

Aspektierung von Thesaurus-Relationen - Öffnung in universale Anwendbarkeit?

Harald H. Zimmermann, Universität des Saarlandes, Saarbrücken

Zusammenfassung

Der Thesaurus wird als Spezialfall eines elektronischen Wörterbuchs gesehen. Es werden zunächst allgemeine Regeln zur Qualitätssicherung eines Thesaurus aufgestellt. Hierzu gehören Transparenz, Flexibilität, Kompatibilität, Konsistenz und (bezogen auf den konkreten Anwender) Individualisierbarkeit.

Mit Blick auf die wachsenden technischen Möglichkeiten und angesichts der zunehmenden Integration in elektronische Informationssysteme wird vorgeschlagen, die Thesaurus-Norm in Richtung auf einen Universalthesaurus (UT) auf der Grundlage einer standardisierten (normseitig gepflegten) Bedeutungsdifferenzierung zu entwickeln, wobei die bestehenden Relationstypen als Basisrelationen neu bestimmt und durch Einführung einer Liste situationsbezogener inhaltlicher Merkmale zur Aspektierung ergänzt werden. Auch bezüglich einer Fachgebietsdifferenzierung (und -hierarchisierung) sollte die Norm eine Vorgabe machen.

Es wird darüber hinaus gezeigt, wie allgemeinere sprachliche Kategorisierungen zur Typisierung und Differenzierung von Thesauruseinträgen systematisch einbezogen werden können.

1. EINFÜHRUNG

Im folgenden wird die gängige, in DIN 1463 beschriebene und auch in der einschlägigen Literatur aufgenommene Definition eines "Thesaurus" im Sinne der Information und Dokumentation nicht grundsätzlich in Frage gestellt (vgl. z.B.. M. Burkart in GpID 1990, S.162).

Ein Thesaurus ist demnach Teil eines Informationssystems. Seine wesentlichen Anwendungen lassen sich verkürzt wie folgt zusammenfassen:

Ein Thesaurus dient der verdichtenden Deskribierung (Erschließung) des wesentlichen Inhalts einer Wissensquelle (diese muss kein Text sein) in Form der sog. Indexierung. Das Indexierungsergebnis ist dem Abstrakt sehr nahe: An die Stelle zusammenhängender natürlichsprachiger Sätze tritt eine Liste natürlichsprachiger Wörter. Diese Wörter sind nicht völlig frei wählbar, sondern müssen nach bestimmten Regeln einem Thesaurus entnommen werden.

Ein Thesaurus dient beim Information Retrieval der Suche nach relevanten Wissensquellen, indem sich der Nutzer dieser indexierten Wörter bedient. Da die Wörter zugleich im Thesaurus miteinander in Beziehung gesetzt sind, kann das Beziehungsgeflecht auch als Suchhilfe (Pfad) genutzt werden.

Es gibt keinen Thesaurus ohne eine zugehörige Wissenssammlung. Weltweit dürfte es mehrere Tausend solcher - auf bestimmte Fachgebiete und/oder Datenbanken ausgerichteter - Thesauri geben: Die neueste Zusammenstellung des "Thesaurus Guide" der KEG (Eurobrokers, Luxemburg 1992) umfasst rund 600 unter dem Aspekt der aktiven Nutzung selektierte Thesauri (in unterschiedlichen natürlichen Sprachen).

Mit der maschinellen Informationsverarbeitung sind einige "technische" Barrieren gefallen, die ein gedruckter Thesaurus notwendigerweise berücksichtigen musste (bzw.. auch in Zukunft muss). Dies betrifft v.a. das mögliche Volumen der einbezogenen Termini, aber auch die Dimension der Differenzierung in der Beschreibung, schließlich die Relationierung der Termini selbst. Rein theoretisch ist beispielsweise der Relationstyp "Vorzugsbenennung" - wenn eine "echte" Synonymie von Wörtern vorliegt - bei der ausschließlich maschinellen Verarbeitung überflüssig (vorausgesetzt, man lässt den Computer bei der Benutzung die "Zuordnungsarbeit" machen), es sei denn, man versteht beispielsweise die Relationierungsart "Vorzugsbenennung" als ein nützliches Instrument zur Vereinheitlichung der Terminologie in den Wissenschaftssprachen seitens des

menschlichen Nutzers. Auf dem Workshop in Weilburg ("Thesauri als Werkzeuge der Sprachtechnologie") habe ich dazu in den "Anmerkungen zur Neugestaltung der DIN 1463 (Thesauri)" weitere Begründungen gegeben, die hier nicht wiederholt werden müssen (vgl.. Literaturverzeichnis).

Es stellt sich nicht erst heute die Frage, das Thesaurus-Konzept so zu erweitern und zu verallgemeinern, dass eine Art "universale" Nutzung möglich wird. Diejenigen, die ein solches Ziel als undurchführbar bezeichnen (ich verweise hier auf eine entsprechende Stelle in dem o.a. Beitrag von M. Burkart), müssen sich umgekehrt fragen lassen, wie sie eigentlich - angesichts der Verflochtenheit der Wissensgebiete - ihr "universe of discourse" gegen andere abgrenzen wollen. Damit ist natürlich nicht ausgeschlossen, dass es "de facto" fach- oder themenbezogene Wissenssammlungen gibt, die sozusagen nur einen Ausschnitt des Weltwissens behandeln. Eine Ausgrenzung sollte jedoch nicht durch Nichtaufnahme eines für die bearbeitete Wissensquelle relevanten Begriffs erfolgen, sondern allein bedingt sein durch Fehlen der Wissensquelle selbst in der Dokumentation, sei es bewusst oder einfach deswegen, weil sie aufgrund der bearbeiteten Materialien nicht bekannt wird.

Wenn es schon schwierig ist, Fach- bzw. Wissensgebiete voneinander abzugrenzen, so ist es noch schwieriger, Fachsprachen und Fachlexika entsprechend zu differenzieren. Die in der Thesaurus-Forschung geführte Diskussion zur Problematik der inhaltlichen Differenzierung von Wörtern (als Repräsentanten von Begriffen) ist ein Ausdruck dieses Phänomens. Hierbei handelt es sich jedoch um ein generelles Problem bei der Verwendung natürlicher Sprache. Thesauri - wie die Repräsentation von Wissenssammlungen durch natürlichsprachige Wörter - sind und bleiben eine "Krücke", aber jeder Wissenschaftler bedient sich dieser Krücke schon bei der Präsentation seines Wissens (in Form von Texten), insofern hat diese Krücke "Thesaurus" als Verdichtungs- und Suchinstrument ihre Berechtigung so lange, bis Wissenschaft und Forschung bereits im originären Bereich andere Repräsentations- bzw.. Präsentationsmethoden anwenden.

In der Informationswissenschaft ist die Vorstellung von der Notwendigkeit der Eingrenzung von Wissensgebieten weit verbreitet. Ich verweise beispielhaft auf die Betrachtun-

gen von Kunz/Rittel, die schon 1972 als wesentliches Merkmal eines Informationssystems herausstellen, dass es stets "jemandes" Informationssystem ist (Kunz/Rittel: Die Informationswissenschaften, 1972, S. 42). Daraus könnte man folgern, dass Entsprechendes auch für die Thesaurusentwicklung gilt und analog formulieren: "Thesauri sind stets jemandes Thesauri als Teil von "jemandes" Informationssystem."

Ich interpretiere im Folgenden diese (berechtigte) Vorstellung jedoch nicht als generelle Notwendigkeit einer "physikalischen" Differenzierung und Individualisierung, sondern als Aufforderung, in Informationssystemen Konzepte und Verfahren vorzusehen, die eine Differenzierung nach spezifischen Zielgruppen (in erster Linie den spezifischen Nutzern von Informationssystemen) ermöglichen, u.a. zunächst auch im Sinne eines diesen Nutzern gemeinsamen zentralen Wissensgebietes und darauf bezogene wissenschaftlich-technische Erkenntnisse als Gegenstandsbereich eines Informationssystems. Wie man dieses "Gebiet" (z.B. ein "Fachgebiet") systematisch und v.a. akzeptierbar von anderen abgrenzt, wie man zu einer signifikanten Gesamtheit von Wissensquellen usf. kommt, ist jedoch schwer zu ermessen. Ich nehme an, dass letztlich jede vorgenommene Abgrenzung im Grunde willkürlich ist. Um so wichtiger ist die Suche nach Möglichkeiten, ein universelles Konzept zugrunde zu legen.

2. ZU KONZEPTION UND HAUPTMERKMALEN EINES UNIVERSELLEN THESAURUS (UT)

Es geht im folgenden nicht um Vorgaben zum Aufbau eines voluminösen "Superthesaurus", etwa dergestalt, dass beispielsweise alle sinntragenden Wörter einer Sprache (bzw. eingeschränkt: aller Fachsprachen) präskribierend und relationierend von vornherein einbezogen werden. Ziel ist vielmehr die Entwicklung eines allgemeinen Konzepts, das über eine technische Kompatibilisierung von Thesauri (wie sie heute etwa anhand von SGML-Beschreibungen durchaus möglich wäre) hinausgeht und dabei die Inhalte (d.h. Begrifflichkeiten) mit einbezieht und ggf. vorschreibt. Dieses Konzept muss also nicht von vornherein die Einbeziehung aller "Fachwelten" voraussetzen.

Eine inhaltliche Normierung im Bereich der natürlichen Sprache ist ziemlich problematisch. Es gibt jedoch keine Alternative, wenn man die Produktivität in der Thesaurusentwicklung steigern, die Vergleichbarkeit (auch qualitativ) erhöhen und die Akzeptanz der Nutzer (in Informationssystemen) steigern will. Niemand wird auch in Zukunft "sanktioniert" werden, wenn er sich nicht oder nicht hinreichend nach den Vorgaben richtet. Die neue Norm sollte aber für die Entwickler und Anwender so attraktiv - und einsichtig - werden, dass sie "freiwillig" den Vorgaben folgen und damit die Durchsetzung der Norm fördern helfen.

Ein Universal-Thesaurus (UT) sollte nach meiner Auffassung folgende wesentlichen Eigenschaften (En) aufweisen:

- (E1) Die Elemente und Anwendungsmöglichkeiten eines UT müssen für den Nutzer transparent sein, d.h. er muss zumindest verstehen, was im Einzelnen gemeint ist. Dies gilt insbesondere, wenn der Thesaurus in Verbindung mit einer Indexierung / Erschließung von Dokumenten verwendet wird, da man beim Retrieval in jedem Falle auf die UT-Konzepte des Entwicklers (bzw. deren Verwendung durch den Indexierer) angewiesen ist.

Entsprechende Anforderungen finden sich heute an jedes Expertensystem; ein UT kann dementsprechend als ein spezifisches Terminologie-Expertensystem verstanden werden.

- (E2) Ein UT muss flexibel sein. Diese Flexibilität muss insbesondere die prinzipielle Möglichkeit der Ausweitung der Wissensgebiete (d.h. "in die Breite") beinhalten, ebenso aber die Möglichkeit der Spezialisierung (d.h. der Erweiterung "in die Tiefe").
- (E3) Ein UT muss konsistent sein. Dort, wo dies nicht möglich ist (dies gilt v.a. bei der Übersetzung von Wörtern in multilingualen Thesauri), müssen Hinweise auf eine ggf. vorliegende Inkonsistenz gegeben werden.

- (E4) Ein (realisierter) UT muss mit einem anderen (realisierten) UT kompatibel sein. Soweit dies nicht der Fall ist, müssen die unterschiedlichen Konzepte im Rahmen eines Abgleichs (Vergleichs) erkennbar sein.

- (E5) Ein Nutzer muss ggf. "seine" Welt bzw. Problemstellung individuell auf den UT abbilden bzw. damit verknüpfen können. Dies fördert die praktische Anwendbarkeit.

Einleitend wurde bereits die generelle Problematik angesprochen, ein universales Konzept in die Praxis umzusetzen. In der Tat wird es kaum möglich (und auch kaum nötig) sein, einen "Ideal-Universalthesaurus" auch praktisch zu realisieren. Zu den Gründen rechnet, dass es auf der Erde verschiedene Kulturen und darauf im Laufe der Entwicklung der Menschen angepasste Sprachen gibt. Nicht alle Vorstellungen von der "Welt" sind zudem deckungsgleich, und dies führt letztendlich auch zur prinzipiellen "Unübersetzbarkeit" von Sprachen.

Trotz dieser theoretisch sicherlich richtigen Erkenntnis kommt man nicht umhin, praktisch-brauchbare Lösungen zu suchen. So wird heute (fast) alles - bis hin zur Belletristik - "brauchbar" und z.T. einfühlsam natürlichsprachig übersetzt, und in der Lexikographie sind trotz dieser grundsätzlichen Problematik Übersetzungswörterbücher entstanden, die die gängigen Terminologien in Thesauri beispielsweise volumenmäßig um ein Vielfaches übersteigen.

3. VORAUSSETZUNG: STANDARDISIERUNG DER BEDEUTUNGSDIFFERENZIERUNG

Es ist bekannt, dass die Präsentationsform von Begriffen in den natürlichen Sprachen nicht eindeutig ist. Wenn man sich also auf die Verwendung natürlichsprachiger Präsentationsformen (d.h. von "Wörtern") einlässt, muss man diese in irgendeiner Weise vereinheitlichen. Diese Problematik wird in der "praktischen" Thesaurusentwicklung häufig mit dem Hinweis überspielt, dass "in einem Fachgebiet" ein allgemeinsprachlich mehr-

deutiges Wort eine eindeutige Interpretation habe; wo dies nicht der Fall ist, wird über "Vorzugsbenennungen" in Verbindung mit Erläuterungen vereindeutigt.

Dieses Verfahren ist bei einem UT (zudem unter der Forderung, dass "Vorzugsbenennungen" entfallen können) nicht (mehr) verwendbar. Daher ist die erste Forderung, dass man sich national wie international - soweit es um Deskriptoren geht - auf eine Standard-Differenzierung von Homonymen und Polysemen verständigt, ehe man überhaupt daran denken kann, Begriffe über ihre oberflächlich-sprachlichen Wörter zueinander in Beziehung zu setzen.

Bei der Indexierung wie beim Retrieval muss dementsprechend ein Zwischenschritt eingeführt werden, der bei Polysemen zunächst die relevante Differenzierung ermittelt. Über entsprechende Techniken und Erfahrungen im multilingualen Bereich werde ich in Verbindung mit einer relationalen Datenbank am 26. Oktober in Weilburg berichten. Hier sollen demgegenüber - bezogen auf die im KTF-Komitee für Klassifikations- und Thesaurusforschung derzeit diskutierten Beispiele - einige grundlegende Vorstellungen entwickelt werden.

Zunächst kann festgehalten werden, dass heute alle großen Allgemeinwörterbücher (z.B. DUDEN - Das große Wörterbuch der deutschen Sprache in 6 Bänden; Brockhaus-Wahrig in 6 Bänden) im Hinblick auf eine mögliche Standardisierung der Differenzierung homonymer und polysemer Allgemeinbegriffe schon eine gute Materialgrundlage bilden. Die für den UT geforderten Eigenschaften E3 und E4 bedingen zudem eine solche Differenzierung. Um einen größtmöglichen Konsens und eine große Akzeptanz zu erreichen, wird vorgeschlagen, im Rahmen der Standardisierung die Differenzierungen fürs Deutsche zu erarbeiten und (etwa als "elektronischer" Anhang zu einer Thesaurus-Norm) zur praktischen Verwendung bereitzustellen. Dass dies ein Prozess ist, der sowohl allgemeinsprachlich als auch fachsprachlich der Pflege bedarf, ist offenkundig.

In einem Beispielmodell (dem Modellthesaurus THES_GEN) habe ich folgende Verfahrensweise gewählt:

- (a) Markierungen von Bedeutungen durch Indexziffern am Ende des Wortstammes
- (b) kurze definitorische Erläuterungen (Transparenz-Anforderung)
- (c) Verweis von der "unindexierten" Form (wortklassenunabhängig) auf die realisierten Differenzierungen.

In nicht-technischer Form liest sich dies dann etwa so:

Haube_1 Kopfbedeckung

Haube _2 Abdeckung eines Geräts

Schleuse_1 Vorrichtung zum Absperren von Wasser

Schleuse _2 aus zwei Toren und dazwischenliegender Kammer bestehende Anlage

Schleuse _3 hermetisch abschließbarer Raum als Durchgangsraum (z.B. zur
Desinfektion)

Die kurzen Erläuterungen sind ad hoc gewählt; dies ist (auch formal) sicherlich noch weiter zu systematisieren. In den "klassischen" Wörterbüchern sind zudem häufig zwischen diesen Varianten selbst ggf. noch Beziehungen hergestellt, die etwa die Herkunft oder die semantische "Nähe" betreffen. Auch dies wäre ggf. noch einzubringen, ist aber im vorliegenden Argumentations-Zusammenhang arbiträr.

4. STANDARDISIERUNG UND FLEXIBILISIERUNG VON FACHGEBIETSMERKMALEN

Fachgebietskennungen haben in allen Bereichen der Sprachverarbeitung eine große Tradition. Dennoch gibt es hierzu m.W. keine allgemeine Regelung. Dies betrifft einerseits die Aufteilung der Wissenschafts- und Technikbereiche selbst, noch mehr aber die praktische Handhabung. In der Praxis (d.h. in "jemandes" Informationssystem) wird meist so verfahren, dass im zentral behandelten engeren Wissensgebiet (ad hoc) eine sehr weitgehende Differenzierung erfolgt, die Randgebiete dagegen nur grob benannt (oder mangels Belegen auch gar nicht behandelt) werden.

Im Zusammenhang mit der Erstellung und Anwendung eines Thesaurus haben Fachgebietsdifferenzierungen meist mehrere Funktionen: Sie gliedern einen Wortschatz allgemein weiter auf, andererseits dienen sie ggf. der Bedeutungs-differenzierung.

Fachgebietskennungen lassen sich im übrigen zueinander hierarchisieren, und relatieren. Um diesen Bereich flexibel zu halten, sollte der Standard hierzu "defaultmäßig" eine Vorgabe machen; die relevante Verarbeitungssoftware könnte dem Nutzer jedoch auch Alternativen bieten.

Ich bin der Auffassung, dass das UT-Konzept eine standardisierte Fachgebietsgliederung einschließen sollte, die jedoch dem Entwickler von Spezialthesauri bestimmte Möglichkeiten der weitergehenden Spezialisierung (oder auch systematisch begründeten Abweichung) eröffnet. Zudem sollte man für Spezialthesauri (die ja der Regelfall sein werden) mindestens eine Unterscheidung nach "Kerngebieten" und "Randgebieten" der Differenzierung einführen, um einer (möglichen) Kompatibilisierung von Spezialthesauri Rechnung zu tragen. Ein Spezialthesaurus sollte dabei nach der Regel verfahren, dass er seine Unterscheidung von Fachgebietsangaben (bei Deskriptoren) in Randgebiete und Kerngebiete im Einzelfall expliziert.

Während man sich bei Bedeutungen eher um Einzelfälle "streiten" wird, ist die Zuordnung einer Fachgebietskennung nicht unproblematisch. Der "Standard" ist hier eher deskriptiv; man muss also damit rechnen, dass verschiedene Spezialthesauri zu einem Terminus (gleicher Bedeutung) unterschiedliche Fachgebietskennungen vergeben (haben). Dennoch sollte man den Versuch wagen, auch diese Komponente für einen der Norm entsprechenden Thesaurus verbindlich festzuschreiben.

5. ZUR ASPEKTIERUNG/DIFFERENZIERUNG VON THESAURUSRELATIONEN

Mit der (im Titel hervorgehobenen) Aspektierung (man könnte auch sagen: Attribuierung) von Thesaurusrelationen (v.a. im Zusammenhang mit der Hierarchierelation) entwickelt sich der Thesaurus weiter in Richtung auf eine "Wissensbank", bleibt aber "lexikalisch" basiert, d.h. wortbezogen. Die Beschreibungs- und Repräsentationsform ist also

weiterhin das "semantische Netzwerk" (Verknüpfung natürlichsprachiger, wenn auch bedeutungsmäßig bereits differenzierter Wörter), die Wörter (als "Knoten") werden jedoch mit "Kanten" verknüpft, die eine Beschriftung (zur Differenzierung der Relation) enthalten.

Die "Aspektierung" bzw. "Attribuierung" wird eingeführt, um die Zahl der Relationen kleiner zu halten. So kann man z.B. ausdrücken (R = Regionalspekt, E = Etymologie-Aspekt):

Abitur (R-DEU,E-LAT) "is-syn-to" Matura (R-AUS)

Abitur (R-DEU,E-LAT) "is-syn-to" Reifeprüfung (R-ALL,E-DEU)

Man könnte natürlich jede dieser Aspectierungen der Relationen als eigenständigen Relationstyp betrachten. Dies ist in der Sache durchaus äquivalent und daher ein Präsentations- und weniger ein Repräsentationsproblem; vielleicht sollte man dieses unter dem Gesichtspunkt der Einfachheit / Komplexität (Verstehbarkeit) entscheiden.

Weitaus wichtiger ist die Frage, ob man für das UT-Konzept zu einem Konsens über einen (vorläufigen) Kanon der Aspekte / Attribute kommen kann. Ich gehe einmal davon aus, dass auch hier von einem fachübergreifenden Basismodell ausgegangen wird, das fach- und aufgabenspezifisch weiter verfeinert werden kann.

Im folgenden gebe ich zunächst die Relationen an und dazu Beispiele für mögliche Aspectierungen (die Abkürzungen sind arbiträr):

5.1 Äquivalenzrelationen:

is-syn-to	allg. Synonym (beide Richtungen), ohne Aspekt
is-syn-to	regionale Variante (beide Richtungen), aspektiert (s.o.)
is-syn-to	Abkürzung / Langform (Aspekt)
is-syn-to	Numerus-bezogene Variante (Aspekte: N-SING bzw. N-PLUR)

is-orth-syn-to orthographische Variante (Schreibvariante): dies ist m.E. besser als eine weitere Relation darzustellen, sie könnte um Aspekte wie "veraltet", "etymologische Schreibung" oder "bevorzugte Schreibung" ergänzt werden.

Als fachspezifische Besonderheiten der Aspektierung der Relation "is-syn-to" könnte man einbringen:

- chemisches Element / Langform
- chemische Formel / Langform
- biologischer Name (lat.) / dt. Bezeichnung
- Fachbegriff / Laienausdruck (Beispiel: Infarkt / Herzschlag)

Natürlich sind die zusätzlichen Merkmale (beispielsweise "Singular", "Plural", "Schweizerisch", "Österreichisch" ...) möglichst umfassend zu standardisieren (d.h. aufzulisten).

5.2 Hierarchische Relation(en) bzw. Aspektierungen

Nach Art der Verfahrensweise im Entity-Relationship-Modell (das argumentativ auch von U. Reimer in seinem Beitrag über "Neue Formen der Wissensrepräsentation", GpID 1990 S. 186 ff. herangezogen wurde) möchte ich - wie oben bereits praktiziert - auch hier "sprechende" englischsprachige Bezeichnungen für die Relationen benutzen. Die Beispiele stammen weitgehend aus den Materialien des "Weilburger" Vorschlags von Schmitz-Esser:

Es wird vorgeschlagen, mindestens die folgenden Aspektgruppen bzw. Relationenbündel einzuführen:

Relationen:

- | | | |
|---------------------|--------------------|---------------|
| (R1) VERWANDTSCHAFT | has-relat (invers) | |
| (R2) TEIL/GANZES | has-as-part | is-part-of |
| (R3) ELEMENT/MENGE | has-as-element | is-element-of |

Aspektierte Relationen:

Da es sich bei der Relationierung stets nur um einen Aspekt der möglichen Merkmale eines Begriffs/Wortes handelt, nenne ich diese Relationen "aspektiert".

(R4) by-comp has-comp

(R5) is-a

(R6) is-comp-of

(A1) FORM Form

(A2) PROC Prozess (Vorgang)

(A3) INST Instrument

(A4) TYPE Typ

(A5) MATE Material

Beispiele (mit expliziter Relationierung und Markierung):

(R 1)	has-relat	(invers).
Mutter_2	has-relat	Tochter
Tochter	has-relat	Mutter_2

(R2)	has-as-part	is-part-of
Auto	has-as-part	Kühlerhaube (?)
Kühlerhaube	is-part-of	Auto (?)
Stadtbefestigung	has-as-part	Stadtter
Stadtter	is-part-of	Stadtbefestigung
Stadt_1	has-as-part	Stadtbefestigung
Stadtbefestigung	is-part-of	Stadt_1 (!)
Kirche_1	has-as-part	Kirchentür
Kirchentür	is-part-of	Kirche
Schleuse_2	has-as-part	Schleusentor
Schleusentor	is-part-of	Schleuse_2

(R3)	has-as-element	is-element-of
Verbindungsmittel	has-as-element	Mutter_1
Mutter_1	is-element-of	Verbindungsmittel

(A1)by-comp FORM	has-comp FORM	is-comp-of FORM
Tor_2		
	by-comp	Bogentor (FORM)
Bogentor	is-a	Tor_2
Bogentor	has-comp	Bogen (FORM)
Bogen (FORM)	is-comp-of	Bogentor

(A2)by-comp PROC	has-comp PROC	is-comp-of PROC
Tor_2	by-comp	Rolltor (PROC)
Rolltor	is-a	Tor_2
Rolltor	has-comp	rollen (PROC)
rollen (PROC)	is-comp-of	Rolltor
Band_1	by-comp	Fließband
Fließband	has-comp	Transport
Fließband	is-a	Band 1
Transport (PROC)	is-comp-o	Fließband

(A3)by-comp INST (has-comp) is-comp-of (is-a)

Eingang_1	by-comp	Haustür (INST)
Haustür (INST)	is-comp-of	Eingang_1
Eingang_1	by-comp	Tor_2 (INST)
Tor_2 (INST)	is-comp-of	Eingang_1

Ausgang_1	by-comp	Tür_2 (INST)
Tür_2 (INST)	is-comp-of	Ausgang_1
Ausgang_1	by-comp	Tor_2 (INST)
Tor_2 (INST)	is-comp-of	Ausgang_1

(A4)by-comp TYPE has-comp TYPE is-comp-of TYPE is-a

Tor_2	by-comp	Stadttor (TYPE)
Stadttor	is-a	Tor_2
Stadttor	has-comp	Stadt (TYPE)
Stadt (TYPE)	is-comp-of	Stadttor

Tür_2	by-comp	Haustür (TYPE)
Haustür	is-a	Tür_2
Haustür	has-comp	Haus_1 (TYPE)
Haus_1 (TYPE)	is-comp-of	Haustür

(A5) by-comp MATE has-comp MATE is-comp-of MATE is-a

Band_1	by-comp	Metallband (MATE)
Metallband	has-comp	Metall (MATE)
Metallband	is-a	Band_1
Metall (MATE)	is-comp-of	Metallband

Band_1	by-comp	Plastikband (MATE)
Plastikband	has-comp	Plastik (MATE)
Plastikband	is-a	Band_1
Plastik (MATE)	is-comp-of	Plastikband

Tür_2	by-comp	Bronzetür (MATE)
Bronzetür	has-comp	Bronze (MATE)
Bronzetür	is-a	Tür_2
Bronze (MATE)	is-comp-of	Bronzetür

Holz_1	by-mater-comp	Holztür
Holztür	has-comp	Holz_1(MATE)
Holztür	is-a	Tür_2

Holz_1(MATE) is-comp-of Holztür

Die Liste der (Standard-)Relationierungen bzw. -aspektierungen lässt sich sicherlich noch fortsetzen. Für die Diskussion der Konzeption sollte das Vorgestellte jedoch ausreichen.

6. SEMANTISCHE MERKMALE

Einige verbale Beschreibungen stellen in sich selbst bereits eine Art Typisierung dar. Beispiele dafür sind Beziehungen zwischen einem Vorgang und der zugehörigen Methode bzw. dem zugehörigen (technischen) Verfahren:

Information	Informationstechnik
Information	Informationsmethode

Es wäre möglich, hierfür eigene Relationen einzuführen:

Information	has-as-meth	Informationsmethode
Informationsmethode	is-meth-of	Information

Information	has-as-tech	Informationstechnik
Informationstechnik	is-tech-of	Information

usf.

Statt dessen - inhaltlich äquivalent dazu - schlage ich vor, semantische Merkmale als Attribute zu den Wörtern einzuführen, die Relation allgemein "is-related-to" zu benennen und hierfür den Terminus "Formatrelation" (m.E. besser als Pertinenzrelation) einzuführen. Die Feindifferenzierung ergibt sich dann unter Auswertung der semantischen Merkmale.

Zu den Merkmalen sind etwa zu rechnen (auch dies ist nur modellhaft zu verstehen und für die Standardisierung bzw. auch fachspezifisch weiter zu systematisieren):

VORG Vorgang

BERF Beruf

METH Methode

TVER technisches Verfahren

HUMN menschliches Wesen

Dies ergibt dann folgende Relationierung:

Information (VORG)	is-related-to	Informationsmethode (METH)
Information (VORG)	is-related-to	Informationstechnik (TVER)
Informationstechnik (TVER)	is-related-to	Information (VORG)
Informationsmethode (METH)	is-related-to	Information (VORG)

Im MESH-Thesaurus findet sich m.W. ein Beispiel für die fachspezifische Merkmalsvergabe bei der Relationierung ("Therapie", "Diagnose" usf.).

7. MORPHOLOGISCH-SEMANTISCHE KATEGORISIERUNG

Ein weiterer Bereich der Merkmalsvergabe sind die rein morphologischen Beziehungen (z.B. Verb-Adjektiv-Substantiv; bei Substantiven und Eigennamen sollte man die Genusangabe mitführen). Hierzu schlage ich vor, ebenfalls die Formatrelation zu benutzen und die üblichen Merkmale zu vergeben:

singen (VRB)	is-related-to	Gesang (SUB)
Haus (SUB)	is-related-to	Häuschen (SUB,DIM)
gelb (ADJ)	is-related-to	Gelb (SUB)

usf.

Dass es sich um eine "Derivationsrelation" handelt, ergibt sich aus der Auswertung der Attribute / Merkmale.

8. NAMEN UND KATEGORIENNAMEN

Unsere Untersuchungen im Rahmen des Projekts MITI haben verdeutlicht (ohne dass dies hier weiter ausgeführt werden kann; ich verweise hierzu wieder auf meinen Beitrag

für Weilburg am 26. Oktober), dass es nützlich sein kann, für den Einstieg in die Thesaurus-Nutzung v.a. beim Retrieval ein "Top-Term" (z.B. "ALLGEMEIN") als Einstiegspunkt zu vergeben. Es erscheint darüber hinaus nützlich, im Thesaurus als Kategoriennamen "ausgezeichnete" Wörter einzuführen. Theoretisch sind die (semantischen) Merkmale und Aspekte selbst solche Kategoriennamen; doch gibt es weitere, bislang nicht behandelte Kategorien wie Eigennamen, geographische Namen, Maße und Gewichte. (Die Kategoriennamen stehen in Großschreibung):

Stuttgart is-a ORT
Harald is-a VORNAME
Deutschland is-a STAAT
Schwarzwald is-a REGION
Kilogramm is-a GEWICHTSMASS

9. AUSBLICK

Die Darstellung sollte zeigen, dass ein Thesaurus heute mehr sein kann (und muss) als ein Bündel von Relationierungen. Die Informationstechnik ist gefordert, die notwendigen Voraussetzungen zu schaffen, die Rahmentchnik für solche "Thesaurus-Expertensysteme" zu installieren. Es ist m.E. jetzt notwendig, möglichst rasch zu einer Neufassung (d.h. einem Vorschlag) der Thesaurus-Norm zu gelangen, um den Software-Entwicklern eine entsprechende Entwicklungssicherheit zu geben.

Folgende Punkte, die hier nicht behandelt werden konnten, erscheinen darüber hinaus für die weiteren Überlegungen wichtig:

- Die Möglichkeit, einen elektronischen Thesaurus zum Zwecke der rein intellektuellen Nutzung in Papierform bereitzustellen, kann durch geeignete Markierungen mit einbezogen werden. Hierbei (nicht im Indexierungsbereich) rechtfertigt sich ggf. auch die Vergabe von Vorzugsbenennungen.
- Die Frage und Möglichkeit der Entflechtung komplexer sprachlicher Ausdrücke (Beispiel: Geldbriefträger -> Briefträger + Geldzustellung), etwa durch die Einbringung von Rollenindikatoren, muss weiter untersucht werden.

- Die Frage der "Generalisierung" des Thesauruskonzepts in Richtung auf ein sprachliches "Relationenwörterbuch" für verschiedene Anwendungen muss weiter untersucht werden.
- Der Zusammenhang zwischen Klassifikation und Thesaurus muss in geeigneter Form einbezogen werden (Beispiel: Die Stich- und Schlagwortverzeichnisse).

Im Ergebnis sollte m.E. nicht "die" Norm für "den" Thesaurus entstehen, sondern ein - allerdings überschaubares - Kompendium an Vorgaben und Differenzierungen, die systematische Teillösungen ebenso möglich machen wie eine Qualitätskontrolle - ganz im Sinne der Hauptthematik dieses Dokumentartags in Jena.

Literatur:

Burkart, Margarete: Dokumentationsprachen. In: Grundlagen der praktischen Information und Dokumentation, hrsg. von M. Buder, W. Rehfeld und Th. Seeger, München 1990, S. 143 - 182.

Kunz, Werner; Rittel, Horst: Die Informationswissenschaften. München 1972.

Reimer, Ulrich: Neue Formen der Wissensrepräsentation. In: Grundlagen der praktischen Information und Dokumentation, hrsg. von M. Buder, W. Rehfeld und Th. Seeger, München 1990, S.183-216.

Thesaurus Guide. Analytical directory of selected vocabularies for information retrieval, prepared by EUROBrokerS. CEC, Brüssel 1993.

Zimmermann, Harald: Wortrelationierung in der Sprachtechnik - Stilhilfen, Retrievalhilfen, Übersetzungshilfen. In: Kognitive Ansätze zum Ordnen von Wissen. Fortschritte der Wissensorganisation Bd. 2, Frankfurt 1992, S. 287 - 286.

Zimmermann, Harald: Anmerkungen zur Neugestaltung der DIN 1563 (Thesauri). In: Kognitive Ansätze zum Ordnen von Wissen. Fortschritte der Wissensorganisation Bd. 2, Frankfurt 1992, S. 313-318.