Harald H. Zimmermann

Nutzbarmachung und Nutzung maschineller Wörterbücher in der Fachinformation und im Büro

10.12.83

1. Einführung

1.1 Elektronische Wörterbücher

Es gibt verschiedene Ansätze im Hinblick auf die Konzeption und Erstellung von Wörterbüchern. Eine sprachwissenschaftliche Betrachtung wird den Zusammenhang zwischen Wörterbuch (Lexikon) und Grammatik (sprachliches Regelsystem) in den Vordergrund stellen. In modernen Grammatiktheorien hat das Wort - hier gleichzusetzen mit einem Worteintrag (engl.: ENTRY) in einem Wörterbuch - dabei nicht mehr die fundamentale Bedeutung als Grundeinheit der Sprache: v.a. ist ein Wort längst nicht mehr die kleinste bedeutungstragende Einheit eines Sprachsystems. LYONS (1981) führte bereits als eine Art Synonym für Wort den Begriff "Morphemkomplex" ein. Es ist hier nicht der Ort, die sprachwissenschaftliche Problematik der Wortdefinition zu erläutern. Hierzu sei auf die einschlägigen Artikel in ALTHAUS/HENNE/WIEGAND 1973 (u.a. "Lexikologie") verwiesen. Mit HESS/BRUSTKERN/LENDERS 1983 (S. 7) sollen im Sinne der Definition von LYONS Wörter, die als "lexikalische Einheiten in Wörterbüchern zusammengestellt" werden, als Kombination von Morphemen betrachtet werden.

Dies ist v.a. dann legitim, wenn man - wie im folgenden - ein Wörterbuch aus der Sicht der Kommunikation (zwischen Menschen bzw. Menschen und Maschinen) und Information behandelt. Hierbei treten grammatisch-formalistische Aspekte in den Hintergrund; im Vordergrund steht der Gesichtspunkt des Wissenstransfer. Wörterbücher stellen - aus dieser Sicht der Informationsverarbeitung - Informationssammlungen dar, die formal oder natürlichsprachlich dargestelltes Wissen zu vorwiegend sprachbezogenen Sachverhalten speichern und verfügbar machen. Ein Wörterbuch bzw. eine Wort-Informationssammlung stellt dabei eine geordnete Menge von Informationseinheiten (den sog. lexikalischen Einheiten) dar. Diese Einheit enthält üblicherweise eine graphische Repräsentation: gleichsam die natürlichsprachige Ausdrucksform (Benennung) für eine (komplexen) sprachlichen Begriff. Mithilfe dieser Benennung läßt sich aufgrund einer Strategie, die die Kenntnis des (alphabetischen) Ordnungsprinzips der Informationssammlung voraussetzt, das gespeicherte Wissen erreichen. Die Benennung dient also (informationstechnisch betrachtet) zugleich als Zugriffsschlüssel zur Information.

Die lexikalischen Einheiten oder (vereinfacht:) Begriffe in einem Wörterbuch stehen nun im allgemeinen nicht völlig isoliert nebeneinander, vielmehr herrschen zwischen ihnen (da sie Ausdruck und Teil des Sprachsystems sind) Beziehungen vielfältigster Art. Diese können explizit gemacht sein (z.B. durch Verweise), sie können sich aber auch immanent in den Merkmalen widerspiegeln, die als Informationen bzw. Erläuterungen dargestellt werden. Sie können daneben auch aus den Begriffen/Wörtern abgeleitet werden, die zur Erläuterung verwendet werden. So bringt beispielsweise das Merkmal "Substantiv neutrum" zu dem Eintrag HAUS zum Ausdruck, dass ein so bezeichneter Begriff morphosyntaktisch in die gleiche Klasse einzuordnen ist wie ein entsprechend gekennzeichneter anderer Begriff des Wörterbuchs, etwa KIND. Die Erläuterung zu GITARRE: "Ein MUSIKINSTRUMENT, das .. " verdeutlicht implizit die Zugehörigkeit des Begriffs (GITARRE) zu einer bestimmten Begriffsgruppe (Ober-/Unterbegriffsrelation, hier MUSIKINSTRUMENTE), auch wenn diese nicht explizit gekennzeichnet sein sollte. In beiden Fällen wird vom Benutzer erwartet, dass er eine entsprechende sprachliche Regel oder Strategie anzuwenden in der Lage ist.

Derartige Vernetzungen bzw. Klassenzuweisungen werden in gedruckten Wörterbüchern - wie erwähnt - in der Regel nur aufgrund gleichartiger Merkmale deutlich. Man muss hier allerdings darauf hinweisen, dass es v.a. in der Fachinformation bereits seit langem Darstellungsformen gibt, die begriffliche Vernetzungen (v.a. semantische Relationierungen) auch in die Anordnung der Einträge übertragen. Es handelt sich um die sog. Thesauri. Einige allgemeinsprachliche Wörterbücher (im Deutschen z.B. WEHRLE/EGGERS) sind ähnlich aufgebaut; hier wird meist über ein zusätzliches alphabetisches Wortregister ein weiterer Zugriff möglich.

Weitergehende Möglichkeiten eröffnet die elektronische Datenverarbeitung (EDV): In elektronischen "Datenbanken" ist es üblich, mehrere Zugriffspfade (sog. Indizes) unabhängig von der physikalischen Anordnung der Daten (die im Prinzip dabei willkürlich sein kann) festzulegen. Schließlich ermöglicht das maschinelle Suchverfahren, das in der Regel tausendfach oder gar millionenfach schneller ist als das menschliche, auch aufgrund geeigneter Prozeduren (z.B. bei den sog. Relationalen Datenbanken) eine beliebige Selektion von Teilen der Informationssammlung aufgrund der Kombination von Merkmalen (d.h. entsprechender "Regeln"). Ein weiterer hier zu nennender grundlegender Unterschied in den Zugangs- und Auswertungsmöglichkeiten, der ebenfalls den prozeduralen Zugriff betrifft, ist darin zu sehen, dass gegenüber herkömmlichen gedruckten Wörterbüchern in der EDV-Bearbeitung nicht nur Identität, sondern irgendein sinnvolles mathematisches Verfahren zur Ermittlung der Distanz von Begriffen (sei es nun morphologisch oder semantisch orientiert) benutzt werden kann, um Informationen zu finden bzw. einen Begriff zu identifizieren.

1.2 Neue Rahmenbedingungen für die Wörterbucharbeit

Mit dem Elektronischen Zeitalter, an dessen Anfang zumindest die sog. Industriestaaten heute stehen, ist in verschiedenen Bereichen ein Wandel der Informationsbedürfnisse und -gewohnheiten in Sicht. Er wird nicht nur Innovationen, sondern mit diesen auch Instabilitäten mit sich bringen. Wenn man z.B. unter diesen Vorzeichen über Wörter-Bücher spricht, so ist dies vielleicht bald nur noch abstrakt zu verstehen: Man kann sich ausrechnen, dass im Jahre 2000 - sei es aus Kostengründen, sei es wegen dann fehlender Vertriebswege - umfangreiche Wörterbücher nicht mehr gedruckt, d.h. auf Papier gebracht werden, sondern per Kabel distribuiert und verfügbar werden, dass z.B. ein neues Enzyklopädisches "Lexikon" nur mehr auf Bildplatte erscheint oder über Breitbandnetze on-line abrufbar wird.

Mit dieser Entwicklung ändert sich auch der Stellenwert, der dem Herstellen eines Wörterbuchs zukommt. War es - wenigstens in großen Teilen - stets eine wissenschaftliche Arbeit, Wörterbücher zu entwickeln, ja ging es vielfach um grundlegende sprachliche Erkenntnisse, die es zu sammeln und zu dokumentieren galt (man denke auch an die vielen historisch-grammatischen Wörterbücher), so ist heute ein erweiterter Bedarf v.a. an Gebrauchswörterbüchern festzustellen, der in dieser Form bislang nur indirekt zum Tragen kam. Natürlich sind der berühmte DUDEN und verwandte Wörterbücher längst Gebrauchswörterbücher geworden, aber sie bestimmen keineswegs den Alltag schriftsprachlicher Kommunikation.

Mit der Entwicklung der Textverarbeitung - im weiteren Sinne: der Bürokommunikation - beginnt sich ein neuer Gebrauchsbereich zu entwickeln. Ähnlich verhält es sich mit der sog. Fachinformation und -kommunikation, bei der u.a. über sprachliche Begriffe ein Zugang zu weltweiten Informationsbanken geschaffen werden muss. Ein dritter Bereich, der sich bereits mit der Einführung von VIDEOTEX (in der BR Deutschland "Bildschirmtext" - BTX) abzuzeichnen beginnt, ist die sog. Publikumsinformation. Im letzten Falle müssen dem Laien, d.h. auch und vor allem dem Schreib-Ungewohnten, (sprachliche) Hilfen gegeben werden, um sich in den entstehenden "Warenhäusern der Information" zuverlässig zurechtzufinden.

Es stellt sich dabei das Problem, inwieweit bestehende Wörterbücher, die zumeist im Zuge der modernen Satztechnik in irgendeiner Form "maschinenlesbar" sind, genutzt werden können. Auch wenn die Orientierung an neuen Märkten sicherlich nicht abrupt vor sich gehen wird - "gedruckte" Wörterbücher wird es auch über das Jahr 2000 hinaus noch geben - so wird es doch für die heutigen Hersteller von kommerziellen Gebrauchswörterbüchern (d.h. die Verlage) zu einer Art Überlebensfrage werden, hierbei den Anschluss zu wahren. Umgekehrt wird es für die Software-Industrie, v.a. im Bereich der Büro-Kommunikation, interessant, sich in diesem Feld - sei es nun durch eigene Entwicklungen oder durch Produkte in Zusammenarbeit mit der Verlagsindustrie - zu betätigen. Man kann jedenfalls davon ausgehen, dass der Bedarf an "intelligenten" Bürosystemen gewaltig ansteigen wird.

2. Anforderungen der Fachinformation an maschinelle Wörterbücher

Unter "Fachinformation" wird im folgenden eingeschränkt derjenige Bereich der Fachinformation und -kommunikation verstanden, bei dem sprachlich "kondensiertes" Fachwissen textuell (schriftlich oder mündlich) ausgetauscht wird. Im vorliegenden Zusammenhang kann weiter eingeschränkt werden auf denjenigen Bereich derartiger Fachinformation, der mithilfe von Computern als Distributionsmedium erfolgt. Da eine Kommunikation in der gesprochenen Sprache, bei der seitens des Computers ein tief gehendes Sprachverstehen vorliegt, trotz aller Anstrengungen der Forschungen im Bereich der sog. "Künstlichen Intelligenz" (KI) für eine breite kommerzielle Anwendung noch nicht in Sicht ist, kann schließlich zumindest die Kommunikation über maschinelle Spracherkennung ausgeklammert werden.

Ein wesentlicher Teil der Fachinformation wird im textuellen Bereich heute über Informationsbanken abgewickelt. Im Mittelpunkt stehen dabei bibliographische Datenbanken, z.B. in der Medizin, in der Technik, im Patentwesen, in der Rechtsprechung. Eine der größten Datenbanken in diesem Bereich sind die Chemical Abstracts: hier werden pro Jahr ca. 500.000 Literaturangaben mit Abstracts gespeichert und online, d.h. im Dialog Mensch-Maschine verfügbar. Allein

im deutschen Patentbereich werden z.Zt. jährlich über 40.000 Kurzfassungen zu Patentanmeldungen (sog. Offenlegungsschriften) gespeichert. Man schätzt die jährlichen Veröffentlichungen (Monographien, Zeitschriftenaufsätze ...) weltweit derzeit auf über 5 Millionen pro Jahr.

Ein wesentliches Problem bei der Distribution von Fachliteratur ist - wie allgemein einsichtig - die sog. Sprachbarriere: Fachlich relevante Texte liegen nur teilweise in der jeweiligen Muttersprache vor; selbst die Kurzfassungen von Primärliteratur (Abstracts) sind - bei starker Orientierung auf das Englische als wissenschaftliche Mittlersprache zumindest in der westlichen Welt - nur in Teilen zugänglich. Von den über 100.000 Patentanmeldungen in Japan wird beispielsweise weniger als die Hälfte mit englischen Kurzfassungen versehen, ähnliches gilt für andere Sprachen und auch für andere Themenbereiche.

Die "sprachliche" Unterstützung eines Informationssuchenden ist bei den weltweit verfügbaren Informationssystemen (Datenbanken-Hosts) äusserst gering. Linguistisch-lexikalische Verfahren kommen so gut wie nirgends zum Einsatz. Die Systeme der größten Zentren für derartige Datenbanken in den einzelnen Ländern, z.B. DIALOG in den USA, QUESTEL in Frankreich, INKA in der Bundesrepublik Deutschland, halten nur minimale Hilfen für die sog. Freitextsuche (d.h. die Suche mit natürlichsprachigen Begriffen im Titel oder der Kurzfassung) bereit. Als eine bescheidene Hilfe steht hier vor allem die sog. "Trunkierung" zur Verfügung. Dabei wird - vereinfacht dargestellt - eine Zeichenkette am Ende mit einer Markierung versehen, um zu verdeutlichen, dass danach noch beliebige oder eine bestimmte Zahl weiterer Zeichen (Buchstaben) folgen können, die bei dem Vergleich der angegebenen Zeichenkette mit der Liste der Einträge in der Datenbank vernachlässigt werden können. So dient die Markierung mit "$" im Beispiel DISTRIBUT$ dazu, eine Reihe von Wortformen (z.B. DISTRIBUTION, DISTRIBUTE, DISTRIBUTED) als "Treffer" zuzulassen. Es ist dabei Aufgabe des Recherchierenden, dies intellektuell vorherzusehen. Während diese Methode bei Sprachen wie dem Englischen (heute) noch erträglich erscheint, ist ein derartiges Verfahren bei stark flektierenden Sprachen, z.B. dem Deutschen, angesichts der Fülle von Unregelmäßigkeiten und v.a. wegen der Kompositumbildung nicht mehr möglich.

Daraus folgt, dass in der Fachinformation die Integration einsprachiger Lexika mit Berücksichtigung morphologischer Merkmale ein besonderes Desiderat sein muss. Um so erstaunlicher ist es, dass bislang in diesem Bereich kein ernsthafter Versuch z.B. seitens der großen Verlage gemacht wurde, vorhandene Daten hierfür einzusetzen. Hierzu einige Beispiele aus der Bundesrepublik Deutschland: Der Brockhaus-Wahrig verfügt über mehr als 200.000 maschinenlesbare Stichwörter mit reichhaltiger flexions-morphologischer Markierung, entsprechendes gilt für den 6-bändigen DUDEN. Es ist zudem relativ einfach, ein Wörterbuchverfahren zusätzlich mit einem Dekompositions- und Derivationsalgorithmus auszustatten, der es erlaubt, themen- und textspezifische Wortzusammensetzungen mit großer Genauigkeit zu ermitteln und somit auch in solchen Fällen Retrievalhilfen (d.h. Suchhilfen beim (Wieder-)Findevorgang) anzubieten, in denen das lexikalische Inventar an sich nicht unmittelbar für eine Identifikation ausreicht. Damit nicht genug: Es liegt inzwischen eine Reihe fachspezifischer Lexika und Terminologien vor (v.a. mehrsprachig, z.B. EURODICAUTOM), die für eine derartige Nutzung eine gute Grundlage bilden könnten.

Es ist müßig, danach zu fragen, warum bislang v.a. seitens der Informationsindustrie hier keine verstärkten Anstrengungen gemacht wurden. Ein wesentlicher Grund dürfte darin liegen, dass im Verlagsbereich - trotz des hohen edv-technischen Kenntnisstandes - nicht genügend Know-How

bezüglich der Nutzbarmachung der Produkte in anderen Bereichen außerhalb der gedruckten Werke vorliegt, ja dass im Gegenteil der neue Markt der elektronischen Fachinformation eher mit Misstrauen betrachtet wird: die Verlagsindustrie erscheint angesichts dieser rasanten Entwicklung weitgehend verunsichert. Diese Unsicherheit verstärkt sich noch aufgrund der Unklarheit bzgl. der Copyright-Fragen in der EDV.

Inzwischen liegen andererseits erste Erfahrungen mit der Anwendung flexionsmorphologischer Wörterbücher in der Fachinformation vor. Sie wurden beispielsweise im Rahmen von Forschungen an der Universität des Saarlandes an deutschsprachigen Daten gesammelt. Das der sog. "Computergestützten Texterschließung" (d.h. der maschinellen Indexierung) zur Fachinformation (CTX) zugrundeliegende maschinelle Wörterbuch baute ursprünglich auf einem lexikalischen Inventar auf, das seitens der Stichwörter weitgehend dem "einbändigen" Lexikon von G. Wahrig in einer Ausgabe zum Ende der 60-er Jahre entsprach. Es ist daher interessant, ausschnitthaft von diesen Erfahrungen zu berichten.

Ziel war es, Wortformen in beliebigen deutschsprachigen Texten mittels eines maschinellen Lexikons und eines darauf bezogenen Algorithmus auf mögliche Grundformen zurückzuführen. Hierzu wurden gemäß der traditionellen Grammatik Wortklassen differenziert, wobei die Wortformen der sog. Funktionswortklassen (Konjunktionen, Präpositionen ...) sowie die reinen Adverbien (z.B. HEUTE, OFT) als solche lexikalisch erfasst wurden, während die Substantive, Verben und Adjektive mit ihrem Stamm (genauer: bei unregelmässiger Wortbildung mit mehreren Stämmen) verzeichnet wurden. Beispiele für derartige Stämme sind HAUS (für HAUS, HAUSE, HAUSES) und HÄUSER (für HÄUSER, HÄUSERN), GEH (für GEH, GEHE, GEHEN, GEHT, GEHST), GING (für GING, GINGST, GINGE, GINGEN) usf. Diese Stämme erhielten Markierungen bezüglich der möglichen Flexionsendungen und ggf. einen Verweis auf die "Grundform".

Inzwischen ist dieses maschinelle Wörterbuch auf rd. 150.000 derartiger "Stämme" angewachsen, wobei Erweiterungen vorwiegend aus den Fachgebieten der behandelten Texte (Recht, Patentwesen, Sozialwissenschaften) stammen. Da für die Ermittlung relevanter Grundformen im Deutschen die Gross-/Kleinschreibung von gewisser Bedeutung ist, wird neben dem Flexionskode auch die Angabe zur Wortklasse mitgeführt. (Sie stellt übrigens bei den Nomina einen Teil der Flexionsmarkierung dar.)

Neben dem flexionsmorphologischen (Basis-)Lexikon, das zudem syntaktische Angaben aufweist, wurden inzwischen weitere Lexika erstellt, die v.a. für das Datenbank-Retrieval von Bedeutung sind: In einem sog. "Derivationslexikon" werden Adjektive, Verben und Nomina, die formal-flexivisch in einem sinnvollen Zusammenhang stehen, einander zugeordnet. Die Zuordnungsmöglichkeit - unabhängig von etwaigen Bedeutungsvarianten - wird partiell bereits beim ersten Auftreten einer Derivation beim maschinellen Lexikonabgleich erkannt, sie wird jedoch intellektuell kontrolliert und expliziert. So wird beispielsweise eine Derivationskette wie SINGEN - GESANG - SANGBAR intellektuell erzeugt, während Ableitungen in derivationsmorphologisch "einfacheren" Fällen (z.B. VERFÜGEN - VERFÜGBAR - VERFÜGBARKEIT - VERFÜGUNG) weitgehend automationsgestützt erfolgen, da das System über einen entsprechenden Algorithmus verfügt. Die Konsequenz der Verfahrensweise ist es, dass ein Rechercheur in einer Textdatenbank unter automatischer Einbeziehung derartiger Relationen eine grössere Trefferquote erreichen kann.

Ähnliche Möglichkeiten ergeben sich, wenn sinnvolle Teilwörter eines Kompositums berücksichtigt werden. Auch hierbei ist die algorithmische Ermittlung auf der Basis des flexionsmorphologischen Lexikons eine wichtige Hilfe. Dennoch zeigte gerade die Praxis, dass eine intellektuelle Kontrolle - v.a. im fachterminologischen Bereich - absolut notwendig ist, um "sinnlose" Zerlegungen (die gegenwärtig in rund einem Drittel (!) der Fälle auftreten) zu vermeiden. Bei der Überprüfung der automatischen Neu-Zerlegungen (beim erstmaligen Auftreten eines Kompositums) durch Fachleute (im vorliegenden Falle z.B. durch Mitarbeiter des Deutschen Patentamts, deren Daten gegenwärtig für Retrievalzwecke testweise aufbereitet werden) sind Kuriositäten (und entsprechende Heiterkeitserfolge) an der Tagesordnung. Gerade die ansonsten kaum (allenfalls mit unvergleichlichem Aufwand) erreichbare Zuverlässigkeit der Dekomposition machte dementsprechend die Einführung eines Kompositum-Lexikons erforderlich. Dies führt zwar einerseits zu erhöhtem Speicherbedarf, reduziert andererseits aber den Analyseaufwand (Rechenzeit) für die aktuelle Texterschließung beträchtlich, da der Dekompositions-Algorithmus (eine komplizierte rekursive Prozedur) nur noch dann verwendet werden muss, wenn über eine Flexionsendungs-Analyse im Basis- bzw. im Kompositum-Lexikon kein Treffer erzielt wurde.

Zusammenfassend seien nochmals die wesentlichen Merkmale aufgeführt, die Lexika aufweisen sollten, welche für Fachinformationszwecke eingesetzt werden:

(1) flexionsmorphologische Angaben/Verweise
(2) derivationsmorphologische Angaben/Verweise
(3) (de-)kompositionsspezifische Angaben/Verweise.

Untersucht man bestehende (gedruckte) Gebrauchslexika unter diesen Gesichtspunkten, so muss man feststellen, dass diese Kriterien nur partiell systematisch erfüllt werden. Dies gilt besonders für den Bereich der Komposition und Derivation: Ein Hauptproblem bildet die streng alphabetisch-lexikalische Anordnung, auf die hin nahezu alle Lexika ausgerichtet sind. Selbst dort, wo eine sog. "Nesterbildung" vorliegt, d.h. dann, wenn unter einem Haupt-Stichwort auch Derivationen und Komposita aufgeführt sind, ist allenfalls der gemeinsame "linke" Bestandteil (Wortanfang) nesterbildend: eine Zusammenstellung unabhängig von der alphabetischen Ordnung ist daher nicht möglich. Dafür ein Beispiel aus dem 6-Bände-DUDEN: Das derivativische "Nest" EINSPRECHEN, EINSPRECHER, ... EINSPRUCH ist durch den Einschub von EINSPRENGEN, EINSPRINGEN, EINSPRITZ ... "gesprengt"; EINGABE ... ist durch EINGANG von EINGEBEN getrennt usf. Selbst dann, wenn - wie im Falle der Erstellung der "Saarbrücker" maschinellen Wörterbücher - ein Gebrauchswörterbuch gleichsam als "Steinbruch" zugrundegelegt wird, ist daher eine eingehende intellektuelle Aufbereitung der Einträge unter den genannten Gesichtspunkten erforderlich.

Ein weiteres Problem stellt die Verwendung herkömmlicher Wörterbücher im Zusammenhang mit einer semantischen Disambiguierung (Bedeutungsdifferenzierung) dar. Eine entsprechende "Vereindeutigung" oberflächig mehrdeutiger Benennungen ist v.a. dann wichtig, wenn über die Morphologie hinausgehende Begriffsrelationierungen vorgenommen werden bzw. wenn mehrsprachige Indexierungen oder Sprachübersetzungen erfolgen sollen. Die vorhandenen Gebrauchslexika sind gemäß ihrer Zielgruppe (d.h. der unmittelbaren Nutzung durch den Menschen) - wenn sie überhaupt Bedeutungsdifferenzierungen vornehmen - eher deskriptiv: verschiedene Bedeutungen werden dabei mehr oder minder präzise unterschieden (aufgelistet) und allenfalls über sprachliche Definitionen und Beispiele erläutert. Auch hierfür ein kleines Beispiel, zufällig unter

vielen ausgewählt: So unterscheidet der Brockhaus-Wahrig bei EINFLÜSTERN 3 Bedeutungen (1. - allgemein - "jemandem etwas flüsternd mitteilen"; 2. - theatersprachlich - "souffl ieren"; 3. jemanden "heimlich beeinflussen"). Der Grosse DUDEN nennt nur 2 Bedeutungen (1. "in flüsterndem Ton ... sprechen"; 2. "heimlich einreden, überreden"). Dies mag für die Benutzung durch den Menschen ausreichen, ist jedoch für eine maschinelle Bearbeitung weitestgehend unzureichend. Nützlich sind allenfalls - soweit vermerkt - syntaktische oder syntakto-semantische Merkmalangaben, etwa bei den Verben zu den Kasus (Valenzen) bzw. zur präpositionalen Attribuierung. Diese "Schwächen" der Gebrauchslexika (sie ließen sich an allen bestehenden Wörterbüchern nachweisen) - bezogen auf eine unmittelbare Nutzung für maschinelle Weiterverarbeitungsprozesse - sind nun keineswegs – dies muss hier deutlich bleiben - auf unzulängliche Bearbeitungen zurückzuführen: sie resultieren teilweise aus der Vagheit der Sprache (dies wird z.B. bei den semantischen Unterscheidungen besonders deutlich), v.a. jedoch aus der (bisher) unterschiedlichen Zielgruppenorientierung. Ein gängiges Gebrauchswörterbuch setzt andere Problemstellungen voraus. Der menschliche Benutzer hat Probleme mit der Rechtschreibung, der Aussprache, der Beugung, der Begriffserklärung (die unter Verwendung von ihm möglichst bekannten Begriffen erfolgt) und sucht in der Tat Anwendungsbeispiele und "wissenschaftlich" allenfalls eine Angabe zur Wortgeschichte (Etymologie); er zieht keinen Nutzen aus formalsemantischen oder auch -syntaktischen Merkmalen. Die Strategien, nach denen heutige maschinelle Sprachanalysesysteme arbeiten, werden daher in den gängigen Wörterbüchern wenig oder gar nicht unterstützt. Solche "maschinellen" Strategien nutzen z.B. Angaben über die Vorkommenshäufigkeit einer Bedeutungsvariante in einem (Teil-)Fachgebiet aus, sie differenzieren Bedeutungen aufgrund kontextueller Begriffe, die mit ihnen in einer bestimmten oder freien (assoziierten) Relation stehen usf.

Weitaus besser erscheinen hierzu - dies gilt zumindest für zentrale Fachbegriffe - bestehende Fachterminologien oder Klassifikationssysteme geeignet. Es handelt sich einmal - wie erwähnt - um die sog. "Thesauri". Ihr (bisheriger) Zweck ist es in der Regel, über ein sog. "Kontrolliertes Vokabular" von (Fach-)Begriffen bei der intellektuellen Indexierung (d.h. der Texterschließung) durch "Deskriptoren" eine größere Einheitlichkeit und Systematik zu erreichen. Bei der intellektuellen Indexierung soll v.a. die in natürlichsprachigen Texten auftretende Mehrdeutigkeit von Benennungen überwunden werden. Der Thesaurus stellt dabei ein "kunstsprachliches" Vokabular dar, das es auch erlaubt, bedeutungsmäßig eindeutig festgelegte natürlichsprachige Benennungen zu verwenden. Man unterscheidet im allgemeinen sog. "Vorzugsbenennungen", die einem Dokument als Deskriptor zugewiesen werden können, und "Nichtdeskriptoren", bei denen auf die "Vorzugsbenennungen" verwiesen wird. Wird beim intellektuellen Indexierungsvorgang ein Nichtdeskriptor in einem Dokument (Text) identifiziert und für relevant (d.h. das Dokument spezifizierend) gehalten, so wird an seiner Stelle die "Vorzugsbenennung" als Deskriptor vergeben. Zur Erleichterung des Retrievalprozesses werden in den Thesauri im allgemeinen die Benennungen zusätzlich über semantische Relationen miteinander verknüpft (vgl. dazu z.B. die Norm DIN 1463).

Derartige Thesauri - die zu einer Reihe von Fachgebieten, etwa der Medizin (MEDLARS / MESH), vorliegen, können somit eine wichtige Grundlage für maschinelle (Indexierungs-)Lexika bilden, v.a. wenn entsprechende Kontext-Strategien zur Bedeutungsdifferenzierung natürlichsprachiger Benennungen zum Einsatz kommen. Ähnliches gilt für Klassifikationssysteme: So kann die im Text vorkommende Benennung SEELE im Bereich der IPC "HO1B" (d.h. der Bereich "Kabel und Leitungen" in der Internationalen Patentklassifikation) nur in der Bedeutung "Kabel-Inneres" verstanden werden.

Für eine zweite (tiefere/verfeinerte) "Ebene" der maschinellen Erschließung von Fachinformation sind dementsprechend zusätzliche Lexika bzw. Lexikon-Markierungen erforderlich, die etwa wie folgt differenziert werden können:

(4) formal-syntaktische Informationen (zur Ermittlung von Mehrwortbegriffen, zur syntaxbezogenen semantischen Vereindeutigung) ;
(5) lexikalische Bedeutungsdifferenzierung (evtl. Definitionen/Erläuterungen zur Erleichterung der intellektuellen Bearbeitung) ;
(6) semantische Relationierung.

Es erscheint durchaus möglich, dass längerfristig auch multifunktionale Lexika entwickelt werden, die soweit strukturiert und differenziert sind, dass sowohl die direkte menschliche Nutzung als auch die maschinelle Auswertung möglich wird. Für die Zwecke der entsprechenden Nutzung sind dementsprechend (maschinelle oder maschinengestützte) Aufbereitungen erforderlich, die ggf. nur Teile des Inventars (etwa für den Druck eines Publikum-Wörterbuchs) heranziehen. Abschließend ist festzuhalten, dass erst eine Kombination aus Angaben in bestehenden Terminologielisten, Thesauri und Gebrauchswörterbüchern eine Grundlage schafft für das Inventar zur Verwendung in der maschinellen Fachinformation. Es ist nicht zu verkennen, dass v.a. für eine tiefergehende computergestützte Indexierung bzw. Übersetzung noch erhebliche Investitionen bezüglich des Aufbaus ausreichend dimensionierter und strukturierter Lexika erforderlich sein werden.

3. Anforderungen der Bürokommunikation an maschinelle Wörterbücher

Bürokommunikation und Fachinformation sind in Bezug auf das hier im Mittelpunkt stehende Thema der Dokumentablage und des Wiederfindens nicht prinzipiell, sondern eher graduell unterscheidbar. Gegenwärtig (dies muss nicht für alle Zukunft so sein) ist die Entwicklung in der Fachinformation v.a. von der Aufgabe her bestimmt, die Daten möglichst (fach-)sprachlich präzise und inhaltlich differenziert zu erschliessen. Wissenschaftliche Fachkommunikation ist in großen Teilen weltweit zu sehen, so dass hier das Problem der Sprachbarriere weit eher zu Buche schlägt als im Alltag der (meist einsprachigen) Bürokorrespondenz. Obwohl "Bürotexte" und "Fachliteratur" sich vielfach überlappen können (man denke z.B. an das Büro eines Wissenschaftsbetriebes, aber auch an das eines Patent- oder Rechtsanwalts), soll diese Unterscheidung hier beibehalten werden, da sich v.a. in Bezug auf die Funktion(en) des Büros als einer betrieblichen Informationsvermittlungsstelle eine Reihe anderer Anforderungen an die Textver- und -bearbeitung stellt.

Zu den typischen textuellen Dokumenten (Textsorten), die in einem Büro (ohne Berücksichtigung fach- oder branchenspezifischer Besonderheiten) bearbeitet werden, gehören der BRIEF, das PROTOKOLL, der BERICHT, die NOTIZ. Vor allem bei der Korrespondenz werden Anforderungen an die Erfüllung bestimmter Normen (Rechtschreibung, Silbentrennung, Zeichensetzung, Stil) gestellt, die zumindest partiell lexikalisch orientiert bzw. bestimmbar sind. Die technischen Rahmenbedingungen unterscheiden sich - jedenfalls heute noch - ebenfalls von solchen der Fachinformation: Während in der Fachinformation große Datenpools (z.T. mit vielen Millionen Dokumenten) auf zentralen (Groß-)Rechnern verfügbar gehalten werden und der Benutzer über Fernleitung mithilfe eines "einfachen" Terminals Zugang erhält, verlangt die

Arbeit im Büro zunehmend "intelligente" Microcomputer bzw. Textsysteme, die lokal, d.h. vor Ort individuell und aktuell spezifische (Büro-)Funktionen verfügbar machen. Diese sollen besonders die Texterstellung (Generierung) und -bearbeitung, zunehmend auch das Ablegen und Wiederfinden (Archivierung) unterstützen. Eine systematische Wörterbuchpflege durch den Benutzer (Sekretärin, Sachbearbeiter, Manager) ist nicht möglich, allenfalls sind in Zukunft über Büro-Kommunikations-Netze einige zusätzliche zentrale Funktionen vorstellbar.

Im Bereich der Korrespondenz sind v.a. die Silbentrennung und die Rechtschreibhilfe zu einem besonderen Desiderat geworden. Es gibt heute kaum mehr ein Textsystem, das nicht über einen sog. SPELLING CHECKER verfügt, bei Sprachen mit besonders "langen" Wörtern (wie dem Deutschen) ist die (halb-)automatische Silbentrennung eine Selbstverständlichkeit.

Bei den automatischen Rechtschreibhilfen werden insbesondere frequenzorientierte Wortformenlisten verwendet; der Benutzer hat dabei die Möglichkeit, diese Listen textbezogen zu ergänzen. Somit werden zunächst "fälschlich" als mögliche Rechtschreibfehler identifizierte korrekte Wortformen nach Eingabe durch den Benutzer später als richtig erkannt, so dass im Laufe der Zeit der Anteil der echten Rechtschreibfehler steigt.

V.a. in den USA hat man darüber hinaus (zum Englischen) inzwischen umfangreiche Wörterbücher in die Rechtschreibhilfe integriert (man vgl. den SPELLING CHECKER von SPERRY oder das Verfahren zum IBM-Textsystem). Dabei werden nicht nur flexionsmorphologische Merkmale herangezogen, sondern auch Regeln zur Aussprache und mathematisch-statistische Verfahren (Distanz-Funktionen), die die "Nähe" eines (nicht direkt identifizierten) Textwortes zu einem Lexikoneintrag ausdrücken. Während dabei die flexionsmorphologische Identifikation eine korrekte Zeichenkette signalisiert, bringen Aussprache- und Distanzfunktionen nur Anhaltspunkte für systemseitige Vorschläge (engl.: GUESSES).

Bezüglich der Silbentrennung sind - in Fortentwicklung von Verfahren, wie sie seit längerem im Zeitungssatz oder Buchdruck verwendet werden - v.a. zwei Methoden zu unterscheiden: Einerseits werden (Schreib-)Silben über ein maschinelles (sprachbezogenes) Regelsystem ermittelt, wobei Ausnahmen in einer Ausnahmenliste verzeichnet sind, die beim Aufruf einer Silbentrenn-Funktion z.B. zum Randausgleich (Justieren) zuvor abgefragt wird. (Es sei hier angemerkt, dass der vorliegende Text nach einem derartigen Verfahren vollautomatisch über "Randausgleich" getrennt wurde) Eine alternative Methode ist es, v.a. in Verbindung mit der Rechtschreibhilfe, Maschinen-Lexika als "Positiv-Liste" heranzuziehen, d.h. alle darin verzeichneten Wörter (intellektuell) "vorzutrennen". Während die erste Methode - angesichts der prinzipiellen Problematik eines sich verändernden Wortschatzes - man denke an die Fremdwörter - in gewisser Weise "fehleranfällig" bleibt (dennoch sei beispielsweise angemerkt, dass die Fehlerquote des hier verwendeten Programms bereits weit unter 1 Prozent - bezogen auf die Trennvorschläge - liegt), ist die zweite Methode - sofern nur lexikalisierte Wörter überhaupt getrennt werden (im Englischen durchaus nützlich, im Deutschen angesichts der Komposita nicht immer anwendbar) völlig unproblematisch. (Allerdings muss meist für die nicht erkannten Wörter eine intellektuelle Trennung erfolgen.)

Inzwischen sind daneben - wiederum ist die USA hier federführend - erste Verfahren zu einer Art "Stilhilfe per Computer" auf dem Markt. Bei diesen Verfahren hat der Autor/Schreiber eines Briefes oder Berichts z.B. die Möglichkeit, sich (über eine Art Synonymrelation) Begriffsvorschläge vom System unterbreiten zu lassen. Zeigt er z.B. auf das im Text vorkommende Wort

NICE, so macht ihm der Rechner - wohlgemerkt beim Vorgang der Textgenerierung - ggf. einen Alternativ-Vorschlag, etwa (im Beispiel) den Vorschlag, das Wort NICE durch BEAUTIFUL zu ersetzen.

Es wird im Bürobereich nicht mehr lange dauern, bis am Arbeitsplatz Wörterbuchfunktionen unterschiedlichster Art - d.h. sowohl zur (sprachbezogenen oder enzyklopädischen) Information des menschlichen Benutzers wie auch zur Unterstützung von Systemfunktionen wie automatische Silbentrennung und Rechtschreibkorrektur - zum "Alltag" gehören werden. Die bestehenden (gedruckten) Gebrauchswörterbücher sind dabei im Bürobereich unter vielen Aspekten nutzbar. So verzeichnen heute alle modernen Wörterbücher zum Deutschen etwa die (Schreib-)Silbentrennung; aufgrund der notierten flexionsmorphologischen Angaben ist weitestgehend auch eine grundform-orientierte "Positiv-Liste" zur Rechtschreibhilfe vorhanden. Bedeutungslexika wie z.B. das Synonymwörterbuch von KNAUR ließen sich durchaus auch als Ausgangspunkt für umfassende und systematische maschinelle Stilhilfen verwenden.

Für das Englische/Amerikanische liegen offensichtlich bereits unmittelbare "Umsetzungen" (maschinenlesbarer) gedruckter Wörterbücher zur Rechtschreibhilfe in Büro- oder Textsystemen vor, wobei die sog. SPELLING CHECKER oft noch über zusätzliche Funktionen, z.B. zur Ermittlung der Wortähnlichkeit verfügen. Eine weitere Funktion, die wenigstens partiell auf den "gedruckten" Vorlagen aufbauen kann, ist der phonetische Vergleich (Aussprache-Ähnlichkeit). Das Interesse der Verlage (z.B. Longman, Merriam-Webster) ist offenbar gegeben, an solchen Entwicklungen im Bürobereich zu partizipieren. Obgleich also in Teilbereichen (v.a. zu flexionsmorphologisch "einfacheren" Sprachen wie dem Englischen) die Voraussetzungen günstig sind, bestehende Gebrauchswörterbücher zu integrieren, ist dies nicht möglich ohne Ergänzungen und Abwandlungen des vorliegenden Materials. Eine Rechtschreibhilfe, die z.B. automatisch "wissensorientierte" Rechtschreibfehler korrigiert, benutzt geradezu "Falscheinträge" (z.B. LYBIEN, SODASS, ATLET), um eine Korrekturgrundlage zu haben (hier: LIBYEN, SO DASS, ATHLET). Man muss fast sagen: "leider" gibt es bislang keine Gebrauchswörterbücher, die systematisch (d.h. erfahrungsorientiert) auch orthographisch fehlerhafte Einträge aufweisen, um den Benutzer zu den "richtigen" Schreibweisen hinzuführen.

Wenn man etwas in die Zukunft blickt, so kann man davon ausgehen, dass für die Textarchivierung und das entsprechende Wiederfinden von Briefen, Berichten, Notizen usf. die Nutzung von Textwörtern an Bedeutung gewinnen wird. Neben "strukturierten" Angaben (z.B. Absendername, Adressat, Datum bei BRIEFEN) wird man im Bürobereich bei entsprechend grossen Datenmengen auch themenorientiert (d.h. mit Textbegriffen) suchen wollen. Auch wenn zunächst die Ansprüche der Benutzer nicht so hoch sein werden wie im Bereich der Fachinformation, so können doch flexionsmorphologische Funktionen (im Deutschen darüber hinaus sicherlich auch Derivations- und Dekompositionsverfahren) hierbei eine gute Unterstützung bieten. Auf längere Sicht werden also - sieht man einmal ab von einer intensiven Lexikon-Pflegekomponente - ähnliche Anforderungen auftreten wie bei der Fachinformation.
Wenn sich die Entwicklung der Telekommunikation, die Miniaturisierung der Computertechnik und die Dimensionierung der Speichermöglichkeiten weiterhin so rasch vollziehen, darf letztlich die Prognose gewagt werden, dass es noch in diesem Jahrhundert - quantitativ gesehen, d.h. bezogen auf die Nutzungsfrequenz - einen totalen Wandel vom Wörterbuch zur Wörter(daten)bank geben wird, der erheblichen Einfluss auf die Herstellung und den Vertrieb von Gebrauchswörterbüchern nehmen wird. Dabei gibt es heute noch eine Chance für Verlage, Soft- und Hardware-Hersteller, sich zu einer Kooperation zu entschließen, die das jeweilige Know-How der Beteilig-

ten optimal für eine angemessene Entwicklung neuer Formen der (elektronischen) Sprachinformation nutzen kann.

Literatur und Quellen

Althaus, H.P.; Henne, H.; Wiegand, H.E.: Lexikon der germanistischen Linguistik. Tübingen 1973

Brockhaus-Wahrig: Deutsches Wörterbuch in sechs Bänden. Wiesbaden 1980 ff. (bislang 1-5)

DUDEN: Das große Wörterbuch der deutschen Sprache in sechs Bänden. Mannheim 1977 ff.

Hess, K.; Brustkern, J.; Lenders, W.: Maschinenlesbare deutsche Wörterbücher. Sprache und Information Bd. 6, Tübingen 1983

Lyons, J.: Introduction to Theoretical Linguistics. Cambridge. (deutsch: Einführung in die moderne Linguistik. München 1971 ff.)

Schaeder, B.: Lexikographie als Theorie und Praxis. Tübingen 1981

Stammerjohann, H. (ed.): Handbuch der Linguistik. Allgemeine und angewandte Sprachwissenschaft. München 1975

Wehrle, H.; Eggers, H.: Deutscher Wortschatz. Stuttgart 1961 ff.