In: W. Gödert; P. Jaenecke; W. Schmitz-Esser (Hrsg., 1992): Kognitive Ansätze zum Ordnen und Darstellen von Wissen. Fortschritte in der Wissensorganisation (INDEKS Verlag) 2 (1992), S. 313-318

Harald H. Zimmermann
Universität des Saarlandes, Saarbrücken

Anmerkungen zur Neugestaltung der DIN 1463 (Thesauri)

1. Einleitung

Auf der Tagung in Weilburg war die Frage der Neugestaltung der DIN 1463 Teil 1 (Erstellung und Weiterentwicklung von Thesauri) ein zentrales Thema. Dr. Schmitz-Esser war bestrebt, neuere Erkenntnisse v.a. unter dem Aspekt der Erstellung, Bereitstellung und Nutzung im Rahmen EDV-gestützter Systeme in die Überlegungen einzubeziehen.

Der folgende Beitrag baut auf Diskussionsbeiträgen des Verfassers auf, die in einigen Punkten systematisiert und ergänzt wurden. Ergänzend wird auf die (anwendungsorientierte) Darstellung von Margarete Burkart (Abschnitt "Thesaurus" im Artikel "Dokumentationssprachen") und den (systematisierenden) Diskussionsbeitrag von Ulrich Reimer "Neue Formen der Wissensrepräsentation" verwiesen. Beide Studien finden sich im Sammelband "Grundlagen der praktischen Information und Dokumentation" (3., völlig neu gefasste Ausgabe, Band 1, ed. M. Buder, W. Rehfeld, Th. Seeger, München 1990). Dort wird auch weiterführende Literatur angeführt.

2. Zielsetzung und aufgabenbezogene Differenzierung

Die Einschränkung des Geltungsbereichs der DIN 1463 auf "Information und Dokumentation" ist heute nicht mehr sinnvoll, zumindest als zu eng anzusehen. Eine Ausweitung dieses "klassischen" Bereichs ist erforderlich, etwa um die betriebliche Information und Kommunikation (bis hin zur Textgenerierung und -ablage) und die Publikumsinformation (Beispiele: Videotext, Bildschirmtext).

Fragen der Wissensrepräsentation und deren sprachlicher Präsentation spielen beim Information Retrieval zunehmend eine Rolle; die bei U. Reimer dargestellten Repräsentationsmethoden sind dementsprechend von zentraler Bedeutung, ohne dass diese hier im Detail weiter betrachtet werden können.

Die EDV-Unterstützung bei der Dokumenterschließung und beim Retrieval macht große Fortschritte, das elektronische Publizieren erleichtert entscheidend die Pflege und Bereitstellung gedruckter Datensammlungen. Es ist heute kaum mehr vorstellbar, dass ein Thesaurus ohne elektronische Basis entwickelt und gepflegt wird. Falsch wäre es allerdings, die "klassische" Bedeutung des Thesaurus und das damit verbundene Themenfeld der *intellektuellen Dokumenterschließung* (d.h. der traditionellen Indexierung) und des *manuellen Retrieval* (d.h. des Aufsuchens von Daten in einer gedruckten Datensammlung) bei einer Neuorientierung außer Acht zu lassen.

Man muss sich bei allen Überlegungen bewusst sein, dass jeder Thesaurus (bzw. die daraus abgeleitete Indexierung über Deskriptoren) nur eine Krücke bei der Informationsvermittlung darstellt, wie die natürliche Sprache selbst nur ein *Instrument* der Wissensvermittlung ist. Es ist darüber hinaus wenig förderlich, Verfahren der *Klassifikation* den Anwendungen von Thesauri bei der

Dokumenterschließung und - wiederfindung entgegenzustellen. Im Gegenteil: In Zukunft sollte man die Vorteile, die beide Erschließungs- und Suchstrategien mit sich bringen, miteinander in Einklang bringen, um dem Nutzer die jeweils bestmögliche Hilfestellung bei der Problemlösung zu bieten. Das Stich- und Schlagwortverzeichnis des Deutschen Patentamts erscheint als ein - noch ausbaufähiges - Beispiel dafür, wie man Brücken zwischen beiden Verfahren schlagen kann.

3. Das Prinzip des Multifunktionalen Thesaurus

Bisher werden Thesauri auf die Anforderungen bestimmter *Zielgruppen* (z.B. Experten in einem besonderen Fachgebiet) und - meist damit verbunden - auf *Ausschnitte* des in der Sprache repräsentierten *Weltwissens* (bezogen auf entsprechende Sachverhalte) bezogen. Dies ist jedoch einerseits problematisch, weil in Datensammlungen ganz selten klar abgrenzbare "Teilwelten" existieren. Meist werden zumindest zwei solcher Teilbereiche praktisch benötigt, etwa die "Rechtswelt" und die "Bauwelt" beim Erstellen eines Vertrags über einen Brückenbau.

Konzeptionell betrachtet ist daher die Herstellung, Bereitstellung und Anwendung eines *multifunktionalen Thesaurus* (Schmitz-Esser nennt dies einen Mehrzweckthesaurus) eine große Herausforderung sowohl an die Wissenschaft wie die Praxis. Einen solchen Thesaurus muss man sich vorstellen als ein Konstrukt, das den Anforderungen verschiedenster Zielgruppen und Anwendungen gerecht wird, ohne dabei trotz der (notwendigen) Differenzierung inkonsistent zu werden. Zugleich muss er für den Nutzer bzw. die Entwickler transparent bleiben und zudem auf die Bedürfnisse der Anwender anpassbar sein (vgl. zur generellen Problematik auch: H. Zimmermann: Multifunctional Dictionaries. In: The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries. LINGUISTICA COMPUTAZIONALE III, 1983, S. 279-288).

Es ist nicht praktikabel, einen solchen Thesaurus "von oben nach unten" zu erstellen. Ausgangspunkt sind - wie bisher auch schon - die Anforderungen, die sich beim praktischen Indexieren und Recherchieren ergeben. Die (neue) DIN kann aber Vorsorge treffen, dass anwender- und anwendungsspezifische Thesauri

- mehreren Präsentationsformen (vereinfacht ausgedrückt: der Papier-Nutzung und der EDV-Nutzung) gerecht werden,
- mehrere Nutzergruppen (den qualifizierten Fachmann wie den Fachanfänger, evtl. auch den sog. "interessierten Laien") ansprechen,
- sich mit anderen Thesauri (aus verschiedenen, sich ggf. auch überlappenden Wissensgebieten) verknüpfen lassen.

Derartige Lösungen verlangen - wie dies im Bereich der Welt der Technik-Normierung selbstverständlich ist - ein hohes Maß an Präzision und Kontrollierbarkeit, in besonderem Maße aber die Akzeptanz der Anwender. Angesichts der Vagheit sprachlicher Begriffe (und in der Konsequenz auch der Vagheit der natürlichsprachigen Benennungen, die schließlich zu eher "unscharfen" Thesauri führen) ist heute nicht absehbar, ob ein solch "universelles" Konzept überhaupt tragfähig ist.

Selbst für den Fall, dass es nicht gelingt, einen Universalthesaurus auf einer solchen Basis zu realisieren, dass es nicht einmal möglich ist, die Multifunktionalität innerhalb eines fachspezifischen Thesaurus sicherzustellen, kann mit einer solchen Vorgehensweise vermutlich viel Doppelarbeit vermieden werden, denn eines ist sicher: Die gegenwärtige Norm ist nicht dazu angetan, diese möglichen Entwicklungen zu unterstützen.

4. Das Konzept einer universellen Bedeutungsdifferenzierung

Das Konzept des *semantischen Netzes* gilt - in unterschiedlichen Bezeichnungen - allgemein als Grundmodell des Thesaurus: Benennungen (für Begriffe) werden als "Knoten" verstanden, die über gerichtete "Kanten", d.h. über spezifische *semantische Beziehungen*, miteinander verbunden sind.

Der Vorteil natürlichsprachiger Benennungen, dass sie - im Gegensatz zu abstrakteren Notationen etwa bei der Klassifikation - beim menschlichen Nutzer aufgrund seiner (fach-)sprachlichen Vorbildung ohne zusätzliche Schulung weitgehend die gleichen Begriffe assoziieren, wird beeinträchtigt durch die Problematik, dass mehrere Bezeichnungen für den gleichen "Begriff' verwendet werden (Synonymie) und oberflächlich gleiche Bezeichnungen / Benennungen verschiedene Begriffe beinhalten (Homonymie, Polysemie). All dies sind bekannte Phänomene. Der "klassische" Thesaurus wird hier z.T. auch als ein Instrument der Sprachnormung verstanden. Diesen Aspekt sollte man jedoch in Zukunft - zumindest ist dies die Meinung des Verfassers - zurückstellen, v.a. angesichts der Möglichkeiten, die die EDV bei der Nutzung bereitstellt.

In der praktischen Anwendung wird die für ein semantisches Netzwerk (den Thesaurus) an sich unbedingt notwendige Bedeutungsdifferenzierung häufig vernachlässigt. Bei der Beschränkung auf ein Fachgebiet bleibt dies vielfach ohne Auswirkungen, da die allgemeine Regel gilt, dass im Fachgebiet nur die fachspezifische Bedeutung repräsentiert ist. In einem solchen Falle wird der Nutzerbereich allerdings auf den fachlich Qualifizierten eingeschränkt (was an sich weniger problematisch ist), eine Vernetzung oder Verknüpfung mit anderen Thesauri ist jedoch erheblich erschwert.

Man kann sich vorstellen, dass gleiche Benennungen für unterschiedliche Begriffe über einen einfachen Index differenziert werden, wie dies in traditionellen Wörterbüchern üblich ist, verbunden mit einer entsprechenden Definition oder Erläuterung. Dies ist für Differenzierungen innerhalb des gleichen Systems (Thesaurus) durchaus ausreichend, löst aber nicht die allgemeine Problematik der Portabilität und Kumulierung.

Als wesentliches Instrument zur Erreichung der Übertragbarkeit wird die Entwicklung und Einrichtung einer neutralen, universalen, terminologisch kontrollierten *Referenzdatenbank* vorgeschlagen. Die Erstellung eines solchen Systems muss demokratisch erfolgen, d.h. in Abstimmung mit den Thesauruserstellern. Sie erfasst und verwaltet differenzierend und kumulativ die in allen Thesauri verwendeten Benennungen und die diesen zugrundeliegenden Begriffe, so dass eine Konsistenz der Daten übergreifend gewährleistet ist, wenn es zu Verknüpfungen kommt.

Die Errichtung eines solchen terminologischen Referenzsystems ist insofern eine übernationale Aufgabe, als zugleich die Übersetzung bzw. Übersetzbarkeit der Benennungen (und der zugrundeliegenden Begriffe) einbezogen werden muss. Das System muss zudem so flexibel sein, dass Differenzen, die sich aus unterschiedlichen Begriffsfeldern bzw. Wissensrepräsentationen erge-

ben (beispielsweise gilt dies in Bezug auf die unterschiedlichen Rechts- oder Schulsysteme), zumindest beschrieben werden können.

Für die einzelnen Thesaurushersteller bedeutet die Berücksichtigung eines solchen Referenzsystems zugegebenermaßen eine gewisse Mehrarbeit, zumal sich die Begriffe in der Anwendung (im Laufe der Zeit) extern wie intern weiter differenzieren können. In aller Regel wird er jedoch von den (entsprechend qualifizierten) Arbeiten der Clearingstelle bzw. eines Kooperationspartners profitieren können. Als Vorbild - und Ausgangspunkt weiterer Überlegungen - kann die Verfahrensweise bei der Internationalen Patentklassifikation betrachtet werden.

5. Formalisierungsfragen

Wie diese Konsistenz *technisch* dargestellt wird, inwieweit dem Nutzer dies auf Papier oder am Bildschirm präsentiert wird, ist im vorliegenden Zusammenhang sekundär und bleibt im Detail dem einzelnen Entwickler und Anwender überlassen. Entscheidend ist, dass diese universelle Differenzierung (die nicht absolut gesehen werden darf, sondern ein Ergebnis einer Verabredung zwischen den Beteiligten darstellt) bei der "internen" Verarbeitung, die ja in aller Regel computergestützt abläuft, beachtet wird und zu ihr eine ein-eindeutige Schnittstelle (= Referenz) hergestellt ist.

Die Frage der äußeren Gestaltungsform, insbesondere einer Transportschnittstelle für die externe Bereitstellung und den Datenaustausch, ist im Grunde ebenfalls sekundär. Es ist allerdings vorstellbar, dass auch hierzu im Sinne der Erleichterung und Ökonomisierung der Portierung Normierungen erfolgen, etwa unter Zugrundelegung der Standard Generalized Markup Language (SGML) bzw. weitergehender Überlegungen, wie sie etwa von der Text Encoding Initiative (TEI) ausgehen (vgl. dazu die Guidelines "For the Encoding and Interchange of Machine Readable Texts", ed. C.M. Sperberg-McQueen, L. Burnard, Chicago/Oxford 1990 ff.).

6. Der Thesaurus als Expertensystem

In der Frage der Neugestaltung von Thesauri und entsprechender Wissensrepräsentationssysteme steht man heute vor einem Neuanfang. Es gilt nicht nur, einen breiteren und zugleich differenzierteren Ansatz zu finden, der die Möglichkeiten der Nutzung erweitert (etwa bis hin zu Stilhilfen bei der technischen wie allgemeinen Kommunikation). Neue Formen der Nutzung müssen erprobt und überhaupt erst ermöglicht werden.

> Als ein kleines Beispiel soll das Modell eines Expertensystems zur Suche nach fotografischen Motiven angeführt werden, das an der Universität Nancy entwickelt wird: Dem Betrachter werden sukzessive Bilder von einer Bildplatte vorgestellt. Nach der entsprechenden Auswahl, die allein vom optischen Eindruck des vorgelegten Bildes bestimmt ist, wird vom System die jeweilige Bilddeskribierung, verbunden mit einem "Hintergrund-Thesaurus", genutzt, um die vorgenommene Selektion zu bewerten und schrittweise solche Fotografie zu selektieren, die den ("unbewussten") Vorstellungen des Betrachters näher kommen.

Die Verwendung des Computers verlangt bekanntlich eine starke Formalisierung. Schwachstellen eines Systems werden sehr rasch deutlich. Dies schafft andererseits die Möglichkeit technisch gestützter Qualitätskontrollen. Die Erfahrung der letzten Jahre hat andererseits gezeigt, dass bei

der Formalisierung die Möglichkeit, die Deklarationen und Strukturierungen *intellektuell* nachvollziehbar zu machen, d.h. Entscheidungen auch einer späteren Verifizierung durch den Menschen zugänglich zu erhalten, nicht vernachlässigt werden darf. Dies erfordert die Einbeziehung einer (einfach nachvollziehbaren, in der Regel natürlichsprachigen oder grafischen) *Erklärungskomponente* in der angestrebten Norm. Beispiele und Ansätze für Erklärungen finden sich in dem o.a. Artikel von U. Reimer.

In gewisser Weise wird der Thesaurus der Zukunft ein auf Sprachdifferenzierung und -strukturierung bezogenes Expertensystem sein, aus dem für die praktische Anwendung Teillösungen selektiert werden können. Ähnlich, wie dies bei der Normierung der Register schon - zumindest ansatzweise - geschehen ist, muss zwischen allgemeinen Kriterien und den Konkretisierungen (und damit verbundenen Applikationen) unterschieden werden. Der "klassische" Thesaurus bleibt bei dieser Betrachtungsweise als spezielle Anwendungsform erhalten, es bleibt aber ausreichend Flexibilität für weitergehende und auch abweichende Applikationsformen.

7. Ausblick

Die Nutzung des "klassischen" Konzepts eines Thesaurus als Instrument der intellektuellen Indexierung (mit *Verdichtung* auf einige wenige Deskriptoren zu einem Dokument) und der entsprechenden, darauf bezogenen intellektuellen Recherche wird sicherlich auch in Zukunft ein wichtiges Anwendungsfeld bleiben, zumal für das Retrieval (online wie auf CD-ROM) die Terme mit ihren semantischen Verknüpfungen zunehmend elektronisch bereitgestellt und strukturiert nutzbar werden.

Verfahren, wie sie speziell von G. Lustig und G. Knorz zur analogen verdichteten Indexierung verwendet wurden (vgl. das Projekt AIR an der TH Darmstadt), zeigen zudem, dass diese "klassische Nutzung" bei der Dokumenterschließung auch maschinengestützt erfolgen kann.

Mit der Kombination von elektronischen Lexika, Thesauri und Klassifikationssystemen werden für diesen angesichts der Publikationsflut nach wie vor nötigen Verdichtungsprozess weitere Lösungen angeboten werden, wobei der Thesaurus v.a. der qualitativen Verbesserung bei der Ablage und der Recherche dienen kann.

Aber auch in der Textgenerierung und beim Freitextretrieval werden Thesauri (in Kombination mit den elektronischen Lexika) an Bedeutung gewinnen. Die anstehenden Normierungsbemühungen müssen daher auch diese Anwendungsbereiche mit ihren spezifischen Anforderungen einbeziehen. Dies wird erheblich zur definitorischen Klarstellung beitragen, da durch das bisher eher oberflächliche allgemeine Verständnis der Bezeichnung "Lexikon" bzw. "Thesaurus" (letztere angelehnt an Produkte wie Roget's Thesaurus, ein nach heutigen Vorstellungen eher undifferenziertes und unkommentiertes Begriffsgruppenwörterbuch) beim Nutzer wie bei manchen Entwicklern eine sehr unsichere Situation hervorgerufen haben.

Der Thesaurus wird in der Sprachindustrie unter den verschiedensten Aspekten erheblich an Bedeutung gewinnen. Voraussetzung ist allerdings, dass es zu Standardisierungen kommt, die sowohl dem Anwender wie dem Entwickler in qualitativer Hinsicht die nötige Sicherheit geben und zugleich angesichts der anstehenden Entwicklungen eine größtmögliche Ausbaufähigkeit und Nutzungsbreite gewährleisten.

(Stand: März 1992 (T26ZH1O))