische Metathesaurus kann dies auf Kosten einer widerspruchsfreien Thesaurusstruktur erreichen.

Mit den beschriebenen automatischen Optimierungsverfahren wird ein der Form nach konsistenter syntaktischer Metathesaurus erzeugt. Dies kann sich jedoch nachteilig auf die Indexierungsspezifität und die Kohärenz der Benennungen auswirken:

- Die Umwandlung von Deskriptoren in Nichtdeskriptoren senkt für die beteiligten Systeme, bei denen diese Benennung bisher Deskriptor war,

die Indexierungsspezifität. Die Bildung von Deskriptorgruppen verringert die Indexierungsspezifität aller beteiligten Systeme.

- Je nach Ausgangssituation kann die Kohärenz der Deskriptorgruppen derart gering sein, dass halbautomatische Verfahren zur Vermeidung unsinniger Benennungen und zur Erhöhung der Indexierungsspezifität für erforderlich gehalten werden könnten: Sind Deskriptorgruppen homogen, wirkt sich dies auf die Indexierungsspezifität nur unwesentlich aus. Bei inhomogenen Deskriptorgruppen kann sich die Indexierungsspezifität jedoch erheblich vermindern.

Es bleibt zu klären, in welchem Umfang beim automatisch optimierten syntaktischen Metathesaurus Indexierungsspezifität und Kohärenz sinken und Polyhierarchien erzeugt werden. Diese Angaben sind wichtig, um eine umfassende Abwägung zwischen den Vor- und Nachteilen eines rein kumulativen und eines automatisch optimierten syntaktischen Metathesaurus zu treffen und damit die Notwendigkeit halbautomatischer Verfahren oder alternativer Konzepte beurteilen zu können.

Automatisches Verfahren zur Erstellung eines syntaktischen Metathesaurus

Grundsätzlich sollen beim syntaktischen Metathesaurus Strukturen der beteiligten Thesauri unverändert erhalten bleiben. Konflikte mit den Anforderungen des Metathesaurus sind zu markieren. Bei Abweichungen zwischen dem Metathesaurus und den beteiligten Thesauri ist es erforderlich, die zugrunde liegenden Konflikte zu dokumentieren.¹³ Dies gewährleistet die für die Pflege des Metathesaurus notwendige Transparenz und Umkehrbarkeit von Strukturveränderungen. Das Zusammenlegen oder Eliminieren von Benennungen und Relationen sowie die Änderung eines Deskriptors in einen Nichtdeskriptor erfolgt daher über zusätzliche Angaben und Markierungen, die das Retrieval und die Ausgabe des Metathesaurus steuern. Bei der Zusammenführung von Benennungen werden die Relationen an die Benennung des Metathesaurus vererbt.

Die oben beschriebenen Verfahren zur Konfliktbeseitigung können wie folgt organisiert werden.¹⁴

¹³ Hierzu reicht eine Interlingua aus, die ausschließlich die unveränderten Daten der beteiligten Thesauri durch zusätzliches Integrationswissen ergänzt und daher von Nikolai [10, S. 80] Extralingua genannt wird.

¹⁴ Einige Verfahrensschritte bauen aufeinander auf und müssen daher zwingend in der vorgegebenen Reihenfolge abgearbeitet werden; andere Abfolgen verringern dagegen nur den Bearbeitungsaufwand. Da z. B. durch

1. Die Normierung von Benennungen erfolgt vor dem Benennungsabgleich (mindestens Angleichung des Genus und der Schreibweisen von Sonderzeichen und Leerstellen).
2. Benennungsdubletten werden eliminiert, indem man identische Deskriptoren zu einem einzigen Deskriptor und identische Nichtdeskriptoren zu einem einzigen Nichtdeskriptor zusammengeführt.
3. Die Nichtdeskriptordominanz (Regel 1) wandelt Deskriptoren in Nichtdeskriptoren derselben Benennung um.
4. Nichtdeskriptorketten (Regel 3) werden abgebaut, so dass alle Nichtdeskriptoren den Deskriptor der Kette als Vorzugsbenennung erhalten (Ausnahme: zyklische Äquivalenzkette, s. u.).
5. Mehrdeutige Nichtdeskriptor-Deskriptor-Relationen (Regel 2) führen zur Bildung einer Deskriptorgruppe der Vorzugsbenennungen.
6. Bei redundanten hierarchischen Deskriptorketten (Regel 5) werden gleichlange Ketten zusammengeführt, und bei unterschiedlich langen Hierarchieketten wird die kürzere Kette entfernt. Die Reihenfolge der Hierarchieebenen richtet sich nach dem Thesaurus mit der höchsten Priorität.
7. Bei zyklischen Hierarchie- oder Äquivalenzketten (Regel 4) bilden alle Benennungen der jeweiligen Kette eine Deskriptorgruppe. Zur Vermeidung von Deskriptorgruppen können die Hierarchierelationen zyklischer Deskriptorketten alternativ in Assoziationsrelationen umgewandelt werden.
8. Bei Bedarf können Polyhierarchien (Regel 6) von gleichlangen Benennungsketten durch die Bildung von Deskriptorgruppen aus Benennungen derselben Hierarchieebene wieder in Monohierarchien verwandelt werden.
9. Selbstverweise werden entfernt (Regel 8).
10. Relationen werden wieder paarweise disjunkt, indem beim Vorliegen mehrerer Relationen zwischen zwei Benennungen überzählige Relationen entfernt werden (Regel 7: Es setzen sich Relationen mit Nichtdeskriptoren und ansonsten die Relation mit der schwächsten Relationsart durch.).

Zusammenführung von Thesauri in Parlamentsinformationssystemen

Anlass für die Überlegungen zu einem syntaktischen Metathesaurus war die Umstellung des bis dahin durch eine Zentraldokumentation geführten Parlamentsspiegels, einer Datenbank zum Nachweis deutscher Parlamentsmaterialien, auf eine dezentrale Anlieferung der Metadaten durch die einzelnen

die Nichtdeskriptordominanz neue Nichtdeskriptorketten entstehen können, darf Regel 3 erst nach Regel 1 abgearbeitet werden.

Landtagsinformationssysteme.¹⁵ Die deutschen Landesparlamente setzen den Parlamentsthesaurus des Bundestages PARTHES in Form eigener Anwenderthesauri (ANTHES) im Rahmen des sog. PARTHES-ANTHES-Verbunds ein, [6, S. 12-13; 11, S. 134-157]. Bei der Umstellung des Parlamentsspiegels auf eine dezentrale XML-Anlieferung wurde aus Kostengründen das Konzept eines gemeinsamen Makrothesaurus der Landesparlamente nicht weiter verfolgt. Stattdessen ist nur die Bereitstellung des PARTHES vorgesehen. Dies ist aus mehreren Gründen problematisch.

Ein Anwenderthesaurus besteht zu einem wesentlichen Teil aus PARTHES-Strukturen. Von diesen kann jedoch im Einzelfall abgewichen werden. PARTHES-Benennungen, die parlamentsübergreifend von Bedeutung sind, werden in der Regel ergänzt um viele nur lokal relevante Benennungen und Relationen.¹⁶ Auch zwischen ANTHES- und PARTHES-Benennungen können Thesaurusrelationen bestehen. Das bisherige Konzept des Parlamentsspiegels würde daher nur einen Teil der Benennungen und Relationen der Anwenderthesauri abbilden. Die Indexierung der Parlamentsdokumentationen basiert jedoch darauf, dass beim Retrieval Thesaurusrelationen zur Verfügung stehen und den Benutzer zur gesuchten Benennung führen. Dies gilt insbesondere für Mehrwortbenennungen (z. B. Benennungen von Institutionen und Rechtsvorschriften).

Da für jede Wahlperiode eines Parlaments ein eigenständiger Anwenderthesaurus mit den benutzten Bezeichnungen und Thesaurusrelationen aufgebaut wird, würde ein gemeinsamer Metathesaurus der Landesparlamente schon nach 10 Jahren etwa 30 Thesauri abbilden (16 Parlamente mit jeweils 2 Wahlperioden). Nur wenn der Aufwand der Thesaurusintegration und die Komplexität der Thesaurusstrukturen begrenzt werden können, ist eine bessere Thesaurusunterstützung des Parlamentsspiegels realisierbar.

Im Rahmen des Projekts zur Entwicklung eines neuen Landtagsinformationssystems Schleswig-Holstein (LIS-SH) wurde zusammen mit der GLOMAS Deutschland GmbH eine wahlperiodenübergreifende 1-Feld-Suche für die STAR-Datenbanken des Landtags realisiert. Auf die Integration des Thesaurus in die Internetsuche und ein besseres Kosten-Nutzen-Verhältnis des Thesauruseinsatzes wurde hierbei besonderen Wert gelegt [7]. Mittlerweile haben vier Landtage ihre Parlamentsinformationssysteme auf LIS-SH auf-

¹⁵ Internetadresse des Parlamentsspiegels, von der aus auch auf die einzelnen Landtagsinformationssysteme verlinkt wird. URL: <http://www.parlamentsspiegel.de>.

¹⁶ Zwei Drittel der Benennungen des schleswig-holsteinischen Anwenderthesaurus der laufenden 16. Wahlperiode stammen aus dem PARTHES (Stand: Januar 2006).

gebaut: Berlin, Brandenburg, Niedersachsen und Sachsen-Anhalt. Mit der Software STAR wäre es problemlos möglich, die bisher getrennt geführten Thesauri dieser Landtagsinformationssysteme für das Retrieval als einen gemeinsamen Metathesaurus in der jeweiligen wahlperiodenübergreifenden Suche darzustellen. Hierzu bietet sich eine Realisierung als virtueller Metathesaurus an, der bei Bedarf sogar um halbautomatische Korrekturmechanismen ergänzt werden könnte. Da auf die einzelnen Anwenderthesauri direkt zugegriffen werden kann und nur in begrenztem Umfang mit Inkonsistenzen gerechnet werden muss, ist hierfür nur ein geringer Entwicklungs- und nahezu kein Personalaufwand erforderlich.

Im PARTHES-ANTHES-Verbund kommen Äquivalenzrelationen und Hierarchierelationen mit den Relationen 1D-nD, 1D-nND und nD-nND vor. So können z. B. Synonyme auch wie Deskriptoren zur Indexierung verwendet werden (1D-nD-Äquivalenzrelation). Außerdem sind Hierarchie- und Assoziationsrelationen möglich, bei denen ein Element nicht zur Indexierung zugelassen wird und daher nur auf andere Deskriptoren hinweisen soll (nD-nND-Äquivalenzrelation oder 1D-nND-Hierarchierelation bzw. -Assoziationsrelation).

Grundsätzlich wäre ein syntaktischer Metathesaurus für den Parlamentsspiegel ausreichend, da alle beteiligten Systeme ein hohes Maß an Übereinstimmungen bei Benennungen und Relationen haben. Unterschiedliche Ansätze nahezu identischer Benennungen könnten sich hier in engen Grenzen halten. Durch das vorgestellte automatische Verfahren zum Aufbau eines konsistenten syntaktischen Metathesaurus sollte ohne erheblichen Zusatzaufwand ein Metathesaurus Parlamentsspiegel eingesetzt werden können. Die notwendigen Thesaurusdaten der Landtagsinformationssysteme müssten hierzu über ein noch festzulegendes einfaches XML-Austauschformat analog zum bisherigen XML-Export von Dokumentationsdaten an den Parlamentsspiegel geliefert werden.

Problembereiche eines Metathesaurus Parlamentsspiegel

Vor Einsatz eines syntaktischen Metathesaurus ist eine stichprobenartige Untersuchung zu den Unterschieden zwischen den ANTHES-Benennungen und Relationen der einzelnen Parlamente und Wahlperioden sinnvoll, um die Notwendigkeit halbautomatischer Korrekturen beurteilen zu können. Das Hauptproblem beim Metathesaurus Parlamentsspiegel dürften die o. g. Hierarchierelationen darstellen, bei denen der Unterbegriff als Nichtdeskriptor geführt werden kann (1D-nND-Hierarchierelationen, die den Äquivalenzrelationen des Metathesaurusmodells entsprechen). Im PARTHES werden diese Unterbegriffe Nichtregistoren genannt, da sie ursprünglich den Druck von

Papierregistern optimieren sollten. Aufgrund der Indexierungsspezifität und pragmatischer Einzelfallentscheidungen des jeweiligen Anwenderthesaurus kann es sein, dass sehr unterschiedliche Vorzugsbenennungen (hier: Oberbegriffe) für identische Nichtdeskriptoren gewählt werden. Es ist daher noch zu prüfen, in welchem Umfang inhomogene Deskriptorgruppen entstehen und wie hoch der Aufwand für halbautomatische Korrekturen ist. Inhomogene Deskriptorgruppen können vermieden werden, indem man die Relationen zyklischer Deskriptorketten vorrangig in Assoziationsrelationen umgewandelt.

Resümee

Thesauri können mit dem vorgestellten Modell eines kumulierenden syntaktischen Metathesaurus, ohne dass Inkonsistenzen zu beseitigen sind, sinnvoll für ein Retrieval eingesetzt und virtuell zusammengeführt werden. Durch Optimierungsverfahren lässt sich stattdessen aber auch ein der Form nach konsistenter Metathesaurus erzeugen. Von semantischen Modellen unterscheiden sich hierbei vor allem die Regeln zur Behandlung von Äquivalenzrelationen und die Möglichkeit zur Beschränkung auf vollautomatische, thesaurusbasierte Methoden. Konflikte zwischen den beteiligten Thesauri werden vorrangig durch die Bildung von Deskriptorgruppen aufgelöst. Die Beseitigung darüber hinausgehender inhaltlicher Inkonsistenzen kann im Nachhinein halbautomatisch erfolgen.

Das beschriebene syntaktische Modell ist weniger leistungsfähig als ein semantisches Verfahren, da es nur bei Thesauri eingesetzt werden kann, die die o. g. Voraussetzungen erfüllen. Dafür ermöglicht es, den Aufwand erheblich zu verringern und auf später durchzuführende Arbeitsschritte zu verlagern. Für einen sicherlich begrenzten Anwendungsbereich wird ein syntaktisches Modell besser geeignet sein als aufwendige semantische Verfahren. Es lässt sich mit einfachen Algorithmen ausschließlich auf der Grundlage von Thesaurusdaten und ohne erheblichen Personalbedarf umsetzen. Damit können „robuste“ Verfahren realisiert werden, die nur geringe Anforderungen an das Importformat des Metathesaurus stellen.

Da EDV-Konzepte häufig an ihrer Komplexität und den finanziellen Rahmenbedingungen scheitern, ist der personelle, organisatorische und finanzielle Aufwand ausschlaggebend bei ihrer Kosten-Nutzen-Bewertung. Erst durch den vergleichsweise geringen Aufwand für die Umsetzung eines syntaktischen Metathesaurus wird eine parlaments- und wahlperiodenübergreifende Zusammenführung parlamentarischer Thesauri möglich. Voraussetzung hierfür ist jedoch, dass es den Parlamentsdokumentationen gelingt, einen vernünftigen Kompromiss zwischen technischen Möglichkeiten und dokumentarisch-archivarischen Anforderungen zu finden und diesen dann auch in den Parlaments-

verwaltungen umzusetzen. Die intensive Mitwirkung der Dokumentationen an der Entwicklung neuer Parlamentsinformationssysteme und ihr Einsatz bei der Weiterentwicklung des Parlamentsspiegels haben gezeigt, dass hierfür die dokumentarischen Voraussetzungen gegeben sind.¹⁷

Literatur und Internetquellen

- 1 AGOGINO, G. A. (2004). *Developing a Learner-Centered Metathesaurus for Science, Mathematics, Engineering and Technology Education (Final Report)*. NSF Award DUE-0121743. URL: http://best.me.berkeley.edu/~aagogino/papers/Final_Report_Metathesaurus.pdf.
- 2 BODENREIDER, O.; NELSON, S., J.; HOLE, W. T. & CHANG, H. F. (o. J.). *Beyond Synonymy. Exploiting the UMLS Semantics in Mapping Vocabularies*. URL: <http://www.nlm.nih.gov/mesh/beyond.html>.
- 3 Deutsches Institut für Normung (1987). *DIN 1463-1. Erstellung und Weiterentwicklung von Thesauri. Einsprachige Thesauri*. Berlin: Beuth.
- 4 DOERR, M. (2001) Semantic Problems of Thesaurus Mapping. *Journal of Digital Information* 1, 8, Artikel 52. URL: <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/>.
- 5 EWERT, G. & UMSTÄTTER, W. (1997). *Lehrbuch der Bibliotheksverwaltung*. Stuttgart: Hiersemann.
- 6 FENSKE, M. (2000). *Planung eines digitalen Parlamentsarchivs unter Kosten-Nutzen-Aspekten*. Institut für Bibliothekswissenschaft der Humboldt-Universität Berlin. URL: <http://www.ib.hu-berlin.de/~kumlauf/handreichungen/h63/>.
- 7 FENSKE, M. (2002). *Integration der Datenbanken des Schleswig-holsteinischen Landtages zu einem Wissensmanagementsystem*. In Schmidt, R. (Hrsg.): *Content in Context – Perspektiven der Informationsdienstleistungen: Proceedings. 24. Online-Tagung der DGI Frankfurt a. M., 3.-5. Juni 2002* (S. 99-108). Frankfurt a. M.: Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis.

¹⁷ An dieser Stelle möchte ich vor allem meiner Familie für ihre Geduld und meinen Kolleginnen und meinem Kollegen in der Landtagsdokumentation für ihre intensive und kritische Unterstützung bei der Entwicklung des Landtagsinformationssystems und bei der Erstellung dokumentarischer Konzepte danken. Ohne ihren Rückhalt hätte ich das beschriebene Modell eines syntaktischen Metathesaurus nicht fertig stellen können.

- 8 HOLE, W. T. & SRINIVASAN, S. (2000). *Discovering Missed Synonymy in a Large Concept-Oriented Metathesaurus*.
URL: http://www.nlm.nih.gov/research/umls/pdf/Missed_Synonymy.pdf.
- 9 MILI, H. & RADA, R. (1988). Merging Thesauri. Principles and Evaluation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10, 2, 204-220.
- 10 NIKOLAI, R. (2002). *Thesaurusföderationen. Ein Rahmenwerk für die flexible Integration von heterogenen, autonomen Thesauri*. Diss. Universität Karlsruhe. URL: <http://www.ubka.uni-karlsruhe.de/vvv/2003/informatik/6/6.pdf>.
- 11 SCHRÖDER, T. A. (1998). *Parlament und Information. Die Geschichte der Parlamentsdokumentation in Deutschland*. Diss. Universität Düsseldorf. Potsdam: Verl. für Berlin-Brandenburg.
- 12 SCHOGER, A. & FROMMER, J. (2000). Heterogen – was nun? Evaluierung heterogener bibliographischer Metadaten. *Zeitschrift für Bibliothekswesen und Bibliographie* 47, 1, 110-128.
- 13 SCHÖNFELDT, R. (1994). Mathematische Eigenschaften für Thesaurusrelationen. *Nachrichten für Dokumentation* 45, 4, 203-212.
- 14 SCHWARZ, I. & UMSTÄTTER, W. (1999). Die vernachlässigten Aspekte des Thesaurus. Dokumentarische, pragmatische, semantische und syntaktische Einblicke. *Nfd. Information – Wissenschaft und Praxis* 50, 4, 197-203. URL: <http://www.ib.hu-berlin.de/~wumsta/thesau.html>.
- 15 SINTICHAKIS, M. & CONSTANTOPOULOS, P. (1997). A Method for Monolingual Mergin. In Belkin, N. J. [u. a.] (Hrsg.): *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference of Research and Development in Information Retrieval. Philadelphia, Pennsylvania, USA, 27.-31. Juli 1997* (S. 129-138). New York: ACM.
- 16 *Standard-Thesaurus Wirtschaft*. URL: <http://www.gbi.de/thesaurus/>.
- 17 STOCK, M. (1999). Standard-Thesaurus Wirtschaft. Ein neuer Standard der Wirtschaftsinformation? *Password*, 1, 22-29.
- 18 UMSTÄTTER, W. (2001). Leistungsgrenzen der Dokumentations-, Informations- Begriffs- und Wissensorganisation. In Schmidt, R. (Hrsg.): *Information Research & Content Management. Orientierung, Ordnung und Organisation im Wissensmarkt. 23. Online-Tagung und 53. Jahrestagung der DGI. Frankfurt a. M., 8.-10. Mai 2001* (S. 463-473). Frankfurt a. M.: Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis. URL: <http://www.ib.hu-berlin.de/%7Ewumsta/infopub/lectures/lectv.html>.

- 19 UMSTÄTTER, W. (1992). Nutzen der Indexierung bei Online-Datenbanken. In Neubauer, W. [u. a.]. (Hrsg.): *14. Online-Tagung der DGD. Proceedings. Frankfurt a. M., 27.-30. April 1992* (S. 403-420). Frankfurt a. M.: Deutsche Gesellschaft für Dokumentation. URL: <http://www.ib.hu-berlin.de/%7Ewumsta/infopub/pub1991f/pub65.html>.
- 20 UMSTÄTTER, W. (2000). *Zur Begründung der Thesaurusrenaissance. Vortrag vom 5.12.2000. Berliner Bibliothekswissenschaftliches Kolloquium*. URL: <http://www.ib.hu-berlin.de/~wumsta/infopub/lectures/lectu.html>.
- 21 VIEGENER, J. (1997). *Inkrementelle domänenunabhängige Thesauruserstellung in Dokumentenbasierten Informationssystemen durch Kombination von Konstruktionsverfahren*. Diss. Universität Karlsruhe. Sankt Augustin: Infix.
- 22 WERSIG, G. (1985). *Thesaurus-Leitfaden. Eine Einführung in das Thesaurus-Prinzip in Theorie und Praxis*. 2., erg. Aufl. München [u. a.]: Saur.

Die zitierten Internetquellen wurden zuletzt am 08.08.2006 aufgerufen.