Siegfried Reball
Forschungsinstitut Manfred von Ardenne, Dresden

Redundante Indexierungssprachen als Abbild natürlicher Sprache

(Redundant Indexing Languages Patterned After the Natural Language)

Reball, S.: **Redundante Indexierungssprachen als Abbild natürlicher Sprache** (Redundant indexing languages patterned after the natural language.)
In: Intern. Classificat. 7 (1980) No. 1, p. 10–12,

The proposal to raise indexing language convenience at the expense of specificity calls for closer orientation to the natural language, which combines utmost convenience with variability, semantic universality, vagueness, redundancy and other, at first sight negative properties. Document titles are cited for comparison to show that these titles adhere more closely in the required sense, to the natural language than customary indexations and may well show the way toward more natural language indexations.
Automatic translation of document titles already being common practice, the prospects for automatic processing of the proposed more convenient indexations are good. (Author, transl.)

1. Über den Gebrauchswert von Indexierungssprachen

Für die Güte einer Indexierungssprache sind zwei Eigenschaften entscheidend, die beide mit dem Einsatz und der Verarbeitung der Indexierungssprache zu tun haben. Die erste Eigenschaft, der Komfort einer Indexierungssprache ist verantwortlich dafür, wie gut bestimmte Dokumenteninhalte in ihr ausgedrückt werden können und die zweite Eigenschaft, der Formalisierungsgrad, ist verantwortlich dafür, wie gut die Indexierungssprache einer automatisierten Verarbeitung unterworfen werden kann. Im Vergleich dazu ist bei der natürlichen Sprache der Komfort der bestmögliche, der Formalisierungsgrad dagegen denkbar schlecht.

Es ist verständlich, daß die Bemühungen vergangener Jahre darauf gerichtet waren, bei Indexierungssprachen gleichzeitig Komfort und Formalisierungsgrad zu verbessern. Leider ohne zu befriedigenden Resultaten zu gelangen.

Die stillschweigende Annahme, daß mit genauen Vorschriften zum Gebrauch von Deskriptoren und für die Verknüpfung von Deskriptoren zwei unterschiedliche Indexierungen zwei einander ausschließende Dokumentmengen, bzw. im Fall der Verwendung von Oberbegriffen eine die andere völlig einschließende Menge kennzeichnen, ist nur bedingt richtig und führt deshalb mit verschärften Forderungen an die Verwendung von Deskriptoren, Relatoren usw. bei steigendem Indexier- und Recherchieraufwand durchaus nicht zu besseren Ergebnissen. Trotz dieser Tatsache (z.B. (1), S. 31, 82 und 163) besteht „doch allgemein Übereinstimmung darüber, daß die Effektivität der Indexierung davon abhängt, ob die Indexierungssprache logisch ist, d.h. in einem vernünftigen Maß klar und systematisch" (1, S. 61).

Vernünftig scheint zum Beispiel die Festlegung über Synonymverweisungen zu sein. Anstelle von vier oder fünf Synonymen der natürlichen Sprache oder der Fachsprache wird durch eine Siehe-Verweisung verbindlich der Gebrauch eines Begriffes, d.h. eines Deskriptors verlangt; damit erfüllt das Regelwerk der Indexierungssprache eine Normalisierungsfunktion im Sinne eines eindeutigen Gebrauchs seiner Sprachelemente eben auch im Sinne größerer Formalisierung. Nun sind aber Synonyme der Sprache nur selten wirklich bedeutungsgleich. Sie drücken ganz im Gegenteil geringe Bedeutungsunterschiede oder Aspektverschiebungen aus, die beim Gebrauch der Sprache bewußt oder unbewußt eine wichtige Rolle spielen. Sie gehen durch die Siehe-Verweisung der Indexierungssprache verloren und vermindern den Komfort der Indexierungssprache.

Es besteht, wie dieses Beispiel zeigt, ein gegenläufiger Zusammenhang von Sprachkomfort oder -ausdrucksfähigkeit und Formalisierungsgrad, der die Frage provoziert, auf wieviel von einem einmal erreichten Formalisierungsgrad zugunsten der Ausdrucksfähigkeit verzichtet werden könnte, denn die Beschreibungsfunktion der Indexierungssprache läuft nicht auf eine einfache Filterwirkung (2) hinaus, sondern auf eine komplexe, teilweise fließende Abgrenzung oder Zuordnung zwischen und innerhalb von Dokumentmengen, so daß häufig ein aktiver Generierungsvorgang des Indexierens über die reine Dokumenteninhaltsbeschreibung hinaus erforderlich ist. Auch muß das Dokument später für unterschiedliche Nutzer zu unterschiedlichen Fragestellungen als relevant gefunden werden, was durch postkoordinativ zusammengestellte Suchbegriffe nur dann erfolgreich ist, wenn beim (intellektuell gesteuerten) Indexierungsvorgang direkt oder indirekt diese zukünftigen Fragestellungen berücksichtigt wurden.

Während von einem Referat oder einem Dokumententitel heute noch berechtigt gefordert werden sollte, nach verbindlichen Regeln abgefaßt zu werden (Positionsreferat, Strukturreferat, Autorenhinweise), um einen höheren Formalisierungsgrad zu erreichen, sollte m.E. das Ziel der Weiterentwicklung der Indexierungssprachen eine verbesserte Ausdrucksfähigkeit auch u.U. auf Kosten des Formalisierungsgrades sein.

Wenn in einer Indexierungssprache z.B. (ausgenommen wirkliche Synonyme) Quasisynonyme zugelassen würden, könnten damit diffizile Suchfragen nach sonst nicht unterscheidbaren Dokumentmengen formuliert werden. Der Preis dafür ist u.a. eine zunehmende Redundanz der Indexierungssprache, die sogar noch die möglichen Abweichungen beim Indexieren, die nachweislich auch bei gut geregelten Indexierungssprachen häufig sind, wenigstens teilweise auffangen muß. Dabei ist der Begriff der Redundanz etwas weitergefaßt und steht hier

auch für andere Eigenschaften einer Sprache wie Abweichungen im Gebrauch (zeitlich, räumlich, sozial), Unbestimmtheiten der Bedeutung von Benennungen, Mehrdeutigkeit von Benennungen, die alle zusammen tatsächlich redundanzvergrößernd wirken, sobald ihre Begriffe nicht disjunkt zueinander sind.

Da bereits mit dem Gebrauch natürlichsprachlicher oder fachsprachlicher Begriffe als Deskriptoren einige der genannten Eigenschaften in die Indexierungssprache hineingetragen werden, obwohl man sich durch Festlegungen für den Gebrauch der Deskriptoren davon lösen möchte, ist die Frage zu stellen, ob eine weniger formalisierte und dafür redundante Indexierungssprache durch ihren größeren Komfort bessere Ergebnisse beim Wiederaufruf bringt, als eine stark formalisierte, redundanzarme Indexierungssprache.

2. Redundante Indexierungssprachen

Die Vorteile einer redundanten, nämlich der natürlichen, Sprache werden beispielsweise beim MEDLARS-System genutzt, indem in einem zweistufigen Prozeß zunächst mittels eines kontrollierten Wortschatzes recherchiert und der Suchraum eingeengt worden ist und danach mit Wörtern der natürlichen Sprache Titel oder Referate durchmustert werden können (3, S. 72). Kann man diesen Vorteil des MEDLARS-Systems noch ökonomischer gestalten, indem bereits die Indexierungssprache und die Retrievalsprache genügend natürlichsprachlich gestaltet werden? Wenn es gelingt, einer Indexierungssprache solche Eigenschaften wie Variabilität, semantische Universalität, Vagheit und Kontextabhängigkeit der natürlichen Sprache, die zwar alle zusammen eine maschinelle Verarbeitung erschweren (4), mitzugeben, so sollten dabei automatisch auch Vorteile entstehen.

Die Variabilität natürlichsprachlicher Begriffe, ihre Homonymität z.B., wird in Indexierungssprachen ebenso wie die Synonymität ausgemerzt. Ganz im Gegensatz dazu werden in Fachsprachen bewußt laufend Homonyme eingebracht, um neue Sachverhalte, Vorrichtungen, Produkte sinnfällig zu bezeichnen (Eimerkettenschaltung, Kaskade, Lichtgriffel, Schaltbaum, Elektronenkanone, Kasette usw.), was die Transparenz (SCHAFF) der Sprache verbessert bzw. ausmacht.

Wenn in den Deskriptorkombinationen

Laser, Entfernungsmessung
Laser, Nachrichtenübertragung
Laser, Uhrensteine
Laser, Netzhautkoagulation

jeweils nur jene Kategorie von Lasergeräten gemeint ist, die für den speziellen Einsatzzweck geeignet ist, so ist dies ein Beispiel für die Kontextabhängigkeit des Deskriptors „Laser". Bekanntlich verläßt sich die UDC nicht darauf und führt für einen Begriff (z.B. „Temperatur") entsprechend viele DK-Zahlen ein.

Diese Kontextabhängigkeit auch von Deskriptoren kann genutzt werden, wenn mit der Stellung des Deskriptors innerhalb einer grammatischen Schablone oder einer Deskriptorkette, eine Bedeutungsverschiebung verbunden ist.

Bei Dokumententiteln, die bekanntlich in vielen Fällen die gleichen Begriffe enthalten, wie die Indexierung (5), (6) dieser Arbeiten, gibt es z.B. signifikante Unterschiede bei der Häufigkeit verwendeter auch „quasisynonymer" Benennungen. Bei im Mittel dreißig Titeln pro Sachbereich ergeben sich folgende in Klammern angegebenen Häufigkeiten:

Sachbereich	Häufigste Begriffe im Titel
A Schmelzstrahlen mit Laserstrahl	Schmelzen (22), Laserstrahl (19), Metall (11), Laser (7), Stahl (7)
B Bohren mit Laser	Bohrung (27), Laserbohrung (11), Laser (11), Metall (11), Laserstrahl (7), Loch (7), Geometrie (7), Geschwindigkeit (6)
C Schweißen mit Laser	Laser (21), Laserschneiden (21), Schweißen (18), Blech (6), Eindringtiefe (6)
D Trennen mit Laser	Laserschneiden (23), Laser (19), Schneiden (19) Trennen (12), Wirtschaftlichkeit (12)
E Lasermeßtechnik	Laser (3), Anwendung (12), Messung (9), Geschwindigkeit (6), Kontrolle (6), Oberfläche (6)

Es zeigt sich, daß sich z.B. das Schmelzen und das Bohren im unterschiedlich häufigen Gebrauch von „Laser" und „Laserstrahl" unterscheiden und daß z.B. das Trennen durch „Wirtschaftlichkeit" und das Schweißen durch „Eindringtiefe" besonders gekennzeichnet sind, obwohl doch „Eindringtiefe" auch beim Bohren und Trennen und „Wirtschaftlichkeit" bei allen Verfahren relevant sein sollte.

Jedenfalls gibt es in Dokumententiteln Begriffe, die eine selektive Funktion besitzen, die in der I u D bewußt genutzt werden könnte. Läßt man aber im Sinne der obigen Ausführungen die Unterscheidung „Laser" und „Laserstrahl" nicht zu, weil beide Benennungen als quasisynonym betrachtet werden, so geht die Selektivität verloren. Diese Selektivität beruht mindestens zum Teil auf einem konventionellen Gebrauch von Begriffen innerhalb einer Fachsprache oder innerhalb einer Sprache, der sich auch auf den Vorgang der Indexierung übertragen lassen sollte, falls dem Indexierer die Freiheit der Wahl der Begriffsbenennungen gelassen wird.

Dieser konventionelle Gebrauch trifft auch auf Syntagmen oder Satzteile von Texten zu, die sich, in Bezug auf den wesentlichen Informationsgehalt, auf eine kleine Anzahl von Mustern reduzieren lassen (7).

In Weiterführung der Arbeit von SEELBACH (8), der mit Schlüsselphrasen der Form „Substantiv – Präposition – Substantiv" indexiert, könnten superpositionierte Nominalphrasen (9), (10) aus einem geringkontrollierten Vorrat, vergleichbar dem Begriffsvorrat in den entsprechenden Dokumententiteln, Indexierungen einer komfortablen, redundanten Indexierungssprache sein.

3. Automatisierbarkeit redundanter Indexierungssprachen

Da eine weniger kontrollierte Indexierungssprache auch weniger leicht automatisierbar ist, soll abschließend das Ergebnis einer Untersuchung an automatisch übersetzten Dokumententiteln, die auch in diesem Fall modellhaft benutzt wurden, mitgeteilt werden.

Die Analyse von automatisch übersetzten Dokumententiteln aus dem Russischen ins Deutsche in „Avtomatičeskaja Svarka" (Hefte 4–6/1978) ergibt folgende Zahlen:

Von 74 Titeln werden
vollkommen richtig übersetzt: 49 = 66 %
falsch übersetzt: 25 = 34 %.
Von 25 falsch übersetzten Titeln waren
40 % noch eindeutig (Inhalt noch erkennbar)
(31 % Druckfehler, 9 % falsche Präposition)
40 % mehrdeutig
(4 % Druckfehler, 24 % falsches Schlagwort, 12 % falsche Präposition)
20 % irreführend
(4 % falsches Schlagwort, 16 % falsche Präposition)
Die gefundenen Fehler sind aufsummiert zu
35 % Druckfehler,
28 % falsches Schlagwort und
37 % falsche Präposition.

Daran sieht man, daß gerade Präpositionen für die Phrasenbildung eine wichtige Rolle inhaltlicher Beschreibung darstellen. (Der Dokumententitel hat in diesem Fall eine der Indexierung vergleichbare Funktion.) Umso bedenklicher ist die Ausdrucksarmut von Indexierungssprachen an Relatoren, Funktoren, usw. (man vergl. z.B. (11)).

Auch die Dynamik fachsprachlicher Weiterentwicklung wird für Indexierungssprachen mit stark reglementiertem Benennungsvorrat zum Problem (12).

Es ist klar, daß mit der Forderung nach mehr Freizügigkeit beim Gebrauch von Indexierungssprachen bezüglich Benennungsvorrat und Bildung von Indexierungen systemintern mehr Aufwand getrieben werden muß. Die Größe und Leistungsfähigkeit heute bereits bereitstehender Systeme sollte diesen Mehraufwand aber durchaus zu bewältigen erlauben.

Quellen:

(1) Sparck Jones, K., Kay, M.: Linguistik und Informationswissenschaft. München: Verlag Dokumentation 1976
(2) Bielicka, L.: Effektivitätskriterium der Informationsrecherchesprachen. In: Aktualne problemy informacji i dokumentacji, Warszawa 22 (1977) No. 5, p. 16–20
(3) Weiss, P., David, H.: Stand und Probleme automatisierter Informationssysteme in der Medizin, insbesondere des IZIS MEDINFORM. X. Koll. Inf. u. Dok., Ilmenau 1977. Schriftenreihe Heft 40/2, p. 65–78
(4) Klein, W.: Organisation des Wissens durch Sprache. In: IBM-Nachr. 27 (1977) No. 234, p. 11–17
(5) Ghosh, J. S.: Content representation in document titles: a case study with prostaglandin literature. In: Aslib. Proc. 26 (1974) No. 2, p. 83–86
(6) Scholz, J.: Die Bedeutung der Überschrift eines Artikels für das Indexieren. Abschlußarbeit TH Ilmenau, Ilmenau 1972.
(7) Hirschmann, L. et al.: Grammatically-based automatic word class formation. In: Inform. Process. Management 11 (1976) Nr. 2, p. 39–57
(8) Seelbach, D.: Computer-Linguistik und Dokumentation. München: Verlag Dokumentation 1975
(9) Reball, S.: On a term of predicate calculus for indexing in information and documentation and on inferences to be drawn with respect to the degree of formalization of indexing languages. In: Intern. Forum Inform. Doc., Moskva, 3 (1978) No. 3, p. 14–17
(10) Vasileva, I. I.: Ob odnom informacionnom jazyki predikatnogo tipa. Naučno-techn. Inform., Ser. 2, Moskva (1976) No. 12, p. 23–27
(11) Steiger, R.: Über erkenntnistheoretische Grundlagen der Syntax einer Informationssprache. In: Z. f. Phonetik Sprachwiss. u. Komm. forschg. 30 (1977) No. 3, p. 297–299
No. 3, p. 297–299
(12) Harris, J. L.: Terminology change: effect on index vocabularies. In: Inform. Process. & Management. 15 (1979) No. 2, p. 77–88