LEXICOGRAPHICA Series Maior

LEXICOGRAPHICA

Series Maior

Supplementary Volumes to the International Annual for Lexicography
Suppléments à la Revue Internationale de Lexicographie
Supplementbände zum Internationalen Jahrbuch für Lexikographie

107

Published in cooperation with the Dictionary Society of North America
(DSNA) and the European Association for Lexicography (EURALEX)

Chancen und Perspektiven computergestützter Lexikographie

Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher

Herausgegeben von
Ingrid Lemberg, Bernhard Schröder
und Angelika Storrer

Max Niemeyer Verlag
Tübingen 2001

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

[Lexicographica / Series maior]
Lexicographica : supplementary volumes to the International annual for lexicography / publ. in cooperation with the Dictionary Society of North America (DSNA) and the European Association for Lexicography (EURALEX). Series maior. – Tübingen : Niemeyer.
Früher Schriftenreihe
Reihe Series maior zu: Lexicographica

107. Chancen und Perspektiven computergestützter Lexikographie. – 2001

Chancen und Perspektiven computergestützter Lexikographie : Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher / hrsg. von Ingrid Lemberg .... – Tübingen : Niemeyer, 2001
(Lexicographica : Series maior ; 107)

ISBN 3-484-39107-3 ISSN 0175-9264

Gedruckt auf alterungsbeständigem Papier.
Druck: Gulde-Druck, Tübingen
Einband: Industriebuchbinderei Nädele, Nehren

Inhaltsverzeichnis

Einführung

Die moderne Wörterbuchforschung beschreibt Lexikographie, speziell die wissenschaftliche Sprachlexikographie, als eigenständige, kulturelle und wissenschaftliche Praxis. Diese Praxis orientierte sich über Jahrhunderte hinweg nahezu ausschließlich an dem Medium Buch. Bücher waren meist die Träger der lexikographischen Quellentexte, aus denen Belege exzerpiert und in Belegarchiven gesammelt wurden. In Büchern wurden die Endprodukte der lexikographischen Arbeit, die Wörterbuchartikel, publiziert und vermarktet. Lange Zeit war das Medium Buch auch das geeignetste Trägermedium für Nachschlagewerke, da es den gezielten, punktuellen Zugriff auf bestimmte Teiltexte besser unterstützt als beispielsweise ein Film, ein Zettelkasten oder eine Papyrusrolle.

Dies hat sich durch die Computertechnik, vor allem durch Datenbank- und Hypertextsysteme sowie durch Computernetzwerke wie das Internet, geändert. In digitalen Wörterbüchern nimmt der Computer den Nutzerinnen und Nutzern das zeitraubende Blättern und Suchen ab, ein Verweis kann durch einen einfachen Mausklick verfolgt werden, das Durchsuchen von mehreren Nachschlagewerken und die Volltextsuche in den gesamten Artikeltexten ist kein Problem. Von diesen Vorteilen können die Nutzerinnen und Nutzer teilweise bereits dann profitieren, wenn elektronische Wörterbücher nur als digitale Kopien der gedruckten Vorläufer publiziert werden. Doch dies ist erst der Beginn eines Prozesses, bei dem gedruckte Wörterbücher sukzessive durch leistungsfähige digitale Nachschlagewerke ergänzt und für viele Benutzungssituationen oder Benutzungsanliegen ersetzt werden. Wer diesen Prozess mitgestalten und dabei die Chancen und Perspektiven des digitalen Mediums in der Lexikographie ausreizen möchte, muss kritisch hinterfragen, inwieweit die am Buch orientierten Traditionen der Wörterbuchgestaltung, z.B. die Makro-, Mikro-, Zugriffs- und Verweisstrukturen, im digitalen Medium noch zweckmäßig sind und inwieweit das neue Medium nicht nach anderen Gestaltungs- und Organisationsprinzipien verlangt. Speicherplatz ist im digitalen Wörterbuch nicht annähernd so teuer wie Druckraum im gedruckten Wörterbuch. Viele Techniken zur Textverdichtung, insbesondere die Textkomprimierung, die in der Printlexikographie notwendig und zum Teil auch sinnvoll waren, um auf möglichst wenig Platz möglichst viel lexikographische Information unterzubringen, werden im digitalen Medium überflüssig. Statt dessen werden Techniken der strukturierten und redundanzfreien Informationsmodellierung benötigt, wie sie z.B. in der Datenbanktheorie verwendet werden, sowie Strategien, um aus dem digitalen lexikalischen Datenpool diejenigen Angaben herauszufiltern und am Bildschirm zu präsentieren, die in einer bestimmten Benutzungssituation benötigt werden. Dadurch wird es möglich, lexikographische Angaben verständlicher zu formulieren und den Aufbau der Wörterbuchartikel flexibel an den aktuellen Informationsbedarf der Nutzerinnen und Nutzer anzupassen.

Das digitale Medium beendet auch die Diskussion um die „richtige" Anordnung des Wortschatzes, die sich wie ein roter Faden durch die Geschichte der Lexikographie zieht. Die alphabetische Anordnung, zurecht als rein formal und nicht den lexikologischen Zusammenhängen entsprechend kritisiert, war bislang effizient in ihrer Funktion, den gezielten Zugriff auf die lexikographischen Informationen zu erleichtern. Diese Funktion entfällt in digitalen Wörterbüchern, in denen die Suche nach Informationen auf Techniken des *Information Retrieval* basiert. Statt dessen wird nun zu diskutieren sein, wie eine adäquate

linguistisch motivierte Modellierung lexikographischer Angaben im digitalen Medium aussehen soll.

Die Computertechnik verändert auch den lexikographischen Arbeitsprozess. Der Computereinsatz in der Wörterbuchwerkstatt erlaubt es bei entsprechender technischer Infrastruktur, Abläufe effizienter und flexibler zu gestalten und damit gerade umfangreiche Wörterbuchprojekte schneller, qualitätsvoller und kostengünstiger abzuschließen. Wenn das lexikographische Korpus und die Belegsammlungen in Datenbanken verwaltet werden, können Lexikographinnen und Lexikographen von ihren Arbeitsplatzrechnern aus darauf zugreifen. Dies macht den langwierigen und kostspieligen Aufbau von Belegarchiven und die Anmietung von Räumen für deren Aufbewahrung künftig überflüssig. Liegt einmal das lexikographische Korpus zusätzlich mit korpuslinguistischen Methoden aufbereitet vor und ist der lexikographische Arbeitsplatz mit leistungsfähigen Suchwerkzeugen ausgestattet, dann verfügen die Lexikographinnen und Lexikographen über Recherche- und Auswertungsmöglichkeiten, von denen ihre Vorgängergenerationen nur hätten träumen können. Die Verwaltung lexikographischer Beschreibungsergebnisse durch ein Datenbanksystem hilft zudem bei der Konsistenzprüfung und der Verweiskontrolle. Die digitale Verfügbarkeit von Wörterbuchbasis und Beschreibungsergebnissen macht es möglich, Arbeitsabläufe zu dezentralisieren, d.h., auf räumlich verteilte Arbeitsstellen zu verteilen. Die Kommunikationsdienste des Internet können dabei nicht nur projektintern, sondern auch zur Außendarstellung und Qualitätskontrolle genutzt werden, indem z.B. Probeartikel getestet und Rückmeldungen künftiger Nutzerinnen und Nutzer bei der Wörterbuchplanung berücksichtigt werden.

Die hier nur grob skizzierten Vorteile des digitalen Mediums sollten laufenden und geplanten Wörterbuchprojekten genug Anlass geben, die Technikhürde zu überwinden, und die Chancen und Perspektiven der Computertechnik für die Lexikographie zu nutzen. Dieser Band enthält Beiträge von Autorinnen und Autoren aus der lexikographischen Praxis und der historischen und computerlinguistischen Sprachwissenschaft, die sich unter verschiedenen Aspekten mit diesen Herausforderungen auseinander setzen. Drei grundlegende Konzepte der computergestützten Lexikographie stehen dabei im Mittelpunkt: die Einsatzmöglichkeiten der Auszeichnungssprachen SGML/XML für die lexikographische Informationsmodellierung, die Nutzung des Internet als Publikationsmedium für lexikographische Produkte und Hypertext/Hypermedia als Organisationsform für lexikalisches Wissen. Die Beiträge im ersten Teil des Bandes führen in diese Konzepte ein und erörtern deren Chancen und Perspektiven für die computerunterstützte Produktion und Publikation von Wörterbüchern. Im zweiten Teil des Bandes werden theoretische und methodische Aspekte diskutiert, die sich beim Einsatz dieser Konzepte in konkreten Wörterbuchprojekten ergeben.

Im ersten Beitrag erläutern *Gregor Büchel* und *Bernhard Schröder* grundlegende informationstechnische Konzepte und Techniken zur Strukturierung und Verwaltung lexikographischer Daten: Datenbanksysteme, Datenmodelle und Verfahren der konzeptionellen Datenmodellierung sowie die Grundlagen der textuellen Informationsmodellierung mit SGML und XML. Die Aspekte der konzeptionellen Datenmodellierung und des Einsatzes von SGML/XML in lexikographischen Produktions- und Publikationsprozessen werden im nachfolgenden Beitrag von *Ingrid Schmidt* und *Carolin Müller* vertieft. Auf der Basis einer kritischen Auseinandersetzung mit den von der Text Encoding Initiative (TEI) vorgeschlagenen Dokumentstrukturgrammatik (Document Type Definition, DTD) für Print-Wörterbücher entwerfen sie ein lexikographisches Modell, das durch eine modulare Inhaltsstrukturmodellierung flexible Sichten auf die Daten ermöglicht und – im Sinne des „multiple

media publishing" – unabhängig von den Randbedingungen und Konventionen eines bestimmten Trägermediums bleibt.

Der flexible und interaktive Zugriff auf lexikographische Daten ist auch eine der zentralen Zielsetzungen des Hypertext-Konzepts von *Angelika Storrer*. Sie erläutert die wesentlichen Merkmale von Hypertext-Wörterbüchern und diskutiert dann in Thesenform, wie das Mehrwertpotential des neuen Mediums für die Erarbeitung innovativer digitaler Wörterbücher optimal ausgeschöpft werden kann. Um die Mehrwerteigenschaften, die speziell das Internet und sein hypermedialer Dienst World Wide Web (WWW) gegenüber dem traditionellen Medium Buch aufweist, geht es in dem Beitrag von *Ingrid Lemberg*. Durch einschlägige Beispiele veranschaulicht sie Chancen und Perspektiven, die das neue Publikationsmedium vor allem für die wissenschaftliche Lexikographie eröffnet. Der Beitrag gibt einen Überblick über den aktuellen Stand der Entwicklungen im deutschsprachigen Raum und greift künftige Forschungsfragen in diesem Bereich auf. Komplementär dazu diskutiert *Annette Klosa* in ihrem Beitrag grundlegende Aspekte der Publikation von Wörterbüchern auf CD-ROM aus der Perspektive der Verlagslexikographie. Unter Bezugnahme auf einschlägige Rezensionen entwickelt sie Kriterien zur Qualitätsbewertung von CD-ROM-Wörterbüchern und erläutert, wie diese in der Praxis der kommerziellen Wörterbuchherstellung umgesetzt werden können und wie sich die medialen Veränderungen auf die Arbeit der Lexikographinnen und Lexikographen in der Verlagslexikographie auswirken.

Im zweiten Teil des Bandes werden theoretische und methodische Aspekte beim praktischen Einsatz von SGML/XML, Internet und Hypertext/Hypermedia am Beispiel konkreter Projekte diskutiert; der Schwerpunkt liegt dabei auf der wissenschaftlichen Lexikographie des Deutschen. Ein umfangreiches lexikalisches Informationssystem zur deutschen Gegenwartssprache, das Projekt LEKSIS bildet den Hintergrund für die im Beitrag von *Ulrike Haß-Zumkehr* angestellten Überlegungen zur Mikrostruktur im Hypertext-Wörterbuch. Sie macht deutlich, welche methodischen Herausforderungen mit der Nutzung des Hypertext-Konzepts verbunden sind und welche neuen und interessanten sprachtheoretischen und metalexikographischen Fragestellungen sich ergeben. Während LEKSIS bei der Datenmodellierung nicht an ein vorgängig vorhandenes Print-Wörterbuch gebunden ist, geht es in den folgenden Beiträgen um Aspekte der digitalen Aufbereitung von bereits gedruckten Wörterbüchern. Der Beitrag von *Thomas Gloning* und *Rüdiger Welter* beschreibt am Beispiel des Goethe-Wörterbuchs, wie die vielfältigen Aspekte der Architektur eines Wortschatzes, die durch die alphabetischen Anordnung der Lemmata oft verdeckt werden, durch eine mehrdimensionale SGML-Annotierung der Print-Vorlage expliziert werden können und inwiefern eine solche Aufbereitung die Nutzerinteressen besser und flexibler bedienen kann als die gedruckte Publikation. Der Beitrag von *Thomas Burch* und *Johannes Fournier* basiert auf den Erfahrungen in einem großen Projekt zur retrospektiven Digitalisierung mittelhochdeutscher Wörterbücher. Der Schwerpunkt liegt dabei auf den Vorzügen, Schwierigkeiten und Nachteilen, die sich im Zuge der Anwendung der bereits im Beitrag von *Schmidt* und *Müller* (s.o.) diskutierten Vorgaben der Text Encoding Initiative (TEI) ergeben haben. Das Mittelhochdeutsche ist auch Gegenstand des Beitrags von *Ralf Plate* und *Ute Recker*, allerdings geht es dabei nicht um die retrospektive Digitalisierung gedruckter Wörterbücher, sondern um die Produktions- und Publikationsprozesse eines neuen mittelhochdeutschen Belegwörterbuchs, das von Beginn an computergestützt erarbeitet wird. Die Autoren diskutieren Komponenten und Verfahren des verwendeten Publikationssystems und beschäftigen sich mit konzeptionellen Fragen der geplanten elektronischen Publikation.

Die Vorteile und Mehrwerte von Hypertext und Hypermedia für die wissenschaftliche Lexikographie werden von *Gerd Richter* am Beispiel der Hypertextualisierung des Hessischen Flurnamenwörterbuchs sehr eindrücklich aufgezeigt. Der Beitrag diskutiert Fragen der Integration digitalisierter lexikographischer Quellen in das Hypertextwörterbuch sowie Strategien der Konversion von Verweisangaben in Hypertext-Verknüpfungen (Links). Der Beitrag von *Krysztof Petelenz* nähert sich dem Thema „Hypertext" aus der Perspektive der zweisprachigen Lexikographie. Zunächst werden das konzeptionelle Datenmodell und die Architektur eines polnisch-deutschen Online-Wörterbuchs skizziert, das mit dem Hypertextsystem SchemaText entworfen und verwaltet wird. Im Detail wird dann die Modellierung von Mehrwortlexemen und Kollokationen sowie der Rolle von Abbildungen im Hypermedia-Wörterbuch erörtert. *Claudia Kunze* und *Andreas Wagner* beschreiben in ihrem Beitrag den Aufbau, die Informationstypen und die Anwendungsmöglichkeiten des lexikalisch-semantischen Netzes GermaNet. Die in Anlehnung an das englische WordNet organisierte und in einem übergreifenden europäischen Projekt (EuroWordNet) multilingual vernetzte Ressource kann nicht nur von Menschen, sondern auch für die maschinelle Sprachverarbeitung genutzt werden – zur semantischen Disambiguierung, zur Verschlagwortung von Texten sowie zur Akquisition von linguistischen Informationen aus Textkorpora. Abschließend verdeutlicht die Fallstudie von *Lothar Lemnitzer*, in der die Zugriffsprotokolle von zwei im WWW publizierten bilingualen Wörterbüchern ausgewertet wurden, dass das Internet nicht nur ein schnelles und kostengünstiges Publikationsmedium für Wörterbücher ist, sondern auch interessante Perspektiven für die Wörterbuchbenutzungsforschung eröffnet.

Die Idee zu diesem Sammelband entstand durch Diskussionen im Anschluss an das Symposium *Computergestützte Produktion und Publikation von Wörterbüchern*, das wir mit der Unterstützung der Heidelberger Akademie der Wissenschaften, der Gesellschaft für linguistische Datenverarbeitung (GLDV) und dem Institut für deutsche Sprache in Mannheim im September 1998 in Heidelberg veranstaltet haben. Unser herzlicher Dank gilt allen Autorinnen und Autoren für die Beiträge zu diesem Band sowie Kurt Thomas für seine verdienstvolle Arbeit am Layout, an der Indexierung und an der Herstellung der reprofähigen Druckvorlage. Bei der Herstellung des Bandes hat uns dankenswerterweise die GLDV finanziell unterstützt. Für die Aufnahme in die Reihe Lexicographica, Series Maior, danken wir den Herausgebern der Reihe.

Heidelberg/Bonn, Oktober 2000

Ingrid Lemberg
Bernhard Schröder
Angelika Storrer

I Grundlagen

Gregor Büchel, Bernhard Schröder

Verfahren und Techniken in der computergestützten Lexikographie

Jedes Vorhaben, bei dem die Erfassung und Verarbeitung von Textdaten eine wichtige Rolle spielt, wird vor den Problemen stehen,

- in welcher Weise die Textdaten erfasst werden sollen,
- in welcher Form die Textdaten gespeichert werden sollen und
- mit welchen Hilfsmitteln das Textmaterial für die weitere Arbeit oder für eine Veröffentlichung des Materials erschlossen werden soll.

Dabei steht nicht mehr die Bewältigung großer Textmengen im Vordergrund der Überlegungen – textuelle Information, die dem Umfang ganzer Bücherregale entspricht, lässt sich heute mit jedem handelsüblichen PC und preisgünstig zu erwerbender oder gar kostenfreier Software verwalten –, sondern der Umgang mit *strukturierter* textueller Information. Was ist damit gemeint?

Für die allermeisten Formen der digitalen Repräsentation von Texten bildet die lineare Abfolge von Basiseinheiten des Textes – i.d.R. Zeichen – die Grundstruktur der Repräsentation. Einfache Textdateien bestehen schlicht aus einer Abfolge von Zeichenrepräsentationen (Zeichencodes). Neben der linearen Strukturierung, die sich in dieser Art der Kodierung wiederspiegelt, weisen Texte immer auch vielfältige andere explizite und implizite Strukturen auf, die hier pauschal als *nichtlineare* Strukturen bezeichnet werden sollen. Zu den nichtlinearen Strukturen in diesem weiten Sinne gehören:

- Formatierungen und Layout von Texten und Textteilen, Hervorhebung von Textteilen durch den Autor
- Metainformation zum Text, z.B. bibliographische Angaben oder andere Angaben zur Textherkunft, Angaben zur Digitalisierung, Angaben zu Korrekturen usw.,

- Textgliederung,
- interpretierende Angaben zum Text oder zu Textteilen (linguistischer, philologischer, historischer, soziologischer usw. Art),
- Verknüpfungen zwischen Texten oder Textteilen (darunter können explizit vom Autor vorgenommene Verknüpfungen wie Fußnoten oder explizite Querverweise sein oder von einem Interpreten vorgenomme Verknüpfungen) und
- Verknüpfungen mit nicht-textuellen Medien, wie statischen oder bewegten Bildern, Klangdaten, interaktiven Anwendungen, Internet-Ressourcen.

Die Verknüpfungsstrukturen werden auch als die in einem engeren Sinne nichtlinearen Strukturen bezeichnet.

Die explizite Kodierung der genannten nichtlinearen Strukturen wird i.d.R. von modernen Textverarbeitungsprogrammen – zumindest in Ansätzen – unterstützt. Die nichtlineare Kodierungsfähigkeit eines Systems wird allerdings oft bei retrospektiven Projekten zur Digitalisierung historischer Dokumente vor besondere Herausforderungen gestellt: Bei der retrospektiven Digitalisierung hat man es nicht selten mit unterschiedlichen Textzeugen und -versionen, Graden an Sicherheit bei der Entschlüsselung, Textlücken, Textfragmenten mit unklarer Anordnung usw. zu tun. Einerseits erfordert die Abbildung dieser Strukturen auf die von Textverarbeitungsprogrammen gebotenen Möglichkeiten viel Phantasie, und entsprechend lässt der Bedienungskomfort der Textverarbeitungsprogramme bei derartigen Anwendungen zu wünschen übrig. Andererseits wirft der Versuch, die erfassten Daten in unterschiedlichen Medien (Druck, WWW, CD-ROM) oder nur selektiv zu publizieren, gerade bei den nichtlinearen Strukturen nur schwer lösbare Probleme auf. Und schließlich scheitert man zumeist an der Aufgabe, nichtlineare Strukturen in andere Anwendungen oder Kodierungsformate zu exportieren.

Man wird aber auch schon bei linearen Strukturen auf Probleme stoßen, wenn man vor der Aufgabe steht, Zeichen zu kodieren, die in den üblichen Zeichensätzen des Betriebssystems nicht verfügbar sind. Die Darstellbarkeit solcher Zeichen auf dem Bildschirm ist nicht immer auch eine Garantie für die Druckbarkeit von Zeichen oder die Darstellbarkeit der Zeichen auf WWW-Seiten. Und den Export in andere Anwendungen überleben solche aus anderen Codes entliehenen Zeichen selten.

Lexikographische Projekte, wie sie in diesem Band geschildert werden, sind mit den aufgeführten Textkodierungs- und -strukturierungsproblemen in besonderer Weise konfrontiert: Zum einen stellen Lexikoneinträge stark nichtlinear strukturierte Texte dar; im gedruckten Wörterbuch wird durch typographische Mittel eine oft sehr filigrane Mikrostruktur erschließbar gemacht. Die implizite Kodierung der Mikrostruktur durch die Typographie reicht aber normalerweise nicht aus, wenn aus dieser Textgrundlage Wörterbuchversionen für das elektronische Medium abgeleitet werden sollen. Der Hauptgrund dafür ist die Polyfunktionalität typographischer Merkmale. Kursivierung kann in einem Wörterbuch unterschiedliche Funktionen haben; im für die digitale Verarbeitung günstigeren Fall, lässt sich die Funktion rein formal durch den Kontext ermitteln, im ungünstigeren Fall kann die Ermittlung der Funktion nur durch eine Interpretationsleistung der Wörterbuchbenutzerin oder des Wörterbuchbenutzers geschehen. Zur Implementierung unterschiedlicher Sichten auf Wörterbuchartikel in einem elektronischen Wörterbuch, die die Ausblendung von Teilen der Mikrostruktur gestatten, ist eine eindeutige explizite Markierung der Mikrostruktur vonnöten.

Zum zweiten weisen Wörterbücher in starkem Maße Verknüpfungsstrukturen auf. Es bestehen eine Vielzahl von Verweisen innerhalb der Artikel, zwischen Artikeln und von

Wörterbuchartikeln auf Belege. Zunehmend werden auch Verweise auf nicht-textuelle Medien integriert.

Und schließlich arbeiten alle Wörterbuchprojekte mit Korpora retrospektiv digitalisierter Texte. Neben den bereits erwähnten Kodierungsproblemen bei diesen Korpora, stellt sich spätestens bei der elektronischen Publikation eines Wörterbuchs die Frage, ob die Korpora oder ausgewählte Belegstellen zusammen mit dem Wörterbuch publiziert werden sollen und welche Verweisstrukturen zwischen dem Wörterbuch und den Korpora explizit gemacht werden sollen.

Um die angesprochenen Strukturierungs- und Kodierungsfragen zu beantworten und die damit verbundenen Erfassungs- und Publikationsprobleme zu lösen, müssen eine Vielzahl inhaltlicher und technischer Probleme gelöst werden; ein weites Spektrum an Fragen und an prinzipiellen und exemplarischen Lösungen wird in diesem Band angesprochen. In diesem Beitrag geht es um die Grundlagen der Kodierung strukturierter textueller Information mit SGML und XML und der Verwaltung dieser Strukturen in Datenbanksystemen.

1 Kodierung strukturierter Texte mit SGML und XML

In diesem Abschnitt wird diskutiert, welche Anforderungen an ein adäquates Kodierungssystem für strukturierte textuelle Information zu stellen sind. SGML und XML werden als Kodierungssysteme vor dem Hintergrund dieser Forderungen bewertet.

1.1 Anforderungen an ein Kodierungssystem

Je nach Projekt wird man mit unterschiedlicher Gewichtung fordern, dass das gewählte Kodierungssystem hinreichend mächtig ist, die zu kodierenden Strukturen eindeutig repräsentiert, gut maschinell zu interpretieren ist, dass die Kodierung nachhaltigen Bestand hat und auf andere Computersysteme portierbar ist, ferner dass benutzerfreundliche Werkzeuge zur Arbeit mit dem Kodierungssystem bereitstehen. Die folgenden Abschnitte sollen die Anforderungen verdeutlichen.

1.1.1 Mächtigkeit

Ein Kodierungssystem muss zunächst für die explizit zu kodierende Information auch Kodierungsmechanismen bereitstellen. Ein Kodierungssystem, das zur Kodierung rein linearer Strukturen dient, ist beispielsweise ungeeignet zur Kodierung von Verweisstrukturen. Das schließt nicht aus, dass man mit zusätzlichen Konventionen, z.B. durch Reservierung bestimmter Zeichensequenzen für die Kodierung von Verweisstrukturen, die Mächtigkeit eines Kodierungssystems erweitern kann. Die Etablierung zusätzlicher Konventionen bedeutet aber nichts anderes als die Implementierung eines neuen mächtigeren Kodiersystems auf der Basis eines weniger mächtigen. Geschieht diese Erweiterung *ad hoc*, so läuft man ein nicht zu unterschätzendes Risiko, dass die übrigen Anforderungen bezüglich des erweiterten Kodierungssystems nicht mehr erfüllt sind.

1.1.2 Eindeutigkeit

Gerade bei *ad hoc* entworfenen Kodierungssystemen kann es leicht passieren, dass die gewählten Kodierungen nicht eindeutig sind. Im trivialen Fall, der bei Texten, die in den Anfangszeiten der linguistischen Datenverarbeitung erfasst wurden, nicht selten anzutreffen ist, wurden in den damaligen Codes nicht verfügbare Zeichen durch Zeichensequenzen ersetzt, die auch sonst im Text anzutreffen waren, so dass eine automatische Erkennung dieser Zeichen im Code nicht immer möglich ist. Ob *zuende* in einem Kodierungssystem, bei dem alle Umlaute als nicht-umgelauteter Vokal + *e* kodiert werden, als zwei- oder dreisilbig zu interpretieren ist, ist nur unter Zuhilfenahme sprachlichen Wissens entscheidbar. Weniger triviale Fälle sind aber auch in moderneren Kodierungssystemen anzutreffen: Bei Kodierungen, die von reservierten Codes eingeleitet werden, ist nicht immer klar, wo sie enden. Kodierungen, die nichtlineare strukturelle Information zu Textteilen beinhalten, sind oft nicht eindeutig, wenn es um die Frage geht, wo diese Textteile beginnen und wo sie enden; beendet eine neue Kodierung desselben Typs immer den Gültigkeitsbereich der vorangehenden? Diese Frage wird nicht von jedem Kodierungssystem eindeutig beantwortet.

1.1.3 Interpretierbarkeit

Auch die formale Eindeutigkeit von Kodierungen muss nicht immer eine günstige maschinelle Interpretierbarkeit bedeuten. Konventionen können von den Bearbeiterinnen und Bearbeitern aufgrund des sprachlichen und außersprachlichen Wissens, über das sie verfügen, sehr leicht zu erlernen und zu interpretieren sein, ohne dass diese Konventionen deshalb auch leicht programmtechnisch umzusetzen sind. Dies ist immer dann der Fall, wenn es sehr komplexe Abhängigkeiten verschiedener Kodierungen voneinander gibt, eine Kodierung A beispielsweise im Kontext einer Kodierung B auf eine Weise, im Kontext einer Kodierung C aber auf eine ganz andere Weise verwendet wird.

1.1.4 Nachhaltigkeit

In diesem Band wird von Langzeitprojekten berichtet, die sich über viele Jahrzehnte erstrecken. Es ist selbstverständlich, dass die in diesen Projekten akkumulierten elektronischen Textressourcen für die Projektdauer und die elektronisch publizierten Ergebnisse möglichst lange darüber hinaus Bestand haben sollen. Es ist nicht allzu verwegen, zu prognostizieren, dass die Anwendungsprogramme, die heute die Benutzerschnittstellen zu Wörterbüchern auf CD-ROM oder im World-Wide Web (WWW) bereitstellen, nicht über Jahrzehnte hinweg benutzbar sein werden, sofern sie auf avancierteren Techniken beruhen als bloßen auf einem Server bereitliegenden HTML-Seiten. Die Analogie zur bisherigen Entwicklung berechtigt zu dieser Annahme. Nicht wenige nur ein Jahrzehnt alte Anwendungsprogramme sind auf heutigen PCs nur dann noch nutzbar, wenn neben aktuellen Betriebssystemsversionen auch alte DOS-Versionen installiert werden. Und manches alte Programm verweigert selbst dann die Zusammenarbeit mit neuer Hardware.

Hard- und Softwarehersteller haben nur ein geringes wirtschaftliches Interesse an der Gewährleistung einer nachhaltigen Abwärtskompatibilität neuer Systeme, also der Möglichkeit, ältere Software und Daten auf neueren Systemen zu nutzen. Die zusätzliche Forde-

rung weitreichender Abwärtskompatibilität verteuert nämlich die Hard- und Softwareentwicklung und verringert den Anreiz, die neuesten Softwareprodukte zu erwerben.

Noch bedrohlicher als die mangelnde Abwärtskompatibilität der Weiterentwicklungen eines bestimmten Systems kann für Anwendungssoftware das vollständige Verschwinden bestimmter Hardware- und Betriebssystemsplattformen vom Markt sein. Es ist schwer, für manche Anwendungssoftware, die auf in den 80er Jahren noch weit verbreiteten Großrechnern lief, heute noch geeignete Rechner zu finden.

Und schließlich entspricht veraltete Anwendungssoftware auch nicht mehr den heutigen Benutzungsgewohnheiten, ihre Bedienung ist oft nur umständlich zu erlernen, sie wird von den potentiellen Benutzerinnen und Benutzern nicht selten abgelehnt.

Kompatibilitätsprobleme stellen sich aber nicht nur bei Anwendungsprogrammen, sondern auch bei den zugehörigen Dateiformaten. Aktuelle Programmversionen von Textverarbeitungsprogrammen haben nicht selten Schwierigkeiten bei der Wiederherstellung aller Formatierungsinformationen aus Dateien, die mit Vorgängerversionen vor etwa fünf Jahren hergestellt wurden, an weiteren nichtlinearen Strukturen in diesen Texten scheitern sie oft genug. Eine Unterstützung von Formaten nicht mehr auf dem Markt befindlicher Textverarbeitungssysteme ist selten anzutreffen.

Vor diesem Hintergrund muss ein Kodierungsschema auch daraufhin untersucht werden, ob man berechtigterweise prognostizieren kann, dass es auch in einigen Jahrzehnten noch ohne großen Eigenaufwand zur Herstellung von Kompatibilität und ohne Informationsverlust zu verwenden ist.

1.1.5 Portierbarkeit

Will man sich in einem Projekt nicht ein für alle Mal auf bestimmte Hard- und Softwareprodukte festlegen, so stellt sich die Frage der Portierbarkeit der Daten. Können die Daten ohne großen Aufwand und ohne Informationsverlust auch auf anderen Systemen verwendet werden? Zwar ist Portierbarkeit meist auch in den Überlegungen zur Nachhaltigkeit eingeschlossen, aber es kommt auch vor, dass sich Nachhaltigkeitsüberlegungen auf die Annahme stützen, dass sich ein bestimmtes technisches System durchsetzen werde und nachhaltig verfügbar bleibe, z.B. aufgrund seiner technischen Überlegenheit oder der Marktmacht des Herstellers. So kommt es bei Projekten mit starkem Bezug zum WWW vor, dass bei der Entscheidung für ein bestimmtes Kodierungssystem den Ausschlag gibt, ob das Kodierungssystem eine Zukunft im WWW hat. Gleichzeitig kann es aber der Fall sein, dass dieses Kodierungssystemen von den konkurrierenden Browsern in sehr unterschiedlicher Weise unterstützt wird, von einzelnen Browsern also möglicherweise derzeit gar nicht unterstützt wird.

1.1.6 Softwareunterstützung

Die Arbeit mit komplexen Kodierungssystemen kann umständlich, ermüdend und fehleranfällig sein. Es ist für eine ergonomisch sinnvolle Erfassung der Daten wichtig, dass geeignete Softwarewerkzeuge, Spezialeditoren, zur Verfügung stehen. Die effiziente Auswertung der Daten kann wesentlich von der Verfügbarkeit von Recherchewerkzeugen abhängen, die sowohl die linearen als auch die nichtlinearen Textstrukturen nutzen. Und schließlich kann geeignete Publikationssoftware die Druck- oder CD-ROM-Versionen der zu veröffentlichenden Daten sehr erleichtern.

In den folgenden beiden Unterabschnitten werden SGML und XML als Formalismen, mit denen sich komplexe Kodierungssysteme definieren lassen, in Grundzügen vorgestellt, um dann zu diskutieren, inwiefern SGML- oder XML-basierte Formalismen den Anforderungen genügen.

1.2 Die Grundkonzepte von SGML und XML

SGML (Standard Generalized Markup Language) und XML (Extensible Markup Language) sind Formalismen,[1] mit denen sich Kodierungssysteme formal definieren lassen. Man kann SGML und XML als Metasprachen auffassen, mit denen spezielle Kodierungssysteme beschrieben werden. Diese speziellen Kodierungssysteme werden gerne als *Markup Languages* (Markierungs-/Annotierungssprachen) bezeichnet. Insofern als diese Bezeichnung suggeriert, dass die Aufgabe SGML- oder XML-basierter Kodierungssysteme grundsätzlich sei, Markierungs- oder Annotierungsmöglichkeiten für vorgegebene Texte zu definieren, ist sie ungenau. SGML- und XML-basierte Kodierungssysteme werden inzwischen auch zu anderen Zwecken als der Repräsentation textueller Strukturen verwendet, mit SGML und XML lässt sich die Syntax von Programmiersprachen ebenso definieren wie Datenaustauschformate für Datenbanken mit beliebigem Inhalt. Der Schwerpunkt von SGML- und XML-Anwendungen liegt jedoch nach wie vor auf der Repräsentation textueller Strukturen. Und nur darauf soll im Folgenden Bezug genommen werden.

SGML ist die ältere der beiden Metasprachen, sie wurde bereits 1986 definiert und als ISO-Norm 8879 standardisiert. Im folgenden soll zunächst überwiegend von SGML die Rede sein, aber bis auf weiter unten ausdrücklich vermerkte Unterschiede ist die Darstellung auch auf XML übertragbar.

Bei den Grundkonzepten von SGML sind die, die sich ausschließlich auf die bereitgestellten formalen Strukturen beziehen, von denen zu unterscheiden, die sich auf den intendierten Umgang mit den formalen Konzepten beziehen.

1.2.1 Die formalen Grundkonzepte

Der Aufbau eines Textes aus Teilen wird in SGML durch *Elemente* modelliert. Terminologisch soll hier zwischen Elementen als Typen von Textteilen und den konkreten Textteilen selbst als Instanzen dieser Elemente, kurz: *Elementinstanzen*, unterschieden werden. Das Element *abschnitt* kann also nach dieser Sprachweise mehrfach in einem Text vorkommen, jedes Vorkommen ist aber eine neue Instanz dieses Elements. Ein Text als ganzer bildet eine Elementinstanz. Jede Elementinstanz kann Textstücke und weitere Elementinstanzen enthalten. Die in einer Elementinstanz unmittelbar enthaltenen Elementinstanzen dürfen einander nicht überlappen.

Die Elementstruktur eines Textes entspricht also einer Baumstruktur, bei der das umfassendste Element die Wurzel, die enthaltenen Elemente die Zweige und die enthaltenen

---

[1] Wenn man sich im WWW über SGML und XML informieren möchte, bietet das Electronic Text Center an der University of Virginia einen guten Ausgangspunkt gerade im Bereich akademischer Anwendungen, s. http://etext.lib.virginia.edu/standard.html. Sehr viel weiterführende Information findet man auf den XML Cover Pages unter http://www.oasis-open.org/cover/sgml-xml.html.

Textstücke die Blätter bilden. Der Beginn und das Ende eines Elementes wird i.d.R. durch ein *Tag*, eine Marke, angezeigt, das durch die Markup-Begrenzer < und > vom umgebenden Text abgegrenzt ist. Im Anfangs-Tag steht direkt hinter dem linken Markup-Begrenzer der Name des Elementes, im End-Tag steht i.d.R. zwischen den Markup-Begrenzern ein Querstrich und der Name des zu beendenden Elementes. Hier ein Teil dieses Abschnitts mit Abschnitt, Sätzen und Nominalphrasen als Elementen:

```
<abschnitt>
<satz><np>Der Aufbau <np>eines Textes</np> aus <np>Teilen</np></np>
wird in <np>SGML</np> durch <np>Elemente</np> modelliert.</satz>
<satz><np>Ein Text als ganzer</np> bildet <np>ein
Element</np>.</satz>
...
</abschnitt>
```

Elemente können in SGML nun Träger von Eigenschaften sein. Zur Kodierung von Eigenschaften werden in den Anfangs-Tags von Elementen hinter dem Elementnamen Attribute mit zugehörigen Werten notiert. Wollen wir beispielsweise im obigen Abschnitt bei den Nominalphrasen auch deren Kasus und Numerus kodieren, so kann das wie folgt aussehen:

```
<abschnitt>
<satz><np kasus="nom" numerus="sg">Der Aufbau
<np kasus="gen" numerus="sg">eines Textes</np> aus
<np kasus="dat" numerus="pl">Teilen</np></np> wird in
<np kasus="dat" numerus="sg">SGML</np> durch
<np kasus="akk" numerus="pl">Elemente</np> modelliert.</satz>
<satz><np kasus="nom" numerus="sg">Ein Text als ganzer</np> bildet
<np kasus="akk" numerus="sg">ein Element</np>.</satz>
...
</abschnitt>
```

Vor der Kodierung der Gliederungsstruktur muss natürlich festgelegt werden, was kodiert werden soll. In einer Dokumenttypdefinition (DTD) geschieht der formale Teil der Festlegung: Es wird im Wesentlichen festgelegt, welche Elemente es gibt – im Beispiel *abschnitt*, *satz* und *np* –, welche Elemente in welchen anderen enthalten sein dürfen oder müssen, ob bestimmte Elemente unmittelbar Text enthalten können oder nur wieder weitere Elemente. Diese Spezifikation bezeichnet man auch als das Inhaltsmodell eines bestimmten Dokumenttyps. In der DTD wird ferner festgelegt, welche Attribute diese Elemente tragen können. Zu den Attributen wird angegeben, welchen Typs die Werte sind.

Je nach Kodierungszweck und Kodierungsstil kann es zu einem Dokument unzählige mögliche DTDs geben. Durch eine DTD ist nur bestimmt, welche Kodierungen zulässig sind, nicht aber, wie sie verwendet werden sollen. Dies sollte in Kommentaren zur DTD oder in einem separaten Kodierungshandbuch möglichst detailliert und praxisnah niedergelegt werden.

Aus dem hierarchischen SGML-Strukturierungsmodell für Dokumente ergibt sich eine nicht unwesentliche Einschränkung: Die gleichzeitige Strukturierung eines Dokumentes in einander überlappende Einheiten ist ausgeschlossen. Ein Dokument, das durch SGML-Elemente, die Abschnitte umfassen, gegliedert wird, kann nicht gleichzeitig durch SGML-Elemente gegliedert werden, die Seiten umfassen, sofern mehrere Abschnitte auf einer Seite vorkommen können, gleichzeitig aber auch ein Seitenumbruch innerhalb eines Abschnitts vorkommen kann. Denn in solchen Fällen kann weder die betreffende Seitenelementinstanz

als Teilelement der Elementinstanz des Abschnitts noch umgekehrt aufgefasst werden. Zwar sieht der SGML-Standard für solche Fälle die Möglichkeit vor, mit konkurrierenden DTDs zu arbeiten, also eine DTD bereitzuhalten, in der die Seitengliederung, nicht aber die Abschnittsgliederung berücksichtigt wird, und eine andere DTD für den umgekehrten Fall. In das Dokument werden speziell markierte Tags für beide DTDs aufgenommen. Leider gibt es jedoch kaum Werkzeuge, die die gleichzeitige Arbeit mit konkurrierenden DTDs unterstützen würden.

Als Ausweg aus dem Dilemma bietet sich an, bei einer der beiden konkurrierenden Gliederungsarten darauf zu verzichten, mit SGML-Elementen die Einheiten dieser Gliederungsebene zu *umschließen,* sondern stattdessen mithilfe leerer, d.h. weder Text noch weitere Elemente umfassender, Elemente nur den Anfang oder nur das Ende oder beides zu markieren. Damit verzichtet man allerdings auf Möglichkeiten, die zulässigen Strukturen dieser Gliederungsart genauer durch eine DTD zu beschreiben und die Gliederungseinheiten in bequemer Weise zu adressieren.[2]

Die bisher behandelten Mittel zur Kodierung von Textstrukturen erlauben, über den Text ein baumartige Gliederungsstruktur zu legen. Zwischen den Elementinstanzen sowie den Elementinstanzen und dem Text bestehen zwei Basisrelationen: Eine Elementinstanz oder ein Textstück kann in einer (anderen) Elementinstanz unmittelbar enthalten sein. In einer Elementinstanz unmittelbar enthaltene Elementinstanzen oder Textstücke stehen in einer linearen Abfolgerelation zueinander: Die unmittelbar enthaltenen Elementinstanzen oder Textstücke stehen eben immer rechts oder links von anderen Elementinstanzen oder Textstücken.

Zu der baumartigen Gliederungsstruktur tritt noch die Möglichkeit, Elementinstanzen mit Attributen und zugehörigen Werten zu versehen. Der Mechanismus der Attribuierung von Elementinstanzen wird in SGML dazu genutzt, weitere beliebige Relationen zwischen den Elementinstanzen eines Dokumentes oder auch zu Elementinstanzen anderer Dokumente herzustellen.

Um beispielsweise einen Verweis von einem Wörterbuchartikel auf einen anderen zu kodieren, kann man ersteren eindeutig benennen, indem man der zu diesem Artikel gehörenden Elementinstanz ein Attribut mit einem Identifikator, einer zur Bezeichnung keiner anderen Elementinstanz verwendeten Zeichenkette zuordnet.

```
<artikel id=hund><lemma>Hund</lemma>
...
</artikel>
...
<artikel id=katze><lemma>Katze</lemma>
siehe auch <verweis idref=hund>Artikel Hund</verweis>
...
</artikel>
```

Die Elementinstanz von *verweis* im Artikel Katze verweist durch den Wert des Attributs *idref* also auf die Elementinstanz von *artikel* mit dem Identifikator *hund.* Führt man zusätzlich noch Konventionen zur eindeutigen Identifikation anderer Dokumente ein, wie es im WWW durch die Uniform Resource Identifiers (URIs) realisiert ist, so können Verweise auf beliebige Elementinstanzen in beliebigen SGML-Dokumenten realisiert werden. Mit-

---

[2] Vgl. hierzu auch den Beitrag von Thomas Burch und Johannes Fournier in diesem Band. Die TEI befasst sich im Kapitel 31 der Guidelines mit diesem Thema.

hilfe der für XML definierten XML Pointer Language (XPointer) kann auch auf Elementinstanzen ohne Identifikatoren referiert werden, indem man einen Pfad durch den Strukturbaum des Dokumentes beschreibt, der zu der gemeinten Elementinstanz führt. Das Verweiskonzept kann auch derart erweitert werden, dass Verweise mit mehreren Zielen zugelassen werden. Auf diese Weise kann von einer Elementinstanz auf mehrere andere verwiesen werden. Die XML Linking Language (XLink) definiert für XML erweiterte Verweiskonzepte, mit deren Hilfe beliebige Relationen zwischen Elementinstanzen von Dokumenten kodierbar sind und Verweis- und Gliederungsstruktur relativ unabhängig voneinander verwaltet werden können.

Nur erwähnt sei hier ein weiteres Sprachelement von SGML: *Entitäten* sind Code-Sequenzen, die einzelne Zeichen oder Zeichenfolgen vertreten. Damit lassen sich insbesondere Bezeichner für Zeichen bilden, die im aktuell benutzten Code nicht vorhanden sind oder die für die Portabilität der Daten hinderlich wären. Dadurch wird die Integration beliebiger Zeichensysteme in SGML möglich.

1.2.2 Das pragmatische Grundkonzept

SGML ist intendiert als ein Kodierungssystem für ein *deskriptives* Markup. SGML-Tags sollen nicht als Befehle für bestimmte Formatierungsoperationen verstanden werden, sondern mithilfe von SGML-Elementen sollen die inhaltliche Gliederung eines Dokumentes markiert werden.

Wie das inhaltlich durch SGML-Tags gegliederte Dokument für unterschiedliche Ausgabemedien zu formatieren ist, wird unabhängig vom SGML-Markup bestimmt. Ob eine Überschrift fett, größer, zentriert erscheint, soll nicht durch SGML-Tags im Dokument kodiert werden, sondern durch eine gesonderte Spezifikation, die Elementen in Abhängigkeit von ihrer Umgebung und Attributen, Formatierungseigenschaften zuweist. Diese Zuweisung geschieht außerhalb des eigentlichen Dokuments. Die so erreichte Modularisierung ermöglicht es, ohne Veränderung des Dokuments, an unterschiedliche Ausgabemedien optimal angepasste Darstellungsformen des Dokuments herzustellen. DSSSL (Document Style Semantics and Specification Language) ist der zu SGML gehörige Standard zur Spezifikation von Darstellungsformen eines Dokuments, XSL (Extensible Style Language) der entsprechende – z.T. noch in Entwicklung befindliche – Standard für XML, der viele Prinzipien von DSSSL aufgreift.

Dass das pragmatische Grundkonzept des deskriptiven Markup bei der Entwicklung von SGML leitend war, schließt jedoch nicht aus, dass SGML auch für ein darstellungsbezogenes Markup verwendet werden kann. HTML (Hypertext Markup Language), die Sprache des WWW, ist eine SGML-basierte Markup-Sprache, die ursprünglich einige wenige Gliederungsmöglichkeiten für Dokumente vorsah. Die Art der Darstellung der Dokumente blieb weitgehend den Betrachterprogrammen überlassen. Der zunehmende Bedarf an graphischen Gestaltungsmöglichkeiten und an kontrolliertem Layout hat zur Aufnahme von Elementen und Attributen zur Steuerung des Layout geführt. HTML wurde mehr und mehr zur Layout-Programmiersprache für das WWW. Erst die Einführung von Cascading Style Sheets (CSSs) zur Bestimmung von Layout-Eigenschaften von WWW-Seiten macht es möglich, den Anteil von nicht-deskriptivem Markup in den HTML-Dokumenten selbst zu verringern, ohne auf ansprechendes Layout zu verzichten.

Bei der retrospektiven Digitalisierung allerdings wird man des öfteren davon Gebrauch machen müssen, dass sich mit SGML prinzipiell auch graphische Eigenschaften von Texten

beschreiben lassen; überall da, wo unklar ist, wie bestimmte graphische Merkmale (z.B. Wechsel des Schrifttyps, der Schriftgröße oder -farbe) zu interpretieren sind, wird man sich bei der Kodierung auf die (typo-)graphischen Eigenschaften zurückziehen müssen.

1.3 SGML vs. XML

Die Grundkonzepte von SGML sind lückenlos auch auf XML übertragbar. Neben einigen rein notationellen Unterschieden schließt XML insbesondere Merkmale von SGML aus, die die automatische Analyse von SGML-Dokumenten erschweren, dazu gehören beispielsweise Möglichkeiten, Start- oder End-Tags, die aus dem Zusammenhang zu erschließen sind, nach entsprechender Deklaration in der DTD wegzulassen. Die Markup-Minimierungsmöglichkeiten von SGML haben zur Folge, dass viele SGML-Dokumente ohne eine DTD gar nicht eindeutig analysiert werden können, dass also nicht immer eindeutig zu ermitteln ist, wie weit sich eine Elementinstanz erstreckt. Fallen die Minimierungsmöglichkeiten, wie in XML, weg, müssen also alle Elementinstanzen explizit durch ein Start-Tag geöffnet und durch ein End-Tag geschlossen werden; so kann man auch ohne eine DTD überprüfen, ob zumindest die Verschachtelung der Elementinstanzen korrekt ist und ob die verwendeten Entitäten bekannt sind. Fällt die Überprüfung positiv aus, so hat man es mit einem *wohlgeformten* Dokument zu tun. Entspricht die Struktur des Dokuments darüber hinaus noch der zugehörigen DTD, so verfügt man über ein *gültiges* Dokument. Die zwei Stufen formaler Richtigkeit erweisen sich in der praktischen Arbeit als oft sehr nützlich. Nicht immer vollzieht sich die DTD-Entwicklung und die Entdeckung neuer Erfordernisse bei der Kodierungsarbeit synchron. In solchen Fällen kann eine Arbeit mit bezüglich der aktuellen DTD ungültigen Dokumenten vonnöten sein, nichtsdestotrotz bleibt deren Wohlgeformtheit überprüfbar.

1.4 SGML und XML als Antwort auf die Anforderungen?

Die in Abschnitt 1.1 genannten Anforderungen lassen sich mit SGML bzw. XML weitgehend erfüllen.

1.4.1 Mächtigkeit

Durch die große Flexibilität bei der Definition unterschiedlichster Dokumenttypen können SGML und XML als geeignete Werkzeuge für die meisten Kodierungsaufgaben gesehen werden. Mithilfe der erweiterten Verweistechniken lassen sich beliebige Relationen zwischen Dokumententeilen kodieren. Einander überlappende Dokumentteile zwingen jedoch häufig zu intuitiv weniger eingängigen Kodierungsmechanismen. SGML und XML sind insgesamt aber mächtig genug, prinzipiell jede Datenstruktur kodieren zu können.

1.4.2 Eindeutigkeit

Mithilfe von SGML oder XML definierte Kodierungssysteme sind formal eindeutig. Eine Verwechslung von Textstücken oder Markup oder Unklarheiten, auf welche Textteile sich ein bestimmtes Markup erstreckt, sind bei geeigneter DTD-Konstruktion und gültigen

SGML-Dokumenten oder wohlgeformten XML-Dokumenten ausgeschlossen. Dass Kodierungen auch von allen Benutzerinnen und Benutzern eindeutig und hinreichend klar verstanden werden, kann natürlich auf formalem Weg nicht sichergestellt werden. Von der Text Encoding Initiative (TEI) werden jedoch in den TEI Guidelines für zahlreiche Kodierungsprobleme detaillierte Empfehlungen für die Verwendung von SGML-Markup ausgesprochen. Diese Empfehlungen können als eine Grundlage für ein detailliertes Kodierungshandbuch dienen, das auch die pragmatisch-inhaltlichen Fragen der Markup-Verwendung klärt.

1.4.3 Interpretierbarkeit

Durch SGML- bzw. XML-DTDs werden kontextfreie Grammatiken spezifiziert. SGML- oder XML-Dokumente stellen an automatische Analyseprogramme damit ähnliche Anforderungen wie der Code von Programmiersprachen. XML bedeutet gegenüber SGML eine deutliche Reduktion des Analyseaufwands.

1.4.4 Nachhaltigkeit

Jede Prognose über zukünftige technische Entwicklungen ist mit Unsicherheit behaftet. Die wachsende Bedeutung von XML im Internet-Bereich[3] und als Grundlage von Industriestandards in Wachstumsbereichen[4] berechtigt zu der Annahme, dass XML auf Jahre hin ein zentraler Standard zur Textkodierung sein wird. Verlage und Redaktionen setzen zunehmend mit langfristigen Investitionen auf den Einsatz von SGML oder XML bei der medienneutralen Texterfassung. Aber Nachhaltigkeit wird nicht zuletzt auch von den Eigenschaften gesichert, die ein hohes Maß an Portierbarkeit garantieren.

1.4.5 Portierbarkeit

SGML- und XML-Dateien sind reine Textdateien, d.h. die enthaltenen Zeichen kann man mit jedem einfachen Texteditor betrachten. Im Allgemeinen beschränkt man sich auf diejenigen Zeichen des ASCII-Zeichensatzes, die auf verschiedenen Systemen gleich interpretiert werden. Damit stellen SGML- und XML-Dateien an Editier- und Betrachterprogramme minimale Anforderungen, sofern keine SGML- oder XML-spezifischen Editier- oder Formatierfunktionen erwartet werden.

1.4.6 Softwareunterstützung

Der Softwaremarkt stellt Spezialwerkzeuge zur Arbeit mit SGML und zunehmend auch mit XML für die unterschiedlichsten Aufgaben zur Verfügung.[5] Die XML-Tauglichkeit neuer

---

[3] Über aktuelle Standardisierungsbemühungen im WWW-Bereich kann man sich auf den WWW-Seiten des World Wide Web Consortium (W3C) informieren: http://www.w3c.org .

[4] WML, die Beschreibungssprache für Internetseiten, die auf Mobiltelefonen angezeigt werden können, ist beispielsweise ein XML-basierter Standard, vgl. http://www.wapforum.org .

[5] Einen sehr gut aufbereiteten Überblick über Software zur Arbeit mit SGML und XML gibt Steve Pepper in seinem Whirlwind Guide unter http://www.infotek.no/sgmltool/guide.htm .

WWW-Browser dürfte stark zu Popularisierung von XML-Werkzeugen beitragen. Nicht alle Spezialwerkzeuge unterstützen den SGML- oder XML-Standard in allen Einzelheiten. Und nicht jede Funktionalität, die man sich für die Arbeit mit diesen Kodierungssystemen wünscht, ist in einer vorkonfektionierten Softwarelösung erhältlich. Spezielle Anwendungen werden deshalb auf die Entwicklung spezialisierter Lösungen setzen. Im Zentrum dieser Lösungen werden meist Datenbanksysteme stehen, in denen die SGML- oder XML-kodierten Daten verwaltet werden. Darauf gehen die folgenden Abschnitte ein.

2 Datenbanksysteme zur Verwaltung strukturierter Textdaten

2.1 Datenbanksysteme: Allgemeiner Aufbau

Unter dem Begriff *Datenbank (DB)* versteht man eine strukturierte Sammlung von Daten, welche für eine Reihe von unter Umständen unterschiedlichen Anwendungssystemen ein Modell eines in der Regel kleinen Teiles der von Menschen wahrgenommenen Welt repräsentiert (Heuer/Saake [1997], Vossen [1994]).

Das *Datenbank-Management-System (DBMS)* bezeichnet die notwendige Sammlung von Software, mit der unabhängig von Anwendungssystemen die lesenden und schreibenden Zugriffe (Einfügen, Ändern, Löschen) auf eine Datenbank durchgeführt und verwaltet werden. Wenn im weiteren Text von „Datenbanksoftware" oder von „Software des Datenbanksystems" gesprochen wird, sollen damit Synonyme zum Wort Datenbank-Management-System gemeint sein.

Der Begriff *Datenbanksystem (DBS)* beschreibt den Verbund einer nicht leeren Menge von Datenbanken mit einem Datenbank-Management-System.

Datenbanksystem	=	Datenbank	+	Datenbank-Management-System

Bei einem Datenbanksystem handelt es sich um eine für Anwendungssysteme weiterentwickelte Spezialisierung der Grundfunktionalität des Dateisystems eines Betriebssystems. Beim Dateisystem eines Betriebssystems (sequentielles Dateisystem) können mehrere Anwender nur unter erhöhtem Programmieraufwand für ein Anwendungssystem (z.B. durch Programmierung des wechselseitigen Ausschlusses von Schreiboperationen auf einen Datenbestand) in einem gleichen Zeitintervall lesend und schreibend auf den gleichen Datenbestand zugreifen. Sie benötigen weiterhin Werkzeuge für elementare Operationen auf Dateien: Editieren (nach Möglichkeit mit strukturierten Editierhilfen), Suchen, Sortieren. Diese Werkzeuge sind in der Regel mit dem Dateisystem alleine nicht gegeben. Die Folge ist ein in der Regel erhöhter Programmier-(bzw. Installations-)aufwand und ein eher ineffizient zeitlich ausgenutzter Datenbestand im Mehrbenutzerbetrieb.

Ein Datenbanksystem ermöglicht es mehreren Benutzern, in einem gemeinsam genutzten Laufzeitintervall mit dem gleichen physischen Datenbestand zu arbeiten. Die Benutzer greifen nicht mehr direkt, sondern über die Datenbanksoftware auf den Datenbestand zu. Die Datenbanksoftware übernimmt die Kommunikation zwischen den Datenbankbenutzern und den Schreib-/Leseoperationen auf die Datenbestände. Sie bietet folgende Grundfunktionalität an:

1. **Persistenz (= dauerhafte Verwaltung von Datenbeständen) und Sekundärspeicherverwaltung:** Die Daten sollen während der Laufzeit *strukturtreu* vom Haupt- in den Hintergrundspeicher geschrieben werden können. Die Datenbanksoftware muss Funktionen bereitstellen, um adressengesteuert Direktzugriffe auf Datensätze im Hintergrundspeicher ausführen zu können. Hinzu kommt Software, die anwendungsunabhängig das Einfügen, Ändern und Löschen von Daten auf Hintergrundspeichern ausführt.
2. **Verwaltung eines Schemakataloges (Data Dictionary):** Das Schema ist die konzeptuelle Beschreibung der Datenbanken eines Datenbanksystems. Es enthält die Datendefinitionen für sämtliche Daten, die durch das Datenbanksystem verwaltet werden. Dieser Schemakatalog, der auch ‚Data-Dictionary' genannt wird, ermöglicht während der Laufzeit des Datenbanksystems Online-Zugriffe auf die Datendefinitionen des Systems. Die in diesem und im vorhergehenden Absatz genannte Software garantiert die Forderung der *Datenunabhängigkeit* an eine Datenbank. Diese Forderung hat das Ziel, eine in der Regel langlebige Datenbank von ständig auftretenden Änderungen der auf sie zugreifenden Anwendungssysteme abzukoppeln.
3. **Interpretation einer Anfragesprache:** Ein Datenbanksystem bietet dem Benutzer eine Anfragesprache an, mit der er, ohne die interne Speicherung der Daten in der Datenbank zu kennen, auf die Datenmengen des Datenbanksystems zugreifen kann. Der Zugriff geschieht in der Regel interaktiv.

Weitere sechs notwendige Merkmale eines DBMS sind: Sicherung der Integrität, Verwaltung von Benutzersichten, Datenschutz, Transaktionsverwaltung, Synchronisation, Recovery/Datensicherung.

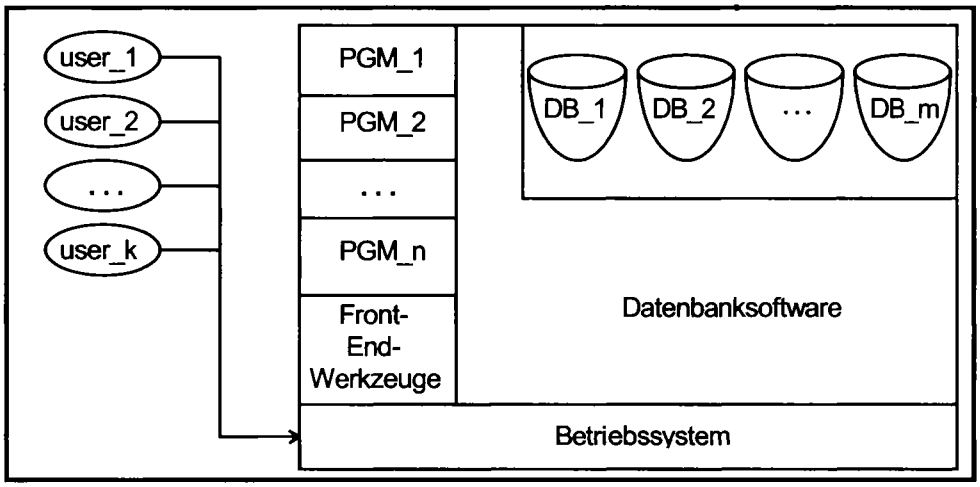


Abb. 1: Allgemeiner Aufbau eines Datenbanksystems

Legende zu Abbildung 1:

user_i ($1 \leq i \leq k$): Anwender des Datenbanksystems, die ihre Schreib- bzw. Leseanforderungen an das Datenbanksystem über einen vernetzten Rechner an den Rechner, auf dem das Datenbanksystem installiert ist, weitergeben.

PGM_i ($1 \leq i \leq n$): DB-Anwendungsprogramm, das durch Schreib- bzw. Leseanforderungen eines Anwenders gestartet wird und mittels der Datenbanksoftware auf Datenbestände des Datenbanksystems zugreift. Jedes DB-Anwendungsprogramm ist in der Regel durch folgenden Schichtenaufbau gekennzeichnet:

Benutzerschnittstelle	Melden und Empfangen von User-Daten
Algorithmische Schicht	Verfahren zur Lösung des Anwendungsproblems
DB-Zugriff	Aufruf von Funktionen der DB-Software

Abb. 2: Schichten eines DB-Anwendungsprogramms

Front-End-Werkzeuge:

- *a) Benutzerschnittstelle* zur interaktiven Verarbeitung von Kommandos der Anfragesprache: Der Benutzer kann hier unmittelbar Kommandos eingeben, die an den Kommandointerpreter weitergeleitet, analysiert und ausgeführt werden.
- *b) Dialogmaskengenerator*: Der Dialogmaskengenerator ist ein Softwaresystem zum Erzeugen von Dialogmaskenprogrammen. Dialogmasken (engl. *forms*) gestalten eine formatierte Eingabe für den lesenden und schreibenden Zugriff auf Datenbanken. Wegen der formatierten Eingabe und der damit verbundenen Möglichkeit der Integritätskontrolle sind Dialogmaskenprogramme für die Bearbeitung von Daten in Lexikonartikeln gut geeignet. Beim lesenden Zugriff ist insbesondere die Unterstützung von *Suchfunktionen* (trunkierte Suche, maskierte Suche) durch Dialogmaskenprogramme hervorzuheben, die inzwischen bei gängigen Datenbanksystemen auch in Form von WWW-Programmen implementiert sind. Die Ergebnisse einer Suchanfrage werden als sortierte Listen ausgegeben. Hierfür stellt das DBMS komfortable, für den Benutzer unaufwendige, in Hinsicht auf das Datenmodell einer Datenbank flexible *Sortierroutinen* zur Verfügung.
- *c) Berichtsgenerator*: Berichtsprogramme sind Programme, die in der Betriebsart Stapelverarbeitung des gegebenen Betriebssystems ablaufen. (Ein Desiderat der computergestützten Lexikographie sind z.B. SGML-Berichtsgeneratoren).

2.2 Datenbankentwurf und Datenbankmodelle

Das Design einer Datenbank entspricht dem Einteilen der Daten eines geplanten Anwendungssystems in verschiedene Entitätenmengen (spätere Tabellen der Datenbank) und dem Bestimmen der zwischen den Entitätenmengen (Entities) bestehenden Beziehungstypen (Relationships). Zum Design von Datenbanken hat sich das Entity-Relationship-Modell als Entwicklungsmethode bewährt. Entity-Relationship-Modelle unterstützen die Abbildung formaler Beschreibungen von Lexikonartikeln, die z.B. in SGML vorliegen, auf Tabellenstrukturen eines einzurichtenden DBS zur Verwaltung lexikographischer Daten. Graphisch werden Entity-Relationship-Modelle durch E/R-Diagramme dargestellt.

Syntaktisch bestehen E/R-Diagramme nur aus zwei Symbolen. Jeweils eines für Entities (Entitytypen) und eines für Relationships (bzw. Beziehungstypen oder Relationstypen). Die Kanten eines E/R-Diagramms können mit Quantitäten (sog. erweitertes E/R-Diagramm) und auch mit einem Durchlaufsinn eingefärbt sein. Es hat sich folgende Darstellungsweise durchgesetzt (Abbildung 3):

Bezeichnung	Symbol	Bemerkung
Entitytypen	Entity-Name	
Relationen	r_1 — Relationstyp — r_2	Die Quantitäten r_1, r_2 bezeichnen Anzahlen sich entsprechender Elemente in den Tupeln des Relationstyps R, der zwischen zwei Entitätenmengen A und B besteht. Es gibt folgende Standardquantitätsangaben: $r_1, r_2 \in \{1, n, m, c, c{\cdot}n, c{\cdot}m\}$ $c \in \{0,1\}; n, m \in \mathbf{N}$.

Abb. 3: Syntax von E/R-Diagrammen

Nachfolgend ist als Beispiel für die Informationselemente der Artikel des „Hegel-Lexikons" von Hermann Glockner (Glockner [1935]) ein Entity-Relationship-Modell angegeben (Abb. 4).

Nach Aufstellung des Entity-Relationship-Modells kann in Hinsicht auf die Merkmale des vorliegenden lexikographischen Datenmodells die Auswahl des Typs des zu benutzenden DBMS (z.B.: relational oder objektorientiert) getroffen werden. Fragen für die Entscheidungsfindung können z.B. sein:

- Sind die Daten vorwiegend als Tupel strukturiert oder sind sie hauptsächlich in Hierarchien bzw. Netzwerken angeordnet?
- Will man die Daten durch eine komplexe oder durch eine einfache Anfragesprache verwalten?
- Hat man zwischen den verschiedenen Entitätenmengen viele oder wenige Eigenschaften, die sich „vererben" lassen?
- Möchte man ein weitverbreitetes DBMS oder kann man auch mit einem weniger verbreiteten DBMS arbeiten?

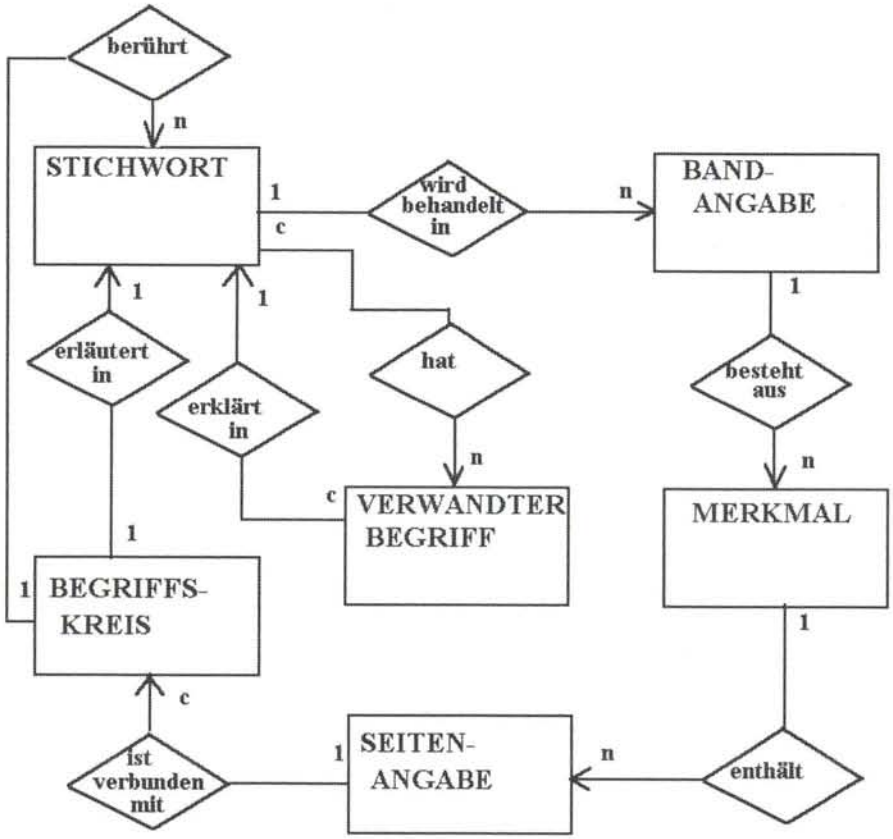


Abb. 4: E / R-Diagramm zu H. Glockner: „Hegel-Lexikon"

2.2.1 Relationale Datenbanksysteme (RDBS)

Das relationale Datenbankmodell wurde 1970 von E.F. Codd entwickelt. Die unmittelbar abbildbaren Datenstrukturen sind Tupel elementarer Datentypen, die in Tabellen zusammengefasst werden. Das relationale Datenbankmodell basiert auf den mathematischen Operationen der relationalen Algebra (Selektion, Projektion, Verbund, Differenz und Vereinigung von Relationen) (Sauer [1994], Schwinn [1992]). Die Relationen (Beziehungen) werden durch zweidimensionale Tabellen dargestellt. Dabei wird die Anzahl der Spalten fest vorgegeben. Die Zeilen der Tabelle enthalten dabei die Datenobjekte (Datensätze), die durch ihren Schlüssel unterschieden werden können. Es kommt also keine Zeile einer Tabelle zweimal vor. Die Spalten (Datenfelder) enthalten die Attributwerte, die in einer Spalte immer vom gleichen elementaren Datentyp sind. Elementare Datentypen sind z.B. INTEGER (ganze Zahl), CHAR(n) (Zeichenkette der Länge n), FLOAT (Gleitkommazahl), usw. Die Reihenfolge der Zeilen und Spalten ist im relationalen Datenmodell gleichgültig.

	SP_1	SP_2	...	SP_k	...	SP_N
Z_1				a_{1k}		
...				....		
Z_M				a_{Mk}		

Abb. 5: Allgemeiner Tabellenaufbau

Legende zu Abbildung 5:

- SP_K: Name der k-ten Spalte (=: Name des Attributes A_k)
- Datensätze sind in den Zeilen $Z_1, Z_2, \ldots, Z_M$ enthalten.
- Tupelstruktur: $Z_i=(a_{i1}, a_{i2}, \ldots, a_{iN}) \in W_1 \times W_2 \times W_3 \times \ldots \times W_N$ (W_j ist Wertebereich zur Spalte Sp_i). Alle Attributwerte des Attributes A_k sind vom gleichen elementaren Datentyp $dtyp(A_k)$.

Mit relationalen Datenbanken ist in kanonischer Weise die Datenbanksprache SQL (SQL = *S*tructured *Q*uery *L*anguage) verbunden (Knebel/Postels [1991], Miesgeld [1991], Petkovic [1995]). SQL basiert auf der Relationenalgebra und dem Tupelkalkül, das dem Tabellenaufbau relationaler Datenbanken zugrunde liegt. SQL ist eine Anfrage-, Datendefinitions- und Datenmanipulationssprache. Mit ihr können Benutzersichten, Dateiorganisationsformen und Zugriffspfade definiert werden. SQL ist eine genormte Datenbanksprache (ANSI/ISO [SQL-92], eine neue Norm [SQL3] steht zur Verabschiedung an).

In einem einfachen lexikographischen Datenbanksystem werden Daten aus Hegel-Registern und Hegel-Lexika verwaltet. Hieraus ist das Beispiel in Abbildung 6 entnommen worden.

```
                                Terminal - averoest
File  Edit  Session  Options  Help
PERFORM:   Query  Next  Previous  View  Add  Update  Remove  Table  Screen  ...
Shows the next row in the Current List.          ** 1: glocktab table**
 +-------------------------------------------------------------------------+
 |                      Glockner : Hegel - Lexikon                         |
 |-------------------------------------------------------------------------|
 |                                                                         |
 |  Schlagwort   [Abstraktion                          ]                   |
 |  Grundform    [Abstraktion                          ]                   |
 |                                                                         |
 |  Schlagwortzeile                                                        |
 |  [(abstrakt,abstrakt-konkret,Allgemeines)                           ]   |
 |  [                                                                  ]   |
 |                                                                         |
 |  Bemerkung 1 : [B          ]                                            |
 |  Bemerkung 2 : [           ]                                            |
 |                                                                         |
 |  Begriffskreise :                                                       |
 |  [Begriff                                                           ]   |
 |  [Dasein                                                            ]   |
 |  [symbolische Architektur                                           ]   |
 |  [indische Religion                                                 ]   |
 +-------------------------------------------------------------------------+
```

Abb. 6: SQL-FORM zur Verarbeitung der GLOCKNER-Tabelle

2.2.2 Objektorientierte Datenbanksysteme (OODBS)

Das Konzept *objektorientierter Datenbanken* vereinigt **zwei** Mengen wesentlicher Eigenschaften (Atkinson [1992]):
(1.) Es beinhaltet ein *objektorientiertes Datenmodell.*
(2.) Es enthält die Speicherungstechniken eines *gewöhnlichen Datenbanksystems* (Persistenz, Sekundärspeicherverwaltung, Verwaltung von Transaktionen, Recovery-Techniken, ...; vgl. Heuer [1997], Hughes [1992], Saake/Türker/Schmitt [1997]).

Auszug aus der Liste der Forderungen zu (1.):

(1.1) Möglichkeit der *direkten Modellierung komplexer Objekte* (abstrakte Datentypen).
(1.2) *Objektidentität (OID).*
(1.3) Jedes Objekt kapselt Struktur und Verhalten *(Kapselung).* Die Struktur eines Objekts wird durch Instanz-Variablen beschrieben. Das Verhalten eines Objekts wird durch Methoden beschrieben.
(1.4) Objekte mit gemeinsamer Struktur und gemeinsamen Verhalten werden in *Klassen* gruppiert. Jedes Objekt ist Instanz einer Klasse.
(1.5) Klassen können als Spezialisierung anderer Klassen definiert werden *(Vererbung).*

OODBS eignen sich zur Speicherung von komplizierten Datenstrukturen, deren Abbildung auf Tabellen eines RDBMS mit Schwierigkeiten verbunden ist. (Eine Tabelle hat eine möglicherweise große, aber immer konstante, den Anwendungsdaten beim Betrieb des DBS vorgegebene Anzahl von Spalten, die während der Laufzeit nicht geändert werden kann!) Zu solchen Datenstrukturen zählen z.B.: Baumstrukturen (Hierarchien) mit vielen Hierarchiestufen, netzwerkartige Strukturen mit unterschiedlichen Kantentypen.

Eine Klasse definiert den Aufbau und beschreibt über die Methoden das Verhalten eines Objektes. Alle Instanzen einer Klasse besitzen somit das gleiche Verhalten und die gleiche Wertestruktur. Durch Definition von Klassen vergrößert sich die Menge der zulässigen Datentypen im Data-Dictionary. Damit bietet sich die Möglichkeit, abstrakte Datentypen zu definieren (List-Typen, Mengentypen). Als Anfragesprachen für OODBS werden höhere Programmiersprachen verwendet, die Klassen als Datentypen zulassen, wie C++, JAVA und SMALLTALK. Wie in der Programmiersprache C++ sind Objekte in einem OODBS verkapselt, d.h. Methoden und Daten eines Objektes bilden eine Einheit. Der Benutzer kann nicht direkt auf die Werte eines Objektes zugreifen, sondern nur über die Methoden, die der Klasse des Objektes bekannt sind. In einem OODBS können dynamische Listen verhältnismäßig einfach als Klassen modelliert werden. OODBS werden besonders zur Realisierung komplexer, datenintensiver Anwendungen im Konstruktions- und Entwurfsbereich (CAD) eingesetzt.

Ein OODBS muss weiterhin das Konzept der Vererbung (Spezialisierung) unterstützen. Beispiel: Die Entität Student ist eine Spezialisierung der Entität Person. Man unterscheidet zwischen Superklassen (hier Person) und sogenannten Subklassen (hier Student). Die Spezialisierungsbeziehung zwischen Klassen führt zu Klassenhierarchien.

Ein Beispiel für Vererbungsbeziehungen zwischen Klassen eines lexikographischen Datenmodells kann in einem Modell einer datenbankorientierten Beschreibung ausgewählter Informationselemente des Artikelaufbaus des „DUDEN – Das große Wörterbuch der deutschen Sprache" (DUDEN [1993]) gegeben werden.[6] Folgende Tabelle (Tab. 1), die in Anlehnung an Lenders (1990) eine „Vorform" einer Artikelbeschreibung darstellt, gibt eine

[6] An dieser Stelle möchten wir Frau Dr. A. Storrer (IDS Mannheim) danken, die uns den Hinweis auf dieses schöne Beispiel einer Vererbungsstruktur gab.

hierarchisch strukturierte Liste von Entitäten (mit Attributen) an, die auf unterschiedlichen Ebenen mit grammatischen Angaben versehen sind.

Tab. 1:

ENTITÄTENMENGEN	LISTENELEMENTE	BEISPIEL	ANM.
LEMMA (Stichwort)		Abstich	(1)
GRAMMATISCHE ANGABE zum LEMMA (=: GRAM_H)		der; -[e]s, -e	
LISTE von LESARTEN	pro Lesart:		(2)
	TITEL der Lesart	das Abstechen	
	ERLÄUTERUNG der Lesart	der A. von Torf, Rasen	
	GRAMMATISCHE ANGABE zur Lesart (optional) (=: GRAM_L)	<o.Pl.>	
LISTE von KOMPOSITA	∅	∅	(3)
Etc.	...	...	...

Anmerkungen zu Tabelle 1:

(1) DUDEN (1993), S. 98.

(2) Das Beispiel weist drei weitere Lesarten auf, wobei die Lesart Nr. 3 zwei Varianten hat:

NR	TITEL der Lesart	ERLÄUTERUNG der Lesart	GRAMM. ANGABE	SACHGEBIET
2)	Art des Kantenverlaufs beim Sakko ...	stark fliehender A.	∅	Schneiderei
3a)	das Abstechen	der A. des [Roh]eisens	<o.Pl.>	Hüttenwesen
3b)	Teil eines Hochofens, ...	die Gießpfanne unter den A. rücken	∅	
4)	das Abstechen, Kontrast	dort erschien sie licht, im A. ihrer nächtlichen Umgebung (Grillparzer, Medea I)	∅	

(3) Das Beispiel enthält keine Komposita.

Im Folgenden wird für die in dieser Tabelle gegebenen Entitätenmengen ein Klassenbeziehungsgraph (Abb. 8) angegeben, der als Instrument des Entwurfes eines OODBS dienen kann. Das Aufstellen eines Klassenbeziehungsgraphen setzt in der Regel eine Entity-Relationship-Analyse voraus.

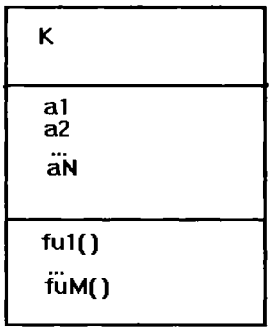


Abb. 7: Klassenbeziehungsgraph

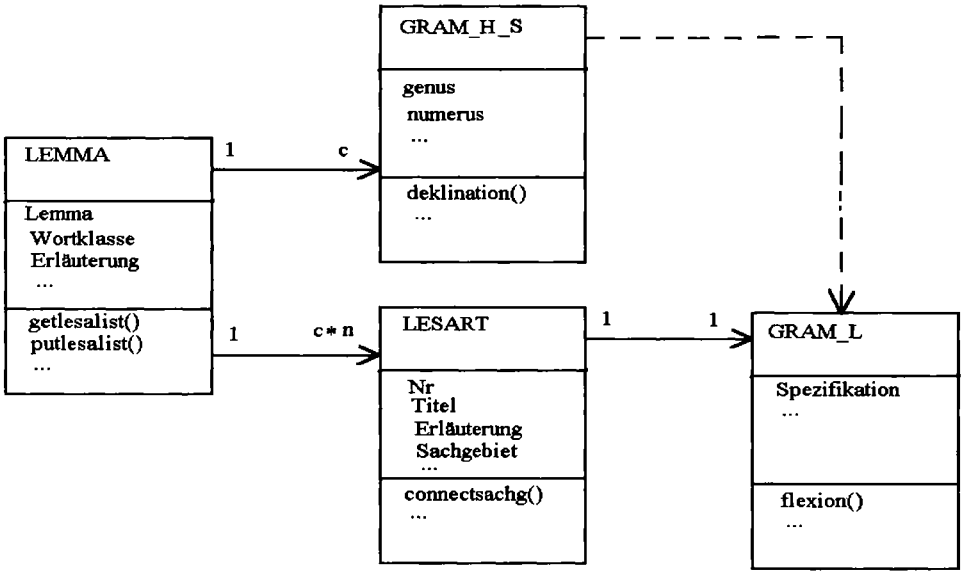
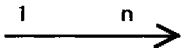
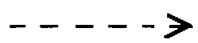


Abb. 8: Klassenbeziehungsgraph

Der (1:n)-Beziehungstyp wird als sog. gerichtete Assoziation mit dem Symbol



dargestellt (Fowler [1995], vgl. auch: Booch [1995], Rumbaugh [1993]). Die Vererbung zwischen zwei Klassen wird hier mit dem Symbol



bezeichnet.

Das Symbol einer Klasse im Klassenbeziehungsgraph (vgl. Abb.6) setzt sich zusammen aus einem Klassennamen K, einer Liste von Attributen a1,..., aN und einer Liste von Methoden fu1(),..., fuM(), die die Daten der Klasse K verarbeiten.

3 Resümee

Dieser Beitrag geht auf Entscheidungskriterien ein, die für die Wahl eines Kodierungssystems für komplexe Textstrukturen relevant sind. In den meisten Fällen dürfte die Wahl eines SGML- oder XML-basierten Kodierungssystems adäquat sein. Da SGML und XML zur Kodierung beliebiger Datenstrukturen verwendet werden können, wird die Verwendung von SGML oder XML immer prinzipiell möglich sein. Die Verwaltung SGML- oder XML-kodierter Daten kann mithilfe von Datenbanksystemen geschehen. Dazu sind die in SGML oder XML kodierten Textobjekte und Beziehungen auf die Ausdrucksmittel des gewählten Datenbanktyps abzubilden. Dies ist in jedem Fall möglich, da in SGML- und XML-kodierten Dokumenten vorhandene Strukturen auf wenige Relationen zwischen Elementinstanzen oder zwischen Text und Elementinstanzen oder zwischen Elementinstanzen, Attributen und Werten abbildbar sind. Welche Datenbank-Modellierung sich unter Gesichtspunkten des Datenzugriffs und der Datenpflege als günstig erweist, hängt stark von der Art der Dokumentstrukturierung und dem Nutzungszweck ab.

4 Literatur

Atkinson, M., et. al. (1992): The Object Oriented Database System Manifesto – In: Bancilhon, F., et al.: Building an Object-Oriented Database System – The Story of O2. – San Francisco, Ca.: Morgan-Kaufmann.

Booch, G. (1995): Objektorientierte Analyse und Design. – Bonn, Albany: Addison-Wesley.

DUDEN: DAS GROSSE WÖRTERBUCH DER DEUTSCHEN SPRACHE in acht Bänden. Hg. G. Drosdowski. Mannheim: Dudenverlag 1993.

Fowler, M. (1997): UML Distilled – Applying the Standard Object Modeling Language. – Reading (Mass.) et al.: Addison-Wesley.

Glockner, H. (1935): Hegel-Lexikon. – In Hegel, G.W.F.: Sämtliche Werke (Jubiläumsausgabe), Bd. 23–26, Stuttgart: Frommann.

Goldfarb, Charles F. (1990): The SGML Handbook. – Oxford: Clarendon Press.

– und Prescod, Paul (1998): The XML Handbook. Upper Saddle River, NJ: Prentice Hall PTR, 1998.

Graham, Ian S., Quin, Liam (1999): XML Specification Guide. New York, NY: John Wiley & Sons.

Hald, A., Nevermann, W. (1995): Datenbank-Engineering für Wirtschaftsinformatiker. – Braunschweig, Wiesbaden: Vieweg.

Heuer, A. (1997): Objektorientierte Datenbanken, Bonn: Addison-Wesley.

– und Saake G. (1997): Datenbanken – Konzepte und Sprachen. – Bonn, Albany: Int. Thomson Publishing Comp.

Hughes, J.G. (1992): Objektorientierte Datenbanken. – München, Wien: C. Hanser (in coedition with Prentice Hall).

Knebel, B., Postels, G. (1991): Einführung in Informix-SQL. – Heidelberg: Hüthig.

Lenders, W. (1990): Semantische Relationen in Wörterbuch-Einträgen – Eine Computeranalyse des DUDEN-Universalwörterbuches. – In: Schaeder, B., Rieger, B. (Hgg.): Lexikon und Lexikographie. Hildesheim: Olms.

– (Hg.) (1993): Computereinsatz in der angewandten Linguistik – Konstruktion und Weiterverarbeitung sprachlicher Korpora. - Frankfurt a.M.: Lang.

Lobin, Henning (Hg.) (1999): Text im digitalen Medium. – Opladen, Wiesbaden: Westdeutscher Verlag.

Misgeld, W. (1991): SQL – Einstieg und Anwendung. – München, Wien: Hanser Verlag.

Möhr, Wiebke, Schmidt, Ingrid (Hgg.) (1999): SGML und XML – Anwendungen und Perspektiven. – Berlin etc.: Springer.

Musciano, Chuck, Kennedy, Bill (1999): HTML – Das umfassende Referenzwerk. 2. Aufl. – Köln: O'Reilly.
Petkovic, D. (1995): Informix 6.0/7.1. – Bonn: Addison-Wesley.
Rumbaugh, J., Blaha, M., et.al. (1993): Objektorientiertes Modellieren und Entwerfen, München, London: Hanser/Prentice-Hall.
Saake G., Türker C., Schmitt I. (1997): Objektdatenbanken. – Bonn, Albany: Int. Thomson Publishing Comp.
Sauer H. (1994): Relationale Datenbanken – Theorie und Praxis. – Bonn, et al.: Addison-Wesley.
Schröder, Bernhard (1998): Pro-SGML: Ein Prolog-basiertes System zum Textretrieval. – In: Gerhard Heyer, Christian Wolff (Hgg.): Linguistik und neue Medien, 205–216. Wiesbaden, DUV.
– und Ostermann-Heimig, Jens (1998): Kants Werke als Hypertext. – In: Angelika Storrer, Bettina Harriehausen (Hgg.): Hypermedia für Lexikon und Grammatik, 233–246. Tübingen, Narr.
Schwinn, H. (1992): Relationale Datenbanksysteme. – München, Wien: C. Hanser Verlag.
Sperberg-McQueen, C. M., Burnard, Lou (1994): TEI Guidelines for Electronic Text Encoding and Interchange (P3). 1.4.2000, http://etext.lib.virginia.edu/TEI.html.
Vossen, G. (1994): Datenbanken, Datenmodelle, Zugriffssprachen. – Bonn: Addison-Wesley.
Wilde, Erik (1999): World Wide Web – Technische Grundlagen. – Berlin, etc.: Springer.

Gregor Büchel, Köln
Bernhard Schröder, Bonn

Ingrid Schmidt, Carolin Müller

Entwicklung eines lexikographischen Modells: Ein neuer Ansatz[1]

> *The data do not speak for themselves. I have been in rooms with data and listened very carefully. They never said a word.*
>
> Milford Wolpoff

Die derzeit am Markt verfügbaren elektronischen Wörterbücher nutzen das Potential ihres Mediums nur sehr begrenzt. Auffällig ist dabei, dass gute Printwörterbücher häufig in qualitativ nicht entsprechende elektronische Versionen umgesetzt werden. Betrachtet man dazu die ständig wechselnden Anforderungen der Medienlandschaft, so ergibt sich die Notwendigkeit neuer Ansätze im Umgang mit lexikographischen Inhalten. Mit diesem Ziel vor Augen gehen wir zunächst auf einzelne Aspekte der folgenden Themenbereiche ein: SGML, Multiple-Media-Publishing, TEI, Metalexikographie. Anschließend diskutieren wir auf dieser Basis einen neuen Ansatz zur Entwicklung eines lexikographischen Modells, das der Schnelllebigkeit der Medien Langlebigkeit entgegensetzt und mit dem gerade deshalb flexibel auf Marktanforderungen reagiert werden kann.

1 Die Grenzen meiner Vorstellungen bedeuten die Grenzen meiner Möglichkeiten

Auf dem Wörterbuchmarkt geht der Trend heute immer mehr dahin, Printwörterbüchern elektronische Versionen beizugeben, oder Wörterbuchprojekte eigens für das elektronische

[1] Für die anregenden Diskussionen danken wir Herbert Ernst Wiegand, Michael Beißwenger, Roman Halstenberg, Boris Körkel, Andrea Martiné und Daniel Strigel sowie Angelika Storrer für wertvolle Hinweise.

Medium zu realisieren. Dabei wird oft der Mehrwert dieser elektronischen Versionen verkündet.

Betrachtet man diese Produkte jedoch genauer, mag man sich fragen, worin dieser Mehrwert eigentlich liegen soll; allein die Publikation im elektronischen Medium ist dafür nicht ausreichend. Allzu oft handelt es sich dabei um eine Widerspiegelung der Printwörterbücher auf „reduziertem Niveau".[2] Widerspiegelung in dem Sinne, dass die Darstellungsweise der Wörterbuchartikel aus der Buchausgabe übernommen wird und bestenfalls noch Verdichtungen aufgelöst und das Layout bildschirmgerecht aufbereitet werden. Manchmal wird das neue Medium dazu durch multimediale Anreicherungen genutzt.[3] Damit gestaltet sich die Benutzung des elektronischen Wörterbuchs letzten Endes aber nicht grundsätzlich anders als die des Printwörterbuchs. Anders sind meist nur die äußeren Zugriffsmöglichkeiten auf die Wörterbuchartikel. Es kann z.B. über ein Suchfeld direkt auf ein bestimmtes Lemma zugegriffen oder eine Volltextsuche über die gesamte Artikelstrecke gestartet werden, statt wie im Buch nur über die alphabetische Anordnung zum Wörterbuchartikel zu kommen. Damit reicht aber die Nachschlagehandlung, ebenso wie beim gedruckten Wörterbuch, lediglich bis zum Artikelanfang, der Artikel selbst muss gelesen werden; dieser Lesevorgang wird höchstens durch Suchzonen[4] erleichtert. Der Bildschirm ist jedoch zum Lesen schlecht geeignet. Das elektronische Medium bleibt, wenn es auf diese Weise benutzt wird, hinter den Vorteilen des Papiermediums zurück.

Ein weiterer Grund für die zum Teil nicht überzeugende Qualität elektronischer Wörterbücher ist darin zu sehen, dass sie oft nicht von den Lexikographen selber, sondern von Softwareherstellern erarbeitet werden. Dagegen gilt es, die Kompetenzen aus dem lexikographischen Bereich mit denen aus dem Bereich neuer Medien zusammenzubringen. Das Ergebnis könnten qualitativ wesentlich bessere elektronische Produkte sein, die es ermöglichen, auf allen in einem Wörterbuch gegebenen Informationen zu recherchieren. Wenn die Kompetenzen aus Lexikographie und Texttechnologie zusammenkommen, können die Inhalte ganz anders gegriffen und ausgenutzt werden. Denkbar wären dann komplexe Suchanfragen wie „alle im 18. Jh. aus dem Französischen entlehnten Substantive". Gerade in solchen gezielten Suchmöglichkeiten liegt der Mehrwert des elektronischen Mediums; nur so erhält das elektronische Wörterbuch seinen mediengerechten Nachschlagecharakter. Es überführt die statische Präsentation aller Angaben im Print in eine dynamische, individualisierte Auswahl. Dem spezifischen Informationsbedürfnis des einzelnen Benutzers kann damit Rechnung getragen werden.[5]

Neben innovativen Benutzungsmöglichkeiten birgt ein neuer Umgang mit elektronischen Medien auch neue Chancen für Lexikographinnen und Verlage. Mit neuartig konzipierten Informationsbasen[6] werden neben unterschiedlichen Präsentationsmöglichkeiten auch andere Handhabungsmöglichkeiten bei der Entstehung oder redaktionellen Bearbeitung von Wörterbüchern möglich. Soll beispielsweise in einem allgemeinen einsprachigen Wörter-

---

2 Feldweg (1997), 110.

3 Dabei kann Multimedia zum reinen ‚Fun-Faktor' werden oder sinnvoll für die Wörterbuchbenutzung eingesetzt sein, wenn damit z.B. ein anderer Zugriff auf die Wörterbuchinhalte geboten wird. Ein Beispiel für letzteres ist die elektronische Version des Collins Cobuild Student's Dictionary (CCSD), in dem über ‚search by pictures' ein onomasiologischer Zugriff auf einen Teil der Wörterbuchartikel geboten wird.

4 Zu Suchbereichsstrukturen im Printwörterbuch vgl. Bergenholtz/Tarp/Wiegand (1999).

5 An neueren Projekten verfolgt z.B. das Projekt LEKSIS des Instituts für Deutsche Sprache Mannheim dieses Ziel (vgl. LEKSIS). Für eine spezifische Benutzungssituation ist auch das Projekt COMPASS entwickelt worden (vgl. Feldweg [o.J.]; Breidt [1998]).

buch der Artikel „Gebäude" verfasst werden, ist es im Hinblick auf die Konsistenz im Wörterbuch wichtig zu wissen, in welchen Bedeutungsparaphrasenangaben Gebäude als genus proximum angegeben wurde. Durch einfache und schnelle Recherchemöglichkeiten in einem Redaktionssystem, könnten alle diese Fälle leicht in der Bedeutungsangabe zu „Gebäude" berücksichtigt werden.

Informationsbasen so zu konzipieren, dass sich aus ihnen sowohl für den Benutzer als für Lexikographen und Verlage völlig neue (Be-)Nutzungsmöglichkeiten ableiten lassen, ist Teil unseres Ansatzes für ein *lexikographisches Modell*. Unter einem lexikographischen Modell verstehen wir ein systematisches Modell zum Umgang mit lexikographischen Inhalten, in dem sowohl die einzelnen Ebenen der Be- und Verarbeitung dieser Inhalte, als auch der ihnen zugrunde liegende Publikationsprozess abgebildet werden. Dabei zielen unsere Überlegungen nicht auf die Umformung eines *einzelnen* Printwörterbuchs in ein *einzelnes* elektronisches Produkt, sondern auf Wörterbuch-Neubearbeitungen oder neue Wörterbuchprojekte, die von Anfang an eine Realisierung in verschiedenen Medien anstreben sowie auf die Aufbereitung schon bestehender Wörterbücher als Informationsbasen. Der sogenannte Mehrwert des elektronischen Produkts darf dabei nicht nur ein Schlagwort bleiben, sondern muss durch einen gut entwickelten Werkzeugcharakter überzeugen.

Mit unserem neuen Ansatz versuchen wir die Grenzen der Vorstellungen auszuweiten, damit die Grenzen der Möglichkeiten nicht weiterhin so eng gesteckt bleiben. Daher ist unser Blick nicht ausschließlich auf die Lexikographie gerichtet, sondern auf Anforderungen der neuen Medien ausgeweitet.

2 Auf dem Weg zu einem neuen Ansatz

Um zu einem neuen Ansatz für ein lexikographisches Modell zu kommen, greifen wir in diesem Kapitel einzelne Aspekte verschiedener Themenbereiche heraus, die dafür nutzbringend sein könnten. Ein solches Modell zu entwickeln heißt heute, die Anforderungen der neuen Medien mit einzubeziehen, ohne die der „alten" Medien zu vernachlässigen. Deshalb betrachten wir in Abschnitt 2.1 den ISO-Standard SGML und stellen ein neues Schema für ein Publikationsmodell für Multiple-Media-Publishing vor. Ein neuer Ansatz muss sich außerdem in Beziehung setzen zu schon bestehenden. In 2.2 beleuchten wir daher die schon vorhandene lexikographische Modellstruktur der TEI unter dem Aspekt, inwieweit sie heutigen Anforderungen noch genügt. Auch erachten wir es für wichtig, die Entwicklung eines lexikographischen Modells theoretisch zu fundieren. Im Gegenstandsbereich der Lexikographie gibt es für Printwörterbücher die von H.E. Wiegand entwickelte Theorie lexikographischer Texte zur Analyse von Wörterbuchstrukturen. Ein Teilbereich daraus wird in Abschnitt 2.3 kurz vorgestellt.

---

[6] Die in dieser Arbeit verwendete Terminologie ist keine rein lexikographische, sondern verwendet auch zahlreiche Termini aus der Informationstechnologie, insbesondere Komposita mit „Information". Es ist ein wichtiges Desiderat, eine eigene und schlüssige Terminologie für den Schnittstellenbereich zwischen Lexikographie und Informationstechnologie zu entwickeln.

2.1 Ein Kern – viele Gesichter: SGML und Multiple-Media-Publishing

Unsere Gegenwart ist geprägt von einer rasanten Veränderung der Medienlandschaft. Den damit einhergehenden ständig wechselnden Marktanforderungen muss bei der Erarbeitung eines Publikationskonzepts Rechnung getragen werden. Wenn ein solches Konzept nicht aktuell veralten soll, müssen zwei zentrale Anforderungen erfüllt sein:

- Langlebigkeit der Datenhaltung und
- Flexibilität hinsichtlich der verschiedenen Präsentationsmedien.

Die Langlebigkeit der Datenhaltung kann mit dem Standard SGML[7] gewährleistet werden. Mit SGML ist man nicht an das proprietäre Format einer Datenbank oder eines Betriebssystems gebunden, sondern verfügt über eine systemunabhängige Schnittstelle. Die Flexibilität hinsichtlich der verschiedenen Präsentationsmedien wird unter das Schlagwort Multiple-Media-Publishing subsumiert. Die Verbindung dieser beiden Anforderungen wird nachfolgend *SGML-basiertes Multiple-Media-Publishing* genannt.

2.1.1 Ein neues Publikationsmodell

Unserem Publikationsmodell liegt ein SGML-basiertes Multiple-Media-Publishing-Konzept zugrunde. Bei seiner Entwicklung gingen wir von der gängigen Sicht auf elektronische Produkte mit ihrer Aufteilung in drei unterschiedliche Ebenen aus. Sie unterscheidet die Ebene der Datenmodellierung von der Präsentations- und der Interaktionsebene. Bei einer Auseinandersetzung damit kristallisierte sich heraus, dass dieses Modell für unsere Zwecke nicht explizit genug ist: Der Muliple-Media-Publishing-Prozess kann damit nicht hinreichend transparent gemacht werden. Um diesen Prozess adäquat greifen zu können, schlagen wir zum einen eine Ausdifferenzierung der Ebene der Datenmodellierung vor; das abstrakte Modell der Datenmodellierung sollte von der Repräsentation der Daten in diesem Modell unterschieden werden. Zum anderen macht eine Aufteilung in Redaktions- und Benutzerebene deutlich, wo in einem Publikationsprozess Präsentation und Interaktion anzusiedeln sind.

In den folgenden Abschnitten erläutern wir das abgebildete Schema eines Publikationsmodells (Abbildung 1). Die Reihenfolge der Darstellung ist an seiner prozeduralen Schicht ausgerichtet; die Ebenen selber werden hinsichtlich ihrer Rolle in diesem Prozess beleuchtet. Schließlich wird der FISCHER WELTALMANACH als illustrierendes Beispiel herangezogen, um zu zeigen, wie mediengerechte und daher unterschiedliche Umsetzungen einer Informationsbasis in ein Buch und eine CD-ROM aussehen können.[8]

---

[7] SGML (Standard Generalized Markup Language), ISO-Standard 8879; eine Untermenge davon ist XML (eXtensible Markup Language), die vor allem im Hinblick auf das Internet heute zunehmend an Bedeutung gewinnt. Aus diesem Grund sollte bei der Anwendung von SGML stets geprüft werden, inwieweit sie mit XML kompatibel ist und an welchen Stellen sie davon abweicht. Eine gute Übersicht zum Thema SGML/XML bietet die SGML/XML-Web-Page von Robin Cover (1999).

[8] Vgl. hierzu auch Kamps/Obermeier/Reichenberger/Schmidt (1999).

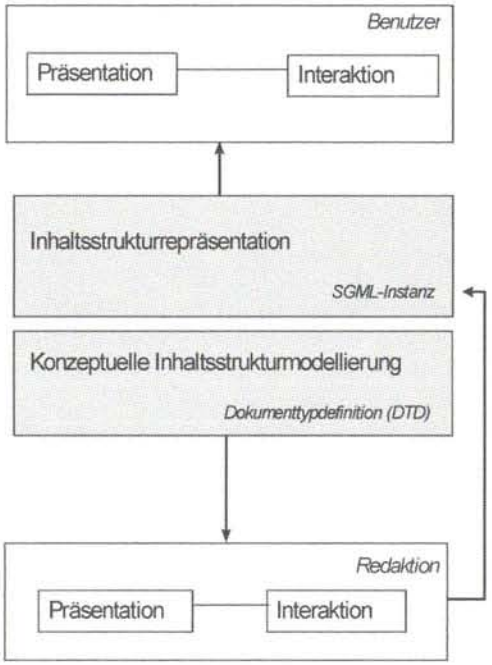


Abb. 1: Publikationsmodell (Schema)

2.1.1.1 Ebene der Inhaltsstrukturmodellierung

Den Anfangspunkt und auch den Kern unseres Publikationsmodells bildet die Ebene der konzeptuellen Inhaltsstrukturmodellierung. Wir haben uns gegen den allgemeinen Terminus Datenmodellierung und für den der Inhaltsstrukturmodellierung entschieden, da wir ein bestimmtes Modellierungskonzept voraussetzen: Die Zeichenketten der Datenbasis werden nach ihrem inhaltlichen Gehalt und ihrem genuinen Zweck als potentielle Informationseinheiten[9] klassifiziert. Diese werden daraufhin in eine Struktur eingebunden, welche ihre Beziehungen untereinander ausdrückt. Verdeutlicht an einem Prozess heißt das, dass man von einer Materialbasis ausgeht, um eine Inhaltsstrukturmodellierung zu entwickeln. Diese Materialbasis kann eine konkrete in Form einer Buchausgabe o.Ä. sein oder eine fiktive, die sich aus dem zu modellierenden Gegenstandsbereich ergibt. Gerade im ersten Fall ist die gedruckte Vorlage lediglich als Basis für einen ersten Analyseschritt zu verstehen. Die Modellierung darf sich nicht auf die Struktur dieser Vorlage beschränken, sondern muss sich von dieser lösen und die inhaltlichen Einheiten und ihre Bezüge untereinander im Hinblick auf die Struktur des Gegenstandsbereiches modellieren. Nur dann wird

[9] Die Informationseinheiten werden im Folgenden verkürzt als Information bezeichnet. Wir verwenden ‚Information' nicht in einem wohldefinierten, sondern im alltagssprachlichen Sinne nach Uszkoreit (1998, 7), der darauf hinweist : „[...] daß wir in unserer Alltagssprache oft den erwarteten Gebrauch von Objekten zur Grundlage ihrer Bezeichnung machen. So wie ein in Acrylharz gegossener Glückspfennig oder Dagobert Ducks erster selbstverdienter Dollar Zahlungsmittel sind [...,] so bezeichnen wir in einem erweiterten Sinn auch potentielle[,] aber ungenutzte Information als Information."

durch die Inhaltsstrukturmodellierung die Voraussetzung dafür geschaffen, dass später verschiedene mediengerechte Umsetzungen möglich sind. Inhaltsstrukturmodellierung heißt damit, dass ausschließlich der inhaltliche Gehalt der einzelnen Informationseinheiten, vollkommen losgelöst von ihrer Anordnung und typographischen Darstellung gefasst wird. In einem SGML-basierten Publikationsprozess entsprechen diese Schritte denen der Dokumentanalyse[10] und der sich daran anschließenden DTD[11]-Entwicklung.

2.1.1.2 Redaktionsebene

Auf der Redaktionsebene wird die Inhaltsstrukturmodellierung für die Eingabe der Inhalte nutzbar gemacht. Präsentation meint, dass die Inhaltsstrukturmodellierung zunächst für die Redakteurin visualisiert wird; unter Interaktion wird das daran anschließende Einfügen der Inhalte in die Struktur verstanden und auch die spätere Arbeit mit den Inhalten auf der Ebene der Inhaltsstrukturrepräsentation. Ein SGML-basiertes Redaktionssystem führt den Redakteur bei der Eingabe durch die Struktur der DTD und unterstützt ihn bei der weiteren redaktionellen Arbeit mit den Instanzen.

2.1.1.3 Die Ebene der Inhaltsstrukturrepräsentation

Ein Ergebnis der redaktionellen Arbeit ist die Inhaltsstrukturrepräsentation, d.h. die eingegebenen Inhalte werden mit ihrer dazugehörigen Struktur abgebildet. Wir verwenden den Terminus Inhalts*struktur*repräsentation mit Bezug auf einen Strukturbegriff, bei dem die Elemente selber unter der Struktur subsumiert sind. In einem SGML-basierten Prozess ist auf der Ebene der Inhaltsstrukturrepräsentation die SGML-Instanz zu finden.

Die Inhaltsstrukturrepräsentation bildet zusammen mit der Inhaltsstrukturmodellierung die *Informationsbasis*, die für weitere redaktionelle Arbeiten ebenso nutzbar gemacht werden kann wie für die Umsetzungen auf der Benutzerebene.

2.1.1.4 Benutzerebene

Die Inhaltsstrukturrepräsentation wird auf der Benutzerebene visualisiert. Dabei werden Programme eingesetzt, die im Idealfall das gesamte Potential der Inhaltsstrukturrepräsentation für die Präsentation und Interaktion ausnutzen. Für eine gedruckte Ausgabe hieße dies, dass ein Satzprogramm die Inhaltsstrukturrepräsentation in das Layout und die Typographie des Buches umsetzt. Für ein elektronisches Produkt sollte das Ziel der Umsetzung ein flexibler Umgang mit der Informationsbasis sein. Flexibel in dem Sinn, dass die Präsentation als benutzerdefinierte, dynamische Schnittstelle zwischen der Informationsbasis und dem Benutzer begriffen wird. Dem Benutzer werden damit individualisierte Interaktions- und Zugriffsmöglichkeiten geboten, die auch entsprechend spezifisch präsentiert werden. Flexibel auch im Hinblick darauf, dass der Benutzer zu seinem eigenen Redakteur

---

[10] Man spricht in der SGML-Terminologie von Dokumentanalyse. Damit ist heute nicht zwingend die Analyse von gedruckten Dokumenten gemeint, sondern die Erschließung der Inhalte und inhaltlichen Bezüge eines Gegenstandsbereichs. In diesem Sinn verstehen wir darunter die Analyse einer Materialbasis mit dem Ziel, eine Inhaltsstrukturmodellierung zu entwickeln.

[11] DTD (Document Type Definition), zu deutsch: Dokumenttypdefinition.

werden kann, indem er beispielsweise Verknüpfungen herstellt oder Notizen hinzufügt, auf denen er anschließend wieder recherchieren kann.

Mit den vier folgenden Abbildungen soll am Beispiel des FISCHER WELTALMANACH verdeutlicht werden, wie flexibel Informationen präsentiert werden können. Sowohl der Buchausgabe[12] als auch der CD-ROM[13] liegt dabei die gleiche SGML-strukturierte Informationsbasis zugrunde. Abbildung 2 zeigt die Präsentation dieser Informationsbasis im Buch anhand eines Ausschnitts mit Informationen zu Argentinien.[14]

Argentinien *Süd-Amerika*
Argentinische Republik; República Argentina – RA (→ Karte VII, B-D 6-9)

Fläche (Weltrang: 8.): 2 780 400 km²

Einwohner (31.): F 1996 35 220 000 = 12,7 je km²

Hauptstadt: Buenos Aires Z 1991: 2 960 976 Einw. (S 1995 A 10,990 Mio.)

Amtssprache: Spanisch

Bruttosozialprodukt 1996 je Einw.: 8380 $

Währung: 1 Argent. Peso (arg$) = 100 Centavos

Botschaft der Republik Argentinien
Adenauerallee 50–52, 53113 Bonn, 02 28/22 80 10

Landesstruktur Fläche: 2 780 400 km² – **Bevölkerung**: Argentinier; (Z 1991) 32 615 528 Einw. – (S) über 90 % Weiße (v. a. europäischer Herkunft, u. a. 36 % italien. und 29 % span. sowie etwa 0,5

Abb. 2: Argentinien (Ausschnitt)

Auf Abbildung 3 ist ein Fenster der CD-ROM[15] zu sehen. Der vom Benutzer definierte Informationsausschnitt ist in einer eng an das Buch angelehnten Darstellung gezeigt. Die linke Seite des Fensters verzeichnet die dazu vom Benutzer ausgewählten Kategorien.

Die Abbildungen 4 und 5 verdeutlichen, dass die gleichen Informationen auf Anfrage des Benutzers auch grafisch visualisiert werden können. Dabei können sowohl die ausgewählten Staaten als auch die ausgewählten Kategorien als Ordnungsprinzip zugrunde gelegt werden.

[12] FWA 99.
[13] FWA 99 CD-ROM.
[14] FWA 99, Spalte 70.
[15] Alle weiteren Beispiele in diesem Abschnitt stammen aus der FWA 99 CD-ROM.

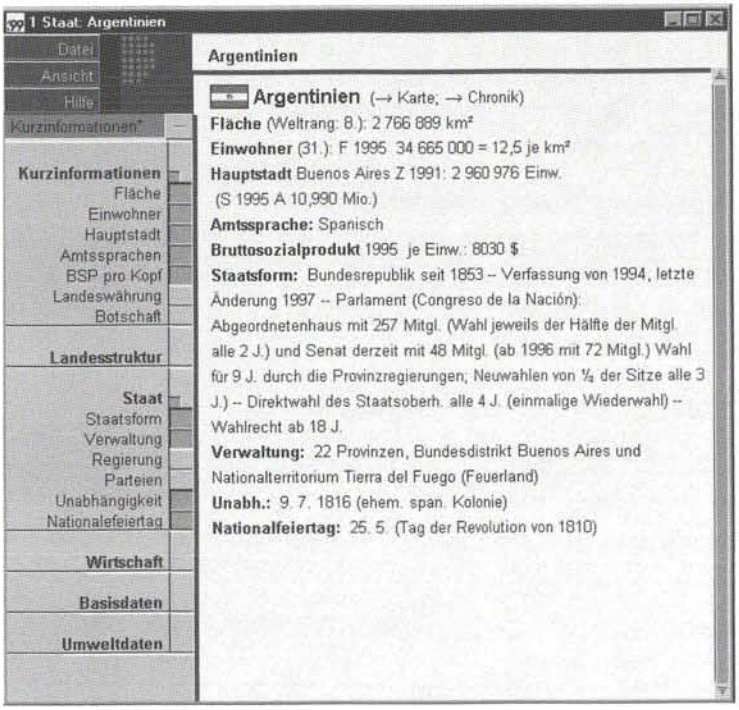


Abb. 3: Benutzerdefinierte Sicht auf Argentinien

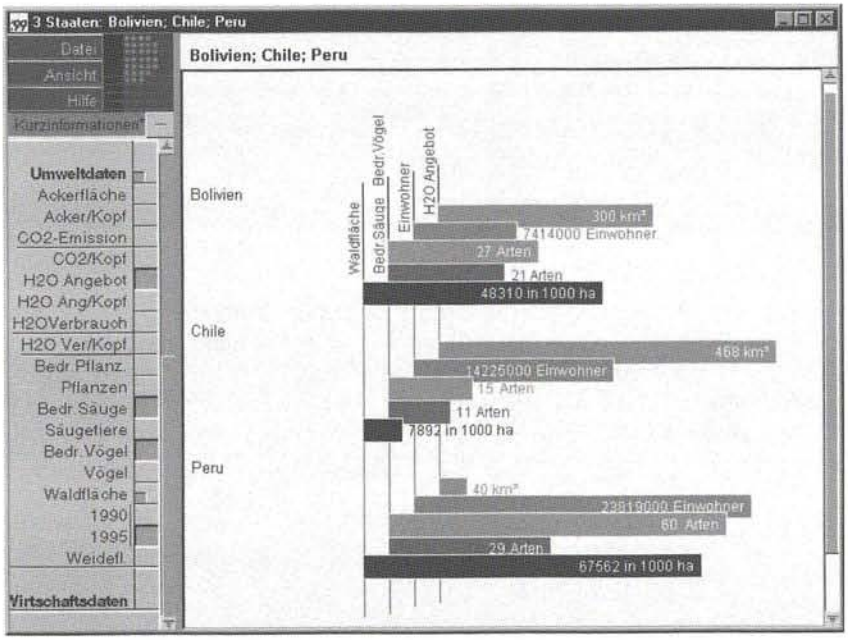


Abb. 4: Grafische Sicht nach Staaten geordnet

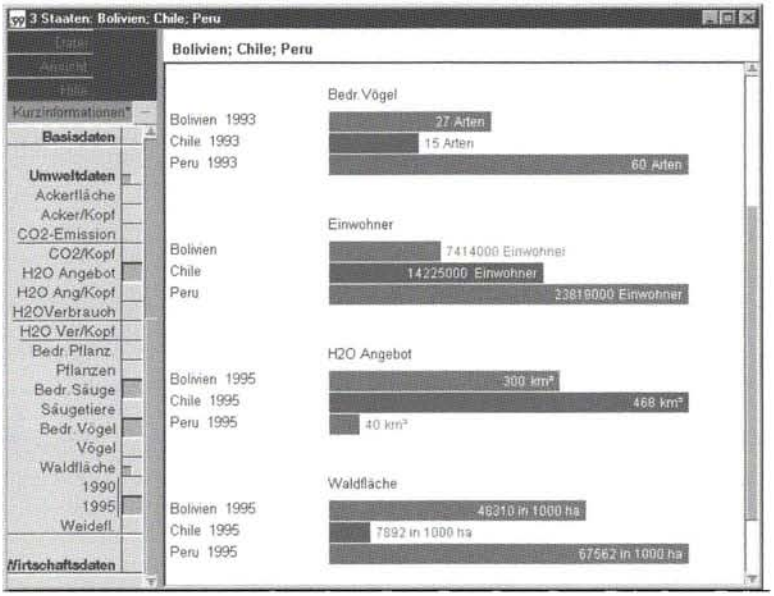


Abb. 5: Grafische Sicht nach Kategorien geordnet

2.1.2 Die zentrale Rolle der Inhaltsstrukturmodellierung

Die Inhaltsstrukturmodellierung legt die Basis für die unter 2.1 formulierten Anforderungen an ein Publikationskonzept und schafft Möglichkeiten, wie sie z.B. durch die Abbildungen 2–5 verdeutlicht werden. Sie sollte den Anspruch haben, möglichst kleine Informationseinheiten – losgelöst von den Anordnungsgesichtspunkten der Präsentation – zu greifen. Dies ist die Voraussetzung für die Bandbreite der Umsetzungsmöglichkeiten auf der Benutzerebene; auch für Umsetzungsmöglichkeiten in Medien, die heute möglicherweise so noch gar nicht denkbar sind. Bildlich gesprochen wird hier die Basis dafür gelegt, dass die Informationen an keinem Medium ‚kleben', d.h. dass die Modellierung weder die spezifischen Charakteristika eines Buches noch die eines elektronischen Produktes abbildet. Auf diese Weise kann Medienunabhängigkeit erreicht werden. Wohlstrukturierte Informationseinheiten sind allein noch kein Garant für ein konsistentes, und – auf elektronische Medien bezogen – leistungsfähiges und flexibles Produkt. Ohne Wohlstrukturiertheit kann jedoch auch durch ein gutes Programm keine flexible Umsetzung stattfinden.

In dem vorgestellten Publikationsmodell wird die Langlebigkeit der Datenhaltung durch den internationalen Standard SGML garantiert. Langlebigkeit heißt dabei auch, dass die Informationsbasis so angelegt sein muss, dass durch neue Medientypen hinzukommende Präsentationsanforderungen nicht in einer Neumodellierung resultieren dürfen.

2.2 Ein Modell für Wörterbücher: Die Text Encoding Initiative (TEI)

Die TEI wurde 1988 als Forschungsprojekt mit dem Ziel ins Leben gerufen, Richtlinien für die Auszeichnung verschiedener Texttypen aus dem geisteswissenschaftlichen Bereich zur Verfügung zu stellen. Da diese auch einen reibungslosen elektronischen Austausch von Do-

kumenten gewährleisten sollten, entschied man sich 1990 für den gerade ein Jahr zuvor verabschiedeten Standard SGML (ISO 8879). Nach insgesamt sechsjähriger Arbeit, an der eine Vielzahl von Fachleuten aus der ganzen Welt beteiligt war, erschienen im Mai 1994 die *Guidelines for Electronic Text Encoding and Interchange*,[16] bekannt als TEI P3 (d. i. TEI Proposal number 3). Dieses 1300 Seiten umfassende Papier beschreibt das von der TEI entwickelte modulare System einer Dokumentenarchitektur und die einzelnen, darin eingebundenen Dokumenttypdefinitionen (DTD). Das modulare System ermöglicht es dem Anwender, die TEI-DTDs den eigenen Bedürfnissen anzupassen.

Die TEI-DTD für Printwörterbücher hat den Anspruch, alle modernen Wörterbücher westlicher Sprachen mittleren Umfangs abzubilden. Sie umfasst Auszeichnungsmöglichkeiten für die Wörterbuchartikel und für die Umtexte; letztere werden auch bei anderen Texttypen der TEI angewandt. Wir beschreiben im folgenden nicht die ganze DTD für Printwörterbücher, sondern greifen charakteristische Aspekte der wörterbuchspezifischen Auszeichnungen heraus, um die damit verbundene Problematik aufzuzeigen.

Bei den wörterbuchspezifischen Auszeichnungen kann man zunächst zwischen hierarchischen und inhaltsbezogenen Elementen unterscheiden. Zu den hierarchischen Elementen werden die verschiedenen Artikeltypen eines Wörterbuchs ebenso gerechnet wie die Auszeichnungen für Homographen und Bedeutungsangaben. Bei den inhaltsbezogenen Auszeichnungen handelt es sich u.a. um Informationen zur Wortform, Grammatik, Verwendung und Etymologie sowie um Definitionen und Beispiele.[17] Des Weiteren werden die Verweiselemente dazugerechnet, die sich auf die Schreibweise oder die Aussprache des Lemmas beziehen.

2.2.1 Flexibilität der hierarchischen Struktur

Die TEI-Wörterbuchstruktur unterscheidet drei Artikeltypen:

`<entry>`	der strukturierte Artikel
`<entryFree>`	der freie Artikel
`<superEntry>`	der gruppierende Artikel

Daneben gibt es eine Sonderform des Artikels:

`<re>`	der verwandte Artikel

Dabei sieht die Struktur des `<entry>` so aus, dass zunächst, in beliebiger Reihenfolge, zwischen Homographie- oder Bedeutungsangaben oder den oben aufgelisteten inhaltsbezogenen Auszeichnungen ausgewählt werden kann und danach erst die jeweiligen Unterstrukturen zur Verfügung stehen. Beim `<entryFree>` wird diese Strukturhierarchie aufgelöst; alle Strukturebenen stehen gleichberechtigt nebeneinander. Der `<superEntry>` stellt die Informationen zur Wortform als optionales Element vor eine Gruppe strukturierter Einträge. Der verwandte Artikel `<re>` ist eine Sonderform des `<entry>` und wird von uns deshalb der hierarchischen Struktur zugerechnet. Er hat gegenüber dem `<entry>` die Ein-

---

[16] Sperberg-McQueen/Burnard (1994); eine ausführliche Beschreibung der Geschichte der TEI findet sich bei Ide/Sperberg-McQueen (1995); eine Auseinandersetzung mit ausgewählten Aspekten der TEI-Wörterbuch-DTD bieten Ide/Véronis (1995); einen guten Überblick gibt die Webpage der TEI.

[17] In den Erläuterungen zur TEI wird auch deren Begrifflichkeit übernommen.

schränkungen, dass er nur innerhalb der drei Artikeltypen vorkommen und keine Homographen verzeichnen kann sowie nicht geschachtelt sein darf.

Als weitere hierarchische Strukturen haben die Elemente zu gelten, die Informationen zu einem Homographen bzw. zu einer Lesart gruppieren:

`<hom>`	Homographengruppe
`<sense>`	Informationen zu einer Lesart

Beide Gruppen können sowohl im `<entry>` als auch im `<entryFree>` vorkommen. Die Homographen- und die Lesartgruppe sind beide wie ein strukturierter Artikel definiert, können jedoch keine weitere Homographengruppe einschließen. Die Lesartgruppe ist ihrerseits rekursiv definiert. Um die so entstehenden Schachtelungsebenen eindeutig voneinander trennen zu können, ist ihr ein entsprechendes Attribut beigegeben. Im folgenden sollen anhand von Beispielen die verschiedenen Hierarchisierungsmöglichkeiten aufgezeigt werden.

Ein strukturierter Artikel, der zwei Lesarten des Lemmas verzeichnet, würde somit folgendem Strukturmuster[18] gehorchen:

```
<entry>
            <!-- Informationen zu beiden Lesarten -->
            <sense n="1">
                  <!-- Informationen zu Lesart 1 -->
            </sense>
            <sense n="2">
                  <!-- Informationen zu Lesart 2 -->
            </sense>
</entry>
```

Verzeichnet ein Lemma hingegen zwei Homographen, denen jeweils zwei Lesarten zugerechnet werden können, teilweise mit Unterbedeutungen, träfe folgendes Strukturmuster zu:

```
<entry>
            <!-- Informationen zu beiden Homographen -->
            <hom n="1">
                  <sense n="1"> ... </sense>
                  <sense n="2"> ... </sense>
            </hom>
            <hom n="2">
                  <sense n="1">
                        <sense n="a"> ... </sense>
                        <sense n="b"> ... </sense>
                  </sense>
                  <sense n="2"> ... </sense>
            </hom>
</entry>
```

Je nach Wörterbuchkonzept könnten die beiden Homographen alternativ als zwei separate Einträge angesehen oder als `<superEntry>` angesetzt werden. Im letzeren Fall müssen sie als zwei strukturierte Artikel gruppiert werden, mit der Möglichkeit, dass die ihnen gemeinsamen Informationen zur Wortform ausgelagert werden können:

---

[18] Die Strukturmuster in diesem Abschnitt sind angelehnt an die TEI-Richtlinien von Sperberg-McQueen/Burnard (1994), Abschnitt 12.2.1.

```
<superEntry>
            <form>
<!--optional ausgelagerte Info.n zur Wortform-->
            </form>
            <entry n="1">
                  <sense n="1"> ... </sense>
                  <sense n="2"> ... </sense>
            </entry>
            <entry n="2">
                  <sense n="1">
                        <sense n="a"> ... </sense>
                        <sense n="b"> ... </sense>
                  </sense>
                  <sense n="2"> ... </sense>
            </entry>
</superEntry>
```

Diese Beispiele zeigen, dass die hierarchischen Eintragsstrukturen sehr flexibel gehandhabt werden können, damit jede mögliche Artikelstruktur von Wörterbüchern abgebildet werden kann. Ein solch hoher Grad an Flexibilität ist jedoch nur dann notwendig, wenn die Modellierung versucht, alle möglichen Darstellungsweisen eines Printwörterbuchs abzubilden. Eine konsistente Informationsbasis sollte jedoch unabhängig von einer einzelnen möglichen Präsentationsform sein und auf die Modellierung inhaltlicher Zusammenhänge fokussieren (vgl. Abbildung 1). Nur so können konsistente Eintragsmuster entstehen, die nach unterschiedlichen Wörterbuchkonzepten und auf unterschiedlichen Medien präsentierbar sind.

2.2.2 Flexibilität der inhaltlichen Struktur

Die inhaltsbezogenen Elemente können sowohl in strukturierten und freien Einträgen wie auch in den Homographie- und Bedeutungsgruppen vorkommen. Dabei ist weder die Reihenfolge noch die Häufigkeit ihres Vorkommens in den jeweiligen Inhaltsmodellen festgelegt. Die inhaltsbezogenen Elemente sind verschieden stark oder auch gar nicht durch weitere wörterbuchspezifische Elemente unterstrukturiert und teilweise rekursiv definiert.

Stark unterstrukturiert sind:

`<form>`	Informationen zur Wortform
`<gramGrp>`	Informationen zur Grammatik
`<etym>`	Informationen zur Etymologie
`<trans>`	Übersetzung in eine Zielsprache

Weniger stark unterstrukturiert sind:

`<eg>`	Beispiel
`<xr>`	Querverweis
`<note>`	Notiz (kein wörterbuchspezifisches Element)

Nicht weiter unterstrukturiert sind:

`<def>`	Definition
`<usg>`	Verwendungsangabe

2.2.2.1 Zum Beispiel: Informationen zur Wortform und Grammatik

Durch die Informationen zur Wortform `<form>` werden alle geschriebenen und gesprochenen Formen des Lemmas und grammatische Angaben gruppiert. Das `<form>`-Element ist rekursiv definiert; die Angaben innerhalb von `<form>` können in beliebiger Reihenfolge und Häufigkeit vorkommen.

Elemente, welche die geschriebene und gesprochene Form des Lemmas fassen, sind:

`<orth>`	Orthographieangabe
`<pron>`	Ausspracheangabe
`<hyph>`	Trennungsangabe
`<syll>`	Silbenangabe
`<stress>`	Betonungsangabe
`<lbl>`	Verwendungsspezifizierung

Elemente, die grammatische Angaben zu einer `<form>` fassen, sind:

`<gram>`	allgemeines Element für Grammatikangaben
`<gen>`	Genusangabe
`<number>`	Numerusangabe
`<case>`	Kasusangabe
`<per>`	Personenangabe
`<tns>`	Tempusangabe
`<mood>`	Modusangabe
`<itype>`	Flektionsparadigma

Das allgemeine Element `<gram>` kann für alle Typen von Grammatikangaben eingesetzt werden. Dabei wird der Angabetyp, beispielsweise Genus, Numerus, Kasus, über ein Attribut bestimmt. Die anderen als spezifische Grammatikangaben ausgewiesenen Elemente können daher als Synonyme des `<gram>`-Elements gelten.

Die Informationen zur Grammatik `<gramGrp>` beziehen sich immer auf das gesamte Lemma und umfassen zum einen die Elemente:

`<pos>`	Wortartangabe
`<subc>`	andere Informationen zur Kategorisierung, wie transitiv/intransitiv, zählbar
`<colloc>`	Kollokationen des Lemmas
`<lbl>`	Verwendungsspezifizierung
`<usg>`	Verwendungsangabe

Daneben stehen die Elemente für grammatische Angaben, wie wir sie schon bei den Informationen zur Wortform aufgelistet haben und die ebenfalls in den etymologischen Informationen vorkommen. Das `<gramGrp>`-Element ist rekursiv definiert; die Angaben innerhalb von `<gramGrp>` können in beliebiger Reihenfolge und Häufigkeit vorkommen.

Rekursive Definitionen ermöglichen eine beliebige Schachtelung eines Elements. Dadurch können beispielsweise bei der Information zur Wortform zwei regional verschiedene Schreibweisen eines Lemmas einer gemeinsamen Ausspracheangabe zugeordnet werden. Da die variante Orthographie mit einer Gebrauchsinformation gekoppelt werden muss, damit sie als geographische Variante ausgewiesen werden kann, wird sie durch ein sepa-

rates <form>-Element geklammert. Die Ausspracheangabe <pron> könnte auch direkt auf die Orthographie des Lemmas folgen, steht aber nach der Orthographievariante, weil die Reihenfolge der Darstellung der im gedruckten Wörterbuch entspricht.[19]

```
<form>
        <orth>colour</orth>
        <form>
                <usg type=geo>U.S.</usg>
                <orth>color</orth>
        </form>
        <pron>'kʌlɛ(r)</pron>
</form>
```

Damit ist eine der Schreibweisen geographisch zugeordnet, die andere nicht. Man kann nun davon ausgehen, dass, wenn keine geographischen Einordnungen vorgenommen werden, es sich immer um britische Schreibweise handelt. Diese Informationen müssen über die Erfassungsrichtlinien klar geregelt werden, damit man zu einem konsistenten Datenbestand kommt; die Struktur selbst bietet keine Kontrolle. Eine mögliche Strukturkontrolle könnte durch ein spezielles Element für orthographische Varianten gegeben werden, das eine geographische Einordnung erzwingt.

Möchte man bei dem Beispielartikel noch die Wortart, die für beide Schreibweisen gilt, festhalten, so kann dies innerhalb des <form>-Elements nur mit dem allgemeinen Element für grammatische Angaben gemacht werden. Über ein type-Attribut wird festgelegt, dass es sich beim Elementinhalt um die Wortart handelt. Auch hier gilt, dass das <gram>-Element vor der Schreibvariante stehen könnte, würde nicht die Struktur des Buches abgebildet.

```
<form>
        <orth>colour</orth>
        <form>
                <usg type=geo>U.S.</usg>
                <orth>color</orth>
        </form>
<pron>'kʌlɛ(r)</phon>
        <gram type="pos">n</gram>
</form>
```

Da sich die grammatische Information innerhalb einer <gramGrp> stets auf das gesamte Lemma bezieht, könnte die Auszeichnung auch wie folgt aussehen:

```
<form>
        <orth>colour</orth>
        <form>
                <usg type=geo>U.S.</usg>
                <orth>color</orth>
        </form>
        <pron>'kʌlɛ(r)</pron>
</form>
<gramGrp>
<gram type="pos">n</gram>
</gramGrp>
```

[19] Das Beispiel in diesem Abschnitt stammt aus dem OALD.

Innerhalb des <gramGrp>-Elements muss nicht auf das allgemeine Grammatikelement zurückgegriffen werden, sondern es gibt es die Möglichkeit, die Wortart durch ein spezifisches <pos>-Element auszuzeichnen anstelle des pos-Attributwerts im vorherigen Beispiel. Die Möglichkeit eines <pos>-Elements gibt es innerhalb von <form> jedoch nicht.

```
<form>
       <orth>colour</orth>
       <form>
              <usg type=geo>U.S.</usg>
              <orth>color</orth>
       </form>
       <pron>'kalɛ(r)</pron>
</form>
<gramGrp>
   <pos>n</pos>
</gramGrp>
```

Diese Beispiele verdeutlichen, dass durch die vielfältigen Interpretationsmöglichkeiten der Strukturelemente schon fast nicht mehr von einer Struktur gesprochen werden kann. Eine Strukturführung für den Benutzer kann nur durch eine Einschränkung der Strukturierungsmöglichkeiten der DTD über TEI-Modifikationen ermöglicht werden. Werden solche DTD-Einschränkungen nicht vorgenommen, müssen Schreibrichtlinien zu einer konsistenten Anwendung beitragen. Die Flexibilität auf Seite der inhaltlichen Auszeichnungen entspricht somit der seitens der hierarchischen Elemente. Auch hier resultiert die Notwendigkeit für eine solche Flexibilität aus dem Anspruch der TEI, alle westlichen Wörterbuchtypen sowie ihre große Bandbreite an Präsentationsmöglichkeiten im Print mit der Strukturierung fassen zu wollen. Die Präsentationssicht zeigt dabei nur die Reihenfolge der Elemente im Print; typographische Realisierungen will sie nicht abbilden.

2.2.2.2 Zum Beispiel: Verdichtung

Vor allem bei Beispielen, Definitionen und etymologischen Angaben ist in Printwörterbüchern das Lemma häufig in verdichteter Form, beispielsweise als Tilde, dargestellt. Diese Art der Verdichtung wird bei der TEI als ein Verweis aufgefasst, da mit der Tilde auf die Schreibung oder Aussprache des Lemmas verwiesen wird. Dabei kann differenziert werden, ob sich die Verweisangabe auf die Grundform oder auf eine flektierte Form des Lemmas bezieht. Das Prinzip soll an zwei Beispielen dargestellt werden, die jeweils auf die Schreibung verweisen.[20]

```
colonel [...] army officer above a lieutenant-~
<entry>
       <form><orth>colonel</orth></form>
       [...]
       <def> army officer above a lieutenant-<oref></def>
</entry>
take [...] The new play really took the public's fancy.
<entry>
```

[20] Die Beispiele sind dem Abschnitt 12.4 der TEI-Richtlinien von Sperberg-McQueen/Burnard (1994) entnommen.

```
        <form><orth>take</orth></form>
        [...]
        <eg>
<q>The new play really <oVar type=pt>took</oVar> the public's fancy.
</q>
</eg>
</entry>
```

Diese Lösung der TEI scheint auf den ersten Blick bestechend. Auf den zweiten Blick zeigt sich aber, dass auch hier aus der Perspektive des Printwörterbuchs modelliert wurde: Dargestellt ist nämlich die verdichtete Form, wie sie für eine bestimmte Präsentationsform erforderlich ist, und nicht die ausformulierte Fassung, die in einer Inhaltsstrukturmodellierung zu finden sein müsste.

2.2.3 Verschiedene Sichten auf Wörterbücher

Die TEI unterscheidet grundsätzlich folgende verschiedene Sichten auf Wörterbücher:

- lexikographische Sicht
  Die lexikographische Sicht beschreibt einen Wörterbuchartikel unabhängig von seiner späteren Darstellung in einem bestimmten Medium.
- redaktionelle Sicht
  Die redaktionelle Sicht befasst sich mit einer Auswahl aus dem Datenbestand für eine ganz bestimmte Ausgabe des Wörterbuchs; sie legt die Reihenfolge der Elemente ebenso fest wie die Prinzipien der Verdichtung und z. T. die typographische Darstellung der Elemente.
- typographische Sicht
  Die typographische Sicht auf ein Wörterbuch beschreibt die zweidimensionale Darstellung der Buchausgabe.

Die TEI-Wörterbuch-DTD stellt Auszeichnungen für die redaktionelle und die lexikographische Sicht von Wörterbüchern zur Verfügung. Dabei gibt es zwei Arten der redaktionellen Sicht: Eine, die alle Zeichen abbildet, die für eine bestimmte typographische Sicht benötigt werden oder eine, die statt dessen festlegt, wie diese sichtspezifischen Zeichen generiert werden sollen. Die redaktionelle Sicht und die lexikographische Sicht können getrennt voneinander ausgezeichnet werden oder in einer Auszeichnung zusammengefasst sein. Im letzteren Fall muss eine der Sichten zur Hauptsicht gemacht werden. Diese bestimmt dann die Struktur des Artikels hinsichtlich der Reihenfolge der Elemente; die Informationen für die andere Sicht werden als Attributwerte mitgeführt. Diese Prinzipien werden an nachfolgenden Beispielen zum Lemma **pinna** verdeutlicht.[21]

Redaktionelle Sicht mit allen Zeichen für die typographische Darstellung:

```
<entry>
   <form><orth>pinna</orth></form>
   <gramGrp><pos>n.</pos>, </gramGrp>
       <form type="infl">
```

[21] Die Beispiele sind teilweise dem Abschnitt 12.5 der TEI-Richtlinien von Sperberg-McQueen/Burnard (1994) entnommen.

```
            <number>pl. </number>
            <orth type="lat" extent="part">-nae</orth> or
            <orth type="std" extent="part">-nas</orth>
        </form>
  [...]
  </entry>
```

Redaktionelle Sicht ohne die Zeichen für die typographische Darstellung:

```
<entry>
   <form><orth>pinna</orth></form>
   <gramGrp><pos>n</pos></gramGrp>
       <form type="infl">
            <number>pl</number>
            <orth type="lat" extent="part">-nae</orth>
            <orth type="std" extent="part">-nas</orth>
       </form>
  [...]
</entry>
```

Werden Zeichen und Text bei der typographischen Umsetzung generiert, so empfiehlt die TEI die Anweisungen dazu im Vorspann innerhalb des Elements `<tagUsage>` festzuhalten.

```
<tagUsage>
  Der Inhalt von <pos> wird durch einen Punkt abgeschlossen.
  Nach dem Element <pos> steht immer ein Komma.
  Bei mehr als einem <orth> steht "oder" zwischen den Elementen.
  [...]
  </tagUsage>
```

Lexikographische Sicht:

```
  <entry>
   <form>
            <orth>pinna</orth>
       <form type="infl">
                 <number>pl</number>
                 <orth type="lat">pinnae</orth>
                 <orth type="std">pinnas</orth>
       </form>
   </form>
   <gramGrp><pos>n</pos></gramGrp>
   [...]
   </entry>
```

Redaktionelle Sicht (Hauptsicht) mit lexikographischer Sicht:

```
<entry>
   <form><orth>pinna</orth></form>
   <gramGrp><pos>n</pos></gramGrp>
       <form type="infl">
            <number>pl</number>
            <orth type="lat" norm="pinnae" extent="part">
                 -nae</orth>
```

```
            <orth type="std" norm="pinnas" extent="part">
                -nas</orth>
      </form>
  [...]
</entry>
```

Auch hier gilt wieder, dass die hohe Flexibilität mit einem großen Risiko hinsichtlich der Datenkonsistenz erkauft wird. Wie unser Publikationsmodell (vgl. Abschnitt 2.1) gezeigt hat, ist es nicht sinnvoll, die Modellierung auf der redaktionellen Sicht aufzusetzen, sondern sie kann nur sinnvoll und konsistent auf der lexikographischen Ebene, d.h. der Inhaltsstrukturmodellierung, geleistet werden. Problematisch ist es zudem, die Modellierung nicht auf eine Sicht festzulegen, sondern sie für mehreren Sichten zu ermöglichen. Zudem ist in der Definition der TEI unserer Ansicht nach die typographische Sicht nicht so deutlich von der redaktionellen Sicht getrennt, wie wir es in unserem Publikationsmodell zwischen Benutzerebene und Redaktionsebene tun. Die TEI zählt vielmehr Präsentationsaspekte der Benutzerebene sowohl zur redaktionellen Sicht, z.B. Verdichtungen und Reihenfolge der Angaben, wie auch zur typographischen Sicht, z.B. typographische Realisierung der Angaben und ihre Anordnung auf der Buchseite.

Die Auseinandersetzung mit der TEI hat gezeigt, dass die TEI zu flexibel ist, wenn es, wie in unserem Publikationsmodell, darum geht, mit einer Inhaltsstrukturmodellierung die Basis für eine Inhaltsrepräsentation festzulegen. Deshalb ist die TEI-Wörterbuch-DTD für das zu entwickelnde lexikographische Modell weitgehend nicht auszunutzen.

2.3 Mikrostrukturen nach H. E. Wiegand

Wie in 2.1.1.1 beschrieben, müssen in einer Datenbasis zunächst einzelne Einheiten identifiziert und klassifiziert werden, damit eine Inhaltsstrukturmodellierung entwickelt werden kann. Die Datenbasis wird bei lexikographischen Projekten aus dem zugrunde gelegten Wörterbuchgegenstand gewonnen.

Für Printwörterbücher bietet die Theorie lexikographischer Texte von H. E. Wiegand eine formalisierte Analysemethode zur Segmentation von Wörterbuchartikeln.[22] Das Ziel dieser mikrostrukturellen Analyse ist die vollständige Segmentierung eines Wörterbuchartikels in *Angaben*. Von Angaben ist dann zu sprechen, wenn der Wörterbuchartikeltext in nicht weiter zerlegbare Textsegmente mit mindestens einem genuinen Zweck aufgeteilt ist, also eine Segmentation in funktionale Textsegmente durchgeführt wurde. Über diesen funktionalen Textsegmenten werden zwei strukturprägende Relationen definiert: eine partitive Relation (Teil-Ganzes) und eine Präzedenzrelation (Vorgänger-Nachfolger). Die Teil-Ganzes-Relation legt die partitive Struktur fest. Die Vorgänger-Nachfolger-Relation ergibt sich aus der Position der Angaben im Artikel und legt damit die präzedentive Struktur fest.

Das Ergebnis einer solchen *funktional-positionalen Segmentation* über einem Wörterbuchartikel ist eine *konkrete hierarchische Mikrostruktur*. Wird von den konkreten Angaben zu Angabeklassen abstrahiert, wird von einer *abstrakten hierarchischen Mikrostruktur* gesprochen.

Diese Analysemethode wurde anhand von Printwörterbüchern entwickelt und legt die einzelnen Artikel im Druckraum zugrunde. Deshalb gilt es zu prüfen, ob sie für eine Analysemethode nutzbar gemacht werden kann, die nicht von der Präsentation in einem

---

[22] Wiegand (1989a); Wiegand (1989b); vgl. auch Storrer (1996).

bestimmten Medium, sondern vom lexikographischen Gegenstandsbereich als solchem ausgeht. Der formalisierte Ansatz der Analysemethode scheint jedoch vielversprechend für ein fundiertes Konzept der Inhaltsstrukturmodellierung in lexikographischen Projekten. Durch formale Methoden werden Regeln für eine nachvollziehbare Modellierung festgelegt. Dies führt zu einer konsistenten Auszeichnung der Inhalte, die ihrerseits eine wichtige Voraussetzung für deren automatisierte Handhabung ist.

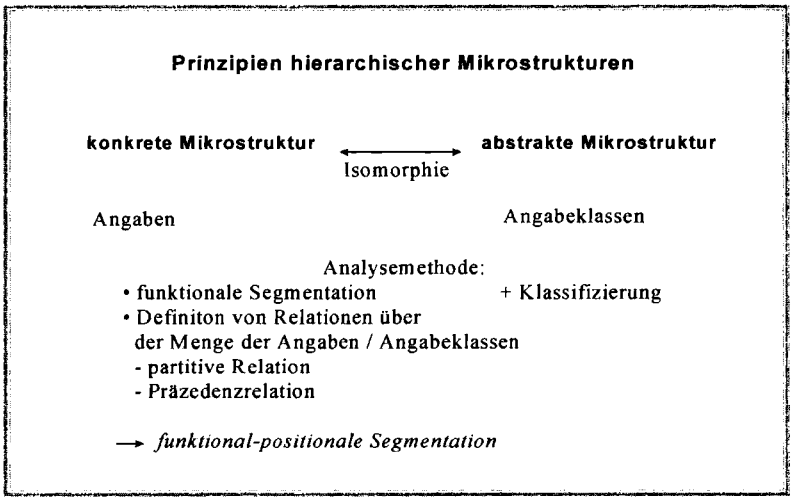


Abb. 6: Prinzipien hierarchischer Mikrostrukturen nach H. E. Wiegand

3 Ein neuer Ansatz

Unser Ansatz versucht Aspekte der unter 2 beschriebenen Bereiche für die Entwicklung eines lexikographischen Modells fruchtbar zu machen. Es sollen ausschnitthaft spezifisch lexikographische Anforderungen angeführt sowie damit verbundene Probleme und denkbare Lösungsansätze aufgezeigt werden. Dabei ist es in diesem Zusammenhang nicht möglich, das gesamte lexikographische Problemspektrum in seiner Komplexität zu erfassen.

3.1 Der Standardisierungs- und Multiple-Media-Aspekt

Die beiden Anforderungen des oben beschriebenen Publikationsmodells, Langlebigkeit der Datenhaltung und Flexibilität hinsichtlich der verschiedenen Präsentationsmedien, sind auch bei lexikographischen Projekten von zentraler Bedeutung. Die Inhalte und ihre Präsentation müssen klar voneinander getrennt, d.h. die Ebenen der konzeptuellen Inhaltsstrukturmodellierung und -repräsentation müssen eindeutig gegen die Benutzer- und Redaktionsebene abgegrenzt sein.

Die Ebene der konzeptuellen Inhaltsstrukturmodellierung ist der Kern, aus dem sowohl elektronische wie gedruckte Wörterbücher entstehen sollen. Diese beiden Präsentationsformen haben sehr unterschiedliche Charakteristika. Dazu zählt die nötige Verdichtung der

Angaben im Printwörterbuch gegenüber ihrer möglichen Auflösung im elektronischen Medium. Die Schwierigkeit besteht darin, dass die gängige Form der Textverdichtung in Wörterbuchartikeln nur in Teilen automatisiert werden kann. Wenn genau diese Form der Verdichtung beibehalten werden soll, müssen Verdichtungen in der Inhaltsstrukturmodellierung mit abgebildet werden. Dies widerspricht jedoch dem Grundprinzip der Inhaltsstrukturmodellierung. Ein Lösungsansatz wäre, dieses Grundprinzip aufzuweichen und eine redundante Datenhaltung der verdichteten und ausführlichen Form zu ermöglichen. Ein anderer wäre, eine mehrschichtige Textannotation anzustreben, welche die verdichteten Formen als separate Schicht über die Inhaltsstrukturmodellierung legt. Damit würde die Trennung zwischen Inhalt und ihrer verdichteten Präsentation bestehen bleiben.[23] Neben diesen Lösungsansätzen gilt es jedoch generell zu überlegen, ob die traditionelle Form der Verdichtung zwingend beibehalten werden muss, oder ob nicht vielmehr eine automatisierbare Form von Verdichtung anzustreben ist.

Weitere Problembereiche sind Skopus- und Adressierungsangaben. Diese sind im Print durch die lineare Anordnung im Druckraum repräsentiert. Diese Art der Repräsentation – wie z.B. der ausgelagerte Formkommentar, der sich auf alle folgenden Lesarten bezieht, in Teilen jedoch bei einzelnen Lesarten eingeschränkt werden kann – eignet sich nicht für eine Inhaltsstrukturmodellierung. Hier müssten bei jeder Lesart alle relevanten Informationen des Formkommentars zu finden sein. Auch für die praktische Umsetzung in ein elektronisches Produkt wird dies wichtig, da durch gezielte Zugriffsmöglichkeiten auf eine einzelne Lesart alle relevanten Angaben zum Formkommentar abrufbar sein müssen. Hier wäre ein möglicher Lösungsansatz, der Inhaltsstrukturmodellierung einen sememspezifischen statt einen polysemistischen Zeichenbegriff zugrunde zu legen. Auf der Benutzerebene wären dann sowohl die Auslagerung des Formkommentars möglich, als auch die Darstellung der spezifischen Formangaben bei der einzelnen Lesart.

3.2 Der Modularitäts- und Flexibilitätsaspekt

Das Prinzip der Modularität, wie es als Grundprinzip in allen TEI-DTDs umgesetzt ist, lässt sich auf Wörterbuchprojekte übertragen. Man wird es immer dann zu einer Anforderung an eine Inhaltsstrukturmodellierung machen, wenn es um umfassende Projekte geht, bei denen eine Struktur in gleicher oder leicht veränderter Form in mehreren Zusammenhängen vorkommt. Wenn sich beispielsweise in einem Wörterbuchnetz die Formangaben zu einer Lesart in den verschiedenen Wörterbuchtypen gleichen, könnte der Formkommentar als Modul definiert werden, das dann von verschiedenen Stellen aus referenziert werden kann. Eine anfallende Änderung in der Struktur des Formkommentars muss dann nur noch an einer Stelle vorgenommen werden, ist aber an mehreren Stellen wirksam. Denkt man dieses hier beispielhaft vorgestellte Prinzip weiter, kommt man zu dem Konzept einer DTD-Bibliothek, in der DTD-Module vorgehalten werden, die bei der Modellierung der unterschiedlichen Wörterbuchtypen verwendet werden können. Dadurch entstehen wörterbuchübergreifend konsistente und gleichzeitig flexible Inhaltsstrukturen.

Für einen Wörterbuchverbund ergibt sich bei der elektronischen Umsetzung daraus der Vorteil, wörterbuchübergreifende Zugriffsstrukturen realisieren zu können. Auf Benutzerebene bedeutet dies eine einheitliche Benutzerschnittstelle und Funktionalität über mehreren Wörterbüchern oder auch Wörterbuchtypen. Dabei ist es unwesentlich, ob die ver-

[23] Zu mehrschichtiger Textannotation vgl. u.a. Alexa/Schmidt (1999).

schiedenen und verschiedenartigen Wörterbücher über die Benutzeroberfläche klar voneinander unterschieden sind, oder ob beim Benutzer das Integrat als völlig neuartiges Wörterbuch erscheint. Die Bandbreite dieser Möglichkeiten ist dabei eng gekoppelt an die Detailliertheit und Art der konzeptuellen Inhaltsstrukturmodellierung.

In einem Wörterbuchnetz kann durch ein modulares System Konsistenz gewährleistet werden. Gleichzeitig verlangen verschiedene Wörterbuchtypen aber auch die Möglichkeit, einzelne Module variieren zu können. Diese Variationsmöglichkeiten müssen daher Teil des modularen Konzepts sein. Ein solches System steht folglich in einem Spannungsverhältnis von Konsistenzanspruch und Flexibilisierungsanforderungen. Dabei sollten die Änderungsmöglichkeiten der Struktur weder für alle Elemente gelten – wie dies bei der TEI der Fall ist – noch so stark eingeschränkt sein, dass ein Modul nur sehr begrenzt eingesetzt werden kann und statt dessen immer wieder neue, aber nur leicht variierte Strukturen entwickelt werden. In beiden Fällen geht dies auf Kosten konsistenter Inhaltsstrukturen. In der Konzeptionsphase ist es daher wichtig, Modularität und Flexibilität in ein ausgewogenes Verhältnis zu setzen.

3.3 Der Aspekt der Inhaltsstrukturanalyse

Für die Entwicklung eines lexikographischen Modells benötigt man auf der Ebene der Dokumentanalyse[24] eine Methode, um die konkreten Daten klassifizieren zu können. Im Hinblick darauf soll die in 2.3 vorgestellte mikrostrukturelle Analysemethode von H. E. Wiegand betrachtet werden.

Da die Theorie lexikographischer Texte und damit auch die Analysemethode der funktional-positionalen Segmentation an Printwörterbüchern entwickelt wurde, ist zu prüfen, ob diese Methode unverändert für eine Inhaltsstrukturmodellierung mit den o. g. Anforderungen übernommen werden kann. Eine der oben ausgeführten Anforderungen ist, dass die Inhaltsstrukturmodellierung unabhängig von der Präsentation sein muss. Der positionale Aspekt darf in der Struktur somit nicht abgebildet werden; damit entfällt die Präzedenzrelation. Darüber hinaus muss der funktionale Aspekt unabhängig von seiner Repräsentation im Druckraum betrachtet werden; er bildet nunmehr ausschließlich die Aufgliederung der Daten in inhaltliche Einheiten ab.

Obwohl die Präzedenzrelation vollständig wegfällt, ist zu bedenken, dass durch sie im gedruckten Wörterbuchartikel nicht nur eine Reihenfolge ausgedrückt wird, sondern auch inhaltliche Zusammenhänge verdeutlicht werden, beispielsweise Skopus- und Adressierungsbeziehungen. Da es bei der Inhaltsstrukturmodellierung jedoch gerade um die Abbildung inhaltlicher Zusammenhänge geht, darf dieser Aspekt der Vorgänger-Nachfolger-Relation nicht verloren gehen, sondern muss durch eine andere Relation ersetzt werden. Diese Relation muss, unabhängig von Präsentationsgesichtspunkten, die inhaltlichen Zusammenhänge abbilden. Unter inhaltlichen Zusammenhängen verstehen wir Beziehungen aller Art, die zwischen inhaltlichen Einheiten bestehen, also beispielsweise der Bezug eines Beispiels auf die dazugehörige Einzelbedeutung oder paradigmatische Relationen zwischen einzelnen Lesarten.

Ein Ansatz für eine Analysemethode ist daher eine funktionale Aufgliederung in Informationseinheiten, die durch typisierte Relationen zueinander in Beziehung gesetzt werden. Diesen vielfältigen Verknüpfungsmöglichkeiten stehen unterschiedliche Relationstypen

---

[24] Vgl. Fußnote 9.

gegenüber, die formal klar voneinander unterschieden werden müssen. Diese Relationstypen können in einer Publikationsumgebung ausgedrückt werden durch:

- hierarchische Schachtelungen der DTD
- Elementnamen in der DTD
- hypertextuelle Verweise
- Objektnetze

Beispielsweise kann durch die hierarchische Schachtelung der SGML-Struktur ausgedrückt werden, welche Informationseinheiten zu einer Einzelbedeutung gehören; die Benennung der Elemente kann zusätzlich noch deutlich machen, in welchen Relationen diese zur Einzelbedeutung stehen. Paradigmatische Relationen sind dagegen als hypertextuelle Verweise oder als Objektnetze[25] denkbar.

Mit der Weiterentwicklung und Formalisierung dieses Ansatzes wäre eine Voraussetzung für eine konsistente und projektübergreifend nachvollziehbare Inhaltsstrukturmodellierung gegeben.

4 Vom Ansatz zum Modell

Die wichtigsten Punkte unseres neuen Ansatzes sind die Entwicklung eines Publikationsmodells, das die Ebene der Inhaltsstruktur klar von der der Redaktion und der des Benutzers trennt und die Herausarbeitung der zentralen Rolle, die dabei der Inhaltsstrukturmodellierung zukommt; sie muss die Komplexität der vorgegebenen Inhalte abbilden. Mit dem Konzept von Inhaltsstrukturmodellierung, welches wir zugrunde legen, geht keine Standardisierung im Sinne von Vereinfachung einher, sondern die konsistente Erfassung der Komplexität der Inhalte. Dadurch können aus einer Informationsbasis verschiedene hochwertige Produkte auf unterschiedlichen Medien entstehen, die den Qualitätsanforderungen des jeweiligen Trägermediums gerecht werden. Printwörterbücher werden schon lange den Anforderungen des Buchmediums gerecht; sie einfach auf das elektronische Medium zu übertragen, negiert die diesem Medium eigenen Qualitätsanforderungen.

Ein langfristig anzustrebendes Ziel ist es, aus diesem Ansatz heraus ein lexikographisches Modell zu entwickeln, das den Anforderungen der heutigen Medienlandschaft gerecht wird.

5 Literatur

Alexa, Melina; Schmidt Ingrid (1999): Modell einer mehrschichtigen Textannotation für die computerunterstützte Textanalyse. In: Möhr, Wiebke; Schmidt, Ingrid (Hg.): SGML und XML. Anwendungen und Perspektiven. Heidelberg, 323–345.

[25] Zu Objektnetzen vgl. u. a. Rosteck/Möhr, Fischer (1994); Kamps/Hüser/Möhr/Schmidt (1996); Topic Maps (1999). Auf eine weitere Beschreibung und Ausdifferenzierung der einzelnen Realisierungsmöglichkeiten muss in diesem Zusammenhang verzichtet werden. Dies wäre ein Thema für einen separaten Beitrag.

Bergenholtz, Henning; Tarp, Sven; Wiegand, Herbert Ernst (1999): Datendistributionsstrukturen, Makro- und Mikrostrukturen in neueren Fachwörterbüchern. In: Hoffman, Lothar; Kalverkämper, Hartwig, Wiegand, Herbert Ernst: Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. 2. Halbband. Berlin, New York 1999, 1762–1832.

Breidt, Elisabeth (1998): Neuartige Wörterbücher für Mensch und Maschine: Wörterbuchdatenbanken in COMPASS. In: Wiegand, H. E. (Hg.): Wörterbücher in der Diskussion III. Tübingen (Lexicographica, Series Maior), 1–26.

CCSD: The COLLINS COBUILD STUDENT'S DICTIONARY Online. http://www.linguistics.ruhr-uni-bochum.de/ccsd .

Cover, Robin (1999): The SGML/XML Web Page by Robin Cover. http://www.oasisopen.org/cover/sgml-xml.html.

Feldweg, Helmut (1997): Wörterbücher und neue Medien: Alter Wein in neuen Schläuchen? Zeitschrift für Literaturwissenschaft und Linguistik 107 (1997), 110–123.

– (1997a): COMPASS: Ein intelligentes Wörterbuchsystem für das Lesen fremdsprachiger Texte. http://www.sfs.nphil.uni-tuebingen.de/Compass/Info-dt.html.

FWA 99: Der FISCHER WELTALMANACH 1999. Hrsg. von Mario von Baratta. Frankfurt 1998.

FWA 99 CD-ROM: Der digitale Fischer Weltalmanach 1999. München 1999.

Ide, Nancy; Véronis, Jean (1995): Encoding Dictionaries. In: Ide, Nancy; Véronis Jean (Hg.): Text Encoding Initiative. Background and Context. (Reprint from Computers and the Humanities, Volume 29, Nos. 1,2 & 3 [1995]), Dordrecht, 167–179.

– und Sperberg-McQueen, C. M. (1995): The TEI: History, Goales, and Future. In: Ide, Nancy; Véronis Jean (Hg.): Text Encoding Initiative. Background and Context. (Reprint from Computers and the Humanities, Volume 29, Nos. 1,2 & 3 [1995]), Dordrecht, 5–15.

Kamps, Thomas; Hüser Christoph; Möhr, Wiebke; Schmidt, Ingrid (1996): Knowledge-based information access for hypermedia reference works: Exploring the spread of the Bauhaus movement. In: Agosti, Maristella; Smeaton, Alan F. (Eds): Information Retrieval and Hypertext. Boston, 225–256.

– und Obermeier, Christoph; Reichenberger, Klaus; Schmidt, Ingrid (1999): SGML für dynamische Publikationen – das Beispiel Fischer Weltalmanach. In: Möhr, Wiebke; Schmidt, Ingrid (Hg.): SGML und XML. Anwendungen und Perspektiven. Heidelberg, 173–192.

LEKSIS: Homepage. http://www.ids-mannheim.de/wiw/.

OALD: Oxford Advanced Learner's Dictionary of Current English. Hrsg. von Hornby, A. S., zusammen mit Cowie, A. P. und Lewis J. Windsor. London 1974.

Rostek, Lothar; Möhr, Wiebke; Fischer, Dietrich (1994): Weaving a web: The structure and creation of an object network representing an electronic reference work. In: Electronic Publishing 6 (1994), 495–505.

Sperberg-McQueen, C. M.; Burnard L.(Hg.) (1994): Guidelines for Electronic Text Encoding and Interchange. Chicago, Oxford.

Storrer, Angelika (1996): Metalexikographische Methoden in der Computerlexikographie. In: Wiegand, H. E. (Hg.): Wörterbücher in der Diskussion II. Tübingen (Lexicographica, Series Maior), 239–255.

TEI: Webpage. http://www.tei-c.org/.

Topic Maps (1999): Topic Maps. ISO/IEC 13250, April 8, 1999 (Final Draft).

Uszkoreit, Hans (1998): Sprachtechnologie für die Wissensgesellschaft: Herausforderungen und Chancen für die Computerlinguistik und die theoretische Sprachwissenschaft (Manuskriptfassung). Erschienen in: Meyer-Krahmer, F.; Lange, S. (Hg.): Geisteswissenschaften und Innovationen. 1999.

Wiegand, Herbert Ernst (1989a): Der Begriff der Mikrostruktur: Geschichte, Probleme, Perspektiven. In: Wörterbücher. Ein internationales Handbuch zur Lexikographie. 1. Teilbd. Hrsg. von Hausmann, Franz Josef; Reichmann, Oskar; Wiegand, Herbert Ernst; Zgusta, Ladislav. Berlin, New York (Handbücher zur Sprach- und Kommunikationswissenschaft 5.1), 409–462.

– (1989b): Arten von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. In: Wörterbücher. Ein internationales Handbuch zur Lexikographie, 1. Teilbd. Hrsg. von Hausmann, Franz Josef;

Reichmann, Oskar; Wiegand, Herbert Ernst; Zgusta, Ladislav. Berlin, New York (Handbücher zur Sprach- und Kommunikationswissenschaft 5.1), 462–501.
- (1996): Über die Mediostrukturen bei gedruckten Wörterbüchern. In: Symposium on Lexicography VII. Proceedings of the Seventh Symposium on Lexicography May 5–6, 1994 at the University of Copenhagen. Ed. by Arne Zettersten, Viggo Hjørnager. Tübingen 1996 (Lexicographica, Series Maior), 11–43.

(Alle Web-Seiten wurden am 15. Juli 1999 zum letzten Mal überprüft.)

Ingrid Schmidt, Heidelberg
Carolin Müller, Heidelberg

Angelika Storrer

Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie

1 Wörterbücher der Zukunft – alte und neue Visionen

„Alles Menschenwerk ist unvollkommen. Zu den unvollkommensten Erzeugnissen des Menschen gehören aber unstreitig die Wörterbücher". So beginnt ein Aufsatz, der 1910 unter dem Titel „Wörterbücher der Zukunft" in der Germanisch-Romanischen Monatszeitschrift erschienen ist (Tiktin [1910, 243]). Der Autor kritisiert darin sehr überzeugend die Unzulänglichkeiten der seinerzeit verfügbaren Wörterbücher und entwirft Konzepte zu deren Verbesserung: Bilder und Illustrationen sollen die Bedeutungserläuterungen anschaulicher und verständlicher machen. Sinngemäß Zusammengehöriges soll auch räumlich beieinander stehen, z.B. sollen bei Nahrungsmitteln alle regional verschiedenen Bezeichnungen aufgeführt werden, die Teile eines Ganzen sollen unter der Bezeichnung für das Ganze zu finden sein. Jede lexikalische Einheit soll bei all den Wortgruppen verzeichnet werden, zu denen sie ihrer Bedeutung nach gehört, sodass man sowohl vom Wort zum Begriff als auch vom Begriff zum Wort gelangen kann. Das Ansinnen, ein kleiner Kreis von Philologen könne für alle Wortschatzbereiche adäquate Bedeutungserläuterungen erstellen, hält Tiktin für ein „Ding der Unmöglichkeit" und empfiehlt deshalb, „das segensreiche Prinzip der Arbeitsteilung" auch für die Lexikographie nutzbar zu machen und fachbezogenen Wortschatz von Lexikographen beschreiben zu lassen, die sich im jeweiligen Fachgebiet auskennen (Tiktin [1910, 248]).

Das von Tiktin konzipierte Wörterbuch der Zukunft gibt es auch im Jahr 2000 noch nicht. Die inzwischen verfügbare Computertechnik bietet allerdings Anlass, bislang als unrealistisch verworfene Konzeptionen für Wörterbücher der Zukunft neu zu prüfen und erneut über die Zukunft der Wörterbücher nachzudenken. Insbesondere die als „Hypertext" bezeichnete digitale Schreib- und Lesetechnologie macht es nun möglich, lexikographische Desiderate und Konzepte zu realisieren, die 1910 nach Tiktins eigener Einschätzung noch einer fernen Zukunft vorbehalten waren. In Hypertexten können Text-, Bild-, Ton- und Videodateien zur anschaulichen Vermittlung lexikalischen Wissens genutzt und die Wörterbuchartikel durch computerisierte Verweise, sog. „Links", verknüpft werden. Lexikalische Daten können so modelliert werden, dass in Abhängigkeit von Nutzerinteressen und Nutzungssituationen die jeweils relevanten lexikographischen Angaben und Verweise herausgegriffen und in ästhetisch ansprechender Weise am Bildschirm dargestellt werden. Das Alphabet als Zugriffsstruktur im gedruckten Medium wird ersetzt durch Suchwerk-

zeuge, mit denen sich solche Datenbanken nach individuell wählbaren Kriterien durchsuchen lassen. Das World Wide Web (WWW), die Hypertext-Plattform im Internet, macht ein solches digitales Wörterbuch nicht nur weltweit zugänglich, sondern unterstützt auch die rasche und unkomplizierte Aktualisierung und Erweiterung der lexikographischen Daten. Die Verbindung von Information und Kommunikation im WWW erleichtert zudem die dezentrale Organisation des lexikographischen Arbeitsprozesses durch verteilt arbeitende Spezialisten.

Nicht zuletzt die Aufsätze in diesem Sammelband zeigen, dass Lexikographie und Metalexikographie beginnen, sich mit den Chancen des Hypertext-Konzepts auseinanderzusetzen.[1] Dennoch sind die derzeit verfügbaren elektronischen Wörterbücher noch stark von den Traditionen der Strukturierung und Gestaltung geprägt, die sich am gedruckten Medium herausgebildet haben. Das Potenzial des neuen Mediums wird erst ansatzweise genutzt. Dies gilt sowohl für die derzeit im WWW verfügbaren Wörterbücher[2] als auch für die auf CD-ROM vermarkteten Produkte[3].

Der Grund dafür liegt darin, dass es sich bei vielen elektronischen Wörterbüchern um Parallelpublikationen zu gedruckten Wörterbüchern handelt und dass es immer noch das Print-Medium ist, an dem die Verlagslexikographie vornehmlich verdient (vgl. Klosa [in diesem Band]). Eine mediengerechte Hypertextualisierung eines Wörterbuchs ist aufwändig und muss sich rechnen; sie kann also unter kommerziellen Randbedingungen nur schrittweise erfolgen. Aufgabe der metalexikographischen Forschung muss es trotzdem sein, Zielvorstellungen für digitale Wörterbücher zu entwickeln, bei denen die Möglichkeiten und Chancen des neuen Mediums optimal genutzt werden. Solche Zielvorstellungen und deren prototypische Implementierung ermöglichen es kommerziellen und wissenschaftlichen Wörterbuchprojekten, die sukzessive Hypertextualisierung ihrer Bestände zu planen bzw. den Aufbau neuer Datenbestände mit den vorhandenen Ressourcen und Rahmenbedingungen möglichst optimal zu gestalten.

Mit den in Abschnitt 3 aufgestellten sieben Thesen möchte ich eine Diskussion über solche Zielvorstellungen anstoßen. Die Thesen greifen Aspekte auf, die in der Literatur zu elektronischen Wörterbüchern bereits diskutiert werden. Die Bündelung und thesenartige Zuspitzung dieser Aspekte erscheint mir aber wichtig zu sein, um künftige Diskussionen zu diesem Thema zu strukturieren, sodass sich die kontroversen Punkte auf Dauer von den weniger kontroversen unterscheiden lassen. Um die Thesen auch für Leser plausibel zu machen, die mit dem Hypertext-Konzept bislang nicht oder nur wenig vertraut sind, werde ich im folgenden Abschnitt kurz die lexikographisch relevanten Facetten des Konzepts erläutern und Hinweise auf weiterführende Literatur geben.

2 Was macht das Hypertextkonzept für die Lexikographie interessant?

Nachschlagewerke und speziell Wörterbücher werden gerne als Vergleichsobjekte herangezogen, wenn es darum geht, die für Hypertext typische nicht-lineare Organisationsform zu erläutern. Das Wörterverzeichnis von Wörterbüchern besteht aus einer Abfolge von Wör-

---

[1] Vgl. Lemberg, Müller/Schmidt, Petelenz, Richter, Haß-Zumkehr (alle in diesem Band).
[2] Vgl. Storrer/Freese (1996), Lemberg (in diesem Band).
[3] Vgl. Feldweg (1997), Storrer (1998).

terbuchartikeln, die in beliebiger Abfolge und partiell rezipiert werden können. Entsprechend besteht auch ein Hypertext aus einer Menge von Modulen (auch „Knoten" oder „informationelle Einheiten" genannt), die in einer vom Benutzer selbst gewählten Abfolge am Bildschirm angezeigt werden. So wie die Wörterbuchartikel durch Verweise miteinander verknüpft sind, so sind auch die Module im Hypertext durch sog. Hyperlinks (im Weiteren „Links" genannt) miteinander verbunden.

Diese Analogie von Wörterbuch und Hypertext hat sicher dazu beigetragen, dass schon sehr früh Wörterbücher in Hypertexte überführt wurden – das 1988 hypertextualisierte OED kann dabei bis heute als Leitprojekt gelten (vgl. Raymond/Tompa [1988]). Bei allen Gemeinsamkeiten dürfen jedoch die zentralen Unterschiede zwischen Print- und Hypertext-Wörterbüchern nicht übersehen werden:

1. Module in Hypertext können im Gegensatz zu den statischen und stabilen Wörterbuchartikeln im Buch bei jeder Nachschlagehandlung neu zusammengesetzt werden, und zwar unter Berücksichtigung von situationsabhängigen Parametern wie der Muttersprache des Nutzers, seines Sprachlevels, dem Typ der Benutzungssituation.[4]
2. Ebenso können Links nicht nur, wie die gedruckten Verweise, vom Autor fest verdrahtet sein, sondern können dynamisch nach Programmanweisung generiert werden und dabei in Abhängigkeit von situationsspezifischen Parametern ihre Eigenschaften verändern.
3. Hypertextsysteme enthalten Such- und Recherchewerkzeuge, mit denen sich große Datenmengen nach selbst gewählten Kriterien im Nu auf eine Weise durchsuchen lassen, die im gedruckten Medium undenkbar oder zumindest sehr zeitaufwändig wäre.
4. In Hypertexte können mediale Objekte wie Ton-, Bewegtbild- und Videodateien eingebunden werden, was das Spektrum der Beschreibungs- und Darstellungsmöglichkeiten für den Wörterbuchnutzer erheblich verbreitert.

Es sind im Wesentlichen drei Merkmale, die Hypertexte von anderen Medien, speziell vom gedruckten Buch, unterscheiden: Hypertexte sind computerverwaltete Texte, Hypertexte erlauben die Mehrfachkodierung von Daten in verschiedenen Symbolsystemen und deren Übermittlung auf mehreren Sinneskanälen, Hypertexte sind nicht-linear organisiert. Im Folgenden möchte ich diese drei Merkmale kurz erläutern und zeigen, was das Hypertext-Konzept für künftige lexikographische Produkte interessant macht.

2.1 Verwaltung durch ein Hypertextsystem

Sowohl zur Produktion als auch zur Rezeption von Hypertexten wird Software benötigt, die man als Hypertextsysteme bezeichnet. Erst Hypertextsysteme stellen die Werkzeuge bereit, um im Netzwerk der Module und Links zu navigieren, Module zu durchsuchen, zu verändern, ggf. vorhandene Tondateien abzuhören und Videosequenzen abzuspielen. Sie ermöglichen es, Module nach den individuellen Vorgaben des Nutzers zusammenzustellen und am Bildschirm anzuzeigen sowie Daten in Abhängigkeit von bestimmten Nutzungssituationen zu filtern. Die Bestimmung von Hypertext als Text, der sich nicht ohne Wertverlust auf Papier ausdrucken lässt, findet sich deshalb zurecht in verschiedensten Hypertext-Definitionen wieder[5]. Das Merkmal ist notwendig, um Hypertexte vom sog. „Text-

---

[4] Vgl. z.B. Petelenz (in diesem Band), Richter (in diesem Band), Thielen/Breidt et al. (1998) und die These 4 in Abschnitt 3.

[5] Z.B. Nelson (1972), Slatin (1991, 56).

design" im gedruckten Medium einerseits, vom computerverwalteten E-Text andererseits abzugrenzen:

- Als *Textdesign* bezeichnet man nach Blum/Bucher (1998) nicht-lineare Gestaltungsformen in gedruckten Zeitungen und Zeitschriften, die inzwischen auch zunehmend in andere Printpublikationen, v.a. Hand- und Lehrbüchern, Einzug halten. Textdesign verfolgt wie Hypertext das Ziel, die partielle und selektive Lektüre zu erleichtern, indem Informationen in kleinere Module zerlegt und auf den Seiten zu Clustern zusammengefügt werden. Dennoch bleiben Textdesign-Produkte stabile, physisch greifbare und begrenzbare Ganzheiten, in denen der Textdesigner über die Zusammenstellung der Module auf den Seiten entscheidet. Die zentrale Idee von Hypertext besteht dagegen genau darin, dass der Nutzer selbst bestimmt, welche Module er sich in welcher Reihenfolge und Anordnung auf dem Bildschirm anzeigen lassen möchte, wobei – wie oben bereits erwähnt – die Module und Links überhaupt nur „virtuell" existieren können, d.h., erst bei Bedarf vom System aus kleineren Textbauteilen zusammengesetzt werden.
- Als *E-Texte (elektronische Texte)*[6] bezeichnet man Texte, die in das World Wide Web eingebunden sind, ohne dass sie die für Hypertext charakteristische Organisationsform aufweisen. *E-Texte* sind häufig Parallel- oder Vorversionen von Print-Publikationen, die v.a. die schnelle und unkomplizierte Publikationsmöglichkeit des Internet nutzen; sie können ohne Wertverlust ausgedruckt und auf Papier gelesen werden.

Im Gegensatz zu E-Texten lassen sich Hypertexte nur mit Hilfe eines digitalen Lesegerätes rezipieren. Hierzu dienen bislang vornehmlich Computerbildschirme, bei denen – beim aktuellen Stand der Technik – erhebliche Abstriche an den Lesekomfort gemacht werden müssen. Allerdings zeigt die wachsende Akzeptanz von sog. „eBooks" und der vorhersehbare Erfolg des internetfähigen Handys, dass man sich nicht voreilig der von Computerskeptikern oft vorgebrachten Auffassung anschließen sollte, Hypertext würde sich wegen des fehlenden Lesekomforts nicht durchsetzen. Erstens hat Bolter (1991, S. 4) aus medienhistorischer Perspektive zurecht darauf hingewiesen, dass Nutzungskomfort nicht das einzige Kriterium für die erfolgreiche Einführung neuer Medium darstellt, dass auch Bücher lange überwiegend auf Lesepulte gelegt und im Stehen rezipiert wurden, dass sich also auch das Medium „Buch" nicht deshalb durchgesetzt hat, weil man es bequem auf dem Sofa lesen konnte. Außerdem ist absehbar, dass hinlänglich kostengünstige und gut transportable, digitale Lesegeräte mit papieräquivalenter Auflösung über kurz oder lang zu erschwinglichen Preisen auf dem Markt verfügbar sein werden. Es bedarf keiner besonderen prophetischen Gabe, um vorherzusagen, dass diese Entwicklungen den Bereich der Nachschlagewerke nachhaltig verändern werden.

Ein Nebeneffekt der Computerverwaltung sollte gerade für die lexikographische Anwendung nicht unterschätzt werden: Weil Druckraum teuer ist, konnte im Printwörterbuch immer nur ein sehr kleiner Teil potenziell verfügbarer Belege zu einem Lemma im zugehörigen Artikel berücksichtigt werden. Speicherplatz ist im Computer hingegen kein vergleichbar teurer Faktor; gerade bei Textdateien kann also ruhig eine Angabe mehr gemacht, ein Beispiel mehr gegeben werden (vgl. dazu auch These 3 in Abschnitt 3). Insgesamt können lexikographische Angaben verständlicher formuliert werden: Techniken zur Abkürzung und Textverdichtung, die in der Printlexikographie sinnvoll waren, um möglichst viele Daten auf möglichst knappem Raum unterzubringen,[7] sind im digitalen Medium überflüssig.

---

[6] Die Bezeichnung E-Text habe ich übernommen von Dieter E. Zimmers Artikelserie „Die digitale Bibliothek", vgl. http://www.zeit.de/digbib/ .

Sie müssen sogar explizit rückgängig gemacht werden, um die lexikographischen Beschreibungen in „intelligente" benutzeradaptive Wörterbuchsysteme zu integrieren (vgl. Breidt [1998], Thielen/Breidt et al. [1998]). Die Anforderung, kurz und prägnant abgefasste Wörterbuchartikel zu schreiben, verschiebt sich dadurch hin zu der Notwendigkeit, im Hypertextsystem Such- und Navigationswerkzeuge bereitzustellen, mit denen man aus einer großen Menge von Angaben diejenigen herausfiltern kann, die in einer aktuellen Benutzungssituation benötigt werden. Während ein kundiger Wörterbuchbenutzer bislang vor allem lernen musste, verdichtete Wörterbuchtexte korrekt zu entschlüsseln und souverän die vorhandenen inneren und äußeren Zugriffsstrukturen zu nutzen, muss der Nutzer digitaler Wörterbücher die Funktionalität seiner Such- und Filterwerkzeuge beherrschen lernen.

2.2 Mehrfachkodiertheit und Synästhetisierung

Hypertextsysteme können verschiedene mediale Objekte (Text-, Bild-, Audio- und Videodateien) verwalten, die in den Modulen kombiniert und durch Links verknüpft werden können.[8] Es kann also auf dem visuellen und dem auditiven Kanal kommuniziert werden und dies unter Verwendung unterschiedlicher Symbolsysteme. Die im gedruckten Medium dominante Schrift lässt sich also nicht nur um Bilder und Grafiken anreichern, sondern auch um Ton- und Videodokumente. Hypertext-Gestaltung heißt, sich bewusst für ein- oder mehrkanalige Informationsvermittlung, für Schrift, Bild, Ton oder Video, zu entscheiden und aus den verschiedenen Elementen ein Ensemble zu flechten, das auf die Rezeption am Bildschirm und deshalb auf eine ganzheitliche Wahrnehmung als Bild hin ausgelegt ist (vgl. Schmitz [1997]). Diese Verflechtung von Schrift, Bild, Ton und Bewegung hat Freisler (1994, 31) als den Synästhetisierungsaspekt von Hypertext bezeichnet.

Der Synästhetisierungsaspekt erlaubt es, Wörterbücher nicht nur zweckmäßig, sondern auch ästhetisch ansprechend zu gestalten. Von lexikographischem Interesse sind aber vor allem die neuen Optionen zur mehrkanaligen Informationsvermittlung.[9] Eine erste, vom Nutzwert aber bereits sehr wertvolle Anreicherung rein textbasierter Wörterbücher besteht darin, phonetische Umschriften um Audiodateien mit vertonten Ausspracheangaben zu ergänzen. Dies bietet sich vor allem für bilinguale und Lerner-Wörterbücher an und kommt dabei den Nutzern entgegen, die phonetische Umschriften nicht oder nur schwer lesen können. Vertonte Ausspracheangaben sind jedoch erst ein erster Schritt hin zu Wörterbüchern, die – ganz im Sinne der in Tiktin 1910 erhobenen Forderung nach besserer Anschaulichkeit – Text-, Bild-, Ton- und Videoelemente nutzen, um die Verständlichkeit und Anschaulichkeit von Bedeutungserläuterungen zu verbessern. Die MS-Bookshelf (1995) hat in ihren Wörterbucheintrag zu **Flamenco** beispielsweise ein Video integriert, das nicht nur die für den Tanz typischen Bewegungen, sondern auch den Rhythmus der dazugehörigen Musik vermittelt. Hypertext-Enzyklopädien wie MICROSOFT ENCARTA vermitteln einen ersten Eindruck, wie Bilder und Videos, aber auch sukzessiv aufgebaute und durch gesprochene Sprache erläuterte Informationsgrafiken, sog. Animationen, zur Erläuterung

---

[7] Zur Textverdichtung im gedruckten Wörterbuch vgl. Wiegand (1987, 37ff), Wiegand (1996).

[8] Da inzwischen fast alle Hypertextsysteme verschiedene mediale Objekte verwalten, verzichte ich auf die früher übliche terminologische Unterscheidung zwischen rein textbasiertem „Hypertext" und multimedialem „Hypermedia".

[9] Vgl. dazu Storrer (1998), Petelenz (in diesem Band); zum Einsatz in der wissenschaftlichen Lexikographie: Lemberg (in diesem Band), Richter (in diesem Band), Schröder (1997).

komplizierter Sachverhalte und Abläufe eingesetzt werden können. Hier eröffnen sich v.a. für die Fachsprachen-Lexikographie interessante Perspektiven.

Ein Online-Wörterbuch, an dem sich der Wert der Mehrfachkodiertheit sehr eindrücklich zeigen lässt, ist das FACHGEBÄRDENLEXIKON PSYCHOLOGIE, das am Zentrum für Gebärdensprache der Universität Hamburg entwickelt wurde.[10] In den Artikeln dieses Wörterbuchs sind schriftliche Bedeutungserläuterungen verknüpft mit Abbildungen von Handformen sowie mit Videos, auf denen die Gebärden vorgeführt werden. Jeder Artikel enthält nicht nur Links zu verwandten Fachbegriffen, sondern verknüpft auch Gebärden, die durch formähnliche Gesten ausgedrückt werden. Der Nutzer kann wahlweise alphabetisch oder über einen Sachindex einstiegen, er kann aber auch nach der Form einer bestimmten Gebärde suchen. Eine schön gemachte Umsetzung von Tiktins Anregung für das „Wörterbuch der Zukunft", dass „bei jedem Tiernamen gesagt würde, wie vom menschlichen Ohre der Schrei des betreffenden Tieres aufgefasst wird, durch welche Lautfolge die menschliche Sprache ihn wiederzugeben pflegt" (Tiktin 1910, 253), ist das multimediale Tiersprachenlexikon SOUNDS OF THE WORLD'S ANIMALS.[11] In ihm ist verzeichnet, wie Tierlaute in verschiedenen Sprachen der Welt verschriftet werden – wobei Proben der Originallaute als Audiodateien abgehört werden können.

2.3 Nicht-lineare Organisationsform

Die Grundidee der nicht-linearen Textorganisation lässt sich folgendermaßen skizzieren[12]: Der Autor eines Hypertextes verteilt seine Daten auf Module, die durch Links miteinander verknüpft sind. Metaphorisch gesprochen entsteht ein Wegenetz, mit den Links als Wegverbindungen zwischen den Modulen als den Orten, an denen Daten gespeichert sind. Die Verweisverfolgung geschieht durch das Aktivieren von Linkanzeigern, die als Schaltflächen, sensitive Wörter oder sensitive Graphiken gestaltet sein können. Ein Mausklick auf einen Linkanzeiger in einem Modul A führt dazu, dass ein damit verbundenes Modul B angezeigt wird.

Die nicht-lineare Organisationsform unterscheidet Hypertexte einerseits von zeitlich-linearen Audio- und Videodokumenten; andererseits von gedruckten Dokumenten, deren Struktur durch Teil-Ganzes- und Vorgänger-Nachfolger-Beziehungen zwischen den Dokumententeilen dominiert wird. Die Vorteile für lexikographische Anwendungen liegen auf der Hand und wurden bereits erwähnt: Während ein gedrucktes Wörterbuch eine sequenzielle Anordnung der Textsegmente im Buch erzwingt[13], erlauben es Hypertextsysteme, verschiedene gleichwertige Zugriffsmöglichkeiten auf die Daten anzubieten. Dies kommt der selektiven und problembezogenen Textrezeption entgegen, wie sie für Wörterbuchbenutzungssituationen typisch ist. Wörterbücher, ob gedruckt oder digital, werden meist ausschnittsweise rezipiert; welcher Ausschnitt selektiert wird, ist abhängig von der Benutzungssituation und dem dabei verfolgten Zweck. Von einem Wörterbuchartikel zu einem Lemma werden oft nur ganz spezielle Angaben benötigt, z.B. die Ausspracheangabe, die Bedeutungserläuterung oder eine grammatische Angabe. Aber auch die nicht-usuelle Ver-

---

[10] Vgl. http://www.sign-lang.uni-hamburg.de/Projekte/PsychLex.html.

[11] Cathy Ball: SOUNDS OF THE WORLD'S ANIMALS, vgl. http://www.georgetown.edu/cball/animals/animals.html.

[12] Zur Diskussion dieses Merkmals vgl. (Kuhlen [1991, 27ff], Freisler [1994, 21f], Storrer [2000]).

[13] Bei Wörterbüchern haben sich dabei verschiedene „Makrostrukturen" genannte Anordnungsformen herausgebildet (vgl. Wiegand [1989], Wiegand [1998]).

wendung eines Wörterbuchs als Lesebuch zielt nicht auf vollständige Textrezeption ab, wie bereits die Empfehlung Jacob Grimms zeigt: „Leser jedes Standes und Alters sollen auf den unabsehbaren Strecken der Sprache nach Bienenweise nur in die Kräuter und Blumen sich niederlassen, zu denen ihr Hang sie führt und die ihnen behagen" (Grimm (1854, XII). Diesen selektiven und punktuellen Rezeptionsformen kommt die nicht-lineare Organisationsform von Hypertext optimal entgegen.

Ein weiterer Vorteil der nicht-linearen Organisationsform liegt darin, dass Textpassagen, die in der Printwelt in verschiedenen Büchern publiziert waren, nun über Links verknüpft werden können. Ein Beispiel hierfür sind digitale Textverbünde auf CD-ROM mit mehreren Nachschlagewerken wie Wörterbuch, Zitatensammlung, Chronologie und Multimedia-Atlas, die durch Links verknüpft sind. Die Links und die ergänzenden Suchfunktionen beschleunigen das gezielte Nachschlagen und fördern das themenorientierte Herumstöbern nach Informationen über die Grenzen einzelner Nachschlagewerke hinweg. Ein Beispiel aus der wissenschaftlichen Lexikographie ist der VERBUND MITTELHOCHDEUTSCHER WÖRTERBÜCHER, der Nachschlagewerke zum Mittelhochdeutschen verknüpft, die häufig im selben Arbeitskontext benötigt werden und auch explizit aufeinander bezogen sind[14]. Speziell zum amerikanischen Englisch gibt es im World Wide Web kostenlos verfügbare Angebote, in denen verschiedene Nachschlagewerke miteinander verknüpft sind; sehr nützlich gerade für Nicht-Muttersprachler sind der WORDSMITH EDUCATIONAL DICTIONARY-THESAURUS[15] und der MERRIAM-WEBSTER ONLINE[16], die beide jeweils ein alphabetisches Wörterbuch und einen Thesaurus integrieren.

3 Sieben Thesen zur Nutzung des Hypertext-Konzepts in der Lexikographie

Die Chancen des Hypertext-Konzepts für die Lexikographie werden zwar zunehmend erkannt, die Diskussion darüber, wie die Mehrwerte des neuen Mediums optimal genutzt werden können, hat aber erst begonnen. Die nachfolgenden Thesen sollen dazu dienen, diese Diskussion zu strukturieren, die einzelnen, vom Medienwandel betroffenen Veränderungen herauszuarbeiten und die kontroversen Punkte von den weniger kontroversen unterscheiden zu helfen[17]. Die Thesen beziehen sich auf ein ideales digitales Wörterbuch der Zukunft, das im Hinblick auf den technischen Stand im Jahr 2000 neu konzipiert werden kann und dabei keinen speziellen ökonomischen und infrastrukturellen Beschränkungen unterliegt. Es ist mir bewusst, dass sich die meisten existierenden Wörterbuchprojekte nicht in einer solchen idealen Ausgangslage befinden, dass gerade die Verlagslexikographie Zwängen unterliegt, die die Verwirklichung solcher Zielvorstellungen nur schrittweise ermöglicht. Ich halte es dennoch für eine wichtige Aufgabe der Wörterbuchforschung, Ideen und Konzepte für digitale Wörterbücher der Zukunft zu entwickeln und

---

[14] Vgl. http://gaer27.uni-trier.de/MWV-online/MWV-online.html und Burch/Fournier (in diesem Band).

[15] http://www.wordsmyth.net/.

[16] http://www.m-w.com/home.htm.

[17] Eine Vorfassung dieser Thesen habe ich mit den Teilnehmern des Workshops „SGML/XML-Einsatz in der Lexikographie" diskutiert, der am 21.9.1999 an der Heidelberger Akademie der Wissenschaften stattgefunden hat. Die Anregungen aus dieser Diskussion haben zu einer Erweiterung von damals fünf auf die nun erörterten sieben Thesen geführt und mich überhaupt dazu ermutigt, einen derart programmatischen Text zu publizieren.

die qualitative Verbesserung existierender Produkte durch Entwicklung von Forschungsprototypen und durch computergestützte Benutzungsforschung voranzutreiben.

Im vorigen Abschnitt wurde bereits gezeigt, dass die zunächst naheliegende Analogie von Wörterbuch und Hypertext einen zentralen Mehrwert des digitalen Mediums außer Acht lässt: Die Art und Weise, wie Daten in einem Datenbank- oder Hypertextsystem strukturiert sind, muss nicht der Art und Weise entsprechen, wie diese Daten dem Benutzer am digitalen Lesegerät (dem Bildschirm, dem Handy, dem eBook) präsentiert werden. Die Zielsetzung bei der Informationsmodellierung besteht vielmehr gerade darin, Daten so zu strukturieren, dass aus ein und demselben Datenpool für verschiedene Anwendungszwecke und Nutzungskontexte die jeweils relevanten Informationen herausgegriffen und in geeigneter Weise präsentiert werden können.[18] Für die Informationsmodellierung gibt es bei digitalen Wörterbüchern zwei strategische Varianten:

1. Die Modellierung orientiert sich vornehmlich an den Bauteilen und Strukturen eines gedruckten Wörterbuchs. Die zentralen Bauteile sind dann lexikographische Angaben, die grundlegenden Strukturen sind Mikro-, Makro-, Artikel- und Verweisstrukturen, wie sie in der Metalexikographie beschrieben sind.[19] Ein Beispiel für eine solche Modellierung ist die Document Type Definition, wie sie von der Text Encoding Initiative (TEI) für die Auszeichnung von Print-Wörterbüchern vorgeschlagen wurde.[20]
2. Die Modellierung orientiert sich vornehmlich an den Einheiten und Strukturen des Wörterbuchgegenstands, in der Sprachlexikographie also v.a. an lexikalischen Einheiten und an den Relationen zwischen diesen. Die zentralen Bauteile sind dann linguistische Einheiten wie Grapheme, Morpheme und Lexeme; grundlegende Strukturen lassen sich beschreiben durch syntagmatische und paradigmatische Relationen sowie durch Type-Token-Beziehungen zwischen linguistischen Einheiten auf der Äußerungsebene und deren Korrelaten auf der Ebene des Lexikons. Ein Beispiel für eine solche Modellierung ist die lexikalische Datenbank WORDNET, die v.a. semantische Relationen zwischen Konzepten und Lexemen berücksichtigt (vgl. Fellbaum [1998]); in der deutschen „Schwester" GERMANET sind aber auch syntagmatische Beziehungen und Wortbildungsregularitäten erfasst (vgl. Kunze/Wagner [in diesem Band]; Hamp/Feldweg [1997]).

Die erste Variante liegt nahe, wenn ein gedrucktes Wörterbuch möglichst rasch ins digitale Medium überführt werden soll. Sie ist auch adäquat, wenn bei Projekten der sog. retrospektiven Digitalisierung vornehmlich angestrebt wird, eine digitale Kopie eines gedruckten Dokuments herzustellen, bei der die Formeigenschaften der gedruckten Vorlage gewahrt bleiben (vgl. Burch/Fournier [in diesem Band]). Wenn man jedoch die in Abschnitt 2 erläuterten Mehrwerte nutzen möchte, die Hypertext dem gedruckten Buch gegenüber aufweist, dann sollte man sich für die zweite Option entscheiden. Meine erste These lautet also:

---

[18] Vgl. hierzu auch Müller/Schmidt (in diesem Band). In der Datenbanktheorie spricht man in diesem Zusammenhang von Datenunabhängigkeit (vgl. Büchel/Schröder [in diesem Band]). Andere Zielsetzungen bei der Datenmodellierung – die Überprüfbarkeit von Integrität und Konsistenz, die Vermeidung von Updateanomalien und die Redundanzfreiheit – spielen im hier behandelten Zusammenhang eine untergeordnete Rolle.

[19] Diese Strukturen sind sehr präzise erforscht und beschrieben, vgl. Wiegand (1989),Wiegand (1991), Bergenholtz/Tarp et al. (1999).

[20] Abrufbar unter http://www.uic.edu/orgs/tei/p3/doc/p3di.txt; vgl. dazu auch Büchel/Schröder, Müller/Schmidt und Burch/Fournier (alle in diesem Band).

These 1: Die konzeptuelle Datenmodellierung für digitale Wörterbücher sollte sich vornehmlich an linguistischen Einheiten und Strukturen ausrichten und nicht an den Strukturen eines Printwörterbuchs.[21]

Diese These möchte ich folgendermaßen begründen: Es wurde immer wieder beklagt, dass die alphabetische Anordnung der lexikalischen Einheiten im Wörterbuch viele wortschatzinterne Bezüge verdeckt, dass semantisch Zusammengehöriges räumlich auseinandergerissen wird.[22] Diesen Nachteil musste man im gedruckten Buch durch Verweis- und Mikrostruktur sowie durch die Erschließung „versteckter" lexikographischer Informationen in Registern kompensieren (vgl. Goebel/Lemberg/Reichmann [1995]).

Das digitale Medium bietet nun die Möglichkeit, die lexikographische Informationsmodellierung sehr eng an den Strukturen des Lexikons auszurichten, die ja in der modernen Linguistik relativ gut erforscht sind.[23] Dies erfüllt nicht nur die immer wieder gestellte Forderung, lexikologische Forschungsergebnisse stärker als bisher in der praktischen Lexikographie zu berücksichtigen, sondern bringt drei weitere Vorteile mit sich. Erstens müssen bei entsprechender Modellierung viele Eigenschaften lexikalischer Einheiten, speziell in den Bereichen Flexion und Wortbildung, nicht mehr bei jedem Lemma einzeln aufgeführt werden, sondern könnten bei Bedarf über generelle Regeln abgeleitet werden. Zweitens lassen sich auf einem linguistisch motivierten Datenmodell flexiblere Sichten und Filter definieren, die sich auf den Informationsbedarf in bestimmten Nutzungskontexten einstellen (vgl. These 4). Drittens erfordert die Beschreibung linguistischer Merkmale und Strukturen für Datenbank- oder Hypertextsoftware ein höheres Maß an Formalisierung als die Erfassung entsprechender Merkmale im gedruckten Wörterbuch.[24] Bei umsichtiger Modellierung kann man deshalb von vornherein eine lexikalische Datenbank aufbauen, aus der Informationen für menschliche Benutzer einerseits, für Anwendungen der maschinellen Sprachanalyse andererseits herausgegriffen und in jeweils adäquater Form präsentiert werden, ganz im Sinne des in Breidt (1998) beschriebenen Wörterbuchs „für Mensch und Maschine".

Auf der Basis einer linguistisch motivierten Informationsmodellierung kann auch das Nachschlagen aus einem Textverarbeitungsprogramm oder einem Internet-Browser heraus komfortabler und flexibler gestaltet werden. Ein einfaches Beispiel: Es ist für flektierende Sprachen wünschenswert, per Mausklick von einer flektierten Wortform im Text zu dem zugehörigen Eintrag im Wörterbuch zu gelangen. Bislang wird eine solche Funktion von keinem verfügbaren elektronischen Wörterbuch des Deutschen angeboten, obwohl die dafür benötigten sprachtechnologischen Werkzeuge vorhanden sind.[25] Die in Klosa (in diesem Band) geäußerte Replik auf meine diesbezügliche Anregung für die PC-BIBLIO-

---

[21] Eine sinngemäß formulierte These hat Andreas Blumenthal bereits 1987 in einem Workshop zu „Prinzipien des Entwurfs lexikalischer Datenbanken" zur Diskussion gestellt; vgl. auch Blumenthal/Lemnitzer/Storrer (1988).

[22] Vgl. Gloning/Welter (in diesem Band) und die Diskussion um integrierte Wörterbücher in dem sehr lesenswerten Sammelband „Nachdenken über Wörterbücher" (Henne [1977, 47f], Wiegand [1977, 102f], Drosdowski [1977, 126f]). Dem Beitrag von Helmut Henne verdanke ich übrigens auch den Hinweis auf den Aufsatz von H. Tiktin.

[23] Im Sinne des in Lang (1983) beschriebenen Verhältnisses zwischen Lexikon und Wörterbuch.

[24] Dass die Beschreibungen in traditionellen Wörterbüchern unsystematisch und wenig konsistent sind, zeigten nicht zuletzt die Versuche, aus ihnen mit halbautomatischen Verfahren maschinenlesbare Lexika für die maschinelle Sprachverarbeitung zu gewinnen (vgl. dazu Boguraev/Briscoe [1989], Heyn [1992], Storrer/Feldweg et al. [1993]).

[25] Vgl. Feldweg (1997), Storrer (1998).

THEK (Storrer [1996]) zeigt, warum ein solches Desiderat so schwer erfüllbar ist: In der traditionellen Printlexikographie werden nicht alle Flexionsformen der Lemmata verzeichnet, sondern – um das Beispiel der Substantive zu nehmen – nur die Formen, die das Paradigma im Deutschen eindeutig identifizieren, d.h. Nominativ und Genitiv Singular sowie Nominativ Plural. Die Angaben sind, um Druckraum zu sparen, meist auch sehr kryptisch formuliert und können in recht unsystematischer Weise um Kommentare erweitert sein. Es ist deshalb nicht ohne Weiteres möglich, die prinzipiell vorhandenen Flexionsinformationen für die automatische Lemmatisierung zu nutzen, wie dies bei einer stärker formalisierten, linguistisch motivierten Datenmodellierung der Fall wäre. [26] Das einfache Beispiel dürfte bereits zeigen, dass eine linguistisch motivierte Datenmodellierung kein Desiderat aus dem lexikologischen oder metalexikographischen Elfenbeinturm ist. Sie schafft vielmehr die Voraussetzungen, um das Nachschlagen in digitalen Wörterbüchern, das ja immer häufiger im Zuge der Rezeption und Produktion von Texten am Bildschirm geschieht, flexibel und effizient an die Bedürfnisse in konkreten Nutzungssituationen anzupassen.

Die zweite These steht in engem Zusammenhang mit der ersten, bezieht sich aber nicht auf die Art und Weise der Datenmodellierung, sondern auf die sich daraus ergebenden Konsequenzen für den lexikographischen Arbeitsprozess.

These 2: Der lexikographische Arbeitsprozess sollte sich an den zu bearbeitenden lexikologischen Phänomenen und nicht am Alphabet orientieren.

Im Wörterverzeichnis gedruckter Wörterbücher muss die Abfolge der Wörterbuchartikel nach einem nachvollziehbaren Kriterium erfolgen, damit die Nutzer auf die Informationen rasch zugreifen können. Hierfür hat sich die alphabetische Anordnung bewährt, auch wenn sie seit langem als rein formal und nicht den lexikologischen Zusammenhängen entsprechend kritisiert wird. Die Kritik wird meist aus lexikologischer Warte formuliert; weitaus seltener werden die Nachteile thematisiert, die daraus entstehen, dass die alphabetische Anordnung im Buch in erheblichem Maße auch den lexikographischen Arbeitsprozess determiniert [27]. Alphabetische Wörterbücher werden traditionell von A–Z nach sog. Buchstabenstrecken abgearbeitet; bei großen Wörterbuchprojekten sind daran mehrere Generationen von Lexikographinnen und Lexikographen beteiligt. Um das Wörterbuch einigermaßen konsistent zu halten, müssen relativ früh bereits der Wörterbuchplan, die Lemmaliste und das Artikelstrukturprogramm festgelegt werden. An diesen Vorgaben lässt sich dann nicht mehr viel verändern, auch wenn zwischen der Bearbeitung von Buchstabenstrecke A und der Bearbeitung von Buchstabenstrecke Z mehrere Jahrzehnte liegen. Dass bei diesem Vorgehen das Lemma „aufgehen" und das Lemma „zugehen" nicht mehr in der angemessenen Einheitlichkeit beschrieben sind, dass Verweise häufiger zu den bereits bearbeiteten Buchstabenstrecken führen als zu den nur geplanten, lässt sich mittlerweile mit elektronischen Auswertungsmethoden auch nachweisen.

Da es im digitalen Wörterbuch keine feste Anordnung der Lemmata, sondern viele Zugriffsmöglichkeiten auf die lexikalischen Informationen gibt, kann der lexikographische

---

[26] Zu der Frage, ob die kryptischen grammatischen Angaben überhaupt von den menschlichen Benutzern problemlos dechiffriert werden können, gibt es m.W. noch keine empirische Studie; meine eigenen Erfahrungen mit Studierenden stimmen mich eher skeptisch.

[27] Die Darstellung ist stark vereinfachend; die Prozesse des lexikographischen Arbeitsprozesses sind detailliert und unter Berücksichtigung des Medienwandels in Wiegand (1999) beschrieben.

Arbeitsprozess nach lexikologisch motivierten Kriterien organisiert werden. Ein solches Vorgehen hat deutliche Vorteile gegenüber dem Abarbeiten von Buchstabenstrecken, weil das methodische Vorgehen bei der lexikographischen Arbeit in Abhängigkeit vom Typ des Lemmazeichens variiert – auf einen sehr einfachen Nenner gebracht: Bei der Beschreibung eines Verbs spielen andere Merkmale und Kategorien eine Rolle als bei der Beschreibung einer Gradpartikel. Wortschatzeinheiten könnten also wesentlich konsistenter bearbeitet werden, wenn der lexikographische Arbeitsprozess an Lemmazeichentypen ausgerichtet wird, wobei sich diese nach Kriterien der syntaktischen und/oder semantischen Zusammengehörigkeit beliebig fein untergliedern lassen. Das Abfassen der Wörterbuchartikel kann dann durch Computerwerkzeuge unterstützt werden, die Angaben zu jedem Lemma eines Lemmazeichentyps auf Konsistenz und auf Kompatibilität mit den Angaben der Lemmata desselben Typs überprüfen. Die Lexikographinnen und Lexikographen können sich jeweils auf die Besonderheiten des aktuell bearbeiteten Lemmazeichentyps konzentrieren und Forscher, die diesen Lemmazeichentyp untersuchen bzw. untersucht haben, an der Beschreibung der betreffenden Einträge mit beteiligen (vgl. These 7).

These 3: Die Verbindung zwischen lexikographischen Beschreibungen und lexikographischen Quellen muss für die Benutzer transparent und nachvollziehbar sein.

Das Quellen- und Belegprinzip gehört zu den wichtigsten Grundsätzen seriöser lexikographischer Arbeit. Die Lexikographinnen und Lexikographen dürfen sich nicht nur auf ihre Sprachkompentenz verlassen, sondern müssen ihre Beschreibungen durch authentisches Sprachmaterial belegen. Zu den wichtigen, aber auch zeit- und kostenintensiven Teilaufgaben lexikographischer Arbeitsprozesse gehört es deshalb, die Quelltexte auszuwählen, zu beschaffen und für die „eigentliche" lexikographische Arbeit, das Abfassen der Wörterbuchartikel, aufzubereiten.[28] Vor dem Einsatz des Computers bei der Wörterbuchproduktion geschah dies im Allgemeinen durch das Exzerpieren von Belegen und deren anschließende Verzettelung und Archivierung in Zettelkästen (vgl. Lemberg [1996]). Inzwischen werden Belege und auch Corpora zunehmend computergestützt verwaltet[29]. Im gedruckten Wörterbuch bekommt der Wörterbuchbenutzer selbst im günstigen Fall nur einen kleinen Bruchteil dieser Arbeit zu sehen, nämlich die Belegangaben, die von den Lexikographinnen und Lexikographen ausgewählt wurden. Bei ein- und mehrsprachigen Gebrauchswörterbüchern fällt die Angabe von authentischen Belegen oft ganz der Erfordernis zum Opfer, die entstehenden Produkte klein, handlich und im Preis erschwinglich zu halten; d.h., in den meisten Einbändern sind Belegangaben durch knappe, selbstkonstruierte Beispielangaben ersetzt.

Im digitalen Medium ist Speicherplatz kein relevanter Kostenfaktor. Deshalb können und sollten in digitalen Wörterbüchern die Wörterbuchartikel verknüpft sein mit den lexikographischen Quelltexten und/oder den daraus exzerpierten Belegen. Dies erlaubt es den Benutzern, den Weg vom Wörterbuch zu den Quellen zurückzugehen, die von den Lexikographen getroffenen Entscheidungen nachzuvollziehen und um eigenständige Recherchen zu ergänzen.[30] Der im lexikographischen Prozess ohnehin zu leistende Aufwand der

---

[28] Für eine detaillierte Beschreibung der hier sehr grob skizzierten Phasen lexikographischer Prozesse mit und ohne Computereinsatz verweise ich auf Wiegand (1999, Kap. 1.5.1. und 1.5.2.).

[29] Vgl. auch Schmidt (1997) und Plate/Recker (in diesem Band). Als Modell für ein Wörterbuch im Übergang vom Zettelkasten auf die Belegdatenbank kann das an der Heidelberger Akademie der Wissenschaften erarbeitete DEUTSCHE RECHTSWÖRTERBUCH gelten (vgl. Speer [1994]).

Quellenbearbeitung kommt dann auch den Nutzern zugute – der Nutzwert der Wörterbücher wird bei gleichbleibenden Kosten erheblich gesteigert.

Der dafür zu betreibende Aufwand ist relativ gering, wenn die Quelltexte und Belegarchive bereits in digitalisierter Form vorliegen und vom Computerarbeitsplatz der beteiligten Lexikographen aus zugänglich sind. Ein computergestützter lexikographischer Arbeitsplatz muss ohnehin Werkzeuge bereitstellen, mit denen Corpus und Belegsammlung nach möglichst flexibel kombinierbaren Kriterien durchsucht und gefiltert werden können. Diese Werkzeuge lassen sich ohne erheblichen Mehraufwand zu Recherchetools für die künftigen Nutzer des Wörterbuchs weiterentwickeln. Bei einer solchen Weiterentwicklung sind zwei Dinge zu beachten: Im Gegensatz zu den Lexikographen, die ja täglich mit den Werkzeugen hantieren, werden die meisten Nutzer des digitalen Wörterbuchs nicht viel Zeit investieren wollen, um sich in die Suchfunktionen einzuarbeiten. Deshalb muss die Oberfläche einfacher gestaltet werden, d.h. man muss zwischen leicht erlernbaren Grundfunktionen und spezielleren Zusatzfunktionen für die „power user" unterscheiden. Wichtig ist es auch, dass der Benutzer bei der Recherche nicht von Belegen überschwemmt wird, sondern den Suchraum nach Bedarf einschränken kann.[31] Voraussetzung hierfür ist die möglichst feinkörnige Annotation der Quelltexte und Belege mit Hilfe von Werkzeugen aus der Texttechnologie, d.h., dass die Quellen des lexikographischen Corpus mit bibliographischen Angaben versehen sind und die lexikalischen Einheiten in Beleg- und Quelltexten auf ihre Grundform zurückgeführt (lemmatisiert) und mit eindeutigen syntaktischen Kategorien versehen werden.[32]

These 4: Die Benutzerschnittstelle von digitalen Wörterbüchern sollte an Typen von Benutzungssituationen adaptierbar sein.

Am Beispiel von zweisprachigen Wörterbüchern wird die Motivation für diese These schnell deutlich: Ein Wörterbuch mit dem Sprachpaar deutsch-englisch wird benutzt von englischen und deutschen Muttersprachlern, zur Hin- und zur Herübersetzung, zum Verstehen und zum Produzieren fremdsprachlicher Texte, zur Suche nach Übersetzungsäquivalenten, nach grammatischen Eigenschaften fremdsprachlicher Einheiten oder nach typischen Kollokationen. Es ist metalexikographisch gut erforscht und beschrieben, dass in Abhängigkeit von solchen Parametern sehr unterschiedliche Informationen benötigt werden.[33] Der Wunsch nach vier bzw. sechs Wörterbüchern pro Sprachpaar ließ sich theoretisch gut begründen, galt jedoch im gedruckten Medium als praktisch nicht umsetzbar.

Anders im digitalen Wörterbuch: Es ist gerade die Stärke von Hypertextsoftware, aus ein und demselben Datenpool die Informationen herauszugreifen und in geeigneter Weise zu präsentieren, die für einen bestimmten Nutzungskontext typischerweise relevant sind. Dadurch wird es möglich, lexikographische Beschreibungen an den Informationsbedarf anzu-

---

[30] Die Vernetzung von Wörterbuch und Belegarchiv ist ein wichtiger Aspekt bei der Hypertextualisierung des DEUTSCHEN RECHTSWÖRTERBUCHS, vgl. dazu Lemberg/Petzold/Speer (1998) und Lemberg (in diesem Band).

[31] Vgl. die „Maxime der optimalen quantitativen Korpusdokumentation" in Richter (in dsm. Band).

[32] Das digitale Wörterbuch der deutschen Sprache des 20.Jahrhunderts (DWDS), das momentan an der Berlin-Brandenburgischen Akademie der Wissenschaften vorbereitet wird (vgl. http://www.bbaw.de/iag/dig_woerterbuch/index.html), wird auf einem derart aufbereiteten lexikographischen Corpus basieren und dieses auch für die künftigen Nutzer des Wörterbuchs recherchierbar machen.

[33] Vgl. Kromann/Riiber/Rosbach (1991) und Breidt (1998) für einen Überblick.

passen, der für eine gegebene Nutzungssituation typisch ist. Statt der statischen Wörterbuchartikel im Printwörterbuch, in denen alle potenziell interessanten Informationen auf engstem Raum komprimiert sind, bekommt ein Nutzer eines digitalen Hypertext-Wörterbuchs Wörterbuchartikel „on demand" mit den Angaben zusammengestellt, die für die aktuelle Wörterbuchbenutzungssituation relevant sind. Erste Prototypen derartiger kontextadaptiver Wörterbücher sind in Petelenz (in diesem Band) und in Thielen/Breidt/Feldweg (1998) beschrieben. Dass es sich bislang nur um Prototypen handelt, liegt weniger daran, dass nicht bekannt wäre, für welche usuellen Benutzungssituationen typischerweise welche Klassen von Angaben relevant werden.[34] Die Ursache liegt vielmehr darin, dass eine kontextadaptive Präsentation lexikalischer Informationen eine linguistisch motivierte und feinkörnige Modellierung der lexikographischen Daten voraussetzt (vgl. These 1). Eine derartige Modellierung erfordert, wenn sie auf der Grundlage eines gedruckten Wörterbuchs erfolgt, einen relativ hohen Auf- und Nachbereitungsaufwand (vgl. Breidt [1998], Feldweg [1997]) und lässt sich deshalb am schnellsten realisieren, wenn ein digitales Wörterbuch unabhängig von einer vorhandenen Printvorlage neu konzipiert werden kann.

These 5: Ein digitales Wörterbuch zur Gegenwartssprache sollte ein Ausbauwörterbuch sein.

In seinem sehr lesenswerten Diskussionspapier zum Computereinsatz in der Dialektlexikographie stellt Martin Schröder das konventionelle gedruckte Abschlusswörterbuch dem neuen elektronischen Ausbauwörterbuch gegenüber: „Während das zum Druck vorgesehene Wörterbuch die Informationsaufnahme notwendig mit der Bearbeitung einer bestimmten Wortstrecke abschließt bzw. hierfür bestenfalls einen Nachtrag bereithält, wird im datenbankorientierten Online-Wörterbuch die Bearbeitung der Wortstrecken grundsätzlich und an allen Stellen offen bleiben." (Schröder 1997, 16). Diese Offenheit bringt nicht nur für die Dialektlexikographie viele Vorteile mit sich, sondern ermöglicht es auch und vor allem der gegenwartsbezogenen Lexikographie, auf Veränderungen im Wortschatz zügig zu reagieren. Die Aktualisierung, die Revision und der Ausbau digital verwalteter lexikographischer Daten wird durch den modularen Aufbau und die computergestützte Verweisverwaltung enorm erleichtert. Dies bringt aber auch neue Probleme mit sich: Es muss auch in einem Ausbauwörterbuch gesichert sein, dass die abrufbaren Informationen verbindlich, verlässlich und zitierbar sind. Ein digitales Wörterbuch, das seine Wörterbuchartikel in vom Benutzer nicht nachvollziehbarer Weise verändert oder löscht, verliert für die Zwecke an Wert, bei denen es auf Verlässlichkeit und Zitierbarkeit ankommt. Werkzeuge und Verfahren der Versionenverwaltung können hier Abhilfe schaffen, wenn die Probleme frühzeitig erkannt und berücksichtigt werden.

These 6: Digitale Wörterbücher sollten die Option der Mehrfachkodierung sinnvoll einsetzen.

Die Argumente für diese These wurden eigentlich bereits in Abschnitt 2.2 genannt und sind auch weitgehend unumstritten: Die Möglichkeit, das bislang textdominierte Wörterbuch durch Bild-, Ton- und Videoobjekte zu ergänzen, kann die Verständlichkeit und Anschaulichkeit lexikographischer Angaben verbessern. Die Chancen für die Dialektlexikographie, die Erklärung fachsprachlicher Termini in ihrem fachsystematischen Kontext und vor allem

[34] Vgl. Kühn (1989) und in Wiegand (1999, Kap. 4.2.2.).

auch für die muttersprachliche und fremdsprachliche Sprachdidaktik liegen auf der Hand. Die neuen Gestaltungsmittel bringen aber auch neue Anforderungen an die Wörterbuchmacher mit sich: Empirische Untersuchungen deuten darauf hin, dass eine schlecht koordinierte Verknüpfung von Text, Bild, Ton und Video die Informationsaufnahme verschlechtert statt sie zu verbessern (Weidenmann [1995]). Mehrfachkodierung zu nutzen, bedeutet deshalb nicht einfach, möglichst viele Text-, Ton-, Bild- und Videoobjekte zu verknüpfen. Wichtig für eine qualitativ hochwertige Anwendung ist vielmehr die Integration der unterschiedlichen Zeichentypen nach semantisch-funktionalen Prinzipien. Hier besteht im Bereich der Lexikographie noch ein erheblicher Forschungsbedarf.[35]

These 7: Online-Wörterbücher sollten die Chancen der Verbindung von Information und Kommunikation gezielt für die Qualitätssicherung nutzen.

Als Online-Wörterbücher bezeichne ich digitale Wörterbücher, die über das World Wide Web zugänglich sind (vgl. auch Lemberg [in diesem Band]). Die Stärke des WWW liegt in der Verbindung von Information und Kommunikation: Mit WWW-Browsern kann man nicht nur Informationen abrufen, sondern auch die Kommunikationsdienste des Internet in Anspruch nehmen, von der elektronischen Post (E-Mail) und den Postverteilern (Mailing-Listen) über die Diskussionsgruppen (Newsgroups), bis hin zu den Online-Konferenzen (Chat). Dies eröffnet Wörterbuchprojekten neuartige Möglichkeiten, in Kontakt mit ihren Nutzern zu treten und diese am Aufbau und an der Pflege des Wörterbuchs zu beteiligen. Die Formen der Partizipation in existierenden Online-Wörterbüchern reichen von der Bitte um Fehlerkorrektur und Rückmeldung, über den Aufruf zur Beteiligung an der Schließung von Lemmalücken, bis hin zu Wörterbüchern, die ganz oder überwiegend von den Beiträgen ihrer „Gäste" leben (vgl. Storrer/Freese [1996] und Storrer [1998, Abschnitt 4.2]).

Ein weiterer Vorteil: Wenn das lexikographische Corpus bzw. die Belegsammlungen eines Projekts in digitalisierter Form über das WWW zugänglich sind, wird die räumliche Nähe der Lexikographen zu einem Belegarchiv unerheblich. Dies erleichtert den arbeitsteiligen Aufbau von Wörterbüchern in räumlich verteilten Arbeitsstellen erheblich. In These 2 wurde bereits erläutert, warum die Arbeitsteilung nicht nach Buchstabenstrecken sondern nach lexikologischen Kriterien erfolgen sollte. Im Idealfall kann bei einem verteilten Wörterbuchprojekt jeder Lemmazeichentyp genau von den Forschern bearbeitet werden, die sich mit dem betreffenden lexikologischen Phänomen bereits beschäftigt haben. Die eingangs zitierte Anregung von H. Tiktin, das Prinzip der Arbeitsteilung auf den lexikographischen Arbeitsprozess zu übertragen, kann also im WWW erstmals effizient und in großem Stil umgesetzt werden. Tiktin hatte dabei vor allem das Problem vor Augen, dass fachsprachlicher Wortschatz im Grunde nur von Lexikographen beschrieben werden kann, die sich im jeweiligen Fachgebiet auch auskennen, dass aber die meist philologisch ausgebildeten Lexikographinnen und Lexikographen häufig in Naturwissenschaft, Handwerk und Technik wenig bewandert sind (Tiktin [1910, 248f]). Das WWW macht es nun möglich, beim Aufbau von allgemeinsprachlichen Wörterbüchern Spezialisten der jeweiligen Fachdisziplinen hinzuzuziehen und somit die Verlässlichkeit und Korrektheit der Erläuterungen zu verbessern. Dieser Aspekt ist wichtig, weil die verständliche Erläuterung von Fachvokabular maßgeblich dazu beiträgt, dass sich die Bürger in der sog. „Informationsgesellschaft" über aktuelle Entwicklungen im wirtschaftlichen, technologischen und wissenschaftlichen

[35] Erste Überlegungen finden sich in Petelenz (in diesem Band) und Hupka (1989).

Bereich ausreichend informieren können, um die politische Auseinandersetzung über die Folgen dieser Entwicklungen nachvollziehen und sich ein eigenes Urteil bilden zu können.

Dass sich der kollaborative Wörterbuchaufbau über das World Wide Web tatsächlich organisieren lässt, zeigen Wörterbuchprojekte, die – bislang weitgehend unbeachtet von der metalexikographischen Forschung und ganz im Geiste der Open-Source-Projekte – Spezialwörterbücher, aber auch allgemeinsprachliche bilinguale Wörterbücher aufbauen und kostenlos zur Verfügung stellen. Ein bemerkenswertes Beispiel für ein Wörterbuch dieser Art zum Sprachpaar Deutsch-Englisch ist LEO (http://dict.leo.org/). 1997 gestartet, erwies es sich bei der in Storrer/Freese (1996) durchgeführten Stichprobe als wenig zuverlässig. Inzwischen haben viele freiwillige Helfer das Wörterbuch deutlich erweitert und verbessert. Die Stärke von LEO liegt in einem sehr reichhaltigen Kollokations- und Phraseologieteil, der viele aktuelle Termini und Wendungen und fachsprachliches Vokabular aus Informationstechnik und Wirtschaft enthält; in diesen Bereichen schlägt es konventionelle bilinguale Printwörterbücher an Abdeckung und Aktualität. Es ist deshalb v.a. für Nutzer wertvoll, die zwar nicht professionell übersetzen, im beruflichen Alltag aber häufig englische Gebrauchstexte verfassen oder verstehen müssen – die Zahl von ca. 200.000 Zugriffen pro Tag (März 2000) zeigt die steigende Beliebtheit der Ressource.

Ein anderer Aspekt der Qualitätssicherung soll zum Abschluss noch angesprochen werden: Die Nutzung von Online-Wörterbüchern über das Internet wird protokolliert, d.h., es ist erstmals möglich, auf einfache Art und Weise herauszufinden, wie viele Nutzer welche Daten abgerufen haben. Auch wenn die in Lemnitzer (in diesem Band) beschriebene Studie eher ernüchternde Ergebnisse lieferte, sollte man das Potenzial nicht unterschätzen, das in der automatischen Protokollierung von Benutzeraktionen für die künftige Wörterbuchbenutzungsforschung liegt. Je erfolgreicher und qualitativ hochwertiger ein Wörterbuch, umso häufiger dürfte es usuell genutzt und nicht nur mit den Lemmata getestet werden, die Nutzern nach der Lemnitzer-Studie offensichtlich spontan zuerst einfallen. Eine entsprechende Studie mit dem Wörterbuch LEO dürfte bereits wesentlich interessantere Ergebnisse zeigen.

Insgesamt bietet die Verbindung von Information und Kommunikation im World Wide Web in bislang unbekannter Weise die Chance, Wörterbücher in Auseinandersetzung mit und unter Beteiligung von den Sprachbenutzern zu erarbeiten und die Qualität der lexikographischen Produkte bereits zu einem frühen Zeitpunkt mit den Nutzern zu testen. Wörterbücher dieser Art herzustellen und in Gang zu halten – schließlich geht es ja gemäß These 5 um Ausbauwörterbücher – wäre „eine Aufgabe, des Schweißes der Edlen wert", wie es Tiktin (1910, S. 253) für seine Vision von den Wörterbüchern der Zukunft formuliert hat.

4 Literatur

Bergenholtz, Henning; Tarp, Sven; Wiegand, Herbert Ernst (1999): Datendistributionsstrukturen, Makro- und Mikrostrukturen in neueren Fachwörterbüchern. In: Hoffman, L.; Kalverkämper, H.; Wiegand, H. E. (Hgg.) (1999): Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung. Berlin/New York, 1762–1832.

Blum, Joachim; Bucher, Hans-Jürgen (1998): Ein Multimedium. Textdesign – ein Gestaltungskonzept für Text, Bild und Grafik. Konstanz.

Blumenthal, Andreas; Lemnitzer, Lothar; Storrer, Angelika (1988): Was ist eigentlich ein Verweis? – Konzeptuelle Datenmodellierung als Voraussetzung computergestützter Verweisbehandlung. In: Harras, G. (Hg.) (1988): Das Wörterbuch: Artikel und Verweisstrukturen. Düsseldorf, 351–373.

Boguraev, Bran; Briscoe, Ted (1989): Utilising the LDOCE grammar codes. In: Boguraev, B.; Briscoe, T. (Hgg.) (1989): Computational Lexicography for Natural Language Processing. London/New York, 85–116.

Bolter, Jay David (1991): Writing Space. The Computer, Hypertext and the History of Writing. Hillsdale NY.

Breidt, Elisabeth (1998): Neuartige Wörterbücher für Mensch und Maschine: Wörterbuchdatenbanken in COMPASS. In: Wiegand, H. E. (Hg.) (1998): Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. Tübingen, 1–27.

Drosdowski, Günther (1977): Nachdenken über Wörterbücher: Theorie und Praxis. In: Drosdowski, G.; Henne, H.; Wiegand, H. E. (Hgg.) (1977): Nachdenken über Wörterbücher. Mannheim, 103–143.

Feldweg, Helmut (1997): Wörterbücher und neue Medien: Alter Wein in neuen Schläuchen? In: LiLi 27/107 (1997), 110–123.

Fellbaum, Christiane (Hg.) (1998): WORDNET: An electronic lexical database. Cambridge, MA.

Freisler, Stefan (1994): Hypertext – eine Begriffsbestimmung. In: Deutsche Sprache 1 (1994), 19–50.

Goebel, Ulrich; Lemberg, Ingrid; Reichmann, Oskar (1995): Versteckte lexikographische Information. Möglichkeiten ihrer Erschließung dargestellt am Beispiel des Frühneuhochdeutschen Wörterbuchs. Tübingen.

Grimm, Jacob (1854): Vorrede zum Deutschen Wörterbuch. In: Grimm, J.; Grimm, W. (Hgg.) (1854): Deutsches Wörterbuch. Leipzig, I–LXVIII.

Hamp, Birgit; Feldweg, Helmut (1997): GermaNet – A lexical semantic net for German. In: Vossen, P.; Calzolari, N. (Hgg.) (1997): Proceedings of the ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid, 9–15.

Henne, Helmut (1977): Nachdenken über Wörterbücher: Historische Erfahrungen. In: Drosdowski, G.; Henne, H.; Wiegand, H. E. (Hgg.) (1977): Nachdenken über Wörterbücher. Mannheim, 7–50.

Heyn, Matthias (1992): Zur Wiederverwendung maschinenlesbarer Wörterbücher – eine computergestützte metalexikographische Studie am Beispiel der elektronischen Edition des „Oxford Advanced Learner's Dictionary of Current English". Tübingen.

Hupka, Werner (1989): Wort und Bild. Die Illustrationen in Wörterbüchern und Enzyklopädien. Tübingen.

Kromann, H.-P; Riiber, T.; Rosbach, P. (1991): Principles of bilingual lexicography. In: Hausmann, F. J.; Reichmann, O.; Wiegand, H. E. et al. (Hgg.) (1991): Wörterbücher. Ein internationales Handbuch zur Lexikographie, 3. Teilband. Berlin/New York, 2711–2728.

Kuhlen, Rainer (1991): Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank. Berlin et al.

Kühn, Peter (1989): Typologie der Wörterbücher nach Benutzungsmöglichkeiten. In: Hausmann, F. J.; Reichmann, O.; Wiegand, H. E. et al. (Hgg.) (1989): Wörterbücher. Ein internationales Handbuch zur Lexikographie, 1. Teilband. Berlin/New York, 111–127.

Lang, Ewald (1983): Lexikon als Modellkomponente und Wörterbuch als lexikographisches Produkt. In: Schildt, J.; Viehweger, D. (Hgg.) (1983): Die Lexikograpie von heute und das Wörterbuch von morgen. Analysen – Probleme – Vorschläge. Berlin, 76–91.

Lemberg, Ingrid (1996): Die Belegexzerption zu historischen Wörterbüchern am Beispiel des Frühneuhochdeutschen Wörterbuchs und des Deutschen Rechtswörterbuchs. In: Wiegand, H. E. (Hg.) (1996): Wörterbücher in der Diskussion II. Tübingen, 83–192.

– und Petzold, Sybille; Speer, Heino (1998): Der Weg des Deutschen Rechtswörterbuchs in das Internet. In: Wiegand, H. E. (Hg.) (1998): Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. Tübingen, 265–289.

Nelson, Theodor H. (1972): As We Will Think. Reprint in: Nyce, J. M. K., Paul (1991) (Hg.): From Memex to Hypertext: Vannevar Bush and the Mind's Machine: 245–259.

Raymond, Darrell R.; Tompa, Frank W.M. (1988): Hypertext and the Oxford English Dictionary. In: Communications of the ACM 31/7 (1988), 871–879.

Schmidt, Hartmut (1997): Plädoyer für eine moderne korpusbasierte deutsche Wortschatzforschung. In: Zeitschrift für Linguistik und Literaturwissenschaft 27/196 (1997), 19–29.
Schmitz, Ulrich (1997): Schriftliche Texte in multimedialen Kontexten. In: Weingarten, R. (Hg.) (1997): Sprachwandel durch den Computer? Opladen, 131–157.
Schröder, Martin (1997): Brauchen wir ein neues Wörterbuchkartell? Zu den Perspektiven einer computerunterstützten Dialektlexikographie und eines Projekts „Deutsches Dialektwörterbuch". In: Zeitschrift für Dialektologie und Linguistik LXIV/1 (1997), 58–66.
Slatin, John M. (1991): Composing Hypertext: A Discussion for Writing Teachers. In: Berk, E.; Devlin, J. (Hgg.) (1991): Hypertext / Hypermedia Handbook. New York et al., 55–64.
Speer, Heino (1994): DRW to FAUST. Ein Wörterbuch zwischen Tradition und Fortschritt. In: Lexicographica 10 (1994). Tübingen, 171–213.
Storrer, Angelika (1996): Wörterbücher zum Anklicken. Ein kleiner Rundgang durch die PC-Bibliothek. In: Sprachreport 2/95 (1996), 9–10.
– (1998): Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In: Wiegand, H. E. (Hg.) (1998): Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. Tübingen, 107–135.
– (2000): Was ist „hyper" am Hypertext? In: Kallmeyer, W. (Hg.) (2000): Sprache und neue Medien. Berlin u. a.
– und Feldweg, Helmut; Hinrichs, Erhard (1993): Korpusunterstützte Entwicklung lexikalischer Wissensbasen. In: Sprache und Datenverarbeitung 17 (1993), 59–72.
– und Freese, Katrin (1996): Wörterbücher im Internet. In: Deutsche Sprache 24/2 (1996), 97–136.
Thielen, Christine; Breidt, Elisabeth; Feldweg, Helmut (1998): COMPASS. Ein intelligentes Wörterbuchsystem für das Lesen fremdsprachiger Texte. In: Storrer, A.; Harriehausen, B. (Hgg.) (1998): Hypermedia für Lexikon und Grammatik. Tübingen, 173–194.
Tiktin, H. (1910): Wörterbücher der Zukunft. In: Germanisch-Romanische Monatszeitschrift II. Jahrgang (1910), 243–253.
Weidenmann, Bernd (1995): Multicodierung und Multimodalität im Lernprozeß. In: Issing, L. J.; Klimsa, P. (Hgg.) (1995): Information und Lernen mit Multimedia. Weinheim, 65–84.
Wiegand, Herbert Ernst (1977): Nachdenken über Wörterbücher: Aktuelle Probleme. In: Drosdowski, G.; Henne, H.; Wiegand, H. E. (Hgg.) (1977): Nachdenken über Wörterbücher. Mannheim, 51–102.
– (1987): Wörterbuchartikel als Text. In: Harras, G. (Hg.) (1987): Das Wörterbuch – Artikel und Verweisungsstrukturen. Jahrbuch 1987 des Instituts für deutsche Sprache. 30–120.
– (1989): Arten von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. In: Hausmann, F. J.; Reichmann, O.; Wiegand, H. E. et al. (Hgg.) (1989): Wörterbücher. Ein internationales Handbuch zur Lexikographie, 1. Teilband. Berlin/New York, 462–501.
– (1989): Aspekte der Makrostruktur im allgemeinen einsprachigen Wörterbuch: alphabetische Anordnungsformen und ihre Probleme. In: Hausmann, F. J.; Reichmann, O.; Wiegand, H. E. et al. (Hgg.) (1989): Wörterbücher. Ein internationales Handbuch zur Lexikographie, 1. Teilband. Berlin/New York, 371–409.
– (1991): Printed Dictionaries and their Parts as Text. An Overview of More Research as an Introduction. In: Lexikographica 6 (1990) (1991), 1–124.
– (1996): Textual Condensation in Printed Dictionaries. A Theoretical Draft. In: Lexikos 6 (1996), 133–158.
– (1998): Altes und Neues zur Makrostruktur alphabetischer Printwörterbücher. In: Wiegand, H. E. (Hg.) (1998): Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. Tübingen, 348–372.
– (1999): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. Berlin/New York.

Angelika Storrer, Mannheim

Ingrid Lemberg

Aspekte der Online-Lexikographie für wissenschaftliche Wörterbücher

1 „Mein gedrucktes Wörterbuch reicht mir"

„Mein gedrucktes Wörterbuch reicht mir. Damit arbeite ich schon seit zwanzig Jahren, da weiß ich, wie ich etwas finde und habe auch selbst viele Randnotizen gemacht. Außerdem steht es auf meinem Schreibtisch griffbereit, und überhaupt finde ich es äußerst unbequem, möglicherweise erst den Computer anzuschalten, um mir dann aus dem Netz ein Wörterbuch zu laden".[1] Diese Gedanken dürften so oder ähnlich wohl fast allen denjenigen Wissenschaftlerinnen und Wissenschaftlern durch den Kopf gehen, zu deren unverzichtbarem Rüstzeug Wörterbücher gehören, wenn sie zum ersten Mal mit Online-Wörterbüchern[2] Bekanntschaft machen. Wozu vertraute Gewohnheiten und Wörterbücher verlassen? Was können Online-Wörterbücher neben oder anstelle von Printwörterbüchern bieten? Eine erste Antwort gibt die Überschrift zu Kapitel 2 dieses Beitrages: Sie bieten mehr.

Mit dem Internet haben wir seit einigen Jahre ein Medium, das in der Lage ist, „sämtliche technisch vermittelten Kommunikationsformen zu integrieren und die mediengeschichtliche Entwicklung auf die bisherige Spitze zu treiben" (Cölfen/Cölfen/Schmitz 1997:261). Daß es ein hervorragend geeignetes Medium für alle Wissenschaften, und damit auch für die Lexikographie ist, dürfte inzwischen außer Frage stehen. Ein Blick in die Linklisten für Wörterbücher bestätigt dies. Allein die Wörterbuchsammlung von Robert Beard mit mehr als 800 Wörterbüchern zu über 160 Sprachen zeigt die außerordentliche

---

[1] So der ungefähre Wortlaut eines Diskussionsbeitrages im Anschluß an den Vortrag „Wörterbücher und lexikographische Informationsmodelle der Zukunft" im September 1998 (Lemberg 1998c).

[2] Bei elektronischen Wörterbüchern unterscheidet man zwischen Online- und Offline-Wörterbüchern, wobei sich die Gruppe der Offline-Wörterbücher weiter unterteilen läßt in elektronische Taschenbücher und PC-Wörterbücher (meist CD-ROM-Wörterbücher). Online-Wörterbücher gehen entweder auf die Printversion eines Wörterbuchs zurück und basieren dann auf Konversion, oder sie sind Neu-Konzeptionen. Sowohl Konversionstypen als auch Neu-Konzeptionen können printorientiert oder innovativ gestaltet sein. Vgl. dazu die Überblicks-Graphik bei Lehr (1996:315).

Attraktion dieses Mediums für Wörterbücher.[3] Und glaubt man den Prognosen zum künftigen Stellenwert des Internet als Publikationsmedium, dann können sich die Wissenschaften eine Beschränkung auf die traditionellen Printmedien künftig auch nicht mehr leisten.[4]

In diesem Beitrag werden einige zentrale Aspekte der Online-Lexikographie, auch in ihrer Abgrenzung und Unterscheidung zum gedruckten Wörterbuch, unter verschiedenen Perspektiven dargestellt und diskutiert. Die Überlegungen zielen auf einsprachige, corpusbasierte, wissenschaftliche Wörterbücher des Deutschen.[5] Keines dieser Wörterbücher ist bislang in einer vollständigen Fassung im Internet abrufbar.[6] Am weitesten fortgeschritten sind die Arbeiten an den mittelhochdeutschen Wörterbüchern von Benecke/Müller/Zarncke und Lexer (vgl. den Beitrag von Burch/Fournier in diesem Band). Das DEUTSCHE RECHTSWÖRTERBUCH hatte im September 1998 eine Probeversion zur Buchstabenstrecke O ins Netz gehängt. Diese wurde im Juni 1999 durch eine stark überarbeitete Fassung zu den Buchstabenstrecken **I**, **N** und **O** ersetzt. Die Publikation der gesamten Wortstrecke von **A** bis **P** ist abschnittweise für die nächsten zwei Jahre vorgesehen. In Angriff genommen wurden die Arbeiten zur Digitalisierung des DEUTSCHEN WÖRTERBUCHS von Jacob und Wilhelm Grimm. Die Online-Fassung dieser Projekte beruht jeweils auf der Konversion bereits bestehender Printwörterbücher. Eine Neukonzeption liegt bislang nur für das LEKSIS-Projekt vor (Fraas/Haß-Zumkehr 1998). Noch nicht endgültig geklärt ist die Publikationsform des DIGITALEN WÖRTERBUCHS DER DEUTSCHEN SPRACHE DES 20. JHS., das an der Berlin-Brandenburgischen Akademie der Wissenschaften in Bearbeitung ist.[7]

Die Überlegungen zur Online-Lexikographie in diesem Beitrag sind also kaum Analyse von bereits Vorhandenem, sondern vielmehr Diskussion von Möglichem, und in vielen Fällen Beschreibung von Entstehendem.[8]

---

[3] Wobei auch hier der Topos gilt, daß Quantität noch nicht Qualität bedeutet. Eine weitere kritische Auseinandersetzung im Sinne von Storrer/Freese (1996) steht dringend an.

[4] „Wenn die Geisteswissenschaften nicht in das Internet gehen, werden sie sehr schnell ganz aus der Welt verschwinden", so lautete das abschließende mündliche Statement des Präsidenten der Mainzer Akademie der Wissenschaften und der Literatur, Clemens Zintzen, auf dem Kolloquium Neue Publikationsformen für geisteswissenschaftliche Akademievorhaben auf CD-ROM und im Internet am 11. April 1997. „Was jetzt im Internet als Wissensbestand und Geltungsanspruch nicht angemessen markiert wird, kann mittelfristig bereits von der Weltkarte der geläufigen Kenntnisse verschwunden sein" (Baasner 1999:Kapitel 0).

[5] Vgl. Deutschsprachige Wörterbücher (2000); „wissenschaftlich" ist hier im Sinn von Wiegand (1989:263) gebraucht.

[6] Bei den exemplarisch in Kap. 6.1 angeführten Dialektwörterbüchern im Internet handelt es sich nicht um wissenschaftliche Dialektwörterbücher des Deutschen, sondern in der Regel um mehr oder weniger umfangreiche Wortlisten, in denen dialektale Äquivalente zu Wörtern und syntagmatischen Verbindungen der neuhochdeutschen Standardsprache gegeben werden oder umgekehrt, und die eher von der Liebe zum Dialekt als von einem echten Informationsanliegen im wissenschaftlichen Sinn zeugen. Sind überhaupt Suchfunktionen vorhanden, so sind sie meist auf die Auswahl des jeweiligen Anfangsbuchstabens beschränkt, der in der Regel dann zu der entsprechenden Wortstrecke im Alphabet führt.

[7] Zu den bibliographischen Angaben und den URLs der Homepages vgl. jeweils Kap. 6.1.

[8] Die von mir dazu angeführten Beispiele für die Umsetzung vom Print in das neue Medium stammen meist aus der lexikographischen Praxis des DEUTSCHEN RECHTSWÖRTERBUCHS, können aber durchaus paradigmatisch für alle Typen wissenschaftlicher Wörterbücher gesehen werden. Die derzeitige Online-Fassung des DRW wurde vom Leiter des DRW, Dr. Heino Speer, erstellt. Ihr liegen zahlreiche konzeptionelle Diskussionen im Kollegenkreis der Forschungsstelle des DRW zugrunde, deren erste Ergebnisse in Lemberg/Petzold/Speer (1998) zusammengefaßt sind.

2 Lexikographische Mehrwerte von Online-Wörterbüchern

Mit lexikographischen Mehrwerten[9] ist ‚das Mehr' an Informationen im Wörterbuch und an Informationsmöglichkeiten für die Wörterbuchbenutzerinnen und -benutzer gemeint, das durch die Publikation in elektronischen Medien ermöglicht wird.
Es handelt sich dabei im Wesentlichen um die folgenden Komponenten:

- Aufhebung des begrenzten Druckumfangs[10]
- Hypertextualisierung (vgl. Kap. 2.1.1)
- multimediale Aufbereitung lexikographischer Daten (vgl. Kap. 2.1.2)
- mehrfache äußere Zugriffsstrukturen und vielfältige Suchmöglichkeiten[11]
- Interaktivität[12]

und als eher internetspezifische Komponenten

- Aufhebung eines statischen zugunsten eines dynamischen Wörterbuchs (Kap. 2.2)
- Kooperation und Interaktion zwischen Lexikograph und Benutzer (Kap. 4.2)

---

[9] Der Begriff des ‚Mehrwerts', auch mit Attribuierungen wie ‚lexikographisch' oder ‚linguistisch', spielt in der Diskussion um elektronische Wörterbücher eine wesentliche Rolle (Fournier 1999; Fraas/Haß-Zumkehr 1998:298; Lemberg/Petzold/Speer 1998:280; Wiegand 1998b:239 sowie Petelenz, Schmidt/Müller und Richter, letzterer ausführlich in Kap. 3, in diesem Band). Es handelt sich um eine Analogiebildung zu dem aus der Informationswissenschaft stammenden Begriff des ‚informationellen Mehrwerts', der in der Phase digitaler Aufbereitung und Präsentation von Wissen und in dem darüber erfolgenden Diskurs eine wesentliche Größe darstellt (Kuhlen 1991:Vorwort; 1995). Gemeint ist damit das Erzeugen von Information durch Informationsarbeit über Wissen. Die drei wesentlichen Schritte zur Erzeugung des Mehrwerts sind 1. die Wissensrekonstruktion, z.B. durch Transformierung in von Rechnern verarbeitbare Formen und den darauf basierenden Recherchemöglichkeiten oder die weitere Verarbeitung zu einer Hypertextbasis, 2. die Informationserarbeitung in Informationssystemen wie Online-Datenbanken und 3. die Informationsaufbereitung. Dabei werden die Methoden zur Informationsaufbereitung häufig als Verfahren zur Erzeugung informationeller Mehrwerte im engeren Sinn bezeichnet. Zu den formalen Verfahren gehören alle Formen der medialen Aufbereitung wie Gestaltung der Bildschirmoberfläche nach kognitiven Prinzipien oder der Einsatz von Animationen zur Verdeutlichung komplexer Prozesse. Zu den pragmatischen Mehrwertleistungen gehören alle Verfahren, durch die Informationen an unterschiedliche Benutzerbedürfnisse, unterschiedliche Informationsverhalten oder unterschiedliche Ziele angepaßt werden können (Zusammenfassung der Ausführungen zur Theorie informationeller Mehrwerte nach Kuhlen 1995:80–94).

[10] Und damit z.B. mehr Wörter oder mehr Belege, aber auch Verzicht auf viele Formen der lexikographischen Textverdichtung (dazu zuletzt Wiegand 1998a). Vgl. dazu auch die Beiträge von Klosa und Richter in diesem Band; speziell zu Formen der Dekomprimierung in den elektronischen Fassungen der mittelhochdeutschen Wörterbücher vgl. Fournier 1999.

[11] Z.B. gezielte Zugriffe über Register (Lemberg 1998d; zur methodischen Grundlegung der Registererschließung vgl. Goebel/Lemberg/Reichmann 1995). Bei den Suchfunktionen, dem sog. Information Retrieval, handelt es sich im wesentlichen um Volltextsuche, Suchen mit Worttrunkierungen, qualifizierende Suchen mit Hilfe von Wildcards oder Booleschen Operatoren oder durch schreibtolerante oder inkrementelle Suchfunktionen. Zu weiteren Formen von Zugriffsmöglichkeiten vgl. Richter, Kap. 3.5, in diesem Band.

[12] Interaktivität bei elektronischen Wörterbüchern meint Beeinflussung des Verhaltens einer Softwareumgebung durch den Benutzer und Anpassung an die jeweils individuellen Bedürfnisse, z.B. das Anlegen eigener Kommentare oder eigener Verschlagwortungen (Storrer 1998:107 u. 122f.). – Vgl. auch Haack 1997 sowie Richter, Kap. 3.4, in diesem Band.

2.1 Der Einsatz von hypermedialen Mitteln

Hypermedia ist eine Zusammenfügung aus *Hyper*text und Multi*media*. Eine ausführliche Darstellung über den Einsatz von hypermedialen Mitteln in der Lexikographie bringt Storrer (1998a; 1998b). In diesem Kapitel werden einige Beispiele für den Einsatz von Hypertext und Multimedia bei wissenschaftlichen Wörterbüchern vorgestellt.

2.1.1 Beispiele für Hypertextualisierungsformen

Hypertexte sind computerverwaltete, nicht-lineare Texte, die die Mehrfachkodierung von Daten in verschiedenen medialen Formen (Schrift, Graphik, Ton und Video) erlauben. ‚Nicht-linear' meint Verteilung der Textdaten auf kleinere Module, die dann über computerisierte Verweise, die sog. Hyperlinks miteinander verknüpft werden, wobei jedes Modul mit mehreren anderen Modulen verknüpft sein kann.[13] Zur Produktion und Rezeption von Hypertexten wird eine spezielle Software, ein sog. Hypertextsystem (vgl. Anm. 25) benötigt. Wörterbücher gelten neben Lexika, Handbüchern und Enzyklopädien als besonders zur Konvertierung geeignete Texte (Kuhlen 1991:175), da sie modularisierte Informationseinheiten besitzen, nach einem einheitlichen Schema aufgebaut sind und durch Querverweise selbst schon hypertextuelle Strukturen aufweisen.[14] Hypertextsysteme bieten für Wörterbücher einen größeren Bedienungskomfort (klicken statt blättern), erweiterte Informations- und Rezeptionsmöglichkeiten und damit eine Fülle von lexikographischen oder informationellen Mehrwerten. Hypertextualisierungsformen lassen sich für verschiedene makro- und mikrostrukturelle Verknüpfungen innerhalb eines Wörterbuchs ebenso gewinnbringend einsetzen, wie für die Verknüpfung der Wörterbücher untereinander.[10]

2.1.1.1 Explizite und implizite Verweise im gedruckten Wörterbuch[15] können im Hypertextwörterbuch als Hyperlinks eingerichtet werden. Die Verweisbefolgungshandlung erfolgt durch Anklicken des jeweiligen Verweislinks statt durch Blättern, ist also wesentlich komfortabler durchzuführen als im gedruckten Wörterbuch. Dies bewährt sich besonders bei Verweisnestern im Wörterbuchartikel, die zehn, zwanzig oder mehr Verweise enthalten können.[16] In diesen Fällen kann die Hypertextualisierung der Verweise auch eine wörterbuchdidaktische Funktion bekommen, da ein Benutzer oder eine Benutzerin möglicherweise das Nachschlagen in einem gedruckten Wörterbuch in mehreren Bänden mit tausenden von Druckspalten scheut, aber andererseits im Online-Wörterbuch

---

[13] Vgl. den Beitrag Storrer in diesem Band. Grundlegend zu Hypertext: Kuhlen 1991.

[14] So z.B. Gabriel (1997:69). Zur Hypertextualisierung in Wörterbüchern vgl. Lemberg 1998b; Storrer 1998a; Weber 1998.

[15] Blumenthal/Lemnitzer/Storrer 1987; Kammerer/Lehr 1996; eine metalexikographisch fundierte und ausführliche Diskussion zur Hypertextualisierung verschiedener Verweisformen im Wörterbuch bringt Kammerer 1998. Er prüft insbesondere, wann eine Übersetzung der Mediostruktur in eine Hyperlinkstruktur möglich und sinnvoll ist, und wann andere Techniken zu bevorzugen sind (S. 155f.).

[16] Es handelt sich dabei meist um Synonymenangaben. Exemplarisch verwiesen sei auf Wörterbuchartikel wie **ausleschen**[2] mit 21 Synonymenverweisen oder **ausmessen**[1] mit 15 Synonymenverweisen im FWB, oder **Notzucht** mit 31 Synonymenverweisen im DRW. Weitere Beispiele bei Mulch (1998:162) und Wiese (1998:152).

durch die Hypertextualisierung der Verweise zu Verweisbefolgungshandlungen geradezu animiert wird.[17]

2.1.1.2 Eine weitere wesentliche Verweisverknüpfung im Hypertextwörterbuch ist die zwischen den Quellensiglen im Belegteil und der Quellendokumentation des Wörterbuchs.[18] Quellensiglen wie z.B.

JJWolff Anf. XVII aus dem SCHWEIZERISCHEN IDIOTIKON, Quellenverzeichnis S. 125,
RepRKG. (Koser) aus dem DRW, Quellenergänzungsheft 4, Sp. 52, oder
DGK aus dem MITTELNIEDERDEUTSCHEN HANDWÖRTERBUCH, Quellenverzeichnis S. 12

dürften auch textkundigen Benutzerinnen und Benutzern die Identifzierung einer Quelle nicht ohne weiteres ermöglichen. Die zur Identifikation einer Quelle erforderliche Nachschlagehandlung im Quellenverzeichnis des gedruckten Wörterbuchs kann im Onlinewörterbuch bei entsprechender Hypertextualisierung durch einen Mausklick ersetzt werden. Abbildung 1 zeigt die derzeitige Realisierung im DRW online.

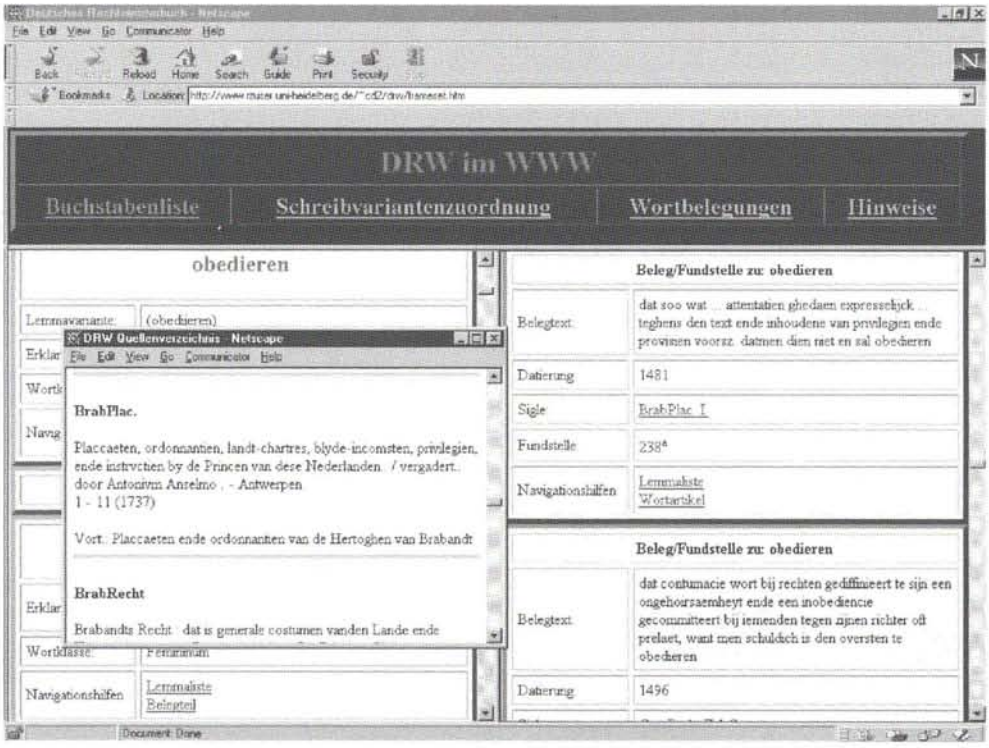


Abb. 1: Hypertextualisierter Zugriff auf das Quellenverzeichnis im DRW

[17] Storrer (2000b:114) spricht in diesen Fällen von der „Appellativ-Funktion" der Links.
[18] Zur Hypertextualisierung von Literaturangaben vgl. Kammerer (1998:161–164).

Die Quellensiglen in den Belegen sind als Hyperlinks eingerichtet. Durch Mausklick auf die Sigle öffnet sich ein Popup-Fenster, in dem man die vollständige Titelaufnahme der jeweiligen Quellensigle einsehen kann. Diese Hypertextualisierungsform für das Quellenverzeichnis wurde z.B. auch in der elektronischen Version der mittelhochdeutschen Wörterbücher gewählt.

2.1.1.3 Belegzitate sind in der Regel aus ihrem Kontext herausgelöste, und möglicherweise durch Auslassungen noch weiter verkürzte Textausschnitte, die den Gebrauch eines Wortes dokumentieren und dem Benutzer sprachliche, semantische und sachliche Informationen zu dem von ihm nachgeschlagenen Wort vermitteln.[19] Es handelt sich um dekontextualisierte und in einen neuen Aussagezusammenhang gebrachte Textteile, die für Wörterbuchbenutzerinnen und -benutzer mit begriffsorientierten Fragestellungen keine hinreichende Informationsquelle darstellen. Nicht zu Unrecht warnt der Historiker Algazi (1996:38) für begriffsgeschichtliche Untersuchungen davor, dem ‚Wörterbucheffekt' zu erliegen, und meint damit die Gefahr, „vom Kontext zu abstrahieren, Textstellen ohne Vor- und Nachgeschichte zu isolieren und zusammenzustellen, um der Deutlichkeit halber separat erscheinende ‚Bedeutungen' herauszuarbeiten".

Den wohl signifikantesten lexikographischen Mehrwert eines Online-Wörterbuchs schafft daher die durch Hypertextualisierung mögliche Verknüpfung von einem Wörterbuch mit seinem Corpus[20], genauer gesagt, vom einzelnen Belegzitat zum jeweils zugehörigen Volltext. Die Verknüpfung von Belegzitat und Volltext bietet für jeden Benutzer und jede Benutzerin die Option, je nach Benutzungsanliegen, eine Rekontextualisierung des Belegzitats vornehmen zu können. Sie ist immer dann möglich, wenn das Corpus (in Teilen oder als Ganzes) in digitalisierter Form, also entweder in Form von elektronischen Faksimiles oder in Form von maschinenlesbaren Texten, vorliegt.[21] Daß die Online-Wörterbücher dabei auch in zunehmendem Maße auf die Ressourcen des WWW, sei es in Form von elektronischen Faksimiles oder sei es in Form von maschinenlesbaren Texten, zurückgreifen können, zeigt das folgende Beispiel aus dem DRW online (abbildung 2).

In diesem Beispiel wurde ein Belegzitat[22] aus der Bambergischen Halsgerichtsordnung von 1507 mit der entsprechenden Faksimileseite der Originalausgabe verknüpft, die von der Universitätsbibliothek Mannheim im Rahmen des MATEO-Projektes[23] online im WWW verfügbar gemacht wurde.

---

[19] Vgl. dazu z.B. Lemberg 1996; Reichmann 1988; Speer 1991.

[20] Hypertextualisierungen zwischen Wörterbuch und Quellentext bringen auch aus der Perspektive der Editionswissenschaft eine völlig neue Qualität, bietet sich doch auch die andere Verweisrichtung von einem digitalen Text zu einem oder mehreren Bezugswörterbüchern an. Eine erste Realisierung bietet die Online-Publikation des Trierer Korpus mittelfränkischer Urkunden des 14. Jhs. (http://gaer27.uni-trier.de/Urkunden/welcome.htm). Zu jeder Urkunde gehört ein lemmatisierter Index. Die Einträge in diesem Index führen als Hyperlinks zu den entsprechenden Lemmata im Online-LEXER (Nachweis in Kap. 6.1).

[21] Zur Online-Konzeption des DRW vgl. Lemberg/Petzold/Speer 1998 sowie die Homepage des Wörterbuchs.

[22] Zu Methoden und Verfahrensweisen der Belegbearbeitung bei der Produktion von Wörterbuchartikeln in der lexikographischen Datenbank des DRW vgl. Lemberg 2000.

[23] MATEO = Mannheimer Texte online: http://www.uni-mannheim.de/mateo/ . – Der im DRW verwendeten Edition der Bambergischen Halsgerichtsordnung, hrsg. von Josef Kohler und Willy

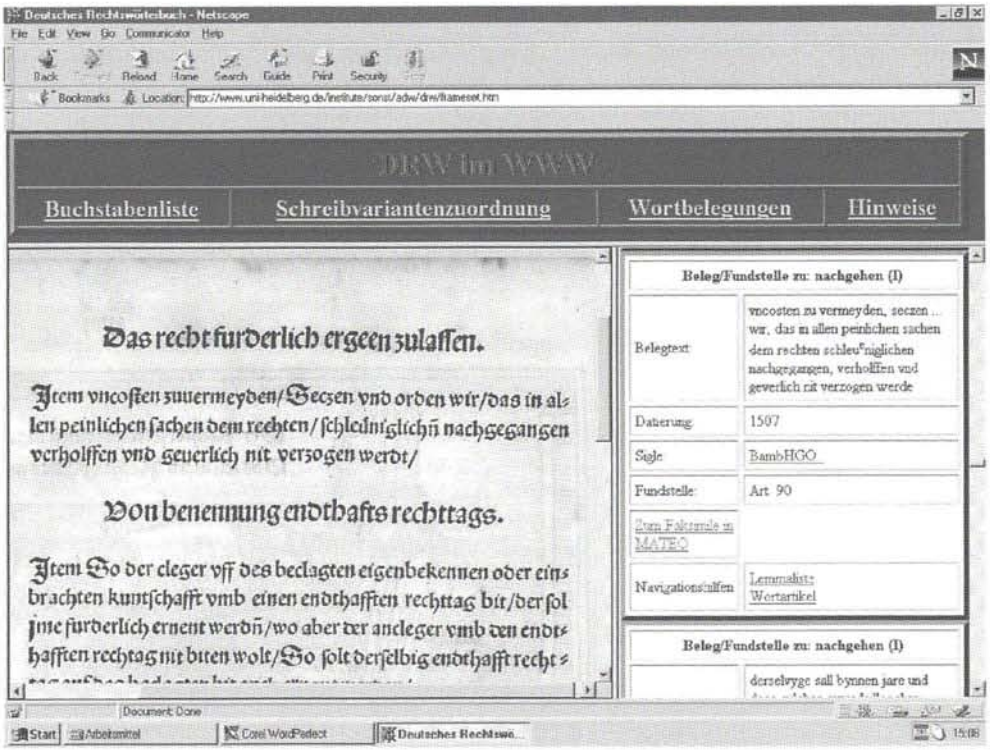


Abb. 2: Belegzitat im DRW (rechtes Fenster) und hypertextualisierter Zugriff auf die faksimilierte Originalseite der Quelle (linken Fenster)

2.1.1.4 Je nach Fragestellung von seiten der Benutzerinnen und Benutzer sind die Informationen eines einzigen Wörterbuchs nicht ausreichend. Weitere Wörterbücher müssen konsultiert werden. Dies kann je nach Gegebenheit einen weiteren Gang zu einem Regal, in einen anderen Raum oder gar in ein anderes Gebäude bedeuten. Daher stellt die Verknüpfung von Online-Wörterbüchern untereinander eine weitere sinnvolle Form der Hypertextualisierung dar. Wie das aussehen kann, zeigt Abbildung 3.

Im Artikelkopf des Wörterbuchartikels **Nase** im Online-DRW stehen die Informationspositionen *Wortklasse*, *sprachliche Erläuterung*, *Navigationshilfen* und *Wörterbücher*. In der Informationsposition *Wörterbücher* befindet sich ein Link auf den Online-LEXER. Weitere Verlinkungen in dieser Position sind denkbar, in Ermangelung von weiteren Online-Wörterbüchern zur Zeit aber nicht realisiert.

Bei Aktivierung des Links im Online-DRW auf den LEXER öffnet sich im rechten Fenster der Wörterbuchartikel zu **Nase** im LEXER, so daß man die beiden Wörterbuchartikel parallel rezipieren kann. Eine weitere Hypertextualisierung ist im Online-LEXER bereits realisiert. In der ersten Zeile des Artikels *Nase* folgt auf die Genusangabe eine in runde Klammern gesetzte Angabe von römischen und arabischen Zahlen, die im gedruckten

---

Scheel (1902), liegt der hier abgebildete Originaldruck zugrunde. Die Abweichungen in der Graphie entsprechen den Editionsprinzipien der Herausgeber (S. XC).

LEXER den Verweis auf das mittelhochdeutsche Wörterbuch von Benecke/Müller/Zarncke darstellt. In der Online-Fassung ist diese Darstellung um einen nach rechts weisenden, roten Pfeil ergänzt. Dieser Pfeil ist ein Link auf die Online-Fassung von BMZ.

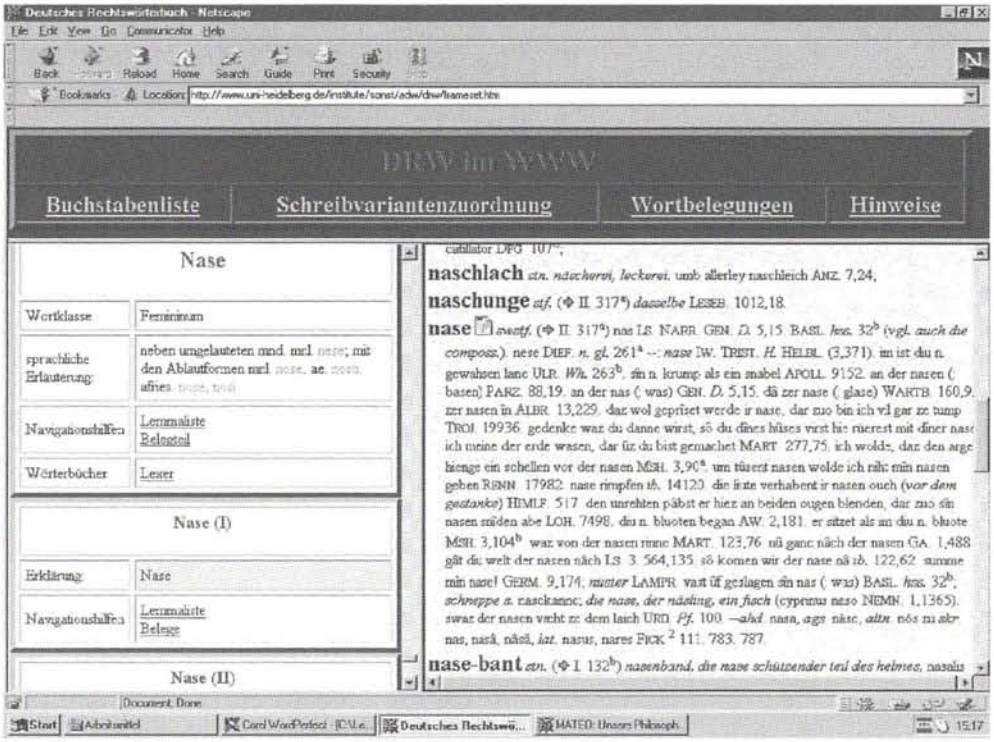


Abb. 3: Link vom DRW (linkes Fenster) zum Online-LEXER (rechtes Fenster)

2.1.1.5 Folgende, für wissenschaftliche Wörterbücher relevante Hypertextualisierungsformen seien noch exemplarisch genannt:

- Verknüpfung mit Hilfstexten, z.B. der Benutzungsanleitung[24] des Wörterbuchs
- Verknüpfung mit anderen Formen der Online-Wissensrepräsentation, die ihren Gegenstand in einer inhaltlich-systematisch geordneten Weise darstellen, also mit Handbüchern, Grammatiken, Sachlexika, kartographischen Werken, Bild- und Textdokumentationen usw.

2.1.1.6 Die Vernetzung von Wörterbüchern bringt uns in der Online-Lexikographie zu völlig neuen Informationstypen. Ein Beispiel dafür ist der Verbund mittelhochdeutscher Wörterbücher im Internet, wie er in Abschnitt 2.1.1.4 kurz angesprochen wurde (ausführlich dazu Burch/Fournier/Gärtner 1998 und Burch/Fournier in diesem Band).

[24] Für die Online-Konzeption des DRW ist eine Hypertextualisierung der Benennungen der einzelnen Informationspositionen vorgesehen (Lemberg/Petzold/Speer 1998:271).

Ein weiteres Beispiel entwickelte Martin Schröder (1997) für die Dialektlexikographie. Seine Grundidee ist die Schaffung eines digitalen DEUTSCHEN DIALEKTWÖRTERBUCHS (DDW), das durch die Vereinheitlichung und Zusammenführung aller dialektalen Sprachdaten des Deutschen die Einzelprojekte der Territorialwörterbücher erstmalig zusammenfaßt, nach Schröders Worten „gewiß eine alte Idee, vielleicht die älteste der Dialektlexikographie überhaupt" (S. 63), die aber durch die computergestützte Produktion der Wörterbücher auf der Basis einer lexikographischen Datenbank und damit der vereinheitlichten Datenstrukturierung eine neue Aktualität erfährt. Voraussetzung ist nach Schröder eine offene Datenbankstruktur, die es zum Beispiel ermöglicht, aus der Gesamtmenge der Daten je nach Fragestellung regionale, lokale oder idiolektale Wörterbücher bzw. Wortschatzzusammenstellungen herauszufiltern. Der Ansatz hat einen kollegialen Gedankenaustausch angeregt, und man darf gespannt sein, ob und wie die Dialektlexikographie diesen Grundgedanken aufgreifen und umsetzen wird.

Ein anderes Gedankenmodell liegt dem Forschungsprojekt LPI (LEXIKOGRAPHISCHES PRODUKTIONS- UND INFORMATIONSSYSTEM) zugrunde, das im Mai 1998 in Heidelberg vom Leiter des DEUTSCHEN RECHTSWÖRTERBUCHS, Heino Speer, konzipiert wurde und seit September 1999 in Heidelberg und Konstanz realisiert wird. Ausgehend von dem konkreten Bedarf, der bei der lexikographischen Arbeit am DEUTSCHEN RECHTSWÖRTERBUCH mit seinen zahlreichen Einzelsprachen und Varietäten von Einzelsprachen an Wörterbüchern und an enzyklopädischen Nachschlagewerken wie z.B. dem HANDWÖRTERBUCH ZUR RECHTSGESCHICHTE oder dem LEXIKON DES MITTELALTERS besteht, hat Speer die Idee eines lexikographisch-enzyklopädischen, datenbankbasierten und internetfähigen Informationssystems entwickelt, zu dem nun in dem LPI-Projekt auf der Basis des Konstanzer Hypertextsystems (KHS)[25] für das DEUTSCHE RECHTSWÖRTERBUCH und das FRÜHNEUHOCHDEUTSCHE WÖRTERBUCH erste Prototypen entwickelt werden. Bestandteile des Systems sind einmal eine Produktionsplattform und zum anderen eine Benutzerplattform im Internet. Wörterbücher können also künftig innerhalb des Systems produziert (vgl. Kap. 3), oder aber nach der Produktion importiert werden, so daß bisherige Produktionstechnologien beibehalten werden können. Das System soll auch für andere Nachschlagewerke als Wörterbücher offen sein, so dass bei einer Integration von Wörterbüchern und enzyklopädischen Nachschlagewerken im Internet eine digitale Basis historischer Kulturforschung entstehen wird. Zugleich sollen die einzelnen Informationseinheiten innerhalb von LPI mit möglichst vielen digitalen Informationen auf verschiedenen Ebenen verknüpft werden, so dass man im Idealfall von einem Belegzitat zu einem Artikel über den Autor, auf eine elektronische Edition und von dort auf Faksimiles der Handschrift rekurrieren könnte – dass zugleich aber die Behandlung dieses Worts in Wörterbüchern anderer Sprachen und Sprachvarietäten sichtbar würde.

Eine Neukonzeption liegt auch mit dem Projekt LEKSIS (vgl. die Angaben in Kap. 6.1) vor, einem lexikalisch-lexikologischen, corpusbasierten Such- und Informationssystem zum

---

[25] Es handelt sich dabei um ein textorientiertes System zur Wissens- und Informationsverwaltung mit integrierten Kommunikationselementen (WITH), das am Informationswissenschaften-Lehrstuhl von Rainer Kuhlen an der Universität Konstanz entwickelt wurde, mit dem Ziel, Konversionsmöglichkeiten von Fachtexten in nicht-lineare Strukturen zu finden. Das Projekt konzentriert sich z.B. auf die Konversion und Integration von Arbeiten mit Online-Datenbanken, von Suchergebnissen aus browsing-orientierten Informationssystemen, von intern gebräuchlichen Textdateien, von der E-Mail-Verwaltung innerhalb des Systems und einiges mehr (vgl. Hammwöhner 1997 sowie online http://hoechst.inf-wiss.uni-konstanz.de:57786/getKHS/WWW-Forschung/WITH).

Wortschatz der deutschen Gegenwartssprache, einem System, „das die sprachwissenschaftlichen Inhalte von vornherein mit Blick auf eine Nutzung in den Mehrwertdiensten des Internet entsprechend auswählt, dimensioniert und strukturiert" (Fraas/Haß-Zumkehr 1999:294 sowie Haß-Zumkehr in diesem Band).

2.1.2 Einsätze von multimedialen Elementen

Unter Multimedia versteht man gemeinhin die durch Hypertextsysteme ermöglichte Integration von Text, Videoelementen (Abbildungen, stehende oder animierte Graphiken und Filmsequenzen) sowie Audioelementen (Laute, Töne, gesprochene Sprache, Geräusche oder Musik). Multimedia gilt als neue Art und Weise der Mediennutzung in Informationsprozessen (Issing/Klimsa 1997:1; Klimsa 1997:9), und ist damit besonders geeignet, lexikalisch-lexikographisches Wissen mit verschiedenen Ausdrucksmitteln zu präsentieren und über verschiedene Sinneskanäle zu vermitteln.[26] Dies gilt nicht nur für die eher populären Wörterbücher und Nachschlagewerke, die häufig nach dem *Infotainment*-Prinzip (eine Zusammenfügung aus *Info*rmation und Enter*tainment*) gestaltet werden, sondern auch für wissenschaftliche Wörterbücher. Im Folgenden seien nur einige wenige Beispiele für die Einsatzmöglichkeiten von Multimedia in wissenschaftlichen Wörterbüchern genannt. Zu multimedialen Elementen im zweisprachigen Wörterbuch vgl. den Beitrag von K. Petelenz in diesem Band.

2.1.2.1 Die Kombination von sprachlicher Beschreibung und jeder Form von Bildern kann gerade in der Nennlexik eine wesentliche Veranschaulichung bringen, und damit zu einer effizienteren Wissensvermittlung beitragen.[27] Dies gilt insbesondere für historische und dialektologische Wörterbücher,[28] die sehr häufig Gegenstände aus einer uns heute nicht mehr vertrauten Alltagswelt beschreiben.[29] Dabei bleiben die rein sprachlichen Beschreibungen in Wörterbüchern häufig unzulänglich, wie die folgenden drei (beliebigen) Beispiele zeigen:

aftergeschir: „Teil des Geschirres, der es dem Zugtier ermöglicht, den Wagen zu bremsen" (FWB)
armbrustnus: „Nuß der Armbrust" (FWB)
Notariatszeichen: „von einem öffentlichen Notar anstelle eines Siegels verwendetes amtliches Zeichen, Notariatssignet" (DRW).

Abbildungen könnten in diesen und unzähligen ähnlichen Fällen eine Veranschaulichung des beschriebenen Gegenstandes und damit oft auch seiner Funktionsweise bringen.

---

[26] Einen umfassenden Überblick über die Integration von Text, Bild, Ton und Video in multimedialen Wörterbüchern gibt Storrer (1998:108–114).

[27] Vgl. Hupka 1989a/b; Weidenmann 1997; zum lexikographischen Mehrwert von Abbildungen vgl. den Beitrag von K. Petelenz in diesem Band (Kap. 7). Zum Synergieeffekt von Text und Bild in wissenschaftlichen Hypertexten vgl. Cölfen/Schmitz 1997.

[28] Wobei gerade Dialektwörterbücher gelegentlich auch in der Printversion mit Abbildungen arbeiten. Für die Dialektlexikographie faßt Martin Schröder (1997:58) zusammen: Datenbanken „werden zu ganz neuen Darstellungsformen führen, etwa zur Integration von akustischen Proben gesprochener Sprache... oder von an jeder beliebigen Stelle zugänglichen Karten und Abbildungen".

2.1.2.2 Die Lexikographie verfügt zwar über erste Untersuchungen über den Stellenwert und die Einsetzbarkeit von Abbildungen im Wörterbuch (Hupka 1989). Daß diese bisher aber eher auf Lernerwörterbücher beschränkt blieben, liegt nicht nur am mangelnden Druckraum, sondern auch an dem für die Produktion eines Wörterbuchs relevanten Zeitfaktor. Andererseits zeichnet sich ab, daß auch Bilddatenbanken das neue Medium Internet als Publikationsmedium nutzen werden. Onlinewörterbücher können sich diese Bilddatenbanken im Internet (Brunschwig 1996/97) zunutze machen und Wort und Gegenstand durch extratextuelle Links miteinander verknüpfen. Dies sieht z.B. die Konzeption für das DEUTSCHE RECHTSWÖRTERBUCH in seiner Online-Version vor (Lemberg/Petzold Speer 1998:273f.).

2.1.2.3 Auch die Kombination von Sprache und auditiven Modulen, z.B. die Einbindung von Tierstimmen im WÖRTERBUCH DER TIERLAUTE oder die Vertonungen schwäbischer Ausdrücke wie *Herrgottsjesuskreizkrabbesacklebdagsjenseitswille* im Wörterbuch SWABIAN INTO ENGLISH (vgl. jeweils Kap. 6.1), muß nicht auf den Infotainment-Bereich beschränkt bleiben. Gerade für die Dialektlexikographie und für die historische Lexikographie zeichnen sich auch hier durch die Einbindung vertonter Beispiele völlig neue Dimensionen ab. Hörbeispiele zu gegenwartssprachlichen dialektalen Ausdrücken oder aber auch z.B. zu mittelhochdeutschen oder mittelniederdeutschen Ausdrücken oder Sätzen würden den Benutzerinnen und Benutzern eines Wörterbuchs eine neue Rezeptionsdimension vermitteln und können damit als wörterbuchpädagogischer Mehrwert gelten.

2.2 Vom statischen zum dynamischen Wörterbuch

2.2.1 Bei der fortschreitenden Bearbeitung eines Wörterbuchs, insbesondere bei Wörterbüchern mit einer längeren Bearbeitungsdauer, wie dies in der Regel bei wissenschaftlichen Wörterbüchern der Fall ist, ergeben sich Veränderungen in der Wörterbuchbasis (z.B. neue Quellentexte, neue Belege) oder den sonstigen Wissenbeständen (wie Neu-Editionen mit Korrekturen zur Datierung oder zu Wortvorkommen im Text). Zudem erhöht sich bei den Lexikographinnen und Lexikographen im Laufe der Bearbeitungszeit auch das Wissen um den Beschreibungsgegenstand und damit die lexikographische Beschreibungskompetenz. Dies zusammen führt fortlaufend zu Korrekturen und Ergänzungen eines Wörterbuchs auch in seinen bereits gedruckten Teilen, und zwar ebenso in allen einzelnen Informationspositionen des Wörterbuchartikels wie im mediostrukturellen und makrostrukturellen Bereich durch die permanent weitergeführte lexikalische Vernetzung oder das Einfügen neuer Wörterbuchartikel aufgrund neuer Belege.

Zwar werden Ergänzungen und Korrekturen in der Regel systematisch gesammelt, dies muß aber nicht zwingend zu einem Nachtragsband nach Abschluß des Wörterbuchs führen. Und wenn es einen Nachtragsband gibt, ist es fraglich, welche Korrekturen er enthalten wird oder auch enthalten kann. So ist es wohl völlig ausgeschlossen, Verweise in der Verweisrichtung von A nach Z zu ergänzen, selbst dann, wenn sie im Produktionsmedium des Wörterbuchs, z.B. einer lexikographischen Datenbank während der laufenden Artikelarbeit, gesetzt werden, wie dies z.B. in der lexikographischen Praxis des DRW der Fall ist (Speer 1995). Nur der Druck einer Neuauflage könnte diese Form der Ergänzung vermitteln, was

[29] Zur Funktion von Abbildungen zur lexikographischen Beschreibung der Rechtssprache und der entsprechenden Online-Konzeption vgl. Lemberg/Petzold/Speer (1998:273f., bes. Anm. 34).

aber bei mehrbändigen Wörterbüchern meist nicht der verlegerischen Realität entspricht. Das gedruckte Wörterbuch ist also in der Regel eine statische Einheit mit nur wenigen Korrekturmöglichkeiten.

Das Publikationsmedium Internet bietet die Chance, von der Statik des Wörterbuchs auf eine Dynamik zu wechseln.[30] Dynamik meint in diesem Fall fortgesetzte Korrektur- und Ergänzungsmöglichkeiten des Wörterbuchs auch in seinen bereits geschriebenen Teilen. Dies gilt für alle Wörterbücher, deren Produktion noch nicht abgeschlossen ist, also sowohl für Wörterbücher, deren Online-Fassung auf Datenkonversion beruht, als auch für Wörterbücher, die von vornherein als Online-Wörterbücher konzipiert werden. Zur Veranschaulichung der Tragweite dient ein Blick in die lexikographische Praxis des DRW.

2.2.2 Das DRW ist bisher in 9 Bänden und 3 Doppelheften des 10. Bandes erschienen, 16 Bände soll es ingesamt geben. Das gesamte Werk wird als Hybridwörterbuch[31] erscheinen, d.h. zum einen als Druckwerk und zum anderen im Internet. Die Produktion der Wörterbuchartikel erfolgt in einer lexikographischen Datenbank (Speer 1995). Beim Schreiben eines Wörterbuchartikels bewegt man sich als Lexikograph mit Recherchefragen innerhalb der Datenbank immer wieder auch in Bereichen bereits gedruckter Artikelteile. Stößt man dabei auf Dateninkonsistenzen oder fehlerhafte Daten (z.B. falsche Quellensiglen, veraltete Datierungen, orthographische Fehler, Verweisfehler oder unzureichende Bedeutungsbeschreibungen), werden diese sofort korrigiert. So findet während des fortschreitenden lexikographischen Prozesses eine rückwirkende Korrektur in Wörterbuchartikeln statt, die bereits in gedruckter Fassung erschienen sind. Auf die Produktionszeit der zu schreibenden Artikelstrecken haben alle diese Korrekturen keinen erwähnenswerten Einfluß.

Verweise auf bereits gedruckte Wörterbuchartikel werden in der Datenbank mit den entsprechenden Gegenverweisen versehen. Welche Auswirkungen dies auf den konkreten Informationsgehalt eines Wörterbuchartikels hat, soll das folgende Beispiel veranschaulichen: der Wörterbuchartikel **Ordal** enthält ein Verweisnest mit insgesamt 18 Verweisen auf Wörterbuchartikel, in denen ein semantischer oder sachlicher Bezug zu Gottesurteilen abgehandelt wird (z.B. auf **Bahrrecht** I, **Hexenprobe**, **Gottesurteil**, **Kesselfang** usw.). Mit dem Setzen der Verweise auf diese einzelnen Wörterbuchartikel wurden jeweils auch Rückverweise auf den Artikel **Ordal** gesetzt, so daß der Informationsgehalt jedes einzelnen dieser Wörterbuchartikel dadurch vergrößert wird, daß durch den Verweis auf **Ordal** jeweils ein Verweisnest erschlossen wird.

Da die Online-Fassung des DEUTSCHEN RECHTSWÖRTERBUCHS auf einem Export der entsprechenden Daten aus der Datenbank beruht und zudem in bestimmten Abständen ein Update erhält, stehen in der Online-Version die gegenüber der Printversion jeweils aktualisierten Artikelfassungen zur Verfügung.

---

[30] Martin Schröder (1997:60) spricht in diesem Zusammenhang vom „Übergang vom konventionellen Abschlußwörterbuch zum neueren Ausbauwörterbuch".

[31] Ich verwende diesen Terminus analog zu dem von Eibl/Jannidis/Willems vorgeschlagenen Terminus ‚Hybrid-Edition' (1999:72). Diese Autorengruppe übernimmt das Bestimmungswort aus der Fachsprache der Technik, in der von Hybrid-Lösungen die Rede ist, wenn zur effektiven Problemlösung zwei verschiedene Technologien miteinander gekoppelt werden.

3 Aspekte der Wörterbuchproduktion

Nicht jedes Wörterbuch entsteht in einer Forschungsstelle oder in einem Verlag, also an einem einheitlichen Ort. So laufen z.B. die Planungen für das neue mittelhochdeutsche Wörterbuch derzeit auf zwei Arbeitsstellen in Trier und in Göttingen hinaus. Ein extremes Beispiel für die dezentrale Erfassung eines Wörterbuchs ist das auf 12 Bände zu je 1000 Seiten konzipierte FRÜHNEUHOCHDEUTSCHE WÖRTERBUCH. Es wurde in Heidelberg von Oskar Reichmann begründet, konzipiert und in den ersten Bänden auch geschrieben. Die weiteren Bände entstehen nun zeitlich parallel in Heidelberg, in Mannheim am Institut für deutsche Sprache, sowie an den Universitäten in Münster, Bonn, Halle, Bochum, Kopenhagen und in Newcastle/GB.[32] Wörterbücher mit dieser Form des Herstellungsprozesses bedürfen einer extrem aufwendigen redaktionellen und logistischen Leistung von Seiten des Herausgebers, um die Homogenität der Artikelstrukturen, die Abstimmung der Lemmaansätze und die mediostrukturellen Vernetzungen auf hohem Niveau zu halten. Hinzu kommen aufwendige Korrespondenzen der einzelnen Bearbeiter.

Mit dem Internet steht der Lexikographie ein völlig neues Produktionsmedium zur Verfügung. So ist auch in dem in Heidelberg angelaufenen LPI-Projekt (vgl. oben, Kap. 2.1.1.6) der Entwurf, die Planung und Realisierung einer datenbankbasierten und internetfähigen Arbeitsoberfläche für dezentral arbeitende Wörterbücher mit Hilfe des Konstanzer Hypertext-Systems vorgesehen. Die Bereitstellung von lexikographischen Datenbanken im Internet ermöglicht die Erstellung von homogenen und vollständig vernetzten Wörterbüchern von weltweit verteilten Standorten aus. Diese neue Herstellungsform eröffnet gerade für Wörterbücher zu Spezialgebieten, zu denen es weltweit nur einige wenige Experten gibt, größere Realisierungschancen (so auch Storrer 1998:124f.).

4 Aspekte der Wörterbuchbenutzung

4.1 Äußerer Benutzungskontext[33]

Wissenschaftliche Wörterbücher zählen nicht immer zu der Ausstattung des Handapparates am eigenen Schreibtisch, sei es im Dienstzimmer, weniger noch zuhause, sondern sind häufig nur in den entsprechenden Fachbibliotheken benutzbar. Diese Bibliotheken wiederum unterliegen bestimmten Öffnungszeiten, so daß die Benutzung eines wissenschaftlichen Wörterbuchs unter diesen Umständen einer starken Reglementierung unterliegt (Wiegand 1998c:521), was paradox anmutet, dienen doch gerade Wörterbücher dazu, akuten Wissensbedarf, wie er in Situationen der Textrezeption (Reichmann 1986:26f.) entsteht, zu befriedigen. Verfügt man über einen Online-Arbeitsplatz, entfallen diese zeitlichen und räumlichen Beschränkungen. Damit wird die „Qualität wissenschaftlicher Forschung insgesamt durch die universelle Verfügbarkeit" (Lemberg/Petzold/Speer 1998:281) eines wesentlichen Hilfsmittels erhöht.

---

[32] Eine Übersicht über den aktuellen Bearbeitungsstand und die Bearbeiterinnen und Bearbeiter findet sich auf der Homepage des Wörterbuchs.

[33] Zum äußeren Benutzungskontext zählen nach Wiegand (1998c:520) die Umstände der Benutzung eines Wörterbuchs wie Ort, Zeit und Dauer sowie die Benutzungszusammenhänge.

4.2 Kooperation von Lexikographen und Benutzern[34]

4.2.1 Ein wissenschaftliches Wörterbuch ist ein Hilfsmittel zum Verständnis von Texten oder zur Beantwortung einzeltexttranszendierender, z.B. begriffsgeschichtlicher Fragestellungen (Reichmann 1986:24ff.). Man kann davon ausgehen, daß die Benutzerinnen und Benutzer wissenschaftlicher Wörterbücher in der Regel in wissenschaftlichen Disziplinen arbeiten und infolgedessen punktuell, nämlich in ihrem jeweiligen Spezialgebiet, ein umfassenderes und zugleich wesentlich detaillierteres Fachwissen aufbringen, als dies jeder noch so erfahrene und kundige Lexikograph vermag. Und im Grunde genommen könnte ein Benutzer oder eine Benutzerin punktuell zu jeder Informationsposition im Wörterbuchartikel detaillierteres und fundierteres Wissen beitragen, weil ihm oder ihr eben an bestimmten Stellen, zu bestimmten Fragen oder Aspekten dieses fundiertere Wissen zur Verfügung steht. Dabei kann es sich ebenso um die Schließung von Datenerhebungslücken – also Hinweise auf neue Wörter, bislang im Wörterbuch nicht belegte Bedeutungen von Wörtern oder Zitaten – handeln, wie um die Identifizierung entstellter Wörter, die Hinweise auf Editions- oder Datierungsfehler oder aber eben um die Korrektur von Bedeutungsangaben, die auf detaillierterem Sachwissen[35] beruhen. Aus diesen Gründen wäre ein Kontakt zwischen Lexikograph und Benutzer eines Wörterbuchs sicher ein hoher Gewinn für die in den Wörterbuchartikeln dargebotenen Informationen. Diese Kontakte finden in der lexikographischen Praxis des Wörterbuchschreibens und der Wörterbuchbenutzung bislang kaum statt. Ein klassisches Print-Wörterbuch wird in einem Institut oder einer Forschungsstelle erarbeitet, geht dann an einen Verlag, wird dort gedruckt und ausgeliefert. Sind die Artikel erst einmal publiziert, wandern neue Informationen dazu, insbesondere Stellungnahmen, Anregungen und Korrekturen von Benutzern, in Archivkästen und Korrespondenzordner, und es bleibt offen, ob sie jemals in einem Nachtragsband erscheinen werden. Da die Benutzer das wissen, sind sie auch wenig motiviert, Verbesserungsvorschläge zu machen. Es gibt also, von einigen Ausnahmen abgesehen, zwischen den Lexikographen und den Benutzern eines Printwörterbuchs keinen Kontakt. Den Lexikographinnen und Lexikographen ist der Benutzerkreis, für den sie das Wörterbuch konzipieren und schreiben, meist völlig unbekannt. Dies spiegeln auch die Vorwörter der Wörterbücher wider.[36]

4.2.2 Eine ganz andere Voraussetzung haben wir, wenn das Wörterbuch im Internet publiziert ist, dem Medium, dessen Grundidee ursprünglich der weltweite, rasche und unkomplizierte Informationsaustausch zwischen Wissenschaftlern war. Und die meisten der bisher im Internet veröffentlichten Wörterbücher suchen auch explizit den Dialog mit ihren Benutzern und ermöglichen verschiedene Formen der Partizipation bis hin zur kollaborativen Wörterbucherstellung (Storrer 1998:124f.).

---

[34] Dazu ausführlich Lemberg 1998e.

[35] Gerade die Bedeutungslexikographie stellt die Lexikographinnen und Lexikographen immer wieder vor Herausforderungen, weil sie in den Wörterbuchartikeln ein sehr breitgefächertes und zugleich sehr detailliertes Fachwissen aufbringen müssen: Artikel wie **lakschauen**, **Moordeich** oder **nachdeichen** erfordern Spezialkenntnisse im Deichwesen, bei der Bearbeitung von Wörtern wie **Lähme**, **Meißelwunde** oder **Nasebreud** ist medizinisches Grundwissen gefragt. Vorstellungen über die Rezeption des Römischen Rechts werden ebenso benötigt, wie das ganz konkrete Wissen über das Bierbrauen, die Schafschur oder die Funktionen eines Mühleisens. – Die Beispiele stammen alle aus dem DRW.

Daß auch die wissenschaftliche Online-Lexikographie diese Kommunikationsmöglichkeit nicht ungenutzt lassen sollte, dürfte außer Frage stehen. Fraglich hingegen ist, ob es wirklich genügt, an irgendeiner Stelle in der Homepage eines Online-Wörterbuchs die E-Mail-Adresse eines Ansprechpartners anzugeben, oder ob man nicht ein Konzept entwickeln sollte, das die Vielschichtigkeit des Mediums Internet besser ausnutzt. Es böte sich z.B. an, daß man ein Korrespondenzforum zum Wörterbuch einrichtet, in dem die Benutzerinnen und Benutzer in verschiedenen Rubriken ihre Kommentare, Ergänzungen oder Fragen abgeben können. Denkbar wäre z.B. eine Rubrik zur Wörterbuchbasis, dann eine weitere Rubrik zur Benutzung des Wörterbuchs und eine dritte Rubrik, die sich gezielt auf die in den einzelnen Wörterbuchartikeln enthaltenen Informationen bezieht. Die Kommentare in letzterer Rubrik könnten dann über Links jeweils zu den entsprechenden Wörterbuchartikeln führen, und man könnte andererseits bei den einzelnen Wörterbuchartikeln einen Kommunikationsicon gestalten, mit dem sich die Benutzer vom Wörterbuchartikel aus ansehen können, ob es zu diesem Artikel bereits Ergänzungen, Kommentare oder ähnliches gibt.

4.2.3 Eine sehr gute Umsetzung für die Integration der Benutzerkommentare bietet das Onlinewörterbuch SWABIAN INTO ENGLISH,[37] von seiner Konzeption her eher ein Ulkwörterbuch, das sich auch als Informations- und Kommunikationsforum für weltweit verstreute Schwaben versteht, und das diesen ein gewisses virtuelles Heimatgefühl – nicht nur über das Forum *Laugebrezl worldwide* und *Talk about* – vermitteln möchte. In der Rubrik *Missing Words* können einmal die Benutzerinnen und Benutzer des Wörterbuchs angeben, für welche schwäbischen Wörter oder Wortverbindungen sie gerne englische Übersetzungen hätten, und zum anderen fragt der Betreiber des Wörterbuchs seinerseits die schwäbischen Spezialisten nach bestimmten Übersetzungen. Geht eine Antwort, also ein Übersetzungsäquivalent ein, wird das Ganze als Eintrag in das Wörterbuch gestellt, und von diesem Nutzerforum aus hypertextualisiert. Wir haben hier ein offenes Wörterbuch, dessen Fortschreibung von seinen Benutzern übernommen wird. Die Kommunikationsforen sind hier in die Makrostruktur des Wörterbuchs integriert und durch die Hypertextualisierung werden Verknüpfungen zwischen den Kommunikationsforen und den Wörterbuchartikeln selbst hergestellt.

4.2.4 Man sollte sich von der humorvollen Beschreibungsintention dieses Beispiels nicht darüber täuschen lassen, daß sich mit der Online-Publikation auch und gerade für die wissenschaftlichen Wörterbücher völlig neue Perspektiven eröffnen – und zu Recht weist Angelika Storrer in ihrem Fazit zur neuen Generation elektronischer Wörterbücher darauf hin, daß sich mit den Kommunikationsmöglichkeiten des Internets „bislang nicht denkbare Formen der kollaborativen Erarbeitung von Wörterbüchern ergeben, die vor allem für die nicht-kommerziell orientierte wissenschaftliche Lexikographie interessant sind" (Storrer 1998:127). Die Motivation von Seiten der Benutzer, sich aktiv zu Wörterbuchinhalten zu äußern, besteht meines Erachtens in dem Wissen, daß der Kommentar nicht mehr in Korrespondenzordnern abgelegt wird, sondern zusammen mit dem Wörterbuch veröffentlicht wird. Hinzu kommt, daß man aus der Benutzungssituation heraus seine Anmerkungen

[36] Vgl. dazu auch Mulch (1997:158).
[37] Vgl. Kap. 6.1; eine ausführliche Beschreibung dieses Wörterbuchs gibt Lemberg 1998e.

weitergeben kann, und dies in dem unkomplizierten, lockeren Stil, der in Emails generell gepflegt wird, auch wenn es um sehr seriöse Inhalte geht.

5 Ausblick

Mit dem Internet haben wir ein multifunktionales Medium, das Publikation und Kommunikation zugleich ermöglicht. So können wir Wörterbuchdaten in völlig neuen Präsentationsformen aufbereiten und das Informationsangebot im Wörterbuch erweitern, indem wir Verknüpfungen mit Korpustexten und anderen Informationssystemen herstellen. Wir haben auch eine Beschleunigung und Globalisierung der Informationsübermittlung und des entsprechenden wissenschaftlichen Diskurses. Die beiden Pole wissenschaftlicher Lexikographie, die Quellenbasis und die Ergebnispräsentation, verschmelzen mit dem dazwischen liegenden Erkenntnis- und Produktionsprozeß.[38] Es entstehen offene, dynamische Wörterbücher oder lexikographische Informationssysteme, denen fortlaufend alle Arten von Daten (neue Quellen, neue Zitate, neue Wörter usw.) hinzugefügt werden können bzw. müssen.[39] Damit wird auch die traditionelle Trennung zwischen Wörterbuch und enzyklopädischen Nachschlagewerken zugunsten eines umfassenden Informationssystems allmählich aufgehoben.[40] Dem außerordentlich gewachsenen Informationsbedarf der modernen Wissenschaften dürfte die hier vorgestellte Publikations- und Kommunikationstechnologie entsprechen können.

---

[38] Eine Entwicklung, die auch in anderen wissenschaftlichen Disziplinen prognostiziert wird. So beschreibt die Frankfurter Allgemeine Zeitung die Zukunftsvision des Historikers Robert Darnton zum Buch der Zukunft: „ein ‚pyramidal' aufgebautes elektronisches Werk, das mit dem herkömmlichen Buch nur noch wenige Gemeinsamkeiten hat. Jede seiner verschiedenen Schichten würde eine bestimmte Wissensebene repräsentieren. In der obersten könnte man eine prägnante Darstellung des Gegenstandes geben, die sich auch für einen Ausdruck anböte. Die unteren Ebenen würden nacheinander Vertiefungen, Dokumentationen und methodische und pädagogische Reflexionen des Autors liefern" (FAZ, 12. Mai 1999, S. N 5).

[39] Online-Wörterbücher bedürfen also auch nach Abschluß der eigentlichen Artikelarbeit einer weiteren Bearbeitung, einer Pflege des Datenbestands, und zwar sowohl in elektronischer als auch in inhaltlicher Hinsicht. Die ständige Weiterentwicklung der elektronischen Produktions- und Publikationstechnologie erfordert eine ständige Pflege der elektronisch konservierten Daten. Sie müssen von Zeit zu Zeit in neue EDV-Systeme konvertiert werden. Dies mag gerade aus der Finanzierungsperspektive auf den ersten Blick aufwendig erscheinen. Andererseits gilt es aber zu bedenken, daß z.B. herkömmliche Belegarchive ebenfalls der Pflege bedürfen. Sie müssen verwaltet, gelagert und zur Konservierung verfilmt werden. Die Filme wiederum müssen, wenn sie verfallen oder wenn es keine entsprechenden Lesegeräte mehr gibt, auf neue physikalische Träger kopiert werden, was wiederum mit mit Kosten verbunden ist.

[40] Ganz im Sinne von Reichmann 1986, 244f.

6 Literatur

Ich danke Heino Speer für zahlreiche, anregende Gespräche. – Die Publikationen im WWW wurden zuletzt geprüft im März 2000.

6.1 Wörterbücher

Beard, Robert: A Web of On-line Dictionaries, 1996, last updated July 1999.
Available: http://www.facstaff.bucknell.edu/rbeard/diction.html .

DWB[1] = Deutsches Wörterbuch von Jacob und Wilhelm Grimm. 16 Bände in 32 Teilbänden. Leipzig 1854–1954.

DWB[2] = Deutsches Wörterbuch von Jacob und Wilhelm Grimm auf CD-ROM und im Internet. Vorstellung des Projektes: http://gaer27.uni-trier.de/GrimmWB/grimmwb.htm .

Dialektwörterbücher im Internet (in Auswahl):
http://www.smo.uhi.ac.uk/saoghal/mion-chanain/Failte_en.html#Deutsch (Linkliste).
http://www.jakob.at/steffen/hess.html (Hessisch).
http://home.t-online.de/home/Holger_Kremb/platt.htm (Bremer Platt).

Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts (in Planung): http://www.bbaw.de/iag/dig_woerterbuch/index.html .

DRW = Deutsches Rechtswörterbuch. Wörterbuch der älteren deutschen Rechtssprache. Hrsg. von der Preußischen Akademie der Wissenschaften, später der Heidelberger Akademie der Wissenschaften. Bisher 9 Bde, sowie Bd. X, Hefte 1–6. Weimar 1914–1999. Homepage und erste Probestrecke: http://www.uni-heidelberg.de/institute/sonst/adw/drw/ .

FWB = Frühneuhochdeutsches Wörterbuch. Hrsg. von Ulrich Goebel und Oskar Reichmann. New York 1989ff. Homepage: http://www.rzuser.uni-heidelberg.de/~d68/html/fwb/fwb1.htm .

LEXER = Mittelhochdeutsches Handwörterbuch von Matthias Lexer. Zugleich als Supplement und alphabetischer Index zum Mittelhochdeutschen Wörterbuch von Benecke-Müller-Zarncke. 3 Bde. Leipzig 1872–1878. Elektronische Fassung: http://gaer27.uni-trier.de/MWV-online/MWV-online.html .

LEKSIS = LEKSIS – Wissen über Wörter. Das lexikalisch-lexikologische Informationssystem: http://www.ids-mannheim.de/wiw/ .

BMZ = Mittelhochdeutsches Wörterbuch. Mit Benutzung des Nachlasses von Georg Friedrich Benecke, ausgearbeitet von Wilhelm Müller und Friedrich Zarncke. 3 Bde., Leipzig 1854–1861. Elektronische Fassung: http://gaer27.uni-trier.de/MWV-online/MWV-online.html .

Mittelhochdeutsches Wörterbuch (in Vorbereitung): http://gaer27.uni-trier.de/MhdWB/ .

Kemmer, Thomas: Swabian into English. 1997. http://www.schwaebisch-englisch.de/ .

Sounds of the Worlds Animals. http://www.georgetown.edu/cball/animals/animals.html .

6.2 Sekundärliteratur

Baasner, Rainer (1999): Digitalisierung – Geisteswissenschaften – Medienwechsel? Hypertext als fachgerechte Publikationsform. In: Zeitschrift für Computerphilologie. Eine elektronische Zeitschrift zum Einsatz des PCs in der Literaturwissenschaft. Veröffentlicht am 11.2. 1999. http://computerphilologie.uni-muenchen.de/jg98/baasner.html .

Blumenthal, Andreas/Lemnitzer, Lothar/Storrer, Angelika (1988): Was ist eigentlich ein Verweis? Konzeptuelle Datenmodellierung als Voraussetzung computergestützter Verweisbehandlung. In: Gisela Harras (Hrsg.), Das Wörterbuch: Artikel und Verweisstrukturen, 351–373. Düsseldorf.

Bolter, Jay D. (1997): Das Internet in der Geschichte der Technologien des Schreibens. In: Münker/Roesler 37–47.

Brunschwig, Colette (1996/97): Die Forschungsstelle für Rechtsgeschichte im Spiegel alter und neuer Medien. In: Geschichte und Informatik, 67–73. Hrsg. von Hannes Schüle, Peter Bär, Gerold Ritter. Vol. 7/8.

Burch, Thomas/Fournier, Johannes/Gärtner, Kurt (1998): Mittelhochdeutsche Wörterbücher auf CD-ROM und im Internet. Der Einsatz von SGML in der Retrodigitalisierung lexikographischer Standardwerke. In: Akademie-Journal. Mitteilungsblatt der Konferenz der deutschen Akademien der Wissenschaften. Heft 2, 17–24.

Cölfen, Elisabeth/Cölfen Hermann/Schmitz, Ulrich (1997): Linguistik im Internet. Das Buch zum Netz – mit CD-ROM. Opladen.

Cölfen, Hermann/Schmitz, Ulrich (1997): Zur Synergie von Text und Bild in wissenschaftlichen Hypertexten. Theoretische und praktische Grundlage der Entwicklung multimedialer interaktiver Lernsoftware. In: Dagmar Knorr/Eva-Maria Jakobs (Hrsg.), Textproduktion in elektronischen Umgebungen, 223–237. Frankfurt (Textproduktion und Medium 2).

Eibl, Karl/Fotis Jannidis/Willems, Marianne (1999): Der Junge Goethe in neuer Ausgabe. Einige Präliminarien und Marginalien, 69–78. In: Roland Kamzelak (Hrsg.), Computergestützte Text-Edition. Tübingen (Beihefte zu editio 12).

Deutschsprachige Wörterbücher (2000): Projekte an Akademien, Universitäten, Instituten. Zusammengestellt in der Arbeitsstelle Göttingen des Deutschen Wörterbuchs von Jakob und Wilhelm Grimm. Göttingen. 2. überarb. Aufl. Elektronische Fassung: http://Grimm.ADW-Goettingen.gwdg.de/wbuecher/ .

Fournier, Johannes (1999): Digitale Dialektik: Chancen und Probleme mittelhochdeutscher Wörterbücher in elektronischer Form. In: Herbert Ernst Wiegand (Hrsg.), Wörterbücher in der Diskussion IV. Vorträge aus dem Heidelberger Lexikographischen Kolloquium (im Druck). Tübingen (=Lexikographica, Series Maior).

Fraas, Claudia/Haß-Zumkehr, Ulrike (1998): Vom Wörterbuch zum lexikalischen Informationssystem. LEKSIS – ein Projekt des Instituts für deutsche Sprache. In: Deutsche Sprache. Zeitschrift für Theorie, Praxis, Dokumentation. 26. Jahrgang, Heft 4, 289–303.

Gabriel, Norbert (1997): Kulturwissenschaften und Neue Medien. Wissensvermittlung im digitalen Zeitalter. Darmstadt.

Goebel, Ulrich/Lemberg, Ingrid/Reichmann, Oskar (1995): Versteckte lexikographische Information. Möglichkeiten ihrer Erschließung, dargestellt am Beispiel des Frühneuhochdeutschen Wörterbuchs. Tübingen (= Lexicographica, Series Maior 65).

Grosse, Rudolf (Hrsg.) (1998): Bedeutungserfassung und Bedeutungsbeschreibung in historischen und dialektologischen Wörterbüchern. Stuttgart/Leipzig. (Abhandlungen der sächsischen Akademie der Wissenschaften zu Leipzig. Philologisch-historische Klasse, Bd. 75, Heft 1).

Haack, Johannes (1997): Interaktivität als Kennzeichen von Multimedia und Hypermedia. In: Issing/Klimsa 150–166.

Hammwöhner, Rainer (1997): Das Konstanzer Hypertextsystem (KHS) im wissenschaftlichen und technischen Kontext. Konstanz (=Schriften zur Informationswissenschaft 32).

Hupka, Werner (1989a): Wort und Bild. Die Illustrationen in Wörterbüchern und Enzyklopädien. Tübingen (= Lexicographica, Series Maior 22).

– (1989b): Die Bebilderung und sonstige Formen der Veranschaulichung im allgemeinen einsprachigen Wörterbuch. In: Wörterbücher 1, 704–726.

Issing, Ludwig J./Klimsa, Paul (Hrsg.)(1997): Information und Lernen mit Multimedia. 2., überarb. Auflage, Weinheim.

Kammerer, Matthias (1998): Hypertextualisierung gedruckter Wörterbuchtexte: Verweisstrukturen und Hyperlinks. Eine Analyse anhand des Frühneuhochdeutschen Wörterbuches. In: Storrer/Harriehausen 145–171.

– und Lehr, Andrea (1996): Potentielle Verweise und die Wahrscheinlichkeit ihrer Konstituierung. In: Wiegand (1996a), 311–354.

Krol, Ed (1995): Die Welt des Internet. Handbuch und Übersicht. Bonn.

Kuhlen, Rainer (1991): Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissenbank. – Heidelberg/New York (= Edition SEL-Stiftung).

– (1995): Informationsmarkt – Chancen und Risiken der Kommerzialisierung von Wissen.
Lehr, Andrea (1996): Zur neuen Lexicographica-Rubrik „Electronic Dictionaries". In: Lexicographica. Internationales Jahrbuch für Lexikographie 12, 310–317.
Lemberg, Ingrid (1996): Die Belegexzerption zu historischen Wörterbüchern am Beispiel des Frühneuhochdeutschen Wörterbuches und des Deutschen Rechtswörterbuches. In: Wiegand 1996a, 83–102.
– (1998a): Lexikographische Erläuterungen im Deutschen Rechtswörterbuch: Gestaltungsmuster in einem Wörterbuch der älteren deutschen Rechtssprache. In: Wiegand 1998d, 135–154.
– (1998b): Hypertextualisierungsformen im Deutschen Rechtswörterbuch. In: Sprache und Datenverarbeitung. International Journal for Language Data Processing. Bd. 22, Heft 1, 44–54.
– (1998c): Wörterbücher und lexikographische Informationsmodelle der Zukunft. Vortrag, gehalten auf dem Symposium Computergestützte Produktion und Publikation von Wörterbüchern, 25. September 1998.
Abstract: http://www.ids-mannheim.de/grammis/abstract.html#lemberg
– (1998d): DRW digital. Neue Wege zur versteckten lexikographischen Information. Available: http://www.rzuser.uni-heidelberg.de/~q63/kopenhag.html
– (1998e): Online-Wörterbücher und ihre Benutzer. Neue Perspektiven für die Wörterbücher und ihre Erforschung. Available: http://www.rzuser.uni-heidelberg.de/~q63/wbbf.html
– (2000): Maschinenlesbare Texte, Faksimiles, Belege: die Erstellung des Deutschen Rechtswörterbuchs in einer lexikographischen Datenbank. In: Maschinelle Verarbeitung altdeutscher Texte V. Tübingen (im Druck). Preprint available:
http://www.rzuser.uni-heidelberg.de/~q63/wuerzbg.html
– und Petzold, Sybille/Speer, Heino (1998): Der Weg des Deutschen Rechtswörterbuchs in das Internet. In: Wiegand 1998d, 262–284.
Mulch, Roland (1998): Probleme der Synonymie in einem großlandschaftlichen Wörterbuch. In: Grosse 157–163.
Münker, Stefan/Roesler, Alexander (Hrsg.) (1997): Mythos Internet. Frankfurt.
Plate, Ralf/Recker, Ute (2000): EDV für Wörterbuchzwecke und neue lexikographische Arbeitsweisen. Erfahrungen beim Aufbau des elektronischen Text- und Belegarchivs für das Mittelhochdeutsche Wörterbuch. In: Maschinelle Verarbeitung altdeutscher Texte V. Tübingen (im Druck).
Putschke, Wolfgang (1994): Überlegungen zur Konzeption eines computerdialektologischen Arbeitsplatzes. In: Klaus Mattheier/Peter Wiesinger (Hrsg.), Dialektologie des Deutschen. Forschungsstand und Entwicklungstendenzen, 244–255. Tübingen.
Recker, Ute/Sappler, Paul (1998): Aufbau des maschinenlesbaren Text- und Belegarchivs für das Mittelhochdeutsche Wörterbuch. In: Grosse, 249–253.
Reichmann, Oskar (1986): Lexikographische Einleitung zum Frühneuhochdeutschen Wörterbuch. Band 1, 10–164.
– (1988): Zur Funktion, zu einigen Typen und zur Auswahl von Belegbeispielen im Historischen Bedeutungswörterbuch. In: Karl Hyldgaard-Jensen/Arne Zettersten (Ed.), Symposium on Lexicography III. Proceedings of the Third International Symposium on Lexicography May 14–16, 1986, at the University of Copenhagen, 413–444. Tübingen (= Lexicographica, Series Maior 19).
Sandbothe, Mike (1997): Interaktivität – Hypertextualität – Transversalität. Eine medienphilosophische Analyse des Internet. In: Münker/Roesler 56–82.
Schröder, Martin (1997): Brauchen wir ein neues Wörterbuchkartell? Zu den Perspektiven einer computerunterstützten Dialektlexikographie und eines Projektes „Deutsches Dialektwörterbuch". In: Zeitschrift für Dialektologie und Linguistik, LXIV. Jahrgang, Heft 1, 57–65.
Speer, Heino (1989): Das Deutsche Rechtswörterbuch. Historische Lexikographie einer Fachsprache. In: Lexicographica. Internationales Jahrbuch für Lexikographie 5, 85–128.
– (1991): Das Deutsche Rechtswörterbuch: Vorstellung des Wörterbuchs und lexikographische Praxis am Beispiel „magdeburgisch". In: Ulrich Goebel/Oskar Reichmann (Ed.), Historical Lexicography of the German Language. Vol. 2, 675–711. Lewiston/Queenston/Lampeter (= Studies in Russian and German 3).

– (1995): DRW to FAUST. Ein Wörterbuch zwischen Tradition und Fortschritt. Lexicographica. Internationales Jahrbuch für Lexikographie 10, 171–213.
– (1998a): s. Lemberg/Petzold/Speer.
– (1998b): Ein Wörterbuch, die elektronische Datenverarbeitung und die Folgen. In: Akademie-Journal. Mitteilungsblatt der Konferenz der deutschen Akademien der Wissenschaften. Heft 2, 11–16.
Storrer, Angelika/Freese, Katrin (1996): Wörterbücher im Internet. In: Deutsche Sprache. 24. Jahrgang, Heft 2, 97–153.
Storrer, Angelika/Harriehausen, Angelika (Hrsg.) (1998): Hypermedia für Lexikon und Grammatik. Tübingen (= Studien zur deutschen Sprache 12).
Storrer, Angelika (1998a): Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In: Wiegand 1998d, 106–131.
– (1998b): Hypermedia in der Lexikographie. Vortrag, gehalten im September 1998. Available: http://www.ids-mannheim.de/grammis/storrer/evortrag/start.htm .
– (2000a): Was ist hyper am Hypertext? In: Kallmeyer, Werner (Hrsg.), Sprache und neue Medien. Berlin u.a.: de Gruyter 2000 (= Jahrbuch 1999 des Instituts für deutsche Sprache).
– (2000b): Schreiben, um besucht zu werden: Textgestaltung fürs World Wide Web. In: Bucher, Hans-Jürgen/Püschel, Ulrich (Hgg.): Die Zeitung zwischen Print und Digitalisierung. Opladen/Wiesbaden: Westdeutscher Verlag, 91–123.
Strzebkowski, Robert (1997): Realisierung von Interaktivität und multimedialen Präsentationstechniken. In: Issing/Klimsa 268–303.
Tergan, Sigmar-Olaf (1997): Hypertext und Hypermedia: Konzeption, Lernmöglichkeiten, Lernprobleme. In: Issing/Klimsa 122–137.
Weber, Heinz J. (1998): Das Homographen-Wörterbuch der deutschen Sprache als Hypertext. In: Storrer/Harriehausen 195–216.
Weidenmann, Bernd (1997): Abbilder in Multimedia-Anwendungen. In: Issing/Klimsa 107–121.
Wiegand, Herbert Ernst (1989): Der gegenwärtige Status der Lexikographie und ihr Verhältnis zu anderen Disziplinen. In: Wörterbücher 1, 246–280.
– (1996a): (Hrsg.), Wörterbücher in der Diskussion II. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. Tübingen. (=Lexicographica, Series Maior 70).
– (1996b): Über die Mediostrukturen bei gedruckten Wörterbüchern. In: Arne Zettersten/Viggo Hjörnager Pedersen (Ed.), Symposium on Lexicography VII. Proceedings of the Seventh Symposium on Lexicography May 5–6, 1994 at the University of Copenhagen, 11–43. Tübingen (= Lexicographica, Series Maior 76).
– (1998a): Lexikographische Textverdichtung. Entwurf einer vollständigen Konzeption. In: Arne Zettersten [...] (Ed.), Symposium on Lexicography VIII. Proceedings of the Eighth Symposium on Lexicography, May 2–4, 1996 at the University of Copenhagen, 1–35. Tübingen (= Lexicographica, Series Maior 90).
– (1998b): Neuartige Mogelpackungen: Gute Printwörterbücher und dazu miserable CD-ROM-Versionen. Diskutiert am Beispiel des Lexikons der Infektionskrankheiten des Menschen. In: Lexicographica. Internationales Jahrbuch für Lexikographie 14, 239–253.
– (1998c): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. Berlin/New York.
– (1998d): (Hrsg.), Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. Tübingen (= Lexicographica, Series Maior 84).
Wiese, Joachim (1998): Zur Darstellung von Synonymengruppen im Brandenburg-Berlinischen Wörterbuch. In: Grosse 151–155.

Wörterbücher (+ Teilband) = Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta (Hrsgg.), Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie. Berlin/New York. 1. Teilband 1989; 2. Teilband 1990; 3. Teilband 1991 (= Handbücher zur Sprach- und Kommunikationswissenschaft 5.1; 5.2; 5.3).

Ingrid Lemberg, Heidelberg

Annette Klosa

Qualitätskriterien der CD-ROM-Publikation von Wörterbüchern

Für Nutzer(innen) wie für Verleger von CD-ROM-Wörterbüchern ist die Frage nach deren Qualität wichtig. Aus praktischen Verlagserfahrungen heraus werden im Folgenden Kriterien für eine qualitätvolle Publikation von Wörterbüchern auf CD-ROM entwickelt und zugleich Chancen und Probleme der elektronischen Publikation thematisiert.

1 Was ist ein Wörterbuch? Was sind „Qualitätskriterien"?

Im GWDS (3954) wird „Wörterbuch" erklärt als „Nachschlagewerk, in dem die Wörter einer Sprache nach bestimmten Gesichtspunkten ausgewählt, angeordnet u. erklärt sind".[1] Diese Definition verzichtet auf das Grundwort „Buch" des Determinativkompositums „Wörterbuch" und führt stattdessen, schon den speziellen Zweck dieser Art von Buch benennend, den übergeordneten Terminus „Nachschlagewerk" ein. Damit kann dieser Wörterbucheintrag problemlos auf verschiedene Publikationsformen von Wörterbüchern bezogen werden, nämlich auf gedruckte wie elektronische. Egal, auf welchem Datenträger die Wörter einer Sprache nach bestimmten Gesichtspunkten ausgewählt, angeordnet und erklärt sind, dieser Datenträger ermöglicht es, in ihm nachzuschlagen.

Entsprechend verweist der Eintrag „Wörterbuch" in DUDEN 8 (829) auf das Lemma „Nachschlagewerk" (505), unter dem neben dem Wörterbuch als sachverwandte Begriffe noch das Lexikon, die Enzyklopädie, das Wörterverzeichnis, das Glossar usw. erscheinen. Dieser Artikel benennt ebenfalls eine Reihe von Spezialwörterbüchern, verschweigt aber auch (gemäß seinem Zweck als Synonymenreihung), dass für alle genannten Nachschlagewerke inzwischen neue, elektronische Publikationsformen existieren.

Zum Stichwort „elektronisches Publizieren" findet man in DUDEN 5 (219) den Eintrag „Veröffentlichung von Informationen online über Computernetze (z.B. Internet)"; die

---

[1] Hier wie im Folgenden wird unter „Wörterbuch" insbesondere ein semasiologisches Wörterbuch verstanden; dieser Wörterbuchtyp ist besonders informationsreich und deshalb für elektronische Publikationsformen besonders interessant.

Publikation von Informationen zu Wörtern auf CD-ROM, um die es in diesem Beitrag geht, wird nicht erwähnt. Im Artikel „dictionary" in OERD (395) heißt es ganz am Ende des Eintrags immerhin: „Many dictionaries are now also published in electronic form."

Die oben zitierte Definition von „Wörterbuch" (im GWDS:3954) arbeitet mit den Verben „auswählen", „anordnen" und „erklären" und bezieht diese Tätigkeiten auf das Objekt „Wörter". – Wörter werden für ein Nachschlagewerk ausgewählt, in ihm angeordnet und erklärt. Ein Wörterbuch ist eine Art Lokalität, dort schlägt man nach. Deshalb arbeitet die Definition mit einem lokalen Relativanschluss: „Nachschlagewerk, in dem [usw.]". Dieser Relativsatz schließt mit einer finiten Verbform im Zustandspassiv; schließlich bildet ein Wörterbuch einen bestimmten Zustand ab, nachdem das Auswählen, Anordnen und Erklären geschehen ist.

Dieses Zustandspassiv der Definition verschweigt zwei weitere Aktantengruppen, die an einem Wörterbuch beteiligt sind, nämlich diejenigen, welche die Wörter auswählen, anordnen und erklären, und diejenigen, für die dies geschieht, weil sie Wörter nachschlagen möchten. Im Rahmen dieses Beitrages interessieren aber auch sie, und zwar im Zusammenhang mit dem anderen Teilaspekt des Themas, der Qualität.

Qualität verstanden als „Güte" entsteht nicht im luftleeren Raum, sondern Qualität ist etwas, was von Menschen, hier nämlich von Lexikograph(inn)en, für andere Menschen, hier nämlich für Nachschlagende, geschaffen wird. Qualität zeigt sich an bestimmten Merkmalen; für ein Wörterbuch wären zwei dieser Merkmale beispielsweise die Plausibilität und die Vollständigkeit der Auswahl der dargestellten Wörter, mit denen ein bestimmtes Nutzerinteresse befriedigt werden soll. Wollen Lexikograph(inn)en qualitätvoll arbeiten, müssen sie sich nicht nur über die Benutzerinteressen im Klaren sein, sondern sie müssen daraus auch eigene Qualitätsstandards ableiten und diese in der täglichen Arbeit sichern.

Damit ist ein erstes Qualitätskriterium für die Publikation von Wörterbüchern fomuliert: Ein gutes Wörterbuch ist eines, das die Erwartungen der Nutzer(innen) erfüllt und ihre Bedürfnisse befriedigt. Diese vielleicht vorwissenschaftliche und lapidare Formulierung leitet von den allgemeinen Überlegungen zu den spezielleren zur CD-ROM-Publikation von Wörterbüchern über und damit zugleich von der Theorie zur Praxis. Aus ebendieser Praxis stammt die folgende Aussage, mit der das erste Qualitätskriterium im Folgenden konfrontiert werden soll:

Auf einer Hauptversammlung des Verlages Bibliographisches Institut und F.A. Brockhaus AG, Mannheim, wurde in Bezug auf den Bereich des elektronischen Publizierens geäußert, die Qualitätsstandards seien diffus und der rabiate Preiswettbewerb gehe oft genug zulasten der Qualität.[2] Mit diesem Dilemma müssen sich alle auseinander setzen, die elektronische Produkte publizieren.

2 Qualitätskriterien von CD-ROM-Wörterbüchern: Abwägen zwischen Technik und Wissenschaft zum Wohle der Nutzer(innen)

Als Einstieg in diesen Abschnitt sollen drei Titel von Rezensionen zu elektronischen Wörterbüchern dienen: „Wörterwucher. Einsprachige Englischwörterbücher auf CD-ROM"

[2] Veröffentlicht im Geschäftsbericht für das Jahr 1997, der 1998 in Leipzig und Mannheim erschienen ist.

(Neeth/Müller 1997), „Wörterbücher zum Anklicken – ein kleiner Rundgang durch die PC-Bibliothek" (Storrer 1995) und „Rechtschreibung auf Knopfdruck: Elektronische Wörterbücher" (Schneider 1993).

Jeder Titel spricht bestimmte Kennzeichen der CD-ROM-Version von Wörterbüchern an: Der erste Titel hebt besonders auf die großen Datenmengen ab, die elektronische Wörterbücher verfügbar machen. Der zweite Titel spielt darauf an, dass CD-ROM-Wörterbücher heute meist aus Büchern hergeleitet werden und mit Begriffen arbeiten, die aus der Welt des Buches vertraut sind, hier mit dem Begriff „Bibliothek". Außerdem benennt er ein technisches Merkmal der elektronischen Wörterbücher – in ihnen kann man Dinge „anklicken" und nicht mehr nachblättern. Der dritte Titel schließlich greift ebenfalls diese technische Komponente auf, doch bezeichnet der Ausdruck „Rechtschreibung auf Knopfdruck" wohl noch mehr: Hier schwingt die Vorstellung mit, dass ein CD-ROM-Wörterbuch vielleicht bzw. hoffentlich mehr ist als ein reines Nachschlagewerk. Warum sollte ein elektronisches Rechtschreibwörterbuch beispielsweise nicht gleichzeitig falsche Schreibungen korrigieren können?

2.1 Wie Qualität von CD-ROM-Wörterbüchern momentan beurteilt wird

An dieser kleinen Auswahl lässt sich bereits zeigen, dass die Frage nach guter Qualität bei CD-ROM-Wörterbüchern mithilfe sehr unterschiedlicher Kriterien sehr unterschiedlich beantwortet werden kann. Man kann Qualität von Wörterbüchern auf CD-ROM an der Quantität der dargebotenen Daten festmachen, am Grad der technischen Umsetzung oder an ihrer Innovativität verglichen mit dem gedruckten Wörterbuch. Grundsätzlich kann man wohl zwischen allgemein-technischen Qualitätsvorstellungen und wissenschaftlich-lexikographischen Qualitätsvorstellungen unterscheiden.

Zur Rubrik „Electronic Dictionaries" in der Zeitschrift „Lexicographica" schreibt Lehr (1996b: 313) aus lexikographischem Selbstverständnis heraus:

> „Nun wäre es zwar ein leichtes, Zeitschriften wie beispielsweise c't, DOS Extra, MAC Welt, PC-Welt und PC-Magazin [...] aufzunehmen, andererseits ist es fraglich, inwieweit dies im Interesse unserer Leserinnen und Leser steht. Der Blick auf die rezensierten Wörterbücher in diesen Zeitschriften ist kein sprachwissenschaftlicher und schon gar kein (meta-)lexikographischer. Gleichwohl sollte nicht übersehen werden, daß die betreffenden Beiträge meist mit sehr viel Sachverstand [...] und einem sicheren Gespür für die Bedürfnisse von Anwenderinnen und Anwendern geschrieben sind."

Auf der anderen Seite stellen Neeth/Müller (1997:206) fest, dass sich aus ihrer Sicht „durch die Erfindung des Computers [...] derzeit eine wahre Revolution der Lexikographie" abzeichne. Einerseits eröffne das neue Medium CD-ROM die Möglichkeit, „riesige Wälzer auf flachen Scheibchen unterzubringen", andererseits erleichtere der Computer auch die lexikographische Arbeit: „Wo bislang Hunderttausende Karteikarten mit Zitaten zu verwalten waren, lassen sich jetzt mehrere Millionen Sätze in Datenbanken archivieren."

2.1.1 Technisch orientierte Qualitätsbeurteilung

Weiterhin nehmen Neeth/Müller (1997:206) an: „Für Benutzer elektronischer Wörterbücher wird vor allem der lästige Akt des Nachschlagens vereinfacht und beschleunigt." Unter dieser Prämisse bewerten sie die rezensierten Wörterbücher. Für „allgemein-technische"

Rezensent(inn)en, wie diese Untergruppe kurz gefasst genannt werden soll, sind deshalb besonders wichtige Qualitätskriterien die Handhabung der CD-ROM (z.B. ihre Installierbarkeit, ihr Platzbedarf auf der Festplatte, ihre Konfigurierbarkeit) und das Angebot an Suchwerkzeugen, also der Ausbaugrad der Software, die das elektronische Wörterbuch erst benutzbar macht. Hier werden z.B. Stichwortsuche, Volltextsuche, fehlertolerante und phonetische Suche und die Geschwindigkeit, mit der das gewünschte Ergebnis angezeigt wird, bewertet.

Nicht unbedingt die Auswahl und die Erklärung der Stichwörter, sondern ihre Anzahl und ihre Präsentation sind ein Qualitätskriterium, außerdem auch Zugaben zum eigentlichen Wörterbuch auf der CD-ROM, etwa integrierte Synonym- oder Zitatensammlungen. Nicht zu vernachlässigen ist schließlich das Preis-Leistungs-Verhältnis. Dieses Qualitätskriterium entscheidet in der Realität übrigens häufig fast alleine über den Kauf.

2.1.2 Inhaltlich orientierte Qualitätsbeurteilung

Als Beispiel für eine eher inhaltlich orientierte Bewertung sei auf Grieser (1998) hingewiesen, eine Rezension zu Englisch-Wörterbüchern, die unter dem Titel „Mut zur Lücke" CD-ROM-Ausgaben bewertet, wobei besonders stark die Qualität des Datenbestandes zählt. Unter „Qualität des Datenbestandes" versteht Grieser (1998:186), „wie treffend die Übersetzungen sind".

Als Anleitung für lexikographische Rezensenten verbindet Lehr (1996b) solche Qualitätskriterien mit eher wissenschaftlichen, anhand deren CD-ROM-Wörterbücher bewertet werden sollen. So ist ein Zeichen für Qualität für sie beispielsweise auch, ob

> „bei dem betreffenden elektronischen Wörterbuch auf Textverdichtungsoperationen verzichtet wurde [...] bzw. ob diese rückgängig gemacht wurden" (1996:317) oder ob „bei der Gestaltung der Wörterbuchartikel an traditionelle lexikographische Formen angeknüpft oder ob ein neuer Weg beschritten wurde" (1996:314).

Ein Hinweis darauf, dass das Preis-Leistungs-Verhältnis interessant sein könnte, fehlt unter den Fragen, die Rezensent(inn)en der Zeitschrift Lexikographica berücksichtigen sollten.

Inzwischen gibt es eine Reihe von Auszeichnungen für gute elektronische Produkte. Leider gibt es wohl keinen Preis, der herausragende CD-ROM- oder Online-Wörterbücher prämiert. Bislang werden Preise für allgemeine Software oder Lernsoftware sowie Multimedia-Produkte verliehen.[3] Sieht man sich die Kriterien an, nach denen solche Preise vergeben werden, wird schnell deutlich, dass es auch in diesem Bereich eine relativ deutliche Trennung gibt zwischen Preisen, die aufgrund allgemein-technischer Pluspunkte verliehen werden (z.B. der Deutsche Multimedia-Award des Deutschen Kommunikationsverbandes[4]), und solchen, die aufgrund eher inhaltlicher Pluspunkte gewährt werden (z.B. der Preis „digita" des Instituts für Bildung in der Informationsgesellschaft e. V.[5]).

---

[3] Neben den oben genannten Auszeichnungen gibt es außerdem folgende Preise für Software und Multimedia-Produkte: den Preis des Metropolitan-Verlages im Jahrbuch „Annual Multimedia", den International EMMA Award, den Preis des Vereins IMPULS – Schule & Wirtschaft e. V. und den Eltern for family-Softwarepreis.

2.2 Wie Qualität von CD-ROM-Wörterbüchern beurteilt werden könnte

Trotzdem sollte die Qualität der CD-ROM-Publikation von Wörterbüchern sowohl an ihrer technischen Umsetzung wie an ihrem Inhalt gemessen werden. Als zweites Qualitätskriterium für die Publikation von Wörterbüchern auf CD-ROM ist deshalb zu formulieren: Ein gutes CD-ROM-Wörterbuch ist lexikographisch solide erarbeitet, arbeitet technisch einwandfrei und schöpft dabei die technischen Möglichkeiten des Mediums CD-ROM in hohem Maße aus.

Eigentlich wird mit dieser Aussage eine Selbstverständlichkeit formuliert, die aber, gemessen an der Praxis, manchmal zumindest teilweise noch Zukunftsmusik ist. Der (zukünftige) Idealfall würde wohl folgendermaßen aussehen:

Ein gutes CD-ROM-Wörterbuch ist ein Wörterbuch, das speziell für dieses Medium konzipiert und verfasst wird. Wird das Wörterbuch nicht direkt für das Medium CD-ROM konzipiert, ist ein gutes CD-ROM-Wörterbuch eines, das seine Papierherkunft „verleugnet". Geht das CD-ROM-Wörterbuch auf ein gedrucktes zurück, muss es seine Erscheinungsform an das neue Medium anpassen. Ein gutes CD-ROM-Wörterbuch ist weiterhin eines, an dessen Erarbeitung Fachleute, nämlich Lexikograph(inn)en, arbeiten, die wissen, wie die Nachschlagebedürfnisse der Nutzer(innen) am besten zu befriedigen sind. Ein gutes CD-ROM-Wörterbuch ist außerdem eines, an dessen Erarbeitung weitere Fachleute, nämlich Programmierer(innen) und Designer(innen), arbeiten, die wissen, wie die Möglichkeiten des Mediums am besten auszunutzen sind.

Ein gutes CD-ROM-Wörterbuch überrascht die Benutzer(innen) mit Möglichkeiten, die sie sich intuitiv schon immer gewünscht haben, z.B. Aktualisierbarkeit des Datenbestandes, Erweiterbarkeit des Datenbestandes um eigenen Wortschatz, Verknüpfbarkeit des Wörterbuchs mit Textverarbeitungssoftware, Durchsuchbarkeit mehrerer Wörterbücher gleichzeitig, Auffindbarkeit von Wörtern, deren Schriftbild man nicht genau kennt, Auffindbarkeit von Wörtern, die in einem Text nur in flektierter Form auftreten, usw.

Ein gutes CD-ROM-Wörterbuch lässt die Benutzer(innen) aus einer Fülle möglicher Informationen diejenigen auswählen, die momentan für sie am wichtigsten sind, z.B. einmal die Orthographie eines Wortes, einmal seine Herkunft, einmal seine Aussprache, einmal sein Vorkommen in einem Zitat von Goethe usw. Ein gutes CD-ROM-Wörterbuch bezieht die Nutzer(innen) also auch durch interaktive Elemente ein. Schließlich ermöglicht ein gutes CD-ROM-Wörterbuch all dies, ohne dass die Nutzer(innen) sich umständlich in seinen Gebrauch einarbeiten müssen. Im Gegenteil: Die Software stellt sich im günstigsten Fall selbst automatisch auf die Bedürfnisse der Benutzer(innen) ein.

Zurückkommend auf die anfangs zitierte Definition von „Wörterbuch" im GWDS könnte man auch sagen: Ein gutes CD-ROM-Wörterbuch ist eines, das die zu erklärenden Wörter gemäß dem Nutzerinteresse auswählt, sie dem Medium entsprechend nicht nur alphabetisch-linear, sondern auch hypertextuell anordnet und sie mithilfe anderer Wörter, Bilder, Filmsequenzen und Aufnahmen gesprochener Sprache erklärt.

---

[4] Der Bewertungsbogen für die Jury des Deutschen Multimedia-Awards 1998 wurde freundlicherweise von Herrn Werner Kierker, Deutscher Kommunikationsverband BDW, Bonn, zur Verfügung gestellt.

[5] Die digita-Kriterien zur Bewertung von Lernsoftware wurden freundlicherweise von Herrn Prof. Dr. Wilfried Hendricks, IBI – Institut für Bildung in der Informationsgesellschaft e. V., Berlin, zur Verfügung gestellt.

3 Aus der Praxis: Einschränkungen in der Realität

Eingeschränkt werden diese Vorstellungen in der Realität durch verschiedene Faktoren, die erklären können, warum es bislang kaum elektronische Wörterbuchprodukte gibt, die den oben entwickelten Qualitätskriterien völlig genügen.

3.1 Das Kostenargument

Warum gibt es momentan für das Deutsche kein umfangreicheres CD-ROM-Wörterbuch, das speziell für dieses Medium entwickelt und verfasst wurde? (Ausgeklammert bleiben hier so genannte elektronische Taschenwörterbücher, die nicht unbedingt einen Papiervorgänger haben müssen.) Der einzige Grund ist wohl, dass eine solche Entwicklung, gemessen an dem erwartbaren Umsatz, den ein Verlag mit elektronischen Wörterbüchern erzielen kann, viel zu teuer wäre. CD-ROM-Wörterbücher erreichen momentan, wenn sie sich gut verkaufen, Auflagenhöhen von wenigen Tausend Stück. Dies ist, verglichen mit manchem gedruckten Wörterbuch, verschwindend wenig. Damit können je nach Entwicklungskosten als Deckungsbeitrag günstigenfalls etwa 5% des Ladenpreises erwirtschaftet werden, wenn die gesamte Auflage verkauft wird. Ohne finanzielle Unterstützung, sei es durch Mischkalkulation mit den gedruckten Werken, durch Mischkalkulation mit anderen CD-ROMs oder mithilfe von Zuschüssen, die Bundeseinrichtungen z.B. für die Publikation wissenschaftlicher Wörterbücher zahlen mögen, ist so selbst eine Zweitpublikation eines schon gedruckten Wörterbuchs auf CD-ROM nicht unbedingt möglich. Elektronisches Publizieren ist im Bereich der Nachschlagewerke in vielen Fällen nach wie vor ein Zuschussgeschäft.

Mit dem Kostenargument könnte man pauschal manches andere der eben genannten Qualitätskriterien erschlagen. Dies wird hier zwar vermieden, doch sollte man bei allen berechtigten Wünschen für ein elektronisches Wörterbuch versuchen, im Hinterkopf zu behalten, dass Wörterbücher Produkte sind, die vermarktet werden müssen. Und dass die meisten Käufer(innen) gewisse Preisvorstellungen haben, von denen sie, egal was ihnen ein Produkt bieten mag, nicht abweichen werden. Berücksichtigen muss man in diesem Kontext auch, dass im Bereich des elektronischen Publizierens (leider) ein heftiger Preiswettbewerb ausgebrochen ist, der unweigerlich auf die Qualität der Produkte Einfluss hat.

3.2 Andere Gründe

Daneben gibt es andere Faktoren, die die Erfüllung der genannten Qualitätskriterien für die CD-ROM-Publikation von Wörterbüchern leider häufig nur in der Theorie einfach erscheinen lassen. Storrer (1995:9) merkt in einer Rezension beispielsweise zu Recht an:

> „Eine Lemmatisierungsoption, mit der man nicht nur nach einer bestimmten Wortform, sondern nach allen Flexionsformen eines Lemmas suchen kann, wäre außerdem wünschenswert [...]".

In der Praxis würde dies voraussetzen, dass entweder bei der Erarbeitung der Wörterbücher, die ja (noch) zunächst auf Papier erscheinen, alle möglichen Flexionsformen mit erfasst werden, oder dass eine Software in die CD-ROM integriert würde, die in der Lage ist, Flexionsformen auf die jeweilige Grundform zurückzuführen (also z.B. in dem Satz *Ich nehme den Hörer endlich ab* den Infinitiv des finiten Verbs richtig als *abnehmen* zu bestimmen). Das eine setzt Zeit (und damit wieder Geld) voraus. Das andere wird bislang wohl deshalb

kaum realisiert, weil die Anbieter entsprechender Software den Qualitätsansprüchen der Verleger noch nicht völlig genügen.

Kammerer (1996:331) bemerkt in einer anderen Rezension:

> „Ein genauerer Blick auf die PC-Bibliothek mit ihren verschiedenen Wörterbüchern aus der DUDENreihe zeigt, daß hier lediglich Printwörterbücher auf eine sehr rigide Art in ein elektronisches Medium überführt wurden. Mit einigen Tools aufgepeppt und einer zugegebenermaßen sehr komfortablen Suchfunktion [...] wurden die Spezifika und die Chancen des Computers weder erkannt noch genutzt."

Aus dieser harschen und zum Teil auch unberechtigten Kritik kann man einen nützlichen Hinweis ziehen: An der Erarbeitung von CD-ROM-Wörterbüchern sollten im Idealfall eben tatsächlich nicht nur Germanist(inn)en, sondern auch Informatiker(innen) beteiligt sein, die die Spezifika des Computers zu nutzen wissen. Im Alltag sieht es häufig allerdings so aus, dass die informatische Fachkompetenz meist von außen eingekauft wird, aber nicht dauerhaft in einer Wörterbuchredaktion verankert ist.

In einer Besprechung zur elektronischen Version eines gegenwartssprachlichen italienischen CD-ROM-Wörterbuches kritisiert Schafroth (1997:325) unter dem Stichwort „Benutzerfreundlichkeit", dass das Blättern von Beleg zu Beleg (und wieder zurück) nicht möglich sei. Außerdem vermisse man eine Funktion zum Abbruch der Recherche. Das Hin- und Herblättern bei der Suche, das Abbrechen des Nachschlagens sind Vorgänge, die aus der Wörterbuchbenutzung vertraut sind. Dass man sie bei der Umsetzung auf das elektronische Medium vergessen konnte, lässt vermuten, dass an der Entwicklung der Software für die elektronische Version keine Lexikograph(inn)en beteiligt waren – und vielleicht der gesunde Menschenverstand zumindest bei diesen Punkten gefehlt hat. Sonst verfügt das von Schafroth (1997) rezensierte Wörterbuch nämlich über ausgezeichnete Suchfunktionen. Auch hier zeigt sich, wie wichtig die enge Zusammenarbeit zwischen Techniker(inne)n und Lexikograph(inn)en ist, wenn ein rundum qualitätvolles CD-ROM-Wörterbuch entstehen soll.

Dieser Bericht aus der Praxis soll aber nicht mit Pessimismus enden. Im Gegenteil: Die Publikation von Wörterbüchern auf CD-ROM nimmt insgesamt zu, und das nicht nur in Deutschland. Die Käuferzahlen wachsen, und es wächst auch die Erfahrung mit dem Medium bei Nutzer(inne)n und Verlagen. Nicht zuletzt entstehen in diesem Bereich Arbeitsplätze. Dies zeigt sich im neuen Berufsfeld des Proofings elektronischer Produkte[6], und den Abteilungen in Verlagen, die den konzeptionellen und herstellerischen Prozess der CD-ROMs betreuen.

4 Müssen Lexikograph(inn)en in Zukunft mehr können?

Ergeben sich durch das neue Medium und die Vorstellungen von Qualität bei CD-ROM-Wörterbüchern auch neue Anforderungen an die Lexikographie und ihre Vertreter(innen) in

---

[6] Mit dem Proofing beschäftigen sich inzwischen spezielle Anbieter, die mit in die Qualitätskontrolle und -sicherung von CD-ROM-Wörterbüchern eingebunden werden. Hier sei auf die Homepage der Firma media supervision, Mannheim, verwiesen, die eine ausführliche Dokumentation der technischen Aspekte der Qualitätssicherung unter http://www.mediasupervision.de im Internet anbietet.

Wörterbuchredaktionen und -arbeitsstellen? Auf diese Frage kann man wohl nur mit einem klaren Ja antworten. Um qualitätsvolle CD-ROM-Wörterbücher machen zu können, müssen Lexikograph(inn)en lernen, sich frei zu machen von der Vorstellung, jedes Wort müsse ausschließlich durch andere Wörter erklärt werden. Das Trägermedium „Papier" sollte ebenso infrage gestellt werden, wie die Anordnung der Lemmata in alphabetischer Reihenfolge.[7] Der Zwang, im Buch vieles durch Abkürzungen und andere Maßnahmen der Platzersparnis (z.B. Ersetzen des Lemmas durch einen Gedankenstrich oder verknappte Bedeutungserläuterungen) darzustellen, fiele weg, sodass möglicherweise auch neue Ansätze zur Formulierung und Gestaltung der Artikel entwickelt werden könnten.

Lexikograph(inn)en müssen also ein Gespür dafür entwickeln, welche Wörter sich z.B. durch Bilder, welche Wörter sich durch Filmsequenzen, welche Wörter sich durch akustische Unterstützung, welche Wörter sich durch Kombinationen dieser Medien mit dem Wort und welche Wörter sich nach wie vor nur durch andere Wörter erklären lassen. Sie müssen ein Gespür dafür entwickeln, wie diese Erklärungen auf dem neuen Medium am besten anzuordnen sind.

Sie müssen die Bereitschaft mitbringen, in elektronischen Redaktionssystemen die Daten noch konsistenter und penibler einzugeben, als das auf Papier nötig war, damit diese Daten für elektronische Publikationen geeignet sind. Und sie müssen bereit sein, sich gewisse technische und didaktische Kenntnisse anzueignen, damit sie mit Vertreter(inne)n anderer Disziplinen gute CD-ROM-Wörterbücher entwickeln können.

Hier zeigt sich, dass sich die Lexikographie insgesamt in einer Phase des Umbruchs befindet. Der Bereich des elektronischen Publizierens entwickelt sich, und damit entwickeln sich auch die Anforderungen an diejenigen, die in diesem Bereich arbeiten.

5 Literatur

Breidt, Elisabeth (1998): Neuartige Wörterbücher für Mensch und Maschine: Wörterbuchdatenbanken in COMPASS. – In: H.E. Wiegand (Hg.): Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. – Tübingen: Niemeyer.

DUDEN 5 = DUDEN BAND 5. FREMDWÖRTERBUCH. Hg. u. bearb. vom Wissenschaftlichen Rat der DUDENredaktion. – Mannheim/Leipzig/Wien/Zürich: Dudenverlag ^{6}1997.

DUDEN 8 = DUDEN BAND 8. SINN- UND SACHVERWANDTE WÖRTER. SYNONYMWÖRTERBUCH DER DEUTSCHEN SPRACHE. Hg. u. bearb. von W. Müller. – Mannheim/Leipzig/Wien/Zürich Dudenverlag ^{2}1997.

Feldweg, Helmut (1997): Wörterbücher und neue Medien: Alter Wein in neuen Schläuchen? – *Zeitschrift für Literaturwissenschaft und Linguistik* 106, 30–43.

Grieser, Franz (1998): Englisch-Wörterbücher. Mut zur Lücke. – *PC Professionell* 9, 180–189.

GWDS = DUDEN. DAS GROSSE WÖRTERBUCH DER DEUTSCHEN SPRACHE. In acht Bänden. Hg. u. bearb. vom Wissenschaftlichen Rat der DUDENredaktion u. den Mitarbeitern der DUDENredaktion unter d. Ltg. von G. Drosdowski. – Mannheim/Leipzig/Wien/Zürich: Dudenverlag ^{2}1995.

Kammerer, Matthias (1996): DUDEN. Das Fremdwörterbuch. Version 1.1 mit integriertem Benutzerwörterbuch. Netzwerkfähigkeit auf Anfrage. Benutzerhandbuch: R. Echtler. Mannheim: Bibliographisches Institut & F. A. Brockhaus AG. – *Lexicographica* 12, 329–331.

---

[7] Vorstellbar wäre z.B. auch eine Anordnung nach Wortfamilien, Synonymengruppen o.Ä. Am besten käme man wohl den Nutzerinteressen entgegen, wenn man die Nutzer(innen) selbst unter verschiedenen Anordnungsformen wählen ließe.

Lehr, Andrea (1996a): DUDEN. Deutsches Universalwörterbuch A–Z. DUDEN Oxford Großwörterbuch Englisch. Englisch-Deutsch/Deutsch-Englisch. Version 1.1 mit integriertem Benutzerwörterbuch und Netzwerkfähigkeit auf Anfrage. Benutzerhandbuch: R. Echtler. Mannheim: Bibliographisches Institut, F. A. Brockhaus AG und Oxford University Press 1990/1994 (PC-Bibliothek). – *Lexicographica* 12, 327–329.

– (1996b): Zur neuen Lexicographica-Rubrik „Electronic Dictionaries". – *Lexicographica* 12, 310–317.

McCorduck, Ed (1997): Collins Cobuild on CD-ROM. HarperCollins Publishers Ltd. and ATTICA Cybernetics, Ltd. 1994. – *Lexicographica* 13, 312–316.

Milan, Carlo (1997): Elektronische Lexikographie. Manuskript zu einem Vortrag an der Otto-Friedrich-Universität Bamberg.

Neth, Hansjörg/Müller, Thomas (1997): Wörterwucher. Einsprachige Englischwörterbücher auf CD-ROM. – *c't* 10, 206–213.

OERD = The Oxford English Reference Dictionary. Hg. Judy Pearsall, Bill Trumble. Oxford: Oxford University Press ^{5}1995.

Osterberg, Jürgen (1996): PC-BIB 2000. Ein Strategiekonzept für die PC-Bibliothek 2.0 und weiter folgende Versionen. Internes Papier. Bibliographisches Institut & F. A. Brockhaus AG, Mannheim.

Schafroth, Elmar (1997): Devoto, Giacomo & Oli, Gian Carlo. Il Dizionario della Lingua Italiana. Edizione su CD-Rom per Windows 3.1. Firenze: Le Monnier 1990 & Editoria Elettronica Editel 1994. – *Lexicographica* 13, 318–326.

Schneider, Rolf (1993): Rechtschreibung auf Knopfdruck: Elektronische Wörterbücher. – *Lexicographica* 9, 220–299.

Storrer, Angelika (1998): Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. – In: H.E. Wiegand (Hg.): Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. – Tübingen: Niemeyer.

– (1995): Wörterbücher zum Anklicken – ein kleiner Rundgang durch die PC-Bibliothek. – *Sprachreport* 2, 9–10.

– und Freese, Katrin (1996): Wörterbücher im Internet. – *Deutsche Sprache* 2, 97–153.

Wiegand, Herbert Ernst (1998): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. – Berlin/New York: Niemeyer.

Annette Klosa, Leipzig

II Anwendungen

Ulrike Haß-Zumkehr

Zur Mikrostruktur im Hypertext-Wörterbuch

In diesem Beitrag werden erste Erfahrungen mit und Überlegungen zu der Aufgabe dargelegt, ein Mikrostrukturenprogramm für ein Hypertext-Wörterbuch zu entwerfen.[1] Zur Hypertextualisierung gedruckter Wörterbücher gibt es inzwischen erste Veröffentlichungen;[2] meist bleibt hier die Bindung an eine gedruckte Vorlage, und sei die Hypertextualisierung noch so konsequent, bestehen.[3] Im Unterschied zu solchen Hypertext-Wörterbüchern gehen nachfolgende Überlegungen von einem vorlagen*un*abhängigen Hypertext aus, dessen allgemeines Ziel es ist, Informationen zum deutschen Wortschatz zu vermitteln. Die hier vorgestellten Erfahrungen und Überlegungen sind an ein konkretes Projekt gebunden: LEKSIS – das lexikalisch-lexikologische Informationssystem des Instituts für Deutsche Sprache, Mannheim[4]. Auf eine (weitere) Projektbeschreibung wird hier aber verzichtet; sie findet sich in Fraas/Haß-Zumkehr (1999), ferner auf der Homepage unter http://www.ids-mannheim.de/wiw. Vor dem Hintergrund dieses Projektes stehen die Bedingungen bzw. lexikografischen Konsequenzen des Mediums Hypertext im Unterschied zum Druck zur Diskussion.

1 Mediale Bedingungen

LEKSIS versteht sich als Informationssystem zum Wortschatz der deutschen Gemeinsprache bzw. zu seinen Elementen; es soll online verfügbar sein und stets aktuell gehalten werden. Die Informationsmenge[5] wird in einer (objektorientierten) Datenbank abgelegt und verwaltet, an deren ‚Füllung' sich mehrere Forschergruppen zu unterschiedlichen (lexikologischen) Gebieten (zunächst des IDS) beteiligen sollen; die Beiträge dieser Gruppen heißen in unserem Zusammenhang (Projekt-)Module.

Im Unterschied zur Planung eines – großen – Wörterbuchs bedeuten zeitliche Nicht-Befristung, ständige Aktualität und frühzeitige Verfügbarkeit, dass es unmöglich ist, fest-

---

[1] Für die sehr fruchtbaren Diskussionen, die den Hintergrund dieses Beitrags bilden, danke ich den Mitgliedern meines Teams (in alphabetischer Reihe) Cyril Belica, Claudia Fraas, Carolin Müller, Sonja Müller-Landmann und Kathrin Steyer.

[2] Da dies kein Forschungsbericht zum Thema Hypertext-(Lexikografie) ist, sei exemplarisch und wegen der dort angegebenen weiterführenden Literatur verwiesen auf Lemberg/Petzold/Speer (1998) und Storrer/Harriehausen (1998). Zu Hypertext vgl. Storrer (2000).

[3] Ein Negativ-Beispiel erläutert Wiegand (1998[99]).

[4] Seit Frühjahr 2000 mit dem Namen *Wissen über Wörter / WiW*.

zulegen, in wie vielen Jahren mit wie vielen Mitarbeiterinnen und Mitarbeitern ein ‚fertiges' Produkt vorliegen kann. Da das System für diverse Projektmodule offen bleiben muss, auch für solche, die es heute noch gar nicht gibt, kann und muss die Mikrostruktur so umfassend und differenziert wie möglich sein.

Die lexikografischen AutorInnen des Systems werden auf die Datenbank zu eigenen und verwandten wissenschaftlichen Zwecken (auch) unmittelbar zugreifen; die eigentlichen Nutzer allerdings werden ausschließlich über ein Hypertextsystem je nach ihren speziellen Interessenlagen und unter Berücksichtigung auch nicht-linguistischer Vorbildung recherchieren können. Das Hypertextsystem soll aus der Datenbank die für eine Rechercheanfrage nötigen Informationen herausholen und mithilfe bestimmter Seitenlayouts im Internet ansprechend präsentieren. So kann der eine Nutzer beispielsweise die Bedeutungsparaphrasen aller Glieder eines onomasiologischen Paradigmas nebeneinander auf dem Bildschirm sehen wollen; die andere Nutzerin interessiert sich hingegen für die Vernetzung des semasiologischen Feldes, der dritte vielleicht für den Vergleich von Kollokations- und Valenzangaben aller Einzelbedeutungen eines Lemmas; und die vierte sucht ausschließlich wortgeschichtliche Informationen zu einem bestimmten Lemma samt Belegen. Daneben gibt es diejenigen Nutzer, die überhaupt ‚erst einmal schauen' wollen, d.h. die einen ersten Überblick über die je Lemma gebotenen Informationen gewinnen und nicht gleich ins Detail gehen wollen. Für sie müssen in geeigneter Weise zusammenfassende und allgemein verständliche, nicht-linguistisch formulierte Angaben sowie Überblicks-‚karten' je Wortartikel vorgehalten werden

Beim gedruckten Wörterbuch hängt die Festlegung des Mikrostrukturenprogramms normalerweise von der Einschätzung des Adressatenkreises ab (vgl. Wiegand 1998, 249) – bei einem Hypertextsystem wie LEKSIS eher nicht mehr: Hier wird die Strukturierung der Daten von ihrer Präsentation getrennt vorgenommen, denn die Ausarbeitung der Präsentationsform(en) wird als eigene lexikografische Handlung geplant, die auf die Erarbeitung der Mikrostrukturen nur zurückgreift. Das Mikrostrukturenprogramm steht also der Datenbank und ihrer Struktur näher als der nutzerzugewandten Hypertextoberfläche. Von allen möglichen Nutzungsprofilen muss das Mikrostrukturenprogramm zunächst insoweit unabhängig bleiben, als in diesem Informationspotential[5] lediglich diejenigen Elemente definiert und geordnet werden, die bei irgend einer Recherche relevant werden *könnten*. Es handelt sich also um eine Art lexikografisches Maximalprogramm,

- das (1.) viel mehr Wörterbuchtypen in sich vereint als im Druckmedium möglich,
- das (2.) ein viel umfassenderes Mikrostrukturenprogramm erfordert als im Druckmedium nötig,
- das (3.) ständig ergänzt und korrigiert werden können muss und
- (4.) dessen interne Struktur nicht die eines primär linear aufgebauten Textes ist, sondern die eines Hypertextes.

Aus Letzterem folgt vor allem, dass die Textsegmente bzw. die auf irgend eine Weise definierten atomaren Elemente der Mikrostruktur unabhängig voneinander rezipierbar, dennoch kombinierbar sein und entsprechend kontextfrei formuliert werden müssen, denn jeder Nutzer, jede Nutzerin rezipiert sie in einer potenziell anderen Reihenfolge oder in einer potenziell anderen typografischen Nähe.

---

[5] Der Ausdruck *Information* wird hier – angesichts der interdisziplinär uneinheitlichen Bestimmung und Abgrenzung von *Information, Wissen* und *Daten* – absichtlich völlig unterminologisch verwendet.

Das Maximalprogramm bedeutet aber nicht, dass zu jedem Lemma die selbe informatorische Dichte vorhanden sein wird – im Gegenteil: der im Prinzip unendliche Prozess des Auf- und Ausbaus eines online-Informationssystems wie LEKSIS wird Nutzer und Nutzerinnen daran gewöhnen, dass nicht zu jedem Lemma die gleiche Menge, Dichte und Tiefe an Information zur Verfügung steht. Dies ist aber kein Nachteil gegenüber umfangreichen gedruckten Wörterbuchprojekten, bei denen jahre- und jahrzehntelang nur ein Teil der Alphabetstrecke benutzbar ist, bei denen Verweise bis zum Schluss vielfach ins Leere gehen und bei denen die ersten Lieferungen veralten, bevor die letzten erscheinen. In einem Informationssystem muss den Nutzern selbstverständlich zu einem sehr frühen Zeitpunkt ihrer Recherche ein Überblick über die vorhandenen und die nicht-vorhandenen Informationen gegeben werden; Hypertextsysteme können dies mithilfe sog. fish-eye-views, bei denen ähnlich dem gleichnamigen Kamera-Objektiv die unmittelbare Informationsumgebung größer oder detaillierter gezeigt wird als die weiter entfernt liegende, und dynamischer site maps verwalten: letztere sind Landkarten ähnliche Darstellungen des Informationsgehalts, die sich je nach Benutzerinteresse verwandeln (lassen). Gedruckte Wörterbücher wie etwa das *Deutsche Wörterbuch* können dies nicht oder erst bei ihrer elektronischen Konversion und entsprechenden Zusätzen.

Der Entwurf des Mikrostrukturenprogramms für ein online verfügbares Hypertext-Wörterbuch hat also von einem maximalen Informationsangebot auszugehen und *alles* zu berücksichtigen, was in irgendeiner Weise gesucht werden oder bei einer Suche einbezogen werden könnte. Damit einher geht die Notwendigkeit, das Informationsangebot weitestgehend zu atomisieren, weitergehend jedenfalls als es bei gedruckten Wörterbüchern üblich und angemessen ist und auch weitergehend als bei Wörterbüchern, die von einer Druck- in eine elektronische Fassung konvertiert werden. Das auf eine Datenbank hin angelegte Mikrostrukurenprogramm umfasst somit textuelle Elemente (z.B. narrativ formulierte Bedeutungsgeschichten, Zusammenfassungen zu den einzelnen Informationsdimensionen wie Semantik, Grammatik, siehe unten), atomare Elemente wie alphanumerische Angaben (z.B. Frequenzangaben, Datierungen) und Attribute wie Genus- oder Valenzangaben, die auf der Nutzeroberfläche erst zu einem lesbaren Teiltext zusammengefügt bzw. ergänzt werden müssen, vorstellbar etwa in Form eines Lückentextes wie

> Ungeachtet der Unterscheidung der [...] Lesarten kommt das Stichwort im Gesamtkorpus [...] mal, im pressesprachlichen Subkorpus [...] mal und im Subkorpus der sprechsprachnahen Texte [...] mal vor.
>
> Etwa um [....] wurde das Stichwort aus dem [...] ins Deutsche entlehnt.
>
> *Oder*: Entlehnung ins Deutsche aus: [...] um [...]

Was hier durch eckige Auslassungsklammern gekennzeichnet ist, muss aus der Datenbank geholt und eingesetzt werden. Voraussetzung ist, dass die jeweiligen Informationselemente dort eindeutig identifiziert sind und eine Form haben, die in solch einen Lückentext „passt".

Aus dem Vorangehenden wird deutlicher, worin die Aufgabe des Entwurfs des Mikrostrukturenprogramms besteht:

1. aus der eindeutig identifizierenden Bestimmung der Elemente oder ‚informationellen Atome',
2. aus der Festlegung ihrer jeweiligen Datenart und Eigenschaften (numerisch, alphanumerisch, standardisierter Text, freier Text, beschreibungssprachlich, objektsprachlich, u.a.),

3. aus der Erarbeitung einer (teilhierarchischen) Elementenstruktur, die relativ unabhängig von Benutzungssituationen ist, und
4. aus der Festlegung möglicher Interrelationen zwischen den Elementen (Verlinkungssystem).

Die Lösung der genannten Aufgaben ist auf computerlinguistische und texttechnologische Kenntnisse ebenso angewiesen wie auf lexikografisch-linguistische – dies alles kann wohl nur in einem Team zusammengeführt werden. Betonen möchte ich an dieser Stelle, dass das Festlegen einer einheitlichen Datenstruktur und das Festlegen eines Datenbankmodells – etwas anders als die Einschätzung in Wiegand (1998, 208) – sich als eine durch und durch lexikografische Tätigkeit erweisen, die allerdings durch die medialen Bedingungen und durch die Teammitglieder mit informatischen Kompetenzen viel stärker in Richtung Entscheidungseindeutigkeit, logische Konsistenz und Redundanzfreiheit verändert wird – eine Veränderung, die man als Verbesserung bezeichnen muss. Von ‚Standardisierung' lexikografischer Tätigkeit spreche ich hier absichtlich nicht, denn damit wird oft implizit auf eine Beschränkung lexikografischer Freiheit und Behinderung von Kreativität Bezug genommen, die im Fall eines dynamischen Hypertextsystems keineswegs gegeben sind. Im Gegenteil, der Zwang, ein umfassendes Mikrostrukturenprogramm zu entwerfen, das in eine SGML/XML-DTD und weiter in eine Datenbankstruktur übersetzt werden soll, hat – wie nachfolgend gezeigt wird – Kreativität freigesetzt.

Es stellt sich die Frage, warum man erst eine Document Type Definition (DTD) für SGML/XML entwerfen sollte, wenn man eigentlich ein Datenbankmodell braucht. Die Antwort lautet: weil eine SGML/XML-Datenstruktur datenbankunabhängig ist. Dies fällt allerdings ins Gewicht, wenn man zu Beginn noch nicht weiß, welche Datenbanksoftware man sich in Zukunft leisten kann, wie zufrieden man mit ihr sein wird, ob nicht gewisse Systemteile einmal als gedrucktes Wörterbuch herausgezogen werden sollen, und: ob nicht einmal externe Kooperationspartner mit anderer Software sich am Ausbau des Informationssystems werden beteiligen wollen. Alles dies ist unproblematisch, wenn die komplette Datenstruktur in einem datenbankunabhängigen Format wie SGML/XML vorliegt.

2 Mikrostrukturelle Grundsatzentscheidungen

In der Metalexikografie wird davon ausgegangen, dass zu jedem Lemmazeichentyp je eine spezifische abstrakte hierarchische Mikrostruktur gehört (Wiegand 1998, 215). Werden die Mikrostrukturen aber in Form einer DTD dargestellt, so ist die Existenz mehrerer paralleler Strukturen problematisch: Die Redundanzen partiell identischer Mikrostrukturen sind Fallen für widersprechende Angaben, so dass ein informatischer Rat zu beherzigen ist: Keine Angabe darf zweimal in den gleichen Wortartikel eingetragen werden (müssen). Und: Es darf keine doppelten Strukturen geben. Die (einzige) Alternative zum Ansatz mehrerer, lemmatypspezifischer Mikrostrukturen besteht im Ansatz einer einzigen, „mehrzweckgeeigneten", in sich modularen Mikrostruktur, die alle lemmatyp-spezifischen Artikelpositionen/Informationsarten enthält und aus der auf der Autorenoberfläche spezifische Untermengen je Lemmatyp ausgegliedert werden. Die Unterscheidung lemmatypspezifischer Informationsprogramme ist dann Aufgabe der Projektmodule bzw. Forschergruppen, die zum Gesamtsystem beitragen. Sie müssen vor allem den semantisch-pragmatischen Lemmatypen gerecht werden, indem sie die obligatorischen Angaben und not-

wendige Eigenschaften der Nutzeroberfläche festlegen. Hier sorgt das Hypertextsystem dafür, dass die nicht relevanten und im System unausgefüllt gebliebenen Angabearten ausgeblendet werden.

Der modular-hierarchische Charakter der Mikrostruktur zeigt sich z.B. schon darin, dass sich die lexikografischen Autoren z.B. *erst* über die Zuordnung zu einer Wortart entscheiden und *danach* die Flexionsangaben machen müssen, für die das System die passenden Kategorien vorschlägt

Die vorläufig komplexeste Klammer der Mikrostruktur ist der Wortartikel, wobei unter „Wort" alles dasjenige fällt, was als Lemmazeichen infrage kommt, unter Umständen also auch unselbstständige Wortbildungselemente und Mehrworteinheiten. Auch außerhalb der „Wortartikel" wird wohl jedes lexikalische Informationssystem noch weitere, den sog. Umtexten analoge Teile enthalten: Benutzungshinweise, Informationen über das Projekt, die Konzeption des Ganzen und diejenige der realisierten Module, bibliografische Daten, ein terminologisches Glossar, ggf. eine Grammatik und ähnliches. Auf diese wird nachfolgend nicht weiter eingegangen.

Wie ist die Mikrostruktur intern gegliedert, wenn nicht (mehr) von einem linearen Rezeptionsprozess beginnend bei der Lemmazeichengestaltangabe und endend etwa bei den Belegangaben ausgegangen werden kann? Eine Trennung von Datenmodellierung und Nutzeroberfläche bedeutet nicht, dass die einzelnen Angaben in gar keinen strukturbedingten Interrelationen mehr zu sehen sind. Sie alphabetisch, nach Datentyp oder nach Umfang zu sortieren, hieße Zusammengehöriges auseinanderzureißen und wortartikelinterne Klammerungen auszuschließen. Es wäre absurd, wollte man die lexikografischen AutorInnen je Lemma der alphabetischen Reihe nach Abkürzungsangabe, Ableitungsangabe... Wortbildungsproduktivitätsangabe, Worttrennungsangabe bis zuletzt die Zitatangabe eintragen lassen. Es sind lediglich partiell *andere* Inhaltsaspekte, die die Datenmodellierung einerseits und die Gestaltung der Nutzeroberfläche andererseits leiten.

Hinzu kommt die Arbeitsteilung, die bei größeren und längerfristigen lexikografischen Projekten unausweichlich und notwendig ist. Zwar wurde oben gesagt, dass ein Informationssystem wie LEKSIS wörterbuchtyp*un*spezifisch ist, das heißt aber nicht, dass die AutorInnen bei ihrer Wortzschatzbeschreibung keine Schwerpunkte analog zu Wörterbuchtypen setzen. Die partiell gleiche Lemmastrecke wird beispielsweise überlappend von einer Forschergruppe Synonymik, von einer Forschergruppe Fremdwortschatz, einer Forschergruppe Neologie und einer Forschergruppe Kollokationen bearbeitet. Jede dieser Gruppen nimmt je spezifische Angaben etwa zu einem Lemma wie *realisieren* vor und muss diese Angaben mit anderen, inhaltlich „benachbarten" kompatibel halten. Die mikrostrukturellen Vorgaben für die Autoren dürfen also in linguistisch-lexikologischer Hinsicht nicht kontextblind sein.

Außerdem ist trotz der erwähnten planerischen Trennung von Datenmodell und Nutzeroberfläche die frühzeitige Beschäftigung mit nutzerseitigen Aspekten nicht obsolet. Die Bestimmung der atomaren Informationseinheiten darf nicht nur aus linguistischer Sicht im Hinblick auf linguistische Nutzer geschehen, sondern muss bestimmte Informationseinheiten auch aus sprachinteressierter Sicht und damit gröber, zusammenfassender definieren. Bei Letzterem kommen die tradierten Erwartungen an die Textsorte Wörterbuch stärker ins Spiel. Sie sollten auf der neuen Entwicklungsstufe der Textsorte, die bei einem Hypertext-Wörterbuch ansteht, aber nicht so dominant werden, dass die Erwartungen, die an den Papierdruck gebunden sind, lediglich elektronisch kopiert werden.

3 Interne Differenzierung und Modularisierung der Mikrostruktur

Auf der Basis der in den vorhergegangenen Abschnitten erläuterten Aufgaben und Grundsatzentscheidungen ließe sich sicherlich mehr als eine Mikrostruktur konstruieren. Entscheidend ist daher, dass sich diejenige, für die sich ein Projekt wie LEKSIS (möglichst zügig) entscheiden muss, gut begründen lässt.

Interne Differenzierung und Ordnung ergibt sich erstens aus der Klassifizierung der elementaren Informationseinheiten, zweitens aus der hierarchischen Setzung weiterer Klammerungen innerhalb der Klammer Wortartikel.

In der Mikrostruktur wird zwischen Klassen von *Angaben, Kommentaren* und *Erläuterungen* unterschieden; alle gemeinsam werden hier als *Angaben* zusammengefasst. Nach Reichmann (1986, 152ff.) unterscheiden sich Angaben und Kommentare dadurch, dass Angaben „das als faktisch Hingestellte" (ebd. 152) und Kommentare darauf bezügliche Gewichtungen, Interpretationsvorschläge und Verständnishilfen enthalten. Diese Unterscheidung wird in LEKSIS übernommen, aber noch strikter gefasst, insofern Kommentare keinerlei ‚Zusatzinformationen' enthalten dürfen, für die in einem informatorischen Maximalprogramm ja anderswo Platz vorgesehen ist.

Kommentare setzen mindestens eine Angabe voraus, auf die sie sich beziehen; sie dürfen keine neuen Informationen bringen und keine Daten enthalten, die für die Datenbank identifizierungsrelevant sind. Kommentare können zu *jeder* Angabe vorgenommen werden; eine Beschränkung auf einige wenige ist bei Wegfall des im Druck gegebenen Platzproblems nicht mehr zu rechtfertigen.

Der Angabetyp *Erläuterung* dient der Strukturierung der Umtexte.

Angabe(klasse)n im engeren Sinne unterscheiden sich in verschiedenen Hinsichten: Aus lexikografischer Sicht gibt es obligatorische und fakultative relativ zum Gesamtsystem und relativ zu einem bestimmten Bearbeitungsmodul. Beispiel: Die Zahl der obligatorischen Angaben wird in einem umfassenden, bündelnden Informationssystem wie LEKSIS eher gering sein und neben der Lemmazeichengestaltangabe vor allem die mittels korpuslinguistischer Verfahren (teil)automatisch zu gewinnenden Angaben enthalten. Für den Teilwortschatz „Neologismen" oder den Teilwortschatz „sprachhandlungsbezeichnende Verben" werden darüber hinaus eigene Obligatoriken entworfen, hier die Angaben zur Entlehnung oder Wortbildung, dort die Angaben zu Synonymie, Antonymie und Pragmatik.

Angabe(klasse)n unterscheiden sich ferner nach ihrem Status als (a) autonomer Text, (b) funktionales Textsegment, (c) tabellarisches Element. Für gedruckte Wörterbücher definiert Wiegand (1989, 464) Angaben als funktionales Textsegment – in einem datenbankbasiertem Hypertextsystem kommen zwei weitere Möglichkeiten hinzu: Als autonome Textsegmente fungieren solche Angabeklassen, die auf der Benutzeroberfläche in genau derselben Form und ohne Zusätze präsentiert werden, wie dies z.B. für Bedeutungs-‚Geschichten' oder für die Beschreibung pragmatisch komplexer Verwendungszusammenhänge erforderlich ist. Falls in solchen ‚Geschichten' identifizierbare Angaben zu Zeit, Raum oder zu Kategorien relativ begrenzter Kategorieninventare (z.B. abgekürzte Bezeichnungen für Herkunftssprachen) enthalten sind, sollten diese im Fließtext markiert, d.h. getaggt werden, damit sie der Datenbank nicht verloren gehen. Als (b) funktionales Textsegment werden solche Angaben klassifiziert, die sich grob gesprochen in einen standardisierten Lückentext (s.o.) einfügen lassen müssen. Solche Lückentexte dienen auf der Benutzerseite dazu, Angaben zu erläutern, in Zusammenhänge einzuordnen und so ‚verdaulicher' zu machen. Geeignet hierfür sind z.B. formgeschichtliche Angaben, Angaben zum

Zusammenhang eines semasiologischen Feldes oder Angaben zur internen Struktur von Kollokationen. Als (c) tabellarische Elemente werden solche Angaben klassifiziert, die keine kotextuellen Erläuterungen erfordern, wie dies exemplarisch bei Genusangaben, Flexionsparadigmen u.ä. der Fall ist. Auf Benutzerseite können hier bei Bedarf minimale Überschriften oder Etikettierungen wie „Flexion: ..." hinzutreten.

Für jede Angabeklasse muss ferner eine Reihe von Eigenschaften oder Attributen festgelegt werden, die weitere Klassifizierungen zuließen. Diese Festlegungen antworten vereinfacht gesagt auf die Frage: Woraus besteht diese Angabe? Aus objektsprachlichen Ausdrücken (wie bei einer Synonymenangabe), aus beschreibungssprachlichem Text (wie bei einer Bedeutungsparaphrasenangabe), aus beidem? Wenn beschreibungssprachlicher Text zu produzieren ist, müssen dabei standardisierte Kategorialausdrücke (für Genus, für Varietätenzugehörigkeit, für semantische Verschiebungen, u.ä.) verwendet werden oder nicht?

Nur eine Kategorie spielt bei der Klassifizierung von Angaben im Hypertext-Wörterbuch *keine* Rolle: die Position im (gedruckten) Wortartikel.

Soweit die Überlegungen zu den Elementen der Mikrostruktur. Nun zum System, das die Angabeklassen bilden.

Der LEKSIS-Wortartikel wird in sechs informationelle Dimensionen bzw. – auf der Ebene der Datenmodellierung: – komplexe Objekte geteilt, die untereinander gleichrangig und nicht-sequenziell sind. Für deren Ansatz ist zunächst eine auf die Nutzerseite bezogene Entscheidung verantwortlich: Nach Wahl eines Lemmas werden den Nutzern kreisförmig angeordnete Schaltflächen zu folgenden Informationsdimensionen geboten, hinter denen sich die Suchwege in Details verzweigen. Die linguistisch systematische Motiviertheit dieser Dimensionen ist unschwer zu erkennen:

- Schreibung & Aussprache/auf die Ausdrucksseite bezogen
- Bedeutung & Verwendung/auf die Inhaltsseite, semantisch und pragmatisch, bezogen
- Grammatik
- Historisches & Sachliches/Formgeschichte, Bedeutungsgeschichte, Enzyklopädisches
- Dokumentation/Belege, Zitate, korpusanalytische Daten, fremdsprachige Äquivalente, Vergleichsdaten u.a.
- Sprachkritisches & Normatives / Sprachreflexives; hier soll die normative Wirkung deskriptiver Angaben kommentierend transparent gemacht, zur Diskussion gestellt und relativiert werden.

Ein Problem, das sich noch nicht bei der Erarbeitung der Datenmodellierung, aber später stellt, ist die Benennung dieser Schaltflächen.

Die klassische Lexikografie braucht wegen des Zwangs zu Knappheit und hoher Textverdichtung kaum jemals explizite Benennungen der Angabe- oder Informationsarten je Artikelposition zu geben; die Nutzer lernen sie durch Übung und anhand der Typografie zu identifizieren und auseinanderzuhalten, seltener auch durch Benutzungshinweise; sie haben in der Regel aber keine Begriffe dafür.

In der Mikrostruktur klassischer Wörterbücher wird der Terminus Bedeutungsidentifizierungskennzeichnung verwendet (Wiegand 1998, 226), um Bedeutungsstellennummern einer strukturellen Kategorie zuzuordnen. Gemeint sind Angaben der Art „zu *Absatz* (3) gehört das Kompositum *Absatzzahlen*". Solche Bedeutungsstellennummern sind in einem Hypertext-Wörterbuch problematisch, wenn nicht gänzlich unbrauchbar, weil ein Wortartikel über mehrere Bildschirmseiten verteilt ist und die Nutzer nicht mehr durch einen Blick in den oberen Druckraum ersehen können, welche Einzelbedeutung die Nummer „(3)" hat. Die semantisch „leeren" Zahlziffern müssen in einem hypertextuellen Informa-

tionssystem durch Identifikatoren ersetzt werden, die die Zuordnung zu einer der Einzelbedeutungen ohne Umwege gestatten. Wir haben dafür eine sehr kurze Form der semantischen Paraphrase gewählt, die als Bedeutungsettikettierungsangabe bezeichnet und immer dann eingesetzt wird, wenn einzelbedeutungsbezogene Angaben gemacht werden müssen. Etwa: „zu *Absatz* in der Bedeutung ‚Verkauf' gehört das Kompositum *Absatzzahlen*"; hier wird *Verkauf* als Bedeutungsettikettierungsangabe getaggt. Selbstverständlich wird von solch einem semantischen Ettikett immer per Link auf die vollständige semantische Paraphrase verwiesen.

Elektronische Informationssysteme zwingen die Nutzer immer wieder zu Informationsselektionen, was von den Autoren des Systems verlangt, die selegierbaren Teile ihres Informationsangebots in den vorgeschalteten Menüs so zu benennen, dass Nutzer die ihren Bedürfnissen entsprechende Wahl treffen können. Es müssen beschreibungssprachliche Etikettierungen oder – was noch schwieriger sein dürfte: non-verbale Icons – gefunden werden, wobei die terminologische Kluft zwischen Linguisten und Sprachinteressierten scharf zutage tritt. Was LinguistInnen hinter „Angaben zur Semantik" oder „Angaben zur Pragmatik" oder „zum Gebrauch" vermuten, ist – wie ein Gespräch mit Studierenden zeigte – nicht selten partiell identisch mit dem, was unter Etymologie („was das Wort ursprünglich heißt") fällt; „Gebrauch" wird leicht mit morphologisch-syntaktischen Normen assoziiert („ob man *wegen* mit Dativ oder Genitiv gebraucht") oder auch mit semantischer Variation („ob das Wort übertragen gebraucht wird"); Belege bzw. Zitate werden als „Beispiele" klassifiziert und entsprechend, d.h. tendenziell als vorbildlich verstanden (!), aber kaum als Nachweis der in den Angaben formulierten Regeln und als Verifikationsinstanz der lexikologischen Methode.

Für die Wörterbuchbenutzungsforschung öffnen sich hier neue Horizonte. Und: lexikalische Informationssysteme können das begriffliche Inventar sprachreflexiven Wissens der Nutzer und damit vielleicht tendenziell der Sprachgemeinschaft verändern. Das Benennungsproblem ist also kein nebensächlicher Aspekt.

Für die Definition von Elementen und damit der Tags für die DTD genügen Abkürzungen von Benennungen, die die AutorInnen verstehen, selbst wenn sie einem projektspezifischen Jargon entstammen. Allerdings sollte man bei längerfristigen elektronischen Wörterbüchern (und welche bezeichnen sich freiwillig als kurzfristig?!) auf die „Nachhaltigkeit" und transparente Dokumentation des Benennungssystems achten, damit nachfolgende Wissenschaftlerinnen und Wissenschaftler Inhalt und Funktion einer Angabeklasse eindeutig identifizieren können. Wichtig ist aber vor allem die begründete Bestimmung der Angabearten selbst, weshalb wir parallel zum Aufbau der Mikrostruktur ein Redaktions- oder Instruktionshandbuch „mitschreiben", in dem die Überlegungen zur Ansetzung und zur Charakteristik jeder Angabeart festgehalten werden können. Dieses wird den AutorInnen später während der Abfassung der Wortartikel online zur Verfügung stehen.

Die o.g. sechs Informationsdimensionen werden formal als komplexe Objekte oder Klammern aufgefasst, die bestimmte Angaben enthalten, die ihrerseits weitere Angaben einklammern, und/oder eine Reihe atomarer, d.h. nicht weiter zerlegbarer Elemente enthalten. Beispielweise enthält die Bedeutungsdimension u.a. ein komplexes Objekt/ eine Klammer „Paradigmatik", in der die Angaben zu Synonymen, Antonymen usw. zusammengefasst sind, wobei man die Antonymenangabe bei Bedarf typologisch weiter differenzieren könnte. Ausschlaggebend für den Differenzierungsgrad ist das antizipierte lexikologische Interesse an den Datenbankinformationen und ihrer Verknüpfung. Auch zukünftige Forschungen sollten, so schwierig dies erscheint, weitmöglichst berücksichtigt werden, etwa indem ein noch undifferenziertes Datenobjekt vorsorglich eingebaut wird. Atomare

Objekte sind „objektsprachlicher Text" und „beschreibungssprachlicher Text", obwohl auch sie genau genommen aus einer Menge Text zuzüglich obligatorischer Autoren- und Datumsangabe bestehen. Letzteres ist für die langfristige Verwaltung und Aktualisierung der Daten unerlässlich.

Einerseits enthalten die übergeordneten sechs informatorischen Komplexe genau die Menge an Angabearten, die linguistisch wünschbar ist, andererseits müssen in ihnen aber auch solche Angabearten enthalten sein, die nicht-linguistische Nutzer mit Bedarf an Überblicks- und zusammenfassender Information nachfragen könnten. Man kann davon ausgehen, dass solche Zusammenfassungen sich gerade nicht schematisch oder automatisch aus den linguistisch detaillierten Angaben ableiten lassen. Diese zu komprimieren, erhöht die Abstraktion noch. Vielmehr werden sie partiell gesondert zu formulieren und an entsprechender Stelle in der Datenbank abzulegen sein. An diesem Punkt zeigt sich ein weiterer Vorteil des Ansatzes der sechs komplexen Informationsobjekte. Diese können neben linguistisch motivierten Angabarten immer auch eine oder mehrere Angaben vom Typ „zusammenfassender Kommentar für sprachinteressierte Nicht-Linguisten" enthalten, auf die das Hypertext generierende System immer dann zurückgreift, wenn zuvor eine entsprechende Benutzungssituation oder ein Benutzertyp ausgewählt worden ist. Es ist sogar denkbar, das diese Arbeit nachträglich oder gesondert von hierzu besonders befähigten Autoren vorgenommen wird.

4 Vom Verweis zur Linktypologie

Was im gedruckten Wörterbuch der Verweis ist, ist im Hypertext der Link. Vor dem Hintergrund möglicher Links, die in eine Mikrostruktur einzubauen sind, wird allerdings sehr schnell klar, dass es viele, sehr unterschiedliche Arten von Verweisen gibt, die im gedruckten Wörterbuch meist in einen Topf geworfen werden. Verweise sind im gedruckten Wörterbuch in der Regel – von den diese bestätigenden Ausnahmen abgesehen – untypisiert und haben einen weiten bzw. undifferenzierten Skopus.

Der entscheidende Unterschied zwischen Verweis und Link (=Verweis im Hypertext) besteht darin, dass Verweise entweder auf einen anderen Wortartikel als Ganzes (externer Verweis des Typs „**Absatz** [...] → Abschnitt") oder auf eine Bedeutungsstellennummer im selben Artikel zielen (interner Verweis des Typs „**Absatzflaute** [...] Flaute im Absatz (3) [...] **absatzweise** [...] in Absätzen (2b)", während ein Hyperlink sowohl auf einen ganzen Wortartikel (extern verweisend) als auch auf einen Wortartikelteil, etwa alle auf eine bestimmte Einzelbedeutung bezogenen Angaben, und auf eine einzelne Angabe in irgendeinem Wortartikel (extern und intern verweisend) zielen kann. Auch der Ort im Datenmodell, wo sich ein Link befindet (Verweisanker), ist klassifizierungsfähig. Außerdem ist ein elektronischer Verweis im Prinzip bidirektional, ein gedruckter nicht oder nur im Ideal- und Ausnahmefall.

Das neue Medium fordert also eine Linktypologie, die außer den Kriterien Anker und Ziel eine Reihe weiterer klassifizierender Kriterien berücksichtigt.

Denkbar ist eine Unterscheidung nach der Funktion eines Links: der eine Link führt von einer Angabe zum zugehörigen Kommentar (nennen wir ihn *Kommentarlink*), ein anderer von einem Fachausdruck zu dessen Erläuterung ins Glossar (nennen wir ihn *glossierender Link*), ein dritter von einem objektsprachlichen Ausdruck zum entsprechenden Lemma

(*Lemma-Link*) oder zur entsprechenden Einzelbedeutung (*Lesarten-Link*), ein vierter, wichtiger führt innerhalb der mikrostrukturellen Hierarchie nach „oben" zu allgemeineren oder verwandten Angaben oder nach „unten" in weitere Details (*Strukturlink*), ein fünfter verknüpft die Elemente eines der semantischen bzw. lexikologischen Netze, die anzusetzen sind, z.B. Handlungsnetze, diverse paradigmatische und syntagmatische Netze, Wortbildungsnetze, u.a. (*Vernetzungslink*), ein sechster Linktyp führt in eine bestimmte Stelle der Umtexte, z.B. auf die Erläuterungen zu einem linguistischen Projekt, dessen Mitglieder für eine Angabe verantwortlich zeichnen (*Projektlink, Autorenlink*), oder auf die Erläuterungen zu einem (Teil-)Korpus (*Korpuslink*). Zuletzt gibt es noch die rein assoziativen Links, die sich (zu einem gegeben Zeitpunkt noch) keiner der anderen Funktionen zuordnen lassen (*Hotlink*).

Welche Linktypen wie identifizierbar gemacht werden und wie in eine Mikrostruktur einzubauen sind, hängt natürlich stark von der Recherche-Software und den Möglichkeiten ihrer Benutzeroberfläche ab, die zur Verfügung steht.

Bei einem potenziellen Maximum an Information, wie LEKSIS es vorsieht, und bei einer primär linguistisch motivierten Mikrostruktur kann es leicht passieren, dass sich als Prinzip einer Verlinkungsstruktur das „Alles-hängt-mit-allem-zusammen" aufdrängt. Nutzer und Autoren wären damit aber sicherlich überfordert; außerdem sind Redundanzen zu erwarten (vgl. Wiegand 1998[99], 247f.). Es gibt vermutlich eine quantitativ bestimmbare Obergrenze der Anzahl von Links, die Nutzer auf einer Web-Seite ‚verkraften' oder überhaupt wahrnehmen können. Gegenüber der Menge dürfte aber die Qualität von Hilfen zur Orientierung im System die größere Rolle spielen. Das heißt, dass für die Benutzerfreundlichkeit die Strukturlinks, ihre Benennung (s.o.) und Präsentation am wichtigsten sind. Hingegen vermitteln die anderen Linktypen, insofern sie die Vernetzung des Wortschatzes repräsentieren, von der schon Hermann Paul gesprochen hat, dasjenige lexikologische Wissen, das über die wörterbuchtypisch lexemzentrierte Betrachtung hinaus greift und das die Form von wortartikelinternen Angaben sprengt.

Grundsätzlich muss, wie oben erwähnt, die Struktur, in der die Nutzer mithilfe von Strukturlinks navigieren, mit der Mikrostruktur, die – auch – die Navigation auf Autorenseite leitet, nicht identisch sein. Die Sicht der Nutzer und die Sicht der Autoren können aber dann identisch sein, wenn beide Gruppen dergleichen Fachdisziplin angehören, d.h. wenn die Herangehensweise an den Gegenstandsbereich bei den Adressaten gleich oder ähnlich strukturiert ist wie bei den Autoren. Im Projekt LEKSIS stellt die sprachwissenschaftliche Nutzung einen von zwei Benutzungssituationstypen dar; der andere ist der inter- und extradisziplinäre Benutzungssituationstyp. Für letzteren ist die Herangehensweise an den Gegenstandsbereich Wortschatz, seine Einheiten und ihren Zusammenhang, aber sehr heterogen und wenig begrifflich gefasst.

Wie die Struktur, über der nicht-linguistisch interessierte Nutzer navigieren, beschaffen sein sollte, muss wenigstens ansatzweise empirisch erprobt werden, etwa durch Fragebogen. Die Erforschung echter Benutzungssituationen kann erst nach Fertigstellung dieses oder eines ähnlichen Systems geschehen. Die von uns in Erwägung gezogene Fragebogen-Methode muss kombiniert werden mit den Ergebnissen allgemeiner Rezeptionsforschung zum Hypertext. In dieser unbefriedigenden Situation mag es verlockend erscheinen, die Situation der Hypertextnutzung an den weit besser bekannten Benutzungssituationen der Printwörterbücher auszurichten und die Nutzer von vorneherein in ihren am Papierwörterbuch geschulten Erwartungen zu bestätigen und lediglich die dort bekannten Mängel – zu starke Textverdichtung und typografische Unübersichtlichkeit – zu beheben.

Doch dieser Weg verschenkt die neuen Möglichkeiten und verschläft sozusagen die unaufhaltsam in Gang befindliche ‚Evolution' der Lexikografie. Wenn wir, die Linguistinnen und Lexikografen, uns jetzt nicht um angemessene Hypertext-Konzepte lexikologischen Wissens kümmern, werden Fachfremde dies tun und dabei Standards schaffen, die wir später kaum wieder werden außer Kraft setzen können.

5 Zwang zu theoretischer Reflexion

Insgesamt gesehen führt die Erarbeitung der Mikrostruktur für LEKSIS wiederholt zur Auseinandersetzung mit einigen sprachtheoretischen Problemen, von denen hier das wichtigste erläutert sei:

> Was ist der zentrale Gegenstand lexematischer Information – die gesamte Inhaltsseite des Lemmazeichens oder die jeweilige Einzelbedeutung?

In gedruckten Wörterbüchern stehen alle formbezogenen und von Einzelbedeutungen unabhängigen, d.h. orthografischen und morphosyntaktischen Angaben im oft „Kopf" genannten, d.h. ausgelagerten Formkommentar der Mikrostruktur; danach erst kommen die mittels Zahlziffern geordneten einzelbedeutungsspezifischen, meist semantisch-pragmatischen Angaben. Dabei kommt es nicht selten zu Problemen mit dem Skopus des Formkommentars. Unmarkierter Weise umfasst der Skopus alle lesartenspezifischen Angaben; im markierten Ausnahmefall enthalten aber auch die semantisch-pragmatischen Angabeklammern ‚unten' Informationen zu morphosyntaktischen und anderen Besonderheiten, wodurch der Skopus des Formkommentars aufgehoben oder eingeschränkt wird. Vor allem grammatische Informationen, z.T. aber auch historisch-diachrone und enzyklopädische werden auf diese Weise in zwei oder mehr Elemente aufgespalten. Außerdem ist diese gängige Praxis, die in gedruckten Wörterbüchern aus Platzgründen ja nirgends explizit erläutert werden kann, Quelle von Schwer- und Missverständlichkeit. Es stellt sich also die Frage, ob die konventionelle Auslagerung des Formkommentars im Hypertext-Wörterbuch nicht aufgehoben bzw. durch eine konsequentere Zuordnung von formbezogener und inhaltsbezogener Information ersetzt werden könnte.

Die durch die Zweidimensionalität des Druckraums erzwungene Strukturierung suggeriert, dass das Lemmazeichen der eigentliche, zentrale Gegenstand ist, über den das Wörterbuch etwas aussagt, obwohl das Lemmazeichen aus zeichentheoretischer, textlinguistischer und auch semantisch-kognitiver Perspektive im Normalfall keine Einheit darstellt. Man hat es in der einen (text- und korpuslinguistischen) Perspektive vielmehr mit potenziell unendlich vielen, durch Kontext monosemierten Wortzeichen zu tun, die abstrahierend-klassifizierend zu Lesarten zusammengefasst werden. Der Unterschied zwischen Glossar und Wörterbuch besteht in eben dieser Abstraktion. Die Lesarten/Einzelbedeutungen[6] besitzen darüber hinaus aber auch im allgemeinen Sprecherbewusstsein eine mehr oder weniger deutliche Identität. In kognitionslinguistischer Perspektive spricht man daher von Klassenkonzepten als den Einheiten des mentalen Lexikons.

---

[6] *Einzelbedeutung* ist die in der Lexikografie gängigste Bezeichnung für das, was in der Semantik(theorie) meist *Lesart* genannt wird. Die beiden Ausdrücke beziehen sich auf das selbe Denotat.

Es kann hier nur angedeutet, nicht alles aufgeführt werden, was dafür spricht, die einzelne Lesart zum zentralen Gegenstand der Mikrostruktur, d.h. zu der auf das Lemmazeichen unmittelbar folgenden Ebene zu machen und jeweils alles, was es an grammatischen, semantisch-pragmatischen, historischen und enzyklopädischen Informationen gibt, lesartenbezogen zuzuordnen. Dafür spricht u.a., dass paradigmatische und syntagmatische Vernetzungen ausschließlich zwischen Lesarten, nicht zwischen Lemmazeichen bestehen. Im Falle eines z.B. 42-fach polysemen Lemmazeichens, bei dem Schreibung, Aussprache und morphosyntaktische Angaben 42mal bzw. 39mal identisch und 3mal spezifisch sind, hieße dies, 42- bzw. 39mal die gleiche Angabe zu speichern oder von 42 bzw. 39 Positionen aus automatisch die gleiche Angabe zu erzeugen.

Auf der anderen Seite hängen die Lesarten einer Wortform meistens in irgendeiner Weise zusammen, so dass von einer Gesamtbedeutung gesprochen werden kann, auch wenn diese gemessen an der Textbedeutung noch stärker konstruiert ist und eine weitergehende Abstraktionsstufe darstellt. Die lesartenzentrierte Mikrostruktur muss demnach besondere Angaben zum Zusammenhang zwischen den Lesarten eines Lemmazeichens enthalten. Dennoch suggeriert diese Struktur eine weitgehende Autonomie jeder Einzelbedeutung. Bei polysemen Wörtern wie *Absatz* scheint dies angemessen, weniger hingehen bei Konzeptfamilien wie *Schule*.

Alternativ könnte man diesen semasiologischen Zusammenhang zum zentralen Gegenstand der Mikrostruktur machen, dessen Einheit durch weitgehend einheitliche Schreibung, Ausprache und Grammatik „erwiesen" schiene, wenn man nicht gar eine – dann hochgradig konstruierte – „Kern"- oder „Grund"-Bedeutung formulieren wollte. Bei den semantisch-pragmatischen Angaben wie bei den diversen semantischen Vernetzungen müsste dann nach Lesarten differenziert werden – und bei zahlreichen weiteren Ausnahmen. Diese Struktur suggeriert demnach eine quasi natürliche, weil in den seltensten Fällen (nämlich nur bei Konzeptfamilien) begründbare, einheitliche Inhaltsseite der Lemmazeichenform.

Nicht nur aus linguistischer Perspektive dominieren die Argumente für ein lesartenbezogenes Datenmodell; auch datenbanktechnologisch scheint diese Alternative die weniger komplizierte zu sein.

6 Literatur

Fraas, Claudia/Haß-Zumkehr, Ulrike (1999): Vom Wörterbuch zum Informationssystem. Über ein neues Projekt des Instituts für Deutsche Sprache. In: Deutsche Sprache 4/1998 (1999), 289–303.

Lemberg, Ingrid/Petzold, Sybille/Speer, Heino (1998): Der Weg des Deutschen Rechtswörterbuchs in das Internet. In: Wiegand, Herbert E. (Hg.) (1998): Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium, Tübingen, 262–284.

Reichmann, Oskar (1986): Lexikographische Einleitung, in: Frühneuhochdeutsches Wörterbuch, hg. von Robert K. Anderson, Ulrich Goebel und Oskar Reichmann, Bd. 1, Berlin, 10–164.

Storrer, Angelika (2000): Was ist „hyper" am Hypertext? In: Werner Kallmeyer (Hg.) (2000): Sprache und neue Medien. Berlin, New York.

– und Bettina Harriehausen (Hrsg.) (1998): Hypermedia für Lexikon und Grammatik. – Tübingen 1998 (Studien zur deutschen Sprache 12).

Wiegand, Herbert Ernst (1989): Arten von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. – Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. Ed. Franz Josef Hausmann, Oskar Reichmann, Herbert E. Wiegand, Ladislav

Zgusta. 1. Teilbd. – Berlin, New York (= Handbücher zur Sprach- und Kommunikationswissenschaft 5.1), 462–501.
- (1998): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilbd. – Berlin, New York.
- (1998[99]): Neuartige Mogelpackungen: Gute Printwörterbücher und dazu miserable CD-ROM-Versionen. Diskutiert am Beispiel des LEXIKONS DER INFEKTIONSKRANKHEITEN DES MENSCHEN. In: Lexicographica 14/1998 (1999), 239–253.

Ulrike Haß-Zumkehr, Mannheim

Thomas Gloning, Rüdiger Welter

Wortschatzarchitektur und elektronische Wörterbücher: Goethes Wortschatz und das Goethe-Wörterbuch

1 Einleitung

Wer die alphabetische Ordnung des Wörterbuchs aufgibt, der versündigt sich an der Philologie, wetterte Jacob Grimm in die Richtung der ‚nach Wurzeln' geordneten Wörterbücher. Die alphabetische Ordnung eines Wörterbuchs geht an den internen Beziehungen im Wortschatz vorbei, sie ist nur praktisch, donnerte Hermann Paul zwei Generationen später und forderte, man möge auch begriffliche und andere Zusammenhänge bei der Wortschatzarbeit berücksichtigen. Zum Beispiel: *donnern* und *wettern* als redekennzeichnende Verben stehen im Wörterbuch an völlig unterschiedlicher Stelle im Alphabet. Aber die beiden Verwendungsweisen gehören in mindestens zweierlei Hinsicht enger zusammen: zum einen sind die beiden Verben eng bedeutungsverwandt, zum anderen sind die beiden Verwendungsweisen zur Redekennzeichnung wohl durch metaphorische Übertragung auf ähnliche Weise entstanden. Will man ein Wörterbuch drucken, dann muß man sich für einen einzigen Gesichtspunkt der Anordnung entscheiden, z.B. den Gesichtspunkt des Alphabets, den Gesichtspunkt der Bedeutungszusammenhänge oder den bisher nur in Registern erfaßten Gesichtspunkt der historisch-semantischen Charakteristik.

Daneben gibt es eine ganze Reihe von weiteren Gesichtspunkten der Wortschatzarchitektur und von Wortschatzzusammenhängen, die zu dokumentieren sich lohnt. Die Frage ist also nicht nur, ob man ein Wörterbuch alphabetisch oder onomasiologisch organisiert, sondern auch, ob und wie es gelingen könnte, eine Mehrzahl von lexikologischen Organisationsprinzipien bei der Dokumentation eines Wortschatzes gleichzeitig zu berücksichtigen. Im gedruckten Wörterbuch geht das nur eingeschränkt: Man muß sich für ein einziges Prinzip der Anordnung entscheiden und kann andere lexikologische Gesichtspunkte allenfalls durch erschließende Register einholen, ein Verfahren, das sich im ganzen als aufwendig, als wenig praktikabel und demzufolge auch als selten praktiziert erwiesen hat.[1]

Elektronische Datenbasen erlauben es nunmehr, auf einen Datenbestand, z.B. eine Wortschatzdokumentation, nach unterschiedlichen Kriterien zuzugreifen und dabei auch ganz unterschiedliche ‚Ansichten' des Datenbestandes je nach den Interessen und Fragestellungen von Benutzern hervorzubringen. Voraussetzung ist, daß die entsprechenden Informationen, z.B. zur lexikologischen Strukturierung, in expliziter Weise im Datenbestand enthalten sind. Ein Wörterbuch als elektronische Datenbasis mit mehrfacher Zugriffsmöglichkeit ändert damit auch seinen Charakter: Das gedruckte Wörterbuch ist im wesentlichen eine Ansammlung von Einzelwörtern bzw. ‚Einzelwortschicksalen' mit eingeschränkten Möglichkeiten, Zusammenhänge im Wortschatz aufzuweisen. Das elektronische Wörterbuch als komplex strukturierte lexikologische Datenbasis kann sehr viel mehr sein: ein Abbild der komplexen Zusammenhänge im Wortschatz und ein Informationssystem für sehr unterschiedliche Benutzerinteressen.

In den folgenden Abschnitten wird zunächst der Grundgedanke einer komplexen und mehrdimensionalen Wortschatzarchitektur erläutert, es folgt eine Fallstudie zum Wortschatz von Goethe und zum Goethe-Wörterbuch. Dabei werden neben einigen technisch-konzeptionellen Grundgedanken vor allem die erweiterten Nutzungsperspektiven eines elektronischen Wörterbuchs skizziert, das als elektronische Datenbasis erfaßt ist.

2 Zusammenhänge im Wortschatz und Aspekte der Wortschatzorganisation

Die grundlegende Einheit der Wortschatzstrukturierung ist die Verwendungsweise eines Wortes bzw. eines Lexems. Zwischen den einzelnen Lexemen bzw. ihren Verwendungsweisen bestehen vielfältige Zusammenhänge, und man kann die Architektur eines Wortschatzes als eine komplexe Ordnung bezeichnen. Die Komplexität der Wortschatzarchitektur beruht zum einen auf der Mehrzahl von lexikologischen Organisationsprinzipien und Strukturierungsgesichtspunkten, zum anderen auf der Tatsache, daß sich diese Organisationsprinzipien in zahlreichen Fällen auch kombinieren lassen. Und nicht zuletzt natürlich auch auf der großen Zahl der Elemente: der Wörter und Verwendungsweisen.

Ein lexikologisches Organisationsprinzip ist ein Gesichtspunkt, bei dessen Anwendung sich ganz bestimmte Zusammenhänge im Wortschatz ergeben. Nimmt man z.B. den Gesichtspunkt der Bedeutungsverwandtschaft, dann ergibt sich eine semantische Strukturierung des Wortschatzes als eine Art onomasiologisches Netz. Nimmt man den Gesichtspunkt der Zugehörigkeit von Wörtern bzw. Verwendungsweisen zu Fachgebieten oder Sachbereichen, dann ist das Ergebnis eine Fachgebiets- bzw. Sachbereichs-Strukturierung des betreffenden Wortschatzes. Wählt man unterschiedliche Gesichtspunkte der historischen Entwicklung, dann ergeben sich z.B. eine Herkunftsgliederung, ein Altersprofil oder eine historisch-semantische Strukturierung des Wortschatzes. Berücksichtigt man weiterhin die unterschiedlichen Gebrauchsaspekte von Wortschatzelementen, dann ergeben sich u. a. regionale oder gruppenbezogene Formen der Wortschatzstrukturierung, aber auch Wortschatzprofile, die sich auf kommunikative Aufgaben wie die Behandlung spezifischer Themen, die Realisierung funktionaler Textbausteine oder die Befolgung von Kommunikationsmaximen wie Präzision oder Originalität beziehen.

---

[1] Vgl. z.B. den Band VII des ‚Deutschen Fremdwörterbuchs' und die Register zu älteren Auflagen des ‚Kluge' (z.B. 1899, 1934), sodann aber auch die Ergebnisse in Goebel/Lemberg/Reichmann 1995.

Zwischen den einzelnen lexikologischen Organisationsprinzipien bestehen auch vielfältige Zusammenhänge und Kombinationsmöglichkeiten. Zum Beispiel: um Kommunikationsmaximen wie Präzision, Komprimierung oder Originalität zu befolgen, gebrauchen Sprecher oder Schreiber sehr oft bestimmte Wortbildungsmuster, etwa *-ung*-Bildungen für Formen der komprimierten Rede. Oder: die Gesichtspunkte der Herkunft und der thematischen Zugehörigkeit lassen sich fruchtbar kombinieren, weil viele Fremdwörter einem bestimmten Themenbereich, einem bestimmten Sach- oder Fachgebiet zugehören. Auch viele funktionale Teilwortschätze lassen sich weiter untergliedern, indem man den funktionalen Gesichtspunkt mit anderen Gesichtspunkten kombiniert (z.B. Personenbezeichnungen, die fremdsprachiger Herkunft sind; Ereignisbezeichnungen aus dem Themenbereich des Militärs; Querverweisausdrücke, die vor 1500 erstmals belegt sind usw.).

Vor dem Hintergrund der Tatsache, daß es zahlreiche Organisationsprinzipien für Wortschätze gibt, die teilweise auch untereinander zusammenhängen, erscheint der alte Gegensatz zwischen alphabetisch und onomasiologisch organisiertem Wörterbuch als zu einfach. Zu jedem lexikologischen Organisationsprinzip und zu jeder fruchtbaren Kombination von Organisationsprinzipien ist eine darauf bezogene Dokumentation in Form eines Wörterbuchs möglich. Daß viele dieser denkbaren Wörterbücher nie gedruckt wurden, hängt vor allem mit Geboten der Sparsamkeit zusammen. Es ist nun eine besondere lexikologische Herausforderung, die elektronische Datenbasis für ein Wörterbuch mit Markierungen derart anzureichern, daß die Auswertung der entsprechenden Markierungen bei der Benutzung unterschiedliche ‚Ansichten' und Profile des dokumentierten Wortschatzes erlaubt. Die alphabetische Zugriffsmöglichkeit bleibt als eine für das rasche Nachschlagen wichtige lexikographische Errungenschaft[2] natürlich erhalten, sie wird aber ergänzt durch zahlreiche weitere Zugriffsweisen, die auf den internen lexikologischen Markierungen beruhen. Die lexikalische Datenbasis bietet den Benutzern somit die Möglichkeit, unterschiedliche Ausschnitte und verschiedene Ansichten des Wortschatzes auszuwählen. Ist der Datenbestand eines Wörterbuchs als lexikologische Datenbasis vorhanden, dann kann ein Benutzer z.B. durch Ein- und Ausblenden der Belege zwischen zwei Wörterbuchtypen hin- und herschalten, die etwa dem kleinen und dem großen LEXER entsprechen. Man kann den Datenbestand nach Bedarf in einer onomasiologisch aufbereiteten Version nutzen, man kann aus dem Gesamtbestand interessante Teilwörterbücher abziehen, zum Beispiel zu einzelnen Wortarten, zu einzelnen Autoren, zu Textsorten, zu Regionalismen, zu den Fremdwörtern, zu einzelnen zeitlichen Schichten, aber auch zu Kombinationen wie z.B. den Regionalismen, die bei einzelnen Autoren belegt sind.

All diese Ansichten und Auswahlen produziert ein Computer nicht ohne weiteres, sondern aufgrund der Markierungen, die kompetente Lexikographinnen und Lexikographen im Datenbestand vorher verankert haben. Zu den Stärken der beteiligten Menschen gehören im besten Fall Intelligenz, Sprachbeherrschung, vielschichtige Bildung, Konsequenz, Sitzfleisch und philologischer Feinsinn, die Stärke des Computers ist es, in kurzer Zeit riesige Datenbestände im Hinblick auf die markierten Zusammenhänge zwischen den Elementen zu durchforsten und die Ergebnisse in geordneter Form auszugeben, so daß die

---

[2] Vgl. zur strengen Alphabetisierung als einer lexikographischen Errungenschaft u. a. Powitz 1988, 211 f. mit Fußnoten 3 und 11. Hier heißt es über Johannes de Janua (Johannes Balbus) und sein ‚Catholicon' (abgeschlossen 1286; erstmals gedruckt 1460): „Zum Erfolg des ‚Catholicon' trug ein weiterer Vorzug bei: die neuartige Organisation des lexikalischen Stoffes. Die Entdeckung und Anwendung des streng alphabetischen Prinzips ist der Stolz des Autors und sein tatsächliches historisches Verdienst".

Menschen wieder intelligent mit den Ergebnissen weiterarbeiten können. So oder so ähnlich kann man die Verteilung der Arbeit Mensch/Computer vielleicht sehen.

Wir wollen die Idee einer mehrdimensionalen Strukturierung und Erfassung von Wortschätzen in elektronischen Wörterbüchern nun anhand von Goethes Wortschatz und anhand des gedruckten Goethe-Wörterbuchs erläutern und dabei auf einige Nutzungsperspektiven und zumindest kurz auf einige Gesichtspunkte der konzeptionellen und technischen Realisierung eingehen.

3 Die Architektur des Goethe-Wortschatzes

Goethes Wortschatz umfaßt über 90000 Stichwörter, von denen viele ein fein differenziertes Spektrum von Verwendungsweisen haben. Dieser Personal-Wortschatz ist in vielfacher Hinsicht aufschlußreich, und für jede der im folgenden genannten ‚Hinsichten' der Wortschatzstrukturierung ist es wünschenswert, Zugriff auf eine Dokumentation des entsprechenden Teilwortschatzes zu haben.

Ein bekanntes Strukturierungsprinzip sind die semantischen Zusammenhänge im Wortschatz. Der Zusammenhang zwischen Quasi-Synonymen und bedeutungsverwandten Wörtern bzw. Verwendungsweisen wird im Goethe-Wörterbuch in der Artikelposition **Syn** verbucht. Zu den semantischen Zusammenhängen im Wortschatz gehören aber auch die im GWb nicht in einer eigenen Rubrik aufgeführten Gegensatzrelationen, die in Goethes Darstellungen vor allem bei Formen der Kontrastierung eine wichtige Rolle spielen. So wird z.B. im Feld der Wahrnehmung ein Kontrast aufgebaut zwischen *scheinbar* und *eigentlich* bzw. *wahr*.

Goethes Wortschatz umfaßt zahlreiche wichtige Fachgebiete und thematisch-sachliche Bereiche der Zeit um 1800. Zu diesen Themenbereichen und Sachgebieten gehören – hier jeweils mit einigen Unterrubriken und Beispielen – unter anderem:

- Anatomie und Morphologie, z.B. mit Bezeichnungen für Körperteile und Körperbestandteile (*Gelenkapophyse*, *Kinnlade*, *Nasenknochen*, *Zwischenknochen*; *Retina,* im Schnittbereich zwischen Farbenlehre und Anatomie, *Hinterhaupt* im Schnittbereich zur Physiognomik) oder vielfältige Bezeichnungen für Aspekte der Gestalt (*Bildung*, *Fortsatz*, *Hervorragung*, *Sutur*, *Wölbung*);
- Dichtungs- und Literaturtheorie, z.B. mit Verben zur Bezeichnung von möglichen Wirkungen der Dichtung (*bewegen*, *rühren*), Bezeichnungen für literarische Gattungen (*Epos*, *Novelle*) oder literarische Darstellungsformen und Prinzipien (*Abstraktion*, *Anschauung*, *Anmut*, *Bild*, *Genauigkeit, Nutzen*);
- Farbenlehre, z.B. mit Bezeichnungen für Farben und Farbeindrücke, deren zugrundeliegende Wortbildungsmuster in besonders eindrucksvoller Weise auch die Formulierungsnöte in diesem Bereich erkennen lassen (*blaßgelb*, *bläulich*, *blutrot*, *dunkelorange*, *gelb*, *gelblich*, *gelbrot/rotgelb*, *graulich*, *hellgelblich*, *hochrot*, *rosenfarb*, *rötlich*, *weiß*), Bezeichnungen für die Manipulation von Farben und Farbsubstanzen (*lasieren*, *mischen*, *diluieren*), für Geräte (*Chromatoskop*) und allerhand optische Erscheinungen und Verhältnisse (*Gegenlicht*);
- Geologie, Bergbau und Mineralogie, z.B. mit Bezeichnungen für unterschiedliche Gesteinsarten (*Alabaster*, *Bufonit*, *Feldspat*, *Frauenglas*, *Glimmer*, *Gneis*, *Gneisgranit*, *Granit*, *Granitgneis*, *Quarz*, *Sandstein*, *Schörl*, *Tuffstein*), Bezeichnungen für Personen

(*Bergbeflissene*, *Bergverständige*), Bezeichnungen für geologische Formationen (*Altar*, *Bank*, *Flözkluft*, *Gang*, *Gebürge*, *Hügel*, *Kluft*, *Steinbruch*, *Versteinerung*), Bezeichnungen für Eigenschaften von Gesteinen (*abgerundet*, *breccienartig*, *porphyrartig*, *sechsseitig*, *tafelartig*, *quarzhaft*), auch dieser Wortschatzsektor deutet teilweise auf Formulierungsnöte im Umgang mit der Vielfalt der Phänomene hin;
- Pflanzenkunde, z.B. mit Bezeichnungen für Arten von Pflanzen, für Teile von Pflanzen, für Eigenschaften und Formen der Entwicklung;
- Theater und Theatermanagement, z.B. mit Bezeichnungen wie *Gruppierung* oder *Entreegeld*;
- verschiedene Bereiche der verwaltungssprachlichen Praxis, z.B. mit Ausdrücken, die vor allem als Mittel der Höflichkeit und der Kennzeichnung von amtlichen Rollenkonstellationen dienen (als *ohnzielsetzlich* oder *ohnmaßgeblich* sind in der Regel die Voten der Räte gekennzeichnet, um dem Fürsten die Entscheidungsfreiheit zu belassen).

Die strukturierte Beschreibung all dieser thematischen Wortschätze mitsamt ihren teilweise fachlichen Bestandteilen, mit ihren Neubildungen und ihrer differenzierten inneren Struktur ist ein wichtiger Beitrag zum Verständnis, wie das Werkzeug ‚Sprache' im Hinblick auf die Bewältigung thematischer Redeaufgaben funktioniert.

Der Goethe-Wortschatz enthält weiterhin sprachliche Mittel, die wesentliche Denkweisen und Auffassungen Goethes und seiner Zeitgenossen widerspiegeln, also vor allem diejenigen Wörter, die der Begründer des Goethe-Wörterbuchs, Wolfgang Schadewaldt, als Goethesche Grund- und Wesenswörter ansah. Kandidaten für diese Rubrik sind z.B. *dämonisch*, *Dasein*, *Erscheinung*, *ganz*, *Gegenwart*, *Geheimnis*, *Genie*, *Geselligkeit*, *handeln*, *heiter*, *Mitstreben*, *Teilnahme*, *Zusammenwirken* und dergleichen. Das Studium des Gebrauchs dieser Wörter führt oft schnell und zielgerecht zu wesentlichen Grundauffassungen Goethes oder seiner Zeitgenossen, zu Wandlungen und Änderungen dieser Auffassungen (z.B. bei *Genie*) oder auch zu kultur- oder wissenschaftsgeschichtlich interessanten Entwicklungen (z.B. bei *Elektrizität* und seinem Wortfeld, das sich auf ein neues Phänomen und einen neuen Wissenschaftszweig bezieht, der Mitte des 18. Jahrhunderts erst neu beschrieben bzw. entwickelt wurde, oder bei den *Dampf*-Komposita).

Sodann enthält Goethes Wortschatz eine Reihe von Euphemismen, Neubildungen und Anwendungen von Wortbildungsmustern. Ein von Goethe verwendeter Euphemismus ist z.B. *den Vaterlandsboden verlassen* für ‚sterben'. Eine Neubildung ist offenbar das Wort *feminisieren* im Sinne von ‚sich bei der sprachlicheln Darstellung einem weiblichen Publikum anpassen', das Goethe in Anlehnung an *popularisieren* verwendet hat. Die Wortbildungen wurden oben im Zusammenhang mit den Formulierungsnöten schon erwähnt, in die man bei der Beschreibung naturkundlicher Phänomene geraten kann. Viele allgemeinsprachliche Wörterbücher würden sich für solche ‚ephemere' und nur kurzfristige Erscheinungen im Wortschatz gar nicht interessieren. Die Neubildungen sind aber zum einen wichtig, um das kreative Potential des Deutschen um 1800 beurteilen zu können. Wenn man weiterhin der Auffassung ist, daß die Anwendung solcher innovativer Verfahren der Ausgangspunkt ist für historisch-semantische Neuerungen, die sich später als Bedeutungswandel etablieren, dann erscheint es als eine wichtige Aufgabe, den vorhandenen Pool dieser Neuerungen zu erfassen, zu beschreiben und zu kennzeichnen, um auf diese Weise beizutragen zu einem besseren Verständnis der historisch-semantischen Prozesse der Innovation, der Selektion und der Verbreitung von sprachlichen Neuerungen (vgl. hierzu Fritz 1998).

In einer ganz anderen Perspektive auf den Wortschatz Goethes und seiner Zeit kann man die Frage stellen, welche der damals geläufigen Verwendungsweisen den heutigen Lesern nicht mehr geläufig sind, deren Beherrschung aber auch für das Verständnis vieler anderer Texte des 18. und 19. Jahrhunderts zentral ist. Fritz Tschirch hat in der ‚Zeitschrift für deutsche Wortforschung' (1960) einen sehr erhellenden Rundgang zu einigen typischen Fußangeln und Fallgruben bei der Lektüre von Texten um 1800 veröffentlicht, der Pflichtlektüre vor jeder Beschäftigung mit älteren Texten der Goethe-Zeit sein sollte. So sind etwa Verwendungsweisen von *gemein*, *geradezu* oder *umständlich* ‚ausführlich' heute nicht mehr gebräuchlich oder zumindest nicht mehr ohne weiteres in ihrer richtigen Lesart verständlich. Eine Kennzeichnung derjenigen Verwendungsweisen aus der Zeit um 1800, die heute nicht mehr ohne weiteres verständlich sind, würde es erlauben, aus dem Datenbestand ein ‚Idiotikon der Goethezeit' herauszuziehen, ein kontrastives Wörterbuch der Art, wie es Huguet zum französischen Sprachgebrauch des 17. Jahrhunderts angelegt hat, insofern dieser Sprachgebrauch vom modernen Sprachgebrauch abweicht. Ein solches Wörterbuch wäre wertvoll, weil es die wesentlichen sprachlichen Gefährdungen und ‚falschen Freunde' enthielte, die bei der Lektüre der Klassiker auf den Leser lauern.

Eine weitere Dimension der Wortschatzstrukturierung ist die Gliederung in funktionaler Hinsicht, also im Hinblick auf die Frage, welchen Beitrag einzelne lexikalische Einheiten leisten, um bestimmte kommunikative Aufgaben zu realisieren. Goethes Wortschatz ist reichhaltig in funktionaler Hinsicht: Er enthält ein breites Repertoire von lexikalischen Mitteln für einzelne kommunikative Aufgaben wie z.B. Zeitangabe, Ortsangabe, Bezugnahme auf Personen und Gegenstände, Kennzeichnung von Zusammenhängen zwischen Äußerungen, Querverweise auf frühere Textstellen, Charakterisierung von Eigenschaften beim Beschreiben usw. Zum Wortschatz der Charakterisierung von Eigenschaften gehören z.B. die schon genannten Farbbezeichnungen wie *gelbrot* oder Konsistenzbezeichnungen bzw. Konsistenzverwendungsweisen wie *fest*, *weich, quammig, quappig* oder *determinabel.* Zu den lexikalischen Mitteln, um Querverweise auf frühere oder spätere Textstellen zu geben, gehören z.B.: *angeregt*, *erwähnt, ersagt*, *gedacht* oder *vorhergehend* und vielfältige Kombinationen wie z.B. *ob gedacht*, *oft ersagt*, *vor angeführt*, *vor angeregt*, *eingangs erwähnt* oder *mehr erwähnt.*

Die zuletzt genannte Gruppe der Querverweise stellt einen Wortschatzsektor mit Wörtern und Verwendungsweisen dar, die besonders aufschlußreich sind für bestimmte Texttraditionen, hier z.B. der Verwaltungssprache. Querverweisausdrücke wie *ob gedacht* oder *mehr ermelt* waren zuerst im Deutsch der mittelalterlichen Urkunden des 13.und 14. Jahrhunderts üblich, sie wurden dann im 15., 16. und 17. Jahrhundert in zahlreiche andere Texttypen übernommen, ihr Gebrauch nimmt dann vor allem seit dem 18. Jahrhundert – mit dem Prinzip der natürlichen Schreibart – stetig ab. Das letzte und konservative Rückzugsgebiet für den variationsreichen Gebrauch dieser Ausdrücke um 1800 ist die Verwaltungssprache. Es ist ein bemerkenswerter textsortengeschichtlicher Befund, wenn Goethe diese Art von Ausdrücken auch in den naturwissenschaftlichen Schriften nicht selten und in bemerkenswerter Variation verwendet.

Quer zu all diesen lexikologischen Gesichtspunkten liegt der Gesichtspunkt der Herkunft eines Wortes bzw. einer Verwendungsweise. Die Fremdwörter, die Goethe gebraucht, sind zum einen interessant im Hinblick auf das Sprachstadium und seinen etablierten Fremdwortschatz, zum anderen aber auch im Hinblick auf die sich wandelnde Stellung Goethes zu den Fremdwörtern (z.B. *frugal*, *epitomieren*, *ebauchieren*).

Manche der hier in erster Näherung skizzierten Organisationsprinzipien des Goethe-Wortschatzes hängen mit einzelnen oder gar mehreren anderen Organisationsprinzipien zu-

sammen. Die Fremdwörter lassen sich z.B. vielfach bestimmten Sachbereichen zuordnen. Der Wortbestand einzelner Sachgebiete läßt sich weiter untergliedern im Hinblick auf verschiedene funktionale Leistungen. Innovative Bildungen lassen sich vielfach im Hinblick auf typische Wortbildungsmuster charakterisieren usw.

Eine Konzeption der Architektur eines Wortschatzes, seiner Organisationsprinzipien, seiner internen Untergliederung und der Zusammenhänge zwischen Wortschatzsektoren ist die Grundlage für alle elektronische Auszeichnung und Markierung. Alle Aspekte der Gliederung des Wortschatzes müssen aus den zu erfassenden Texten selbst erarbeitet werden. Es verbietet sich dabei, apriorische Systeme der Wortschatzstrukturierung wie die von Dornseiff oder Hallig/v. Wartburg zu verwenden. Dafür gibt es vor allem, aber nicht nur, drei Gründe. Erstens lassen diese Systeme keinen Raum für die Offenheit und die Zusammenhänge zwischen Strukturierungsgesichtspunkten: ein Ausdruck steht an einem Ort des Systems, obwohl er im Hinblick auf einen weiteren Strukturierungsgesichtspunkt auch an einem weiteren Ort des Systems stehen könnte oder sogar stehen müßte. Zweitens sehen diese Systeme in der Regel nicht vor, daß ein Ausdruck unterschiedliche Verwendungsweisen haben kann und mit jeder Verwendungsweise in eine andere Rubrik der Wortschatzstrukturierung gehören kann. Jede Form der Wortschatzstrukturierung, die Polysemie nicht oder ungenügend berücksichtigt, ist wirklichkeitsfremd, nicht nur im Hinblick auf Goethe. Und drittens ist in den Standardversionen dieser Systeme kaum Platz für die zeit- und kulturtypischen Besonderheiten eines Wortschatzes, die sich in der Regel erst aus der genauen Analyse des Aufbaus einzelner Wortschätze und ihrer Sektoren ergeben. Nimmt man die z.B. bei Hallig und v. Wartburg zwar formulierte, aber in der Praxis wenig befolgte Forderung ernst, das System den Erfordernissen des jeweiligen Wortschatzes anzupassen, dann endet man schnell bei der Aufgabe, ein neues System der Wortschatzarchitektur zu entwerfen. Und das ist auch kein bedauerlicher Schaden, sondern die eigentliche Aufgabe der Lexikologen.

Wir kommen nun zur Frage, welchen Nutzen und welche erweiterten Benutzungsperspektiven es bieten würde, die hier skizzierten Gesichtspunkte der Wortschatzstrukturierung in einem elektronischen Goethe-Wörterbuch als einer lexikologischen Datenbasis mit multiplem Zugriff zu kennzeichnen und explizit zu markieren.

4 Ein elektronisches Goethe-Wörterbuch: Einrichtung und Nutzungsperspektiven von lexikologischen Datenbasen

Die Hauptgesichtspunkte, die für ein elektronisches Goethe-Wörterbuch in Form einer lexikologischen Datenbasis mit mehrfachem Zugriff sprechen, ergeben sich zum einen aus der skizzierten Natur der Wortschatzarchitektur, zum anderen aber auch aus den möglichen Benutzerinteressen und möglichen weiterführenden Fragestellungen von Benutzern. Ein elektronisches Wörterbuch als lexikologische Datenbasis sollte zum einen der Komplexität des Gegenstandes gerecht werden, zum anderen für möglichst viele Benutzerinteressen und möglichst viele Arten von Fragestellungen vorbereitet sein. Als eine passende Umgebung bietet sich derzeit eine SGML-kodierte Datenbasis an. Bevor wir auf erweiterte Nutzerperspektiven eingehen, möchten wir einige Grundgedanken dieser Erfassungsumgebung darlegen.

4.1 Mehrfachnutzung von SGML-Daten: gedruckte und elektronische Produkte

Ein Grundgedanke der Textauszeichnung mit Hilfe von Markup-Systemen wie SGML (Standard Generalized Markup Language) ist es zum einen, den gezielten Zugriff auf einzelne Textelemente und darauf beruhende Auswertungen zu ermöglichen, zum anderen können aus ein und derselben Datenbasis unterschiedliche Endprodukte für unterschiedliche Nutzerinteressen hergestellt werden. Dabei sind zunächst zwei Hauptgruppen von Produkten zu unterscheiden: gedruckte Produkte und elektronische Produkte. In beiden Hauptgruppen lassen sich weiterhin Produkte mit unterschiedlicher Nutzungsausrichtung und Dokumentationstiefe erstellen.

Zunächst zu den gedruckten Fassungen. Ein sehr einfacher Artikel sieht in der gedruckten Fassung des Goethe-Wörterbuchs folgendermassen aus:

> **entschmeicheln** *durch Schmeicheleien entlocken* [*Chor:*] Wer die Schönste für sich begehrt,|..Schmeichelnd wohl gewann er sich|Was auf Erden das Höchste;|Aber ruhig besitzt er's nicht:|Schleicher listig e. sie ihm Faust II 9488 ***Syn** entlocken*

Die einzelnen Textelemente wie Lemma, (semantische) Leitbemerkung, Belegtext, Stellenangabe oder Synonym weisen jeweils eigene typographische Kennzeichen auf. In einer elektronischen SGML-Datenbasis werden den einzelnen Textelementen strukturelle Kennungen zugewiesen, mit deren Hilfe man (i) auf die unterschiedlichen Textteile zugreifen und (ii) ihre Eigenschaften manipulieren kann. Ein einfaches Beispiel ist folgender Artikel:

```
<art>
<le>entschmeicheln</le>
<sem>durch Schmeicheleien entlocken</sem>
<bel><komm>[Chor:]</komm><belt>Wer die Schönste für sich
begehrt,|&ausl;
Schmeichelnd wohl gewann er sich|Was auf Erden das Höchste;|Aber
ruhig besitzt er's nicht:|Schleicher listig e. sie
ihm</belt><belst>Faust II 9488</belst></bel>
<syn>entlocken</syn>
</art>
```

Die strukturellen Kennungen in den Spitzklammern haben hierbei folgende ‚Bedeutung':

`<art></art>`	Anfang und Ende eines Wortartikels;
`<le></le>`	Anfang und Ende des Lemmas;
`<bel></bel>`	Anfang und Ende eines Belegteils;
`<komm></komm>`	Anfang und Ende eines Kommentars im Beleg;
`<belt></belt>`	Anfang und Ende des Belegtexts;
`<belst></belst>`	Anfang und Ende einer Beleg-Stellenangabe;
`<syn></syn>`	Anfang und Ende der Angabe von Synonymen.

Diese Art der strukturellen Kennzeichnung von Textelementen ist noch unabhängig von typographischen Entscheidungen im Hinblick auf ein Endprodukt. Will man einen solchen Text drucken, dann sind die strukturellen Kennungen per Programm in typographische Steuercodes zu überführen. Die Herstellung von Satzdaten aus einer SGML-Datenbasis ist eine Aufgabe, die z.B. mit Programmen wie FrameMaker oder Tustep, aber auch in den Satzabteilungen graphischer Betriebe gelöst werden kann. Eine solche Umsetzung beinhaltet etwa, daß die Kennung für ein Lemma (`<le>freilich</le>`) per Programm umgesetzt

wird in typographische Befehle für den Anfang bzw. das Ende von Fettdruck oder daß eine eindeutige Kennung für Auslassungen im Belegtext (`&ausl;`) per Programm entsprechend den redaktionellen Vorgaben ausgetauscht wird (zwei Punkte, drei Punkte, drei Punkte mit runder Klammer, drei Punkte mit eckiger Klammer usw.).

Man kann derzeit leider nicht prognostizieren, daß solche Umsetzungen ganz ohne Anstrengungen und Reibungsverluste vonstatten gehen, aber die Tatsache, daß SGML sich im Bereich der Technischen Dokumentation durchsetzt, wo es oft um mehrere Tausend Seiten und viele Hunderttausend Mark oder Dollar geht, deutet darauf hin, daß die Anwendung dieses Standards sich langfristig ‚lohnt' und daß dieser Standard auch im graphischen Gewerbe zunehmende Verbreitung findet. Im Bereich der Geisteswissenschaften sind es vor allem die weitreichenden Koordinationsvorschläge und die Vorgaben der TEI (Text Encoding Initiative), die auf SGML zurückgreifen.

Für den wissenschaftlichen Einsatz ist die Nutzung einer lexikalischen SGML-Datenbasis, z.B. auf CD-ROM, besonders interessant. Vor allem die Erfahrungen mit dem Oxford English Dictionary auf CD-ROM zeigen, daß die Auswertungs- und Nutzungsmöglichkeiten der elektronischen Version einen qualitativen Sprung darstellen. Zwar werden die Bibliotheken und die Liebhaber des gedruckten Buches (zu denen wir uns selbst mit Entschiedenheit zählen) kaum auf die gedruckte Fassung verzichten wollen, aber für die Zwecke der Auswertung und der Bearbeitung wissenschaftlicher Fragestellungen ist die SGML-Fassung auf CD ein unschätzbares Hilfsmittel.

Aus einer SGML-Datenbasis lassen sich – ebenfalls per Programm, z.B. mit Tustep oder durch Verwendung geeigneter ‚Filter' – auch eine oder mehrere Internet-Fassungen des Datenbestandes erstellen. Die Möglichkeiten des differenzierten Zugriffs und der komplexen Abfrage sind derzeit bei Internet-Materialien zwar nicht besonders gut ausgebaut, dennoch sind vereinfachte Internetversionen ein gutes Mittel, um Material einem weiteren Benutzerkreis kostenfrei oder kostengünstig zugänglich zu machen und gleichzeitig Informationen über höherwertige ‚Vollversionen', die z.B. auf CD-ROM vertrieben werden, zu verbreiten.

4.2 Erweiterte Benutzungsperspektiven eines elektronischen Goethe-Wörterbuchs

Eine elektronische SGML-Datenbasis mit den Daten des Goethe-Wörterbuchs bietet zum einen differenzierte Nutzungsmöglichkeiten für Benutzer(innen) mit unterschiedlichen Interessen, zum anderen erlaubt sie auch den Einbau von Informationen, die eine Grundlage sind für erweiterte Forschungsmöglichkeiten. Differenzierte Nutzungsmöglichkeiten sind zunächst ‚nur' ein Aspekt der Bequemlichkeit und der Zugriffsökonomie, wogegen erweiterte Forschungsmöglichkeiten – z.B. durch Erschließung der Wortschatzarchitektur – einen echten Mehrwert darstellen.

4.2.1 Dokumentationstiefen und Informationstypen

Neben den erweiterten Auswertungs- und Abfragemöglichkeiten bietet eine SGML-Datenbasis den Vorteil, daß sich die Auswahl der Informationstypen und die Dokumentationstiefe von Benutzerseite variieren lassen:

– ein(e) Benutzer(in) kann wählen zwischen dem elektronischen Gegenstück der gedruckten Fassung mit Belegzitaten, einer Fassung ohne Belegzitate, aber mit vollständigem Set lexikographischer Definitionen, im GWb ‚Leitbemerkungen' genannt, und z.B. einer

Version, die nur das wesentliche semantische Gliederungsgerüst wiedergibt, indem man nicht nur auf alle Belege, sondern auch auf alle nicht mit einer eigenen Gliederungsmarke versehenen Leitbemerkungen und Unterleitbemerkungen verzichtet, oder, noch radikaler reduziert, indem man sich nur die Grunddifferenzierungen der höchsten Gliederungsebene(n) darstellen läßt;

– ein(e) Benutzer(in) kann bestimmte Typen von Einträgen einblenden oder ausblenden, z.B. Verweise auf das Grimmsche Wörterbuch, auf Synonyme, auf andere lexikalische Referenzinstanzen oder auf bestimmte Gewährsleute für Goethes Wortgebrauch (z.B. Herder, Schiller, Humboldt);

– die SGML-Version bietet die Möglichkeit, Eintragstypen zu ergänzen, die bislang im GWb nicht vorgesehen waren bzw. nicht konsequent gehandhabt wurden, z.B. Sachgebietskennzeichnungen wie ‚numismatisch', thematische Kategorisierungen wie ‚kosmogonisch', Hinweise zur Herkunft wie ‚Fremdwort' oder pragmatische Kennzeichnungen wie ‚Schimpfwort'. Ein feindifferenziertes Netz solcher Kennzeichnungen ist ein bedeutendes Hilfsmittel, um die spezifische Architektur des Goethe-Wortschatzes zu überschauen.

4.2.2 Vernetzung des elektronischen Goethe-Wörterbuchs mit anderen Dokumenten

Das GWb verweist auf andere einschlägige Informationsmittel bisher nur sporadisch, in problematischen Einzelfällen, etwa wenn für eine prima facie eigenwillige Lemmaansetzung oder sonderbar anmutende Bedeutungserklärung auf entsprechende Vorgänger in lexikalischen Standardwerken hingewiesen werden soll, aber auch wenn es gilt, sich von nur vordergründig plausiblen Erläuterungen, etwa in gängigen Kommentaren oder neuesten Editionen, ausdrücklich abzusetzen. Ein systematischer lexikalischer Abgleich mit anderen Wörterbüchern, die den Zeitraum um 1800 abdecken, ist zwar als Vorarbeit der Artikelautoren obligatorisch, nicht aber als Rubrik im gedruckten Artikel. Gleichwohl wäre es immerhin denkbar, in einem elektronischen Goethe-Wörterbuch Verweise auf solche Informationsmedien vorzusehen, z.B. auf die vorhandenen Kommentare, auf die allgemeinsprachlichen Wörterbücher, die den Zeitraum abdecken (Adelung, Campe, DWb), auf entsprechende fachsprachliche Wörterbücher (zur Anatomie, Bergbaukunde, Botanik, Chemie, Kameralistik, Mineralogie, Ökonomie und Önologie, bis hin zur Zoologie), schließlich auch auf den Sprachgebrauch von Zeitgenossen, von Klopstock über Ignaz von Born bis Schopenhauer. Vor einem solchen Hintergrund des Sprachüblichen wie auch des mehr oder minder bekannten Besonderen und Innovativen ließe sich deutlicher erkennen, was allgemeiner Sprachgebrauch der Zeit und was spezielle Sprachschöpfung, individuelle Eigenart Goethes – oder Herders oder Schillers – ist. Selbstverständlich ist eine solche Vernetzung kein Projekt, das man, vom GWb ausgehend, systematisch quer durch den gesamten Wörterbuchbestand vorantreiben könnte oder sollte. Aber es ließe sich in einer elektronischen Datenbasis immerhin die Möglichkeit vorsehen, Arbeitsergebnisse, die von den Bearbeiterinnen und Bearbeitern sowieso ermittelt werden, auch zu dokumentieren. In der gedruckten Fassung können diese Informationen problemlos unterdrückt werden, um beim eingeführten Aufbau und Erscheinungsbild zu bleiben, ebenso in den Standard-Ansichten der elektronischen Version. Aber: für diejenigen Benutzer, die Angaben dieser Art verwenden können und wollen, wären sie ebenso leicht zuzuschalten.

4.2.3 Aktualisierbarkeit und Revidierbarkeit

Eine SGML-Datenbasis bietet gegenüber der gedruckten Fassung eine weitere wichtige Eigenschaft: Sie ist in allen Wörterbuchstrecken laufend aktualisierbar und den Benutzern leichter und schneller zugänglich zu machen. Auch hierzu ein Beispiel: Sollte die Redaktionskonferenz des GWb etwa aufgrund der Hinweise eines Rezensenten zu der Auffassung gelangen, daß unter den synonymischen Verweisen zur Verwendungsweise 2b des Verbs *bewegen* auch das Verb *rühren* erscheinen sollte, dann könnte man einen solchen Eintrag in einer Datenbasis leicht ergänzen. Ein solcher Eintrag käme den Benutzern von elektronisch aktualisierten Versionen schneller zugute als den Beziehern gedruckter Nachtragsbände.

Die Frage der beschleunigten Aktualisierung ist vor allem im Bereich der synonymischen Vernetzung interessant, die besonders schwer zu überblicken ist und bei der sich oft erst im Lauf der weiteren Bearbeitung von Wörterbuchstrecken neue Einsichten ergeben. Denkbar wäre weiterhin, daß – zumindest in den Grenzen der verfügbaren Arbeitskapazität – auch ältere Wörterbuchstrecken an neue redaktionelle Richtlinien angepaßt und revidiert werden könnten. Die zum Teil fehlenden oder wenig expliziten Bedeutungsbeschreibungen des ersten Bandes zum Beispiel könnten auf diese Weise im Lauf der Zeit und nach Maßgabe der vorhandenen Arbeitskapazität den aktuellen Richtlinien angepaßt werden. Allerdings: angesichts knapper Mittel und reduzierter Mitarbeiterzahlen in den Redaktionen des GWb sind dies derzeit bloße Wunschträume.

4.2.4 Ein elektronisches Goethe-Wörterbuch als kostenlose Internet-Version

Zu den elektronischen Produkten, die aus einer SGML-Datenbasis per Programm erstellbar sind, gehören auch Internet-Versionen von Wörterbüchern. Wir plädieren hier dafür, solche Versionen im Internet kostenlos anzubieten, vor allem wenn ein Wörterbuch-Unternehmen mit beträchtlichen öffentlichen Mitteln gefördert wird. Eine Internet-Version des Goethe-Wörterbuchs ist zunächst ein Zugeständnis an diejenigen Steuerzahler, die keinen Zugang haben zu Bibliotheksexemplaren des GWb oder zu GWb-Exemplaren in Privatbesitz. Um nur zwei Beispiele zu nennen: Journalisten oder Politiker, die dauernd oder zeitweise keinen Zugriff auf eine größere öffentliche Bibliothek haben, die aber auf lexikalischem Weg nach einem passenden Goethe-Zitat suchen, könnten über das Internet in einer entsprechend aufbereiteten, vereinfachten Datenbasis recherchieren. Oder: Schüler, die keine Schulbibliothek mit Goethe-Wörterbuch haben – so etwas soll es tatsächlich noch geben! –, könnten in einem entsprechenden Kurs lernen, (1) wie man das Internet für Zwecke der Recherche benutzt, (2) wie und wofür man das Goethe-Wörterbuch nutzen kann (z.B. die bei der Lektüre älterer Texte durch Bedeutungswandel allenthalben lauernden Mißverständnisse vermeiden lernen; Zitate finden). – Wir nennen den lexikalischen Zugriff auf „passende Zitate" hier deshalb, weil es eine interessante und zunächst auch erschütternde Erfahrung für den englischen Computerphilologen Roy Wisbey war, daß seine monumentale Shakespeare-Konkordanz vor allem für die Suche nach Shakespeare-Zitaten verwendet wurde. Wir meinen aber, man muß realistischerweise auch mit diesem Benutzerinteresse rechnen.

Die angedeuteten unentgeltlichen Zugriffsmöglichkeiten im Internet kollidieren prima facie mit dem Verlagscopyright und den damit verbundenen finanziellen Interessen. Auf der anderen Seite, bei genauerer Überlegung, kann ein „GWb light" im Internet als attraktive Werbung für die Druckfassung oder für eine elektronische Vollversion mit erweiterten Nutzungsmöglichkeiten auf CD-ROM fungieren.

Im Vordergrund sollten allerdings die kulturpolitischen Aspekte einer Internet-Version und die Verpflichtung gegenüber ‚dem Steuerzahler' stehen. Das Unternehmen ‚Goethe-Wörterbuch' hat die öffentliche Hand in den vergangenen Jahrzehnten viele Millionen Mark gekostet. Es ist einer der Widersprüche unserer Zeit, daß viele Leistungen, die mit öffentlichen Geldern erarbeitet wurden, heute ausschließlich den Verlagen ‚gehören' und daß über wichtige kultur- und forschungspolitische Fragen letztlich in den Marketingabteilungen der Verlage entschieden wird. Es ist an der Zeit, daß Verlage und die Repräsentanten der öffentlichen Hand (z.B. die Akademien) neue Formen vereinbaren, wie solche Leistungen einem erweiterten Kreis von Nutzern zugänglich gemacht werden können.

4.2.5 Architektur und onomasiologische Erschließung des Goethe-Wortschatzes

Der eigentliche Mehrwert einer lexikalischen Datenbasis besteht in der Möglichkeit, die vielfältigen Aspekte der lexikalischen Organisation eines Wortschatzes zu erfassen und in einen Zusammenhang zu bringen. Alphabetische Wörterbücher, so lautet ein auch nach Hermann Paul immer wieder vorgebrachter und völlig berechtigter Kritikpunkt, tragen nichts zur Kenntnis der Architektur der dokumentierten Wortschätze bei. Daß sicherer Zugriff durch alphabetische Dokumentation und Erschließung der Architektur nicht unverträglich sind, zeigt z.B. das „Wörterbuch zu Goethes West-östlichem Divan" von Christa Dill: hier ist die alphabetische Abfolge der Wortartikel ergänzt durch ausführliche Darstellungen zum Divan-Wortschatz und zur Divan-Bildwelt sowie durch eine umfassende tabellarische Darbietung der „Wortfelder und Sinngruppen" in Goethes ‚Divan'. Wie läßt sich diese Wortschatz-Erschließung von einem einzelnen Werk übertragen auf die Weite der gesammten Goetheschen Sprachwelt, die, wie erwähnt, über 90 000 dokumentierte Wörter mit einem Vielfachen an Verwendungsweisen umfaßt? Die elektronische Dokumentation von Wortschätzen erlaubt es zunächst, mehrere lexikologische Organisationsprinzipien nebeneinander zu verwenden und von vornherein verschiedene Zugriffs- und Abfragemodi vorzusehen. Voraussetzung dafür ist jedoch eine durchgängige Vorstrukturierung des Wortschatzes, z.B. durch eine konsequente Merkmalsmarkierung – „What you mark is what you get!"

Zu den Organisationsprinzipien, die alternativ bzw. komplementär zur alphabetischen Anordnung Berücksichtigung verdienen, gehören:

- Textsortenzuweisungen (z.B. Dichtung, Privatbrief, amtliches Schreiben);
- funktional-pragmatische Kategorien (etwa vom bedeutungsentblößten Funktionsverb oder rein verstärkenden Adjektiv über wertende Verwendungen bis zum derben Schimpfwort);
- semantische Kategorien und Gebrauchsbereiche (z.B. Konsistenzbezeichnungen; Farbwörter im Zusammenhang der Farbenlehre; sprechhandlungsbezeichnende Ausdrücke im Bereich der Rechtssprache);
- Sachgebiete und Themenbereiche (z.B. Bergbau, Botanik, Ästhetik);
- Neologismen und Hereditäten (genuin Goethesche Wortschöpfung oder nachweisbare Übernahme von anderen Autoren, geistigen Bewegungen, z.B. Pietismus, Kantianismus, Freimaurerei, aber auch aus Dialekten);
- zeitgebundene Verwendung (z.B. nur beim jungen Goethe, nur nach 1815);
- sprachliche Mittel, die in besonders engem Zusammenhang mit Goethes Weltsicht stehen (z.B. Verben der sinnlichen Wahrnehmung und kognitiven Anschauung).

Eine konsequente Kennzeichnung des Goethe-Wortschatzes gemäß solchen Gesichtspunkten, mit den entsprechenden Kennungen in der elektronischen Aufbereitung, würde auch das Herausfiltern von Teilwortschätzen und Kreuzklassifikationen unter Fragestellungen wie den folgenden ermöglichen:

- Welche sprachlichen Mittel verwendet Goethe bei der Behandlung geologischer Gegenstände? Woher stammen sie?
- In welchen Texttypen/ In welcher Zeit kommen Querverweisausdrücke wie *obgemeldet* oder *vorgenannt* bevorzugt vor?
- Welche Schimpfwörter verwendet Goethe? Herrschen dabei Dialektausdrücke vor?
- Welche Ausdrücke bzw. Verwendungsweisen aus dem Sachgebiet der Kunst sind aus der ‚Italienischen Reise' belegt?

4.3 Probleme und Hindernisse

Auf dem Weg zu einer SGML-Datenbasis des elektronischen Goethe-Wörterbuchs begegnen eine Reihe von offenen Fragen und Problemen. Sie liegen auf ganz unterschiedlichen Ebenen: auf der Ebene der Texterfassung und der Datenkonvertierung, auf der juristisch-wirtschaftlichen Ebene und auf der Ebene der bisher befolgten Prinzipien der Textgestaltung.

Ein erstes Problem ist die Frage der Texterfassung und der Datenkonvertierung. Das Goethe-Wörtebuch, das gerade seinen dritten Band vorlegen konnte, ist Mitte der Achtziger Jahre, also erst während der Erarbeitung des zweiten Bandes, zur elektronischen Datenerfassung übergegangen, und zwar mithilfe gängiger Schreibprogramme (Word, Winword), die vom graphischen Betrieb in Satzdaten konvertiert werden. Will man diese Daten verwerten, dann stellen sich u. a. folgende Fragen: Wie aufwendig ist die Umwandlung in SGML-Daten? Wer macht das? Wer bezahlt das? Und: Was tun mit den überhaupt nicht EDV-erfaßten Lieferungen von Band I und II? Das Scannen würde nur ASCII-Daten ohne bzw. ohne zureichende Strukturinformationen liefern, also einheitlich den reinen Wortlaut ohne zureichende Differenzierung zwischen den einzelnen lexikographischen Textelementen.

Ein weiteres Problem stellen die juristisch-wirtschaftlichen Aspekte eines elektronischen Goethe-Wörterbuches dar. Die Rechte am gedruckten GWb liegen nicht bei den Akademien, sondern ausschließlich beim Verlag, der CD- und Internetversionen, nach eigenem Bekunden, nur dann zustimmt, wenn sie ihn keinen Pfennig kosten und wenn der Absatz des gedruckten Werkes dadurch nicht beeinträchtig wird. Die wirtschaftlichen Erwägungen hängen natürlich aufs engste mit den Fragen der Einrichtung einer Datenbasis und dem damit verbundenen Aufwand zusammen.

Ein dritter Problembereich ergibt sich aus den bisher befolgten Prinzipien der Textgestaltung. Von Anfang wird das GWb begleitet von der Losung, nicht schon *im* GWb das leisten zu wollen, was man *mit* dem GWb leisten können soll. Von daher erhebt sich die Frage, inwiefern die aus den erweiterten Benutzungsperspektiven resultierenden Postulate, wie sie soeben angerissen wurden, nicht doch schon ein Stück weit Nutzung, ergebnisorientierte Aufbereitung, in die semantische Arbeit des GWb hineintragen. Insbesondere eine weitergehende Standardisierung der Kennzeichnungssystematik, womit Ansatzpunkte für eine multiple Organisation des Wortschatzes geschaffen würden, erscheint als Eingriff in die primäre semantische Aufbereitung des Belegmaterials in den Redaktionen des GWb. Die Grenze zu einem ‚neuen GWb' wäre rasch überschritten. Wäre das schlimm, vor allem

angesichts der vielfältigen Möglichkeiten für die Erforschung der Goethezeitsprache, die wir oben umrissen haben? Schwer zu sagen. Ganz abgesehen davon, daß es stets mißlich ist, an einer bewährten Konzeption herumzubasteln, sind auch substantielle Verluste nicht auszuschließen: Das zupackende Etikettieren, die festlegende Verbuchung unter einfachen Kategorien und Rubriken, wie sie die Prästrukturierung des Materials für spätere Selektionsprozesse per Mausklick schließlich erfordert, ist im GWb bisher bewußt unterblieben. Standardisierte Kennzeichnungen wie ‚bergmannssprachlich', ‚rechtssprachlich', ‚metaphorisch', ‚polemisch' kommen zwar vor, werden jedoch mit Zurückhaltung eingesetzt, denn das Besondere von Goethes Sprache besteht nicht zuletzt in Grenzüberschreitungen und Ambivalenzen.

Hinzu kommt im Bereich der Textgestaltung, daß vor allem in den älteren Bänden ein sehr enger Zusammenhang zwischen unterschiedlichen Textelementen besteht, der mit dem Prinzip der elektronischen Modularisierung selbständiger Textelemente nur schwer vereinbar ist. So führt etwa die enge Korrelation zwischen Leitbemerkungen und Belegen dazu, daß man in den älteren Lieferungen die Belege nicht ohne weiteres ausblenden kann, wenn man die Leitbemerkungen verstehen will. In anderen Fällen gehören etwa eine Leitbemerkung und die entsprechenden Unterleitbemerkungen so eng zusammen, daß man nicht ohne weiteres einen der beiden Bestandteile ausblenden kann.

4.4 Prototypen

Ein erster Arbeitsschritt beim Versuch, ein Teilstück eines elektronischen GWb herzustellen, bestand darin, die Verwendbarkeit der vorhandenen Winword-Daten zu überprüfen. Die Ergebnisse deuten darauf hin, daß mit Hilfe der verwendeten typographischen Auszeichnungen wie ‚doppelt unterstrichen' und mit Hilfe textsyntaktischer Kombinationsregeln wie ‚Fettdruck unmittelbar nach Neuer Absatz' immerhin Textelemente wie Leitbemerkungen, Belege oder Lemmata strukturell markiert werden können. Hinderlich dabei ist allerdings, daß typographische Auszeichnungen nicht eindeutig und anhand von textlogischen Kriterien verwendet wurden, sondern anhand des Erscheinungsbildes in den späteren gedruckten Artikeln (z.B. Fettdruck nicht nur beim Lemma, sondern auch beim Kürzel **Syn**, mit dem die Synonyme eingeleitet werden). Man muß also versuchen, den mehrdeutigen Gebrauch von typographischen Auszeichnungen aufgrund der Abfolge und der Stellung von Textteilen im Artikel aufzulösen. Wir haben die Winword-Dateien hierfür zunächst in WordPerfect-Dateien umgewandelt, diese dann mit Hilfe des Tustep-Hilfsprogramms *konvert* in ASCII-Dateien, in denen typographische Steuercodes explizit gemacht sind und per Programm manipuliert werden können (z.B. Zeichenketten wie *#{fett_ein}*). Diese typographischen Markierungen wurden dann mit den Textmanipulationsmöglichkeiten von Tustep in SGML-Kennungen umgewandelt und, soweit möglich, textstrukturell aufgelöst. Manuelle Nachbearbeitung und Überprüfung der Ergebnisse wird wohl nicht zu umgehen sein, schon weil die Prinzipien der Texterfassung im Lauf der Bearbeitung des GWb nicht ganz einheitlich waren.

Für den auf diese Weise gewonnenen SGML-Text und seine Textstrukturmuster, die in einer bisher sehr elementaren Document Type Definition (DTD) erfaßt sind, haben wir in Panorama Pro 1.5 versuchsweise unterschiedliche ‚Ansichten' eingerichtet: z.B. eine Version, die der gedruckten Vollversion entspricht, und eine Version, bei der nur das extrahierte Skelett der semantischen Leitbemerkungen sichtbar ist und bei der nach Art des kleinen LEXER die Belege ausgeblendet sind.[3] Zu den nächsten Arbeitsschritten gehören vor

allem die Konzeption unterschiedlicher Internet-Darstellungen aufgrund der SGML-Datenbasis und die detaillierte Konzeption eines Markierungssystems, mit dem die spezifische Architektur des Goethe-Wortschatzes erschlossen werden kann.

5 Zusammenfassung und Ausblick

Wir sind ausgegangen vom Grundgedanken, daß ein elektronisches Wörterbuch als komplex strukturierte lexikologische Datenbasis erweitere Nutzungsperspektiven in zweierlei Hinsicht bietet. Zum einen kann eine solche Datenbasis ein Abbild der komplexen Zusammenhänge im Wortschatz sein, zum anderen kann sie ein Informationssystem sein, das bestimmte Benutzerinteressen besser und flexibler bedient als eine gedruckte Version. Am Beispiel des Goethe-Wortschatzes und des gedruckten Goethe-Wörterbuchs haben wir mit einigen wenigen Überlegungen den Grundgedanken eines mehrschichtig organisierten Wortschatzes erläutert und anzudeuten versucht, wie sich eine solche komplexe Organisation in elektronischer Form erfassen läßt und welche Arten des Erkenntniszugewinns sich daraus ergeben.

Ein wichtiger kultur- und forschungspolitischer Vorschlag war es, aufgrund der Stammdaten des elektronischen Wörterbuchs eine kostenlos angebotene Internet-Version herzustellen, um den Nutzerkreis des Goethe-Wörterbuchs zu erweitern. Vergleichbare Forderungen sind an alle Wörterbuch-Unternehmen zu stellen, die im wesentlichen mit öffentlichen Mitteln gefördert werden. Es steht zu hoffen, daß in absehbarer Zukunft nicht nur die technisch-konzeptionellen Aufgaben und Schwierigkeiten, sondern auch Fragen der Verwertungsrechte in ersprießlicher Weise gelöst werden.

Die Erschließung der Goetheschen Wortschatzarchitektur wird man sich jedenfalls als einen langen Weg der zunehmenden Verfeinerung vorstellen dürfen: „Wenn man denckt fertig zu seyn, gehts erst recht an".[4]

6 Literatur

Coombs, J.H./Renear, A.H./DeRose, S.J. (1993): Markup systems and the future of scholarly text processing. In: Landow, G.P./Delany, P. (eds.): The digital word. Text-based computing in the humanities. – Cambridge, Mass./ London, 85–118 (Zuerst: Communications of the ACM 30, 1987, 933–947).

Deutsches Fremdwörterbuch (1988). Begonnen von H. Schulz, fortgeführt von O. Basler, weitergeführt im Institut für deutsche Sprache. Siebenter Band: Quellenverzeichnis, Wortregister, Nachwort. Hg. von A. Kirkness. – Berlin/ New York: de Gruyter.

Dill, Ch. (1987): Wörterbuch zu Goethes West-östlichem Divan. [Mit einer Einführung zu Wortschatz und Bildwelt und einem Anhang zu Wortfeldern und Sinngruppen.] – Tübingen: Niemeyer.

Fritz, G. (1998): Historische Semantik. – Stuttgart/Weimar: Metzler (= Sammlung Metzler).

---

[3] Eine ausführlichere Darstellung der technischen Realisierung dieser Konzeption eines elektronischen Goethe-Wörterbuchs soll in der Zeitschrift ‚Sprache und Datenverarbeitung' erscheinen.

[4] Goethe, WA, Abt. Briefe I 113,15.

Goebel, U./Lemberg, I./Reichmann, O. (1995): Versteckte lexikographische Information. Möglichkeiten ihrer Erschließung dargestellt am Bespiel des Frühneuhochdeutschen Wörterbuchs. – Tübingen: Niemeyer (= Lexicographica, Series Maior 65).

Goldfarb, Ch.F. (1990): The SGML handbook. – Oxford.

Ide, N./Véronis, J. (eds.)(1995): Text encoding initiative. Background and context. – Dordrecht.

Kluge, F. (1899): Etymologisches Wörterbuch der deutschen Sprache. Sechste verbesserte und vermehrte Auflage. – Straßburg: Trübner.

– und Götze, A. (1934): Etymologisches Wörterbuch der deutschen Sprache. Elfte Auflage. – Berlin: de Gruyter.

Powitz, G. (1988): Das „Catholicon" – Umrisse der handschriftlichen Überlieferung. In: Litterae medii aevi. Festschrift für Johanne Autenrieth zu ihrem 65. Geburtstag. Hg. von M. Borgolte und H. Spilling. – Sigmaringen: Thorbeke, 209–223.

Schadewaldt, W. (1946): Das Goethe-Wörterbuch. – Berlin 1946. Wieder in: Goethe, Jahrbuch der Goethe-Gesellschaft 11, 1949, 293–305.

Tschirch, F. (1960): Bedeutungswandel im Deutsch des 19. Jahrhunderts. Zugleich ein Beitrag zum sprachlichen Verständnis unserer Klassiker. In: Zeitschrift für deutsche Wortforschung 16, N.F. 2, 7–24.

Welter, R. (1998): Zwischen Bedeutung und Benutzer. Zur Mikrostruktur des Goethe-Wörterbuchs. In: Grosse, R. (Hg.): Bedeutungserfassung und Bedeutungsbeschreibung in historischen und dialektologischen Wörterbüchern. – Stuttgart/Leipzig, 145–149 (= Abhandlungen der Sächsischen Akademie der Wissenschaften zu Leipzig. Philologische-historische Klasse, 75/1).

Thomas Gloning, Marburg
Rüdiger Welter, Tübingen

Thomas Burch, Johannes Fournier

Zur Anwendung der TEI-Richtlinien bei der Retrodigitalisierung mittelhochdeutscher Wörterbücher

1 Retrodigitalisierung als Aufgabe

Die wichtigsten derzeit vorhandenen Wörterbücher zur mittelhochdeutschen Sprache sind noch im vorigen Jahrhundert entstanden und müssen – dieser Zwang ergibt sich nicht nur aus dem enormen Zuwachs an Texten, die seit dem Ende des 19. Jahrhunderts durch neue Editionen erschlossen worden sind – dringend durch ein neues, großes mittelhochdeutsches Wörterbuch ersetzt werden. Dieser Mißstand der deutschen Lexikographie ist häufig genug beklagt worden, und seit fünf Jahren beschäftigen sich zwei Arbeitsstellen in Göttingen und Trier mit dem Aufbau eines elektronischen Text- und Belegarchivs, auf dessen Grundlage ein neues Handwörterbuch zum Mittelhochdeutschen ausgearbeitet werden soll.[1] Bis zum Abschluß dieses auf vier Bände angelegten Werkes in etwa 20 Jahren werden die Altgermanistik und die mit mittelhochdeutschen Texten befaßten Disziplinen jedoch auf die älteren mittelhochdeutschen Wörterbücher angewiesen bleiben.

Diese Wörterbücher ihrerseits sind so eng aufeinander bezogen, daß im Grunde genommen kein Wörterbuch ohne das andere benutzt werden kann. Das ergibt sich aus der Geschichte dieser Nachschlagewerke, die hier kurz vorgestellt werden muß. Das älteste und nach wie vor wichtigste mittelhochdeutsche Wörterbuch stammt von Georg Friedrich Benecke, Wilhelm Müller und Friedrich Zarncke (BMZ). Es erschien in den Jahren 1854 bis 1866 und umfaßt vier Bände mit ca. 40.000 Stichwörtern, es zeichnet sich aus durch die differenzierte Systematik der Bedeutungsangaben und einen großen Reichtum an Belegen. Doch ist die Benutzung des BMZ nicht einfach. Der Wortschatz ist nicht nach dem Al-

[1] Zur Notwendigkeit eines neuen mittelhochdeutschen Wörterbuchs vgl. die Beiträge in Bachofer (1988) und die Vorträge von Gärtner, Grubmüller und Nellmann auf dem VIII. Internationalen Germanisten-Kongreß in Tokyo (Begegnung mit dem „Fremden" 1991). Über die Tätigkeit der Arbeitsstellen informieren z.B. Plate/Recker (im Druck); s. ferner Gärtner/Grubmüller (im Druck).

phabet, sondern nach Wortstämmen angeordnet. Ableitungen und Zusammensetzungen sind jeweils ihrem Grundwort zugeordnet; der Benutzer findet ein Lemma somit nur über das Grundwort und den diesem übergeordneten Wortstamm des Wortes,[2] was das Nachschlagen für philologisch weniger geschulte und mit Wortbildungsregeln nicht vertraute Benutzer erschwert.

Schon kurze Zeit nach der Vollendung dieses Wörterbuchs regte sich der Wunsch nach einem rein alphabetischen Index, der das Nachschlagen im BMZ erleichtern sollte. Diesen Index arbeitete Matthias Lexer von 1872 bis 1878 aus, beschränkte sich aber nicht auf die Indexierung des nur wenig älteren Wörterbuchs, sondern ergänzte zugleich das im BMZ gesammelte Material um ca. 34.000 neue Stichwörter, aber auch um weitere Belege zu schon im BMZ verbuchten Lemmata. Zugleich sollte Lexers Werk ein Handwörterbuch für das Mittelhochdeutsche sein, also einen überschaubaren Umfang behalten. Aus diesem Grund entschied sich Lexer dafür, die bereits im BMZ aufgeführten Belege in seinen Artikeln nicht erneut zu zitieren, sondern allein durch Siglen auf diese Belege zu verweisen. Insofern müssen die bei Lexer gedruckten Artikel immer ergänzt werden um die entsprechenden, nur im BMZ vermerkten Informationen. Darüber hinaus verfaßte Lexer zu seinem eigenen Handwörterbuch Nachträge, die vor allem die Artikel der Strecken A bis M, aber auch die der restlichen Alphabetstrecken betreffen.[3] In diesen Nachträgen bucht Lexer zum einen gänzlich neue Wörter, trägt aber auch neue Formen, Bedeutungen und Zitate zu bereits im Hauptteil behandelten Stichwörtern nach.

Zahlreiche Texte wurden erst nach dem Abschluß des Handwörterbuchs durch Editionen erschlossen. Viele dieser Editionen sind mit Glossaren ausgestattet, die einen ersten Zugang zum Wortschatz der Quellen gewähren. Diese Glossare wiederum wurden zwischen 1986 und 1992 von Trierer Altgermanisten im ‚Findebuch zum mittelhochdeutschen Wortschatz' kompiliert, dessen Lemmaansätze eng auf diejenigen Lexers bezogen sind. Nun ist das FINDEBUCH kein eigentliches Wörterbuch, sondern ein Wegweiser zu den Wortverzeichnissen und Glossaren im Anhang von Ausgaben. Im FINDEBUCH finden sich keine Belegzitate und in aller Regel auch keine Bedeutungsbeschreibungen. Doch die Verbreitung und Bezeugung mittelhochdeutscher Wörter kann mit seiner Hilfe zuverlässiger beurteilt werden als allein anhand der im BMZ und im LEXER gebuchten Belege.

BMZ, Lexers Handwörterbuch, seine Nachträge zum Handwörterbuch und das eng auf LEXER bezogene FINDEBUCH, diese vier Wörterbücher müssen also als regelrechter Wörterbuchverbund angesehen werden, dessen stark ausgeprägte Verweisstruktur sich in geradezu idealtypischer Weise für die Abbildung in eine Hypertextstruktur eignet.[4] Mit der Idee zum Aufbau eines elektronischen Wörterbuchverbundes ging zugleich die Vorstellung einher, diesen für Recherchen und näher spezifizierte (lexikographische) Abfragen aufzubereiten, wie sie aus zahlreichen Wörterbüchern in elektronischer Form inzwischen bekannt sind.[5]

---

[2] Lemmata wie **entvarn, unervarn, verge, vart, hôchvart, vertec** usw. finden sich z.B. allesamt in einem Artikel mit ihren jeweiligen Grundwörtern unter dem in Majuskeln gesetzten und den Wortstamm repräsentierenden Hauptlemma *VARN* bzw. in der Lemmaform 1. Sg. Präs. *VAR* und den Stammformen *VUOR, GEVARN.* Bei den Verben ist als Lemma immer die Form der 1. Sg. Präs. angesetzt, z.B. (*ich*) *bir* zu *bërn, biuge* zu *biegen, schiche* zu *geschëhen, gihe* zu *jëhen, schol* zu *suln,* teils werden gar rekonstruierte Formen wie **dinke* zu *denken, *kinne* zu *kunnen, *liube* zu *lieben* angesetzt.

[3] Über verschiedene Verfahren Lexers, Nachträge in das Handwörterbuch einzuarbeiten vgl. Gärtner (1993, 121–124). Dort finden sich auch genauere Angaben zur Quantität des Nachgetragenen.

[4] Vgl. Storrer (1998, 115f.).

Diese beiden Ziele und darüber hinaus eine langfristige und plattformunabhängige Datenhaltung lassen sich am ehesten mit Hilfe der *Standard Generalized Markup Language* (SGML) realisieren. Im folgenden soll deshalb zunächst dargelegt werden, welche Vorteile mit dem Einsatz von SGML verbunden sind (vgl. Abschnitt 1). Anschließend soll erörtert werden, aus welchen Gründen die *Document Type Definitions* (DTD) der *Text Encoding Initiative* (TEI) als SGML-Applikation zur Auszeichnung der mittelhochdeutschen Wörterbücher herangezogen werden (Abschnitt 2). Der Hauptteil dieses Beitrags behandelt ausgewählte Vorzüge, Schwierigkeiten und Nachteile, die sich durch die Orientierung an den TEI-Richtlinien ergeben, und zwar sowohl was die TEI-konforme Auszeichnung der relevanten Wörterbuchelemente (Abschnitt 3) als auch die Publikation und Auswertung des elektronischen Wörterbuchverbundes (Abschnitt 4) angeht. Ein Resümee über die bisherigen Erfahrungen mit dem Einsatz der TEI-Richtlinien für die Retrodigitalisierung der mittelhochdeutschen Wörterbücher schließt diesen Beitrag ab.[6]

2 Die TEI-DTD für Wörterbücher als favorisierte Lösung

2.1 SGML als standardisierte Beschreibungsmethode

SGML ist eine Markierungssprache, mit deren Hilfe genau festgelegt werden kann, welche Arten von Markierungen erlaubt sind, welche Markierungen unbedingt angegeben werden müssen und wie sich die Markierungen vom eigentlichen Text unterscheiden. Dabei ist SGML keine Erfindung der letzten Jahre. Es basiert auf einem Vorläufer namens GML, der im Jahre 1969 von Charles Goldfarb innerhalb eines IBM Forschungsprojektes entwickelt wurde. Anstelle einer einfachen Auszeichnung durch Tags führte GML erstmals das Konzept formal definierter Dokumenttypen ein, die explizit geschachtelte Strukturen erlauben. Auf der Basis von GML wurde, veranlaßt durch das *American National Standards Institute* (ANSI), die Beschreibungssprache SGML entwickelt und 1986 von der *International Organisation for Standardisation* (ISO) als Standard veröffentlicht.

Die Grundidee von SGML besteht in der klaren Trennung von Inhalt (zu vermittelnde Information), Struktur (Abfolge der Information) und Layout (Darstellung der Information in verschiedenen Medien). Ein SGML-Dokument enthält Inhalt und Struktur, nicht aber dessen Layout. Durch diese Trennung wird erreicht, daß ein und dasselbe Dokument in höchst unterschiedlichen Formen präsentiert werden kann.[7]

---

[5] Die Recherchemöglichkeiten des *Oxford English Dictionary* (OED), des ‚Klassikers' unter den elektronischen Wörterbüchern, schildert Jucker (1994); s. ferner Storrer (1998) über Hypermediawörterbücher.

[6] Konzeption, Vorgehen und technische Umsetzung des Wörterbuchverbundes führen Burch/Fournier/Gärtner (1998) und Fournier (1998) genauer aus. Über Perspektiven künftiger Nutzung reflektiert Fournier (2000). Eine erste Version des Verbundes kann unter der Internet-Adresse http://gaer27.uni-trier.de/MWV-online/MWV-online.html eingesehen werden.

[7] *A Gentle Introduction to SGML* ist unter der Internet-Adresse http://www.uic.edu/orgs/tei/sgml/teip3sg/SG.htm zugänglich. Alschuler (1995) informiert ausführlich über Produkte und Werkzeuge und erörtert außerdem, wann SGML-Anwendungen relevant sein können. Deutschsprachige Einführungen in SGML sind z.B. Rieger (1995) und Szillat (1995), die u.a. mit Hilfe von Übungsaufgaben Verständnis für eine sachorientierte Auszeichnung wecken wollen.

Durch die Standardisierung von SMGL ist gewährleistet, daß Dokumente mit jeder Software verarbeitet werden können, die diese Norm unterstützt. Daraus folgt die Unabhängigkeit von Hard- und Softwareherstellern. Aus dieser Unabhängigkeit folgt weiter die Langlebigkeit von in SGML kodierten Dokumenten, da keine aufwendigen Konvertierungen beim Wechsel von Soft- oder Hardware durchgeführt werden müssen.

Ein typisches SGML-Dokument besteht dabei aus drei Abschnitten: der SGML-Deklaration, der Dokumenttyp Definition und einer Dokumentinstanz. Die SGML-Deklaration ist ein formaler Teil eines jeden SGML-Dokumentes, in welchem festgelegt wird, welche Zeichen und Trennsymbole benutzt werden dürfen. Normalerweise ist diese Deklaration allen Dokumenten einer bestimmten SGML-Anwendung gemeinsam. Sie kann explizit im Dokument kodiert sein, wird aber in der Regel durch eine Standardspezifikation gegeben. In ihr sind insbesondere die Zeichen definiert, die die Markierungen vom eigentlichen Text trennen, üblicherweise spitze Klammern (<, >) bzw. Schrägstrich für die Endemarkierungen.

Während die SGML-Deklaration selten von der vorgegebenen abweicht, bildet der zweite Abschnitt das eigentliche Kernstück einer SGML-Anwendung. Hier wird der Dokumenttyp definiert, d.h. es wird eine Menge von Regeln festgelegt, durch die eine Klasse von Texten charakterisiert ist. Die sogenannte DTD definiert die Struktur eines Dokumentes, sie wird für verschiedene Dokumentklassen wie beispielsweise Briefe, technische Dokumentation, Gesetzestexte oder auch Wörterbücher eigens spezifiziert. Ihre Beschreibung erfolgt selbst wieder in SGML, sie wird in der Regel außerhalb des eigentlichen Dokumentes abgelegt.

Die dritte Komponente bildet die Dokumentinstanz, d.h. der eigentliche Text. Dieser enthält die Daten, die durch Markierungen gemäß der vorgegebenen DTD ausgezeichnet sind, sowie einen Verweis auf die zugrundeliegende DTD, falls diese nicht explizit ins Dokument eingefügt wurde. Man spricht hier von einer Instanz eines Dokumentes, weil es sich um eine konkrete Anwendung der in einer DTD spezifizierten Regeln handelt.

2.2 Eine DTD zur Auszeichnung mittelhochdeutscher Wörterbücher

Die zuvor allgemein formulierten Regeln und Strukturprinzipien von SGML müssen auf die mittelhochdeutschen Wörterbücher angewendet werden, wenn diese als SGML-konforme Dokumente in einem elektronischen Wörterbuchverbund publiziert werden sollen. Also müssen Regeln definiert werden, die die Struktur dieser Wörterbücher exakt beschreiben, und zwar so, daß auch Klammerungen, Reihungen und Wiederholungen einzelner Elemente in einem Wörterbuchartikel genau ausgezeichnet werden können. Diese Arbeit setzt die sorgfältige Analyse der Wörterbuchartikel voraus und kann aufwendig und mühsam sein:[8] Zwar dürften die wesentlichen Elemente eines Wörterbuchartikels bekannt und in Wörterbüchern zur gleichen Sprache oder Sprachstufe nicht allzu verschieden besetzt sein, doch können Reihenfolge und Beziehungen zwischen diesen Elementen von Wörterbuch zu Wörterbuch recht unterschiedlich gestaltet sein.[9] Z.B. weist der BMZ als ein auf oberster Ebene alphabetisch nach Wortstämmen geordnetes Wortfamilienwörterbuch eine andere

[8] Das gilt jedenfalls für die Retrodigitalisierung, bei der immer nachvollzogen werden muß, welche Strukturen ein früherer Bearbeiter eigentlich intendierte. Wird ein neues Wörterbuch erarbeitet, kann die Modellierung der Wörterbuchartikel von Anfang an zur Spezifikation einer entsprechenden DTD herangezogen werden.

Makrostruktur auf als Lexers ‚Handwörterbuch', dessen Lemmata sämtlich initialalphabetisch angeordnet sind.

Für die Artikel des ‚Handwörterbuchs' wurde im Rahmen einer Magisterarbeit eine Strukturbeschreibung angefertigt; [10] sie kann als Grundlage einer DTD-Modellierung herangezogen werden. Damit die Besonderheiten eines jeden Wörterbuchs adäquat abgebildet werden könnten, hätten Strukturbeschreibungen auch für jedes weitere Wörterbuch im Verbund entwickelt werden müssen. Die jeweils wörterbuchspezifischen Analysen müßten in einem zweiten Schritt auf ein übergeordnetes, generalisierendes Modell projiziert werden, um eine DTD für den eigentlichen Wörterbuchverbund zu konstruieren.

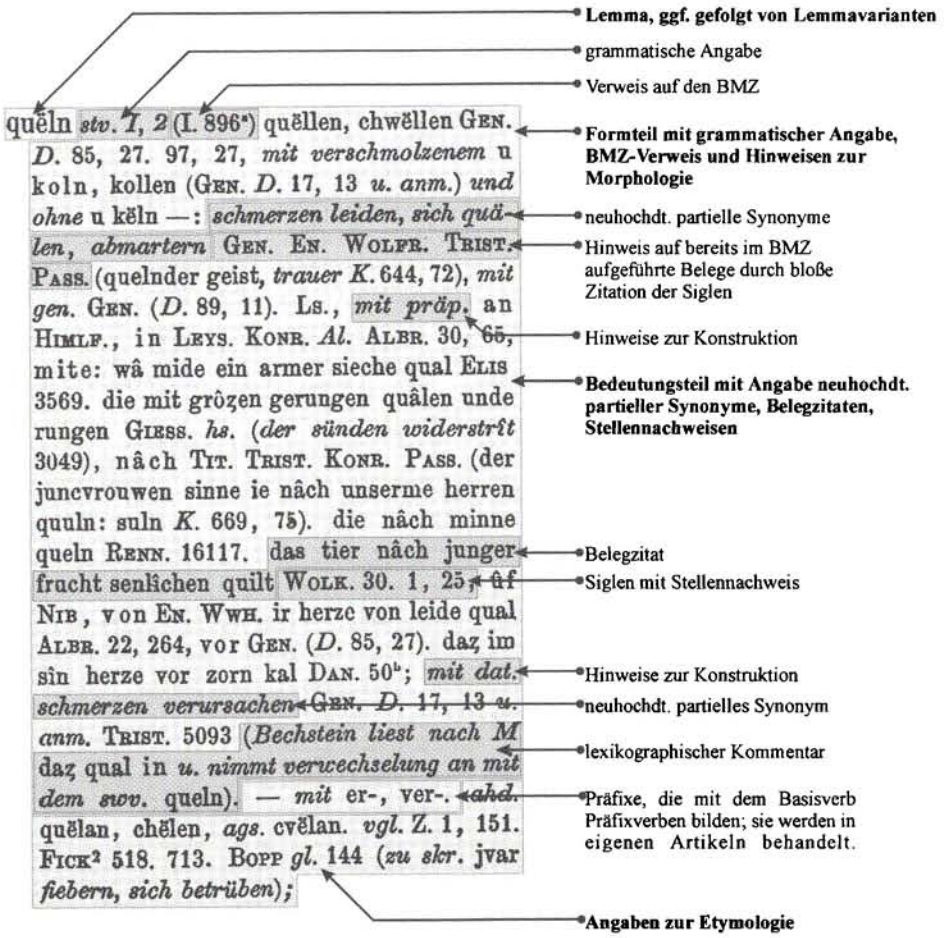


Abb. 1: Artikel *quëln* (Lexer II, 321)

---

[9] Vgl. Hausmann/Wiegand (1989); Wiegand (1989a); Wiegand (1989b).

[10] Rösler (1998). Eine elektronische Version dieser Arbeit ist unter http://gaer27.uni-trier.de/MWV-online/MWV-online.html zugänglich.

Eine ungefähre Vorstellung über die Komplexität der erforderlichen Modellierung soll die Übersicht zur Artikelstruktur des LEXER (Abbildung 1) vermitteln.[11] Daß die Beziehungen zwischen den Wörterbuchelementen noch schwieriger zu beschreiben sind, wenn auch die Strukturen der Nachträge LEXERs, des BMZ und des FINDEBUCHS in das Schema integriert werden, kann man sich leicht vorstellen.

Die Entwicklung derartiger Strukturbeschreibungen ist zeitintensiv. Sehr viel Zeit ist auch vonnöten, solche Strukturbeschreibungen in DTDs umzusetzen und zu testen, ob die jeweiligen Analysen die Bedingungen für eine digitale Umsetzung hinreichend genau erfüllen. Das erfordert nämlich eine längere Erprobungsphase, während der erste Versionen der DTDs modifiziert werden müssen, um die fehlerfreie Umsetzung von Dokumenten in Dateien sicherzustellen, die der jeweiligen DTD konform sind.

2.3 *Encoding Dictionaries:* Kapitel 12 der TEI-Richtlinien

Gerade aufgrund der langwierigen Analyse- und Testphase gibt es Bestrebungen, mehrfach verwendbare DTDs zu definieren, die für viele verwandte Anwendungen eingesetzt und entsprechend zugeschnitten werden können. Eine derartig konfigurierbare DTD wird z.B. von der *Text Encoding Initiative* zur Verfügung gestellt.[12] Die TEI-DTD beruht zwar im wesentlichen auf der Analyse neusprachlicher Wörterbücher zur englischen, französischen und spanischen Sprache,[13] nicht auf Analysen der zu digitalisierenden mittelhochdeutschen Wörterbücher. Doch sind die Fragmente der TEI-DTD ganz bewußt möglichst allgemein gehalten, um eine Anwendung auf unterschiedlich konzipierte Wörterbücher zu ermöglichen. Die Autoren der TEI-Richtlinien waren nämlich der Ansicht, daß eine möglichst ‚weit geschnittene' DTD vielen Forschern den Zu- und Umgang mit SGML-konformer Auszeichnung erheblich erleichtern würde:

> Since the skills needed for modifying the document grammar seem more likely to be found among researchers who want to exploit SGML's document validation powers to the full than among researchers who happen to be working with eccentric document structures, it is clearly preferable for the TEI to err by overgenerating, rather than by undergenerating.[14]

Daher stand zu erwarten, daß die TEI-DTD eine relativ problemlose Auszeichnung auch der mittelhochdeutschen Wörterbücher ermöglichen würde. Darüber hinaus – und das ist ein ganz entscheidender Vorteil der TEI – dürfte das zukünftige Einbeziehen weiterer Wörterbücher in den Wörterbuchverbund leicht möglich sein; eine Eigenentwicklung hingegen erforderte unablässige Erweiterungen und Modifikationen. Endlich ist die TEI-DTD kein rein theoretisches Konstrukt, sondern seit einigen Jahren in vielfältigen praktischen Anwendungen erfolgreich erprobt,[15] so daß langwierige Test- und Modifikationsphasen entfallen können.

---

[11] Rösler (1998, 111) versucht auch, die Beziehungen zwischen den Elementen eines LEXER-Artikels in einer DTD-ähnlichen Notationsweise zu beschreiben.

[12] Die TEI ist eine Vereinigung verschiedener geisteswissenschaftlicher Forschergruppen, die sich das Ziel gesetzt hat, SGML-basierte Applikationen für ganz unterschiedliche geisteswissenschaftliche Projekte zu entwickeln. Dazu gehören u.a. eine DTD für die SGML-konforme Beschreibung von Wörterbüchern (vgl. Guidelines 1990–1994, Kap. 12). Näheres zur TEI unter http://etext.virginia.edu/TEI.html, ferner Jannidis (1997) und Schmidt (1997).

[13] Vgl. die in Kapitel 12.2.2. *Groups and Constituents* der *Guidelines* angeführten Wörterbücher; s. auch die Liste bei Ide/Véronis (1995, 178, Anm. 4).

[14] Sperberg-McQueen/Burnard (1995, 21).

Aus den gerade angeführten Gründen versprach die Anwendung der TEI-Richtlinien ein zügiges Voranschreiten der Retrodigitalisierung mittelhochdeutscher Wörterbücher. Vor dem Erstellen der ersten TEI-konformen Dateien mußte jedoch eine weitere, sehr wichtige Entscheidung getroffen werden. Jedes Wörterbuch kann unter zwei verschiedenen Aspekten betrachtet werden:[16] Es kann einerseits als eine Art Datenbank über Sprachmaterial betrachtet werden, und es kann andererseits ebensogut als historisches Dokument untersucht werden, wenn z.B. sein Layout und seine typographische Gestaltung zum Objekt der (bibliothekarisch-bibliographischen) Forschung werden. Im letzten Fall wäre – auch wenn das anfangs paradox klingen mag – der Inhalt des Dokuments sein Layout.

Die *Guidelines* der TEI halten tatsächlich Mechanismen bereit, um beide Sichtweisen zugleich zu kodieren. Normalerweise wird durch SGML die logische Struktur von Dokumenten kodiert und nicht deren Layout. Will man dennoch derartige Aspekte wie beispielsweise Zeilenwechsel oder Seitenumbrüche in der SGML-Kodierung berücksichtigen, steht man vor dem Problem, daß diese Informationen der logischen Textstruktur entgegen stehen und ihre hierarchische Gliederung aufbrechen. SGML bietet für diesen Fall im wesentlichen zwei Lösungsmöglichkeiten: Einerseits kann man mit konkurrierenden DTDs arbeiten, d.h. das Dokument wird auf zweierlei Weise innerhalb einer Datei beschrieben, also eine hierarchische Auszeichnung der inhaltlichen Strukturen und eine ‚flache' Auszeichnung der Zeilen- und Seitenwechsel. Der Einsatz konkurrierender DTDs erhöht allerdings den Kodierungsaufwand, da zu jeder Markierung notiert werden muß, aus welcher DTD sie stammt. Andererseits kann man die Layoutinformation durch sogenannte EMPTY-Tags in die Dokumenthierarchie einfließen lassen. Die DTD muß dabei gewährleisten, daß diese Tags innerhalb eines jeden anderen Elementes auftreten dürfen. Dies ist möglich durch die Angabe sogenannter ‚inclusive rules', d.h. Regeln, die ein Element in andere Elemente einschließen.

Eine solch aufwendige Kodierung wäre einem zügigen Fortschreiten des Projekts nicht gerade förderlich gewesen. Deshalb haben wir uns dafür entschieden, allein die Datenbankperspektive mit Hilfe TEI-konformer Auszeichnungen festzuhalten. Für die digitale Fassung eines mittelhochdeutschen Wörterbuchverbunds bringt diese Sichtweise den entscheidenden Mehrwert über die zugrunde liegenden Druckwerke hinaus; erst aus der Datenbankkomponente ergibt sich nämlich die Möglichkeit des stichwortunabhängigen systematischen Zugriffs auf die Wörterbücher.

Die Darstellung der Wörterbücher auf dem Bildschirm entspricht dennoch der in den Druckwerken zugrunde liegenden Typographie, allein Zeilenfall und Seitenumbruch sind nicht berücksichtigt. Damit auch diese jederzeit abrufbar sind und z.B. für die Zitation eines Artikels herangezogen werden können, wählten wir eine andere, weniger aufwendige Art ihrer elektronischen Reproduktion. Aus Dateien im TUSTEP-Format – sie bilden ohnehin die Grundlage für die TEI-konforme Aufbereitung der elektronischen Wörterbücher – werden mit Hilfe des TUSTEP-Satzprogramms PostScript-Files der Wörterbuchseiten hergestellt, die den genauen Zeilenfall und Spaltenumbruch der Wörterbücher simulieren. Eine Verknüpfung dieser Dateien mit den entsprechenden Wörterbuchdaten – sie kann über die in der TEI-konformen Wörterbuchversion bei jedem Lemma in einem Attribut enthaltenen Referenz hergestellt werden – ermöglicht ein Nebeneinander von Datenbankperspektive und ‚historischer' Sichtweise, freilich um den Preis, daß allein die ‚tiefenstrukturelle' Perspektive auf das Wörterbuch in SGML-Auszeichnungen festgehalten worden ist.

---

[15] Vgl. Ide/Sperberg-McQueen (1995).

[16] Zum folgenden Ide/Véronis (1995, 167f.).

3 Kodierung mittelhochdeutscher Wörterbücher nach den TEI-Richtlinien

Der Hauptteil dieses Beitrags zeigt, welche Vorzüge und welche Probleme eine TEI-konforme Auszeichnung der mittelhochdeutschen Wörterbücher mit sich bringt. Doch kann nicht immer klar geschieden werden, ob die im folgenden erörterten Probleme durch SGML als solche, durch die spezifische, in den DTDs der TEI vorliegende besondere Form von SGML, oder generell aus dem Versuch resultieren, nur gering standardisierte Wörterbücher durch strikt definiertes Markup auszuzeichnen.

3.1 Probleme der Standardisierung

Die Artikel des BMZ sind nicht streng standardisiert, sondern zeichnen sich durch einen eher diskursiven Wörterbuchstil aus; die geringe Stringenz der Artikelstruktur macht sich sowohl in den Relationen der Artikelteile als auch bei Elementen innerhalb dieser Artikelteile bemerkbar. Die Angaben zur Morphologie, zur Bedeutung und zur Etymologie, die als Hauptkonstituenten eines Wörterbuchartikels betrachtet werden müssen, werden nicht immer in einer bestimmten Reihenfolge geboten.[17] Vielmehr ist zu beobachten, daß diese Hauptkonstituenten nicht nur an beliebigen Stellen eines Artikels vorkommen können, sie treten u.U. auch mehrmals innerhalb desselben Artikels auf.

Eine derart freie Abfolge der Hauptkonstituenten im Artikel dürfte viele nicht streng standardisierte Wörterbücher kennzeichnen. Dementsprechend ist die TEI-DTD auch so formuliert, daß die Elemente `<form>`, `<gramGrp>`, `<sense>` und `<etym>` in einem `<entry>` ohne zwingend vorgeschriebene Reihenfolge wiederholt vorkommen dürfen,[18] wie das auf der nächsten Seite abgebildete *content model* von `<entry>` zeigt:

---

[17] Innerhalb gewisser Grenzen können jedoch Grundtypen unterschieden werden, nach denen die Elemente eines BMZ-Artikels angeordnet worden sind. Bei einem ersten Typ folgt unmittelbar auf das Stichwort die entsprechende althochdeutsche Wortform oder andere Hinweise zur Herkunft und Verwandtschaft des Wortes (in runden Klammern), dann eine grammatische Angabe, ein oder mehrere partielle(s) Synonym(e), ein Etymologieteil, schließlich ein oder mehrere Bedeutungsteile (vgl. aus Band I die Artikel **abec** ‚verkehrt' 3^b 29, **balke** ‚balke' 79^b 36, **bol** ‚werfe, schleudere' 118^a 45, **bitel** ‚der freier' 171^a 15 oder **brünne** ‚schutzwaffe' 270^a 14). Wesentliche Informationen zur Morphologie werden oft unmittelbar nach der grammatischen Angabe erörtert, sind zuweilen aber auch zwischen Etymologie- und Bedeutungsteil eingeschoben. Die Belegreihen innerhalb der Bedeutungsteile beginnen insbesondere bei größeren Artikeln häufig mit Glossenbelegen. Die unmittelbar auf das Stichwort folgenden etymologischen Angaben in runden Klammern finden sich von Band II^a an seltener als noch im ersten Band. Weniger umfangreiche Einträge folgen oftmals einem zweiten Typ, bei dem auf das Stichwort die grammatische Angabe, ein partielles Synonym (selten ein durch partielle Synonyme eingeleiteter Bedeutungs- und Belegteil) und ein Etymologieteil folgen (in dieser Ausprägung ist der Typ v.a. bei Artikeln zu Stammwörtern belegt, vgl. aus Band I z.B. ÂWESEL ‚kraftlos' 74^a 21, DËHSE ‚beil' 311^a 20, HUNT ‚hundert' 727^b 25, KANZ ‚rand' 786^a 13 und **volleiste** 962^b 32); noch häufiger freilich folgen auf das Stichwort allein ein Form- und ein Bedeutungsteil (vgl. wiederum aus Band I **kristâbent** 4^b 7, **adelhaft** ‚adelmäßig' 8^a 42, ADMIRÂT ‚titel des kalifen' 10^a 31, **âlûne** ‚mache leder mit alaun gar' 27^a 21 oder **unbërhaftic** ‚unfruchtbar' 140^b 23). Diese Grundmuster variieren insofern, als nicht immer alle Artikelteile tatsächlich vorhanden sind.

[18] Ide/Véronis (1995, 171).

```
<!ELEMENT entry - - (hom | sense | def | eg | etym | form | gramgrp
| note | re | trans | usg | xr)+ >
```

Schwierig ist die korrekte Auszeichnung solcher Artikel also nicht aufgrund der erforderlichen DTD-Konformität, sondern aufgrund der Tatsache, daß die derart wechselnd angeordneten Artikelteile – da klare und eindeutige Strukturmarker in aller Regel fehlen – mit Hilfe automatisierter Prozeduren nur sehr fehlerhaft und unvollständig ausgezeichnet werden können, so daß eine korrekte Auszeichnung in mühevoller Handarbeit vervollständigt werden muß.

Problematisch wird die korrekte und TEI-konforme Auszeichnung allerdings dann, wenn weder bei der automatisierten noch bei der nachträglichen manuellen Auszeichnung genau festgestellt werden kann, ob und wie bestimmte Teile eines Wörterbuchartikels voneinander getrennt werden müßten. Relativ häufig kann nämlich nicht zweifelsfrei festgestellt werden, wie z.B. die Angaben zur Morphologie von der eigentlichen Bedeutungsbeschreibung und der Wortgeschichte getrennt worden sind; oftmals läßt sich die Bedeutung eines Stichworts oder sein Gebrauch nur in bestimmten Formen allein aus der Wortgeschichte erklären; in einigen Fällen werden im Wörterbuch regelrechte Forschungskontroversen nachgezeichnet:[19]

15RÎM *stm. Wackernagel hält dies wort für dasselbe mit ahd.* hrîm, rîm *Graff* 2, 506 = *series, numerus, ags.* gerîm *computus, calendarium, in letzterem sinne noch alln.* rîm (*vgl. Schmeller* 3, 86) *u. erklärt mhd.* rîm *für vers, insofern er nach der zahl (der silben oder accente), nicht nach der quantität gebaut ist. aber obwohl sich diese bedeutung wohl so hätte entwickeln können, so möchte ich doch Schmeller beistimmen, der mhd.* rîm *für verkürzung aus* rhythmus *hält, vgl. a. a. o. dass wenigstens* rhythmus *im mittellat. in der bedeutung völlig zusammenfällt mit* rîm *ist bekannt. die reimzeile, der vers.* mit behendeclîchen rîmen; wie kan er rime lîmen, als op si dâ gewahsen sîn *Trist.* 47, 14. [...]

Abb. 2: Artikel *RÎM* (BMZ II[a] 703[b] 15)

Die korrekte Markierung solcher Strukturen ist problematisch. Da das Prinzip der hierarchischen Einbettung eine Grundidee von SGML ist, hält die TEI keinen unmittelbaren Mechanismus bereit, mit dessen Hilfe sich überlappende Strukturen ausgezeichnet werden könnten. Überlappende Strukturen widersprechen einem hierarchischen Auszeichnungs-

schema. Sie lassen sich daher nur durch Hilfskonstrukte in SGML darstellen, die die überlappenden Abschnitte in konsekutive Teilstücke aufteilen und diese separat markieren. Eine mögliche Vorgehensweise wird auch hier durch den Einsatz von EMPTY-Tags gegeben, etwa in der in Abbildung 3 gezeigten Form. In diesem Beispiel soll der Bereich B sowohl mit der Markierung M_1 als auch mit M_2 ausgezeichnet werden. Diese Überlappung läßt sich folgendermaßen ausdrücken:

```
Xxx <mark name="M1" type="start">A<mark name="M2"
type="start">B<mark name="M1" type=end">C<mark name="M2"
type="end">xx x
```

Da es sich beim Tag <mark> um ein EMPTY-Tag handelt, werden keine Ende-Tags gesetzt. Der Nachteil dieser Art von Markierung besteht darin, daß der eigentliche Inhalt außerhalb der Markierungen steht, da man nur Anfangs- und Endepositionen von Bereichen in den Text kodiert hat. Man hat keine hierarchische Auszeichnung, sondern eine flache Struktur.

Eine weitere Möglichkeit besteht darin, in der DTD ein Vorkommen der überlappenden Markierungen gegenseitig zuzulassen, d.h. für obiges Beispiel würden wir ein Auftreten von M2 innerhalb von M1 erlauben. Die Auszeichnung hätte dann folgende Form:

```
Xxx <M1>A<M2>B</M2></M1><M2>C</M2>xx x
```

In diesem Fall hat man eine echte hierarchische Kodierung. Der Nachteil besteht allerdings darin, daß sämtliche Fälle von Überlappungen in der DTD berücksichtigt werden müssen, d.h. jede Markierung M_i muß in M_j zugelassen werden, falls zwischen diesen eine Überlappung auftreten kann.

Wir begnügen uns daher mit einer einfacheren Variante der Auszeichnung, obwohl diese der eigentlichen Wörterbuchlogik nicht ganz adäquat ist. Und zwar zeichnen wir entweder den ganzen, durch Überlappungen gekennzeichneten Passus als nur ein Element aus, wobei

---

[19] Wie grammatisch-morphologische, semantische und etymologische Informationen oft eng miteinander verquickt sind, zeigen einige Kurzzitate wohl deutlicher als bloße Hinweise auf Stichwörter des BMZ: „*das geschlecht dieses wortes schwankt sehr, was sich aus den verschiedenen ahd. Formen nur theilweise erklärt. vgl. ahd.* gadingî *stf.* [...]" zu **gedinge** ‚zuversicht' (I 339[b] 21); „*mit dieser specialisirung der bedeutung hängt auch wohl die änderung der form zusammen, die verkürzung des* i *u. die verdoppelung des* t" unter **rîtaere** (II[a] 739[a] 3); „*doch muß* gat *ursprünglich einen weitern umfang gehabt haben; es führt auf ein verlorenes ahd. stv.* gitu, gat, *welches wahrscheinlich die bedeutung* ‚jungere' *hatte"* zu GAT *stn.* (I 487[b] 15); „*was die ursprüngliche form des wortes war, und wie sich aus dieser seine bedeutung herleiten läßt, muß fürs erste auf sich beruhen"* zu BILWIZ ‚eine art elbe' (I 127[a] 4); „*Schmeller* [...] *nimmt 2 verschiedene worte an,* reiten = *zählen mit goth.* raþjan, *mhd.* reden *zusammenstellend,* reiten = *zurüsten aber mit goth.* ráids; *doch* raþjan *und* reiten *liegen lautlich fernab voneinander, und füglich können beide bedeutungen aus derselben grundbedeutung erwachsen sein, die* = series, ordo *war"* zu REITE ‚zählen, rechnen; zurüsten, bereiten' (II[b] 667[a] 2). – Bei der Verschränkung semantischer und etymologischer Information spielt die Frage nach der ‚ursprünglichen' Bedeutung eines Wortes, seiner ‚Grundbedeutung' eine wichtige Rolle (vgl. [1]DWB I, Vorrede, S. XLV und S. XIf. mit Jacob Grimms Kritik an den etymologischen Ansätzen des BMZ). Die Wiedergabe von Forschungskontroversen versteht sich aus der Tatsache, daß die Lexikographie des Mittelhochdeutschen zu Beneckes, Müllers und Zarnckes Zeiten eine noch junge Wissenschaft war: Das Wörterbuch eröffnete somit die Möglichkeit, auch die breitere Fachöffentlichkeit an der Diskussion der Spezialisten teilhaben zu lassen.

genau das Element gewählt wird, dessen Sichtweise im fraglichen Abschnitt als dominant erscheint. Oder es werden tatsächlich verschiedene Artikelteile markiert, auch wenn dieses Verfahren zuweilen etwas gewaltsam scheinen mag. Denn obgleich im letzten Fall ein genauerer Zugriff auf eine Datenbank als möglicher Vorteil ins Feld geführt werden kann, führt dieses Verfahren zu einer künstlich herbeigeführten, starken Aufsplitterung, die der engen Korrelation zwischen den Elementen eines Artikels eigentlich nicht gerecht wird.

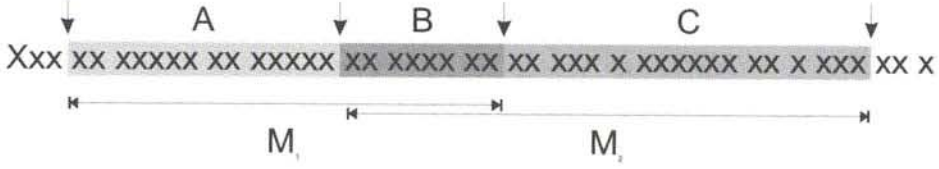


Abb 3: Kodierung überlappender Strukturen M_1 und M_2

Durch den diskursiven Wörterbuchstil kann auch die Auszeichnung der im Wörterbuch zitierten Literatur zum Problemfall werden. Die bibliographischen Angaben wie z.B. Siglen müssen schon deshalb gesondert ausgezeichnet werden, weil der Zugriff auf die Lemmata über die zu ihnen zitierten Texte eine der am häufigsten benutzten Abfragemöglichkeiten darstellen dürfte. Auch hier ist nicht die TEI-konforme Auszeichnung als solche, sondern wiederum die computergestützte Auszeichnung problematisch. Denn im BMZ fehlen eindeutige typographische und anderweitige Strukturanzeiger, mit deren Hilfe die zitierte Literatur fehlerfrei markiert werden könnte. Auch die elektronische Version des Quellenverzeichnisses von Eberhard Nellmann kann nur sehr bedingt verwendet werden, um die Siglen maschinell auszuzeichnen. Die Varianz der im Wörterbuch zitierten Siglen wird in Nellmanns Verzeichnis nämlich nicht vollständig erfaßt.[20] Das ist für einen menschlichen Benutzer völlig unproblematisch, macht eine Auszeichnung durch ein Computerprogramm, das sich allein auf die im Verzeichnis aufgelisteten Siglen stützte, allerdings ineffektiv.

Aus diesem Grund mußten zunächst einige recht unspezifische und vage Regeln formuliert werden, mit deren Hilfe die Siglen in ihrer tatsächlichen graphischen Varianz erfaßt werden konnten. Explizit formuliert lauteten diese Regeln z.B.: „Markiere als Siglen alle Vorkommen von Zeichenketten aus bis zu höchstens zwei Nichtblanks, denen eine Stellenangabe folgt". Die Folge solcher Auszeichnung mit Hilfe von notwendig unspezifischen Regeln war natürlich eine längere Korrektur von Hand. Dabei zeigte sich, daß die mar-

---

[20] Als BMZ-Sigle für „Des Landgrafen Ludwigs des Frommen Kreuzfahrt" führt Nellmann allein *Ludw. kreuzf.* an. Daneben finden sich im BMZ *Ludw. kr., Ludw. krzf.* und *Ludwig kr.* Für das „Buoch von guoter spise" nennt Nellmann die Siglen *b. v .g. sp.* und *b. von guter speise.* Tatsächlich zitiert BMZ diesen Text auch mit folgenden Siglenvarianten: *b. v. g. speise, b. v. guter speise, buch v. g. sp., buch v. g. speise, buch v. gut. sp., buch v. gut. speise, buch v. guter sp., buch v. guter speise, buch von guter speise.* Extreme Varianz zeigen auch die Siglen von Lorenz Diefenbachs ‚Vergleichendem Wörterbuch der gotischen Sprache', das nach Nellmann im BMZ als *Diefenb. g. wb.* firmiert. Tatsächlich erscheint diese Sigle in mindestens 12 Varianten, nämlich als *Diefenb. g. w., Diefenb. g. wb., Diefenb. g. wtrb., Diefenb. g. wrtbch., Diefenb. g. wtrbch., Diefenb. g. wörterb., Diefenb. goth. w., Diefenb. goth. wb., Diefenb. goth. wtrbch., Diefenb. goth. wörterb., Diefenbach g. wb., Diefenbach goth. wtrbch.* und *Diefenbach goth. wörterb.* Versehentlich fehlende Abkürzungspunkte erhöhen diese Varianz, die nicht nur für die drei angeführten Siglen charakteristisch ist, sondern nahezu alle im BMZ, auch viele der im LEXER angeführten Siglen betrifft.

kierten Passagen nicht allein mittelhochdeutsche Texte betrafen, sondern außerdem Kommentare zu diesen Texten und drittens weiterführende wissenschaftliche Literatur aus Monographien und Aufsätzen.[21] Diese verschiedenen Typen zitierter Literatur müssen sinnvollerweise ebenfalls mit Hilfe eines Attributes unterschieden werden. Über die Kennungen `type="sigle"`, `type="kommentar"` und `type="forschlit"` können die zitierten Titel dann jeweils eigenen Datenbankfeldern zugeordnet werden; die Abgrenzung zwischen Kommentaren und Forschungsliteratur dürfte allerdings nicht immer eindeutig zu treffen sein. Auch leuchtet aus den in Anm. 21 genannten Beispielen unmittelbar ein, daß diese verschiedenen Attribute nur sehr bedingt automatisiert vergeben werden können.

Ein weiteres Problem bei der Auszeichnung der im Wörterbuch zitierten Literatur resultiert daraus, daß die Referenz nicht in allen Fällen auf die Sigle folgt, sondern ihr sogar relativ häufig vorangeht. Nach den Empfehlungen der TEI zur *recoverability* (s.u. 4.1) sollte eine Umstellung der Referenz hinter die Sigle im markierten Text vermieden werden, um den Wörterbuchtext durch einfaches Entfernen aller Marken leicht und ohne Veränderung der Reihenfolge zwischen Elementen wiederherstellen zu können. Aus diesem Grund wurden in derartigen Fällen bisher nur die Siglen selbst, nicht aber die zugehörigen Referenzen markiert.

Wie überall stellen auch bei der Auszeichnung von Siglen und Literatur die nur implizit gegebenen Informationen die größte Hürde für ein maschinelles Markup dar, das hier vollends versagen muß. Daß sich bibliographische Hinweise wie „vgl. meine Ausgabe" oder „darnach meine Ausgabe" nur einem ‚kontextsensitiven' menschlichen Bearbeiter erschließen und nur von ihm richtig markiert werden können, liegt auf der Hand.[22] Allerdings ist auch hier eine genaue Auszeichnung unverzichtbar, um den direkten Zugriff auf sämtliche zitierte Literatur zu gewährleisten.

3.2 Probleme der Hierarchisierung

Im vorigen Abschnitt wurde bereits ausgeführt, daß SGML sehr schwerfällig und kaum geeignet ist, wenn überlappende Artikelstrukturen adäquat abgebildet werden sollen. Nach der eigentlichen Philosophie von SGML müssen nämlich Elemente niederer Ordnung immer in solche höherer Ordnung eingebettet sein; in diesem Sinne sind SGML-Dokumente streng hierarchisiert. Die Anwendung der TEI-Richtlinien auf die mittelhochdeutschen Wörterbücher zeigt jedoch mindestens zwei Stellen auf, für die die Hierarchie weniger streng definiert worden ist, als ein Benutzer es zunächst erwarten könnte.

Innerhalb von <sense>-Elementen werden relativ häufig grammatische Angaben zitiert, um z.B. eine Reihe von Belegen, in denen ein Substantiv stark flektiert wird, von einer weiteren Belegreihe zu trennen, die die schwache Flexion des gleichen Substantivs zeigt.

---

[21] Zur ersten Gruppe gehören gängige Siglen wie *Boner, MS., Nib., Parz.* oder *Ulr. Wh.;* auf die Kommentare verweisen z.B. die Formulierungen *Ettmüller zu Frl., Grimm zum gr. Rud., Haupt zur Winsbekin, Lachmann zu Iw., Sommer zu Flore* oder *v. d. Hagen im wb. zu Tristan*; in die dritte Gruppe gehören Untersuchungen wie *H. Jacobson kirchenrechtliche versuche, H. Schreiber die feen in Europa, Karajan beiträge zur geschichte der landesfürstl. münze wiens, Leo in Raumers histor. taschenb.* oder *Rochholz Schweizersagen aus dem Aargau.*

[22] Die Hinweise beziehen sich im ersten Fall auf Zarnckes Ausgabe des deutschen Cato (vgl. BMZ II^{a} 22^{b} 5f.), im zweiten auf Zarnckes Nibelungenlied (BMZ II^{a} 781^{b} 20). Daneben finden sich auch einige Hinweise auf „meinen commentar zum narrenschiff" (z.B. BMZ II^{a} 15^{b} 16f.; 26^{b} 19f.; 49^{b} 40f.).

Wohlgemerkt, die Wortbedeutung ist in beiden Belegreihen gleich.[23] Demnach gehören die grammatischen Angaben zur Bedeutungsbeschreibung der Substantive, also in einen <sense>-Teil. Nun führt die Auszeichnung durch ein <gram>- oder <pos>-Element innerhalb von <sense> beim Validieren des Dokuments zu einer Fehlermeldung. Diese Fehlermeldung läßt sich beheben, wenn <gram> oder <pos> in <gramGrp>-Markierungen eingeschlossen werden. Dieses Markup ist möglich, ohne daß der <sense>-Teil unmittelbar vor dem Beginn von <gramGrp> beendet wird. Nun werden sowohl <sense> als auch <gramGrp> als Hauptkonstituenten eines Wörterbuchartikels betrachtet. In den TEI-Richtlinien ist <gramGrp> allerdings rekursiv definiert, so daß die Hauptkonstituente <gramGrp> als Teil der Hauptkonstituente <sense> verwendet werden darf. Die Definitionen in den Richtlinien der TEI basieren auch hier auf tatsächlich zu beobachtenden Strukturen von Wörterbuchartikeln:

```
<!ELEMENT sense - - (sense | def | eg | etym | form | gramgrp | note
| [...] | handshift | #pcdata)* >
```

In einem anderen Fall ist es nicht möglich, die im Wörterbuch vorgegebene Hierarchie durch die Auszeichnung genau abzubilden, ohne daß es zu Konflikten mit der TEI-DTD kommt. Im Etymologieteil oder in Verweisen zitierte Wortformen werden häufig durch grammatische Angaben disambiguiert.[24] Da die grammatische Angabe unmittelbar zur zitierten Wortform zu rechnen ist, sollte sie nach unserem Dafürhalten in den jeweiligen <lang> oder <ref>-Tag eingebettet werden (a). Doch führt gerade diese, der Logik des Wörterbuchs entsprechende Auszeichnung zu Fehlermeldungen des Parsers, während das nicht streng hierarchisierte Markup (b) als TEI-konform validiert wird.

```
(a)    <xr><ref target="LB01079" n="s. oben">belle <gram
       type="stf"></ref></xr>
       <lang rend="gt">marikreitus <gram type="stm"></lang>

(b)    <xr><ref target="LB01079" n="s. oben">belle</ref> <gram
       type="stf"></xr>
       <lang rend="gt">marikreitus</lang> <gram type="stm">
```

Hier ist es auch nicht möglich, die <gram>-Elemente in ein <gramGrp>-Element einzubetten, da <gramGrp> innerhalb von <ref> und <lang> nicht verwendet werden darf. Die Fehlermeldungen ließen sich in TEI-konformer Weise nur dann beheben, wenn <gram> durch ein Element umschlossen werden könnte, das einerseits zu den ‚Eltern' von <gramGrp> gehörte und andererseits als ‚Kind' von <lang> und <ref> definiert worden wäre.

---

[23] Vgl. z.B. die Artikel **ahe** ‚fluss, wasser', **ërtstam** ‚baumstrunk' und **gîr** ‚geier' im ersten Band LEXERs. Auch Belegreihen zu Substantiven, die in verschiedenen Genera gebraucht werden, lassen sich hier anführen, vgl. die Artikel zu **schipfe** ‚schaufel' oder **schor** ‚schroffer fels, felszacke' in LEXER II. Hier ist der Gebrauch der Genera spezifisch für bestimmte Schreibsprachräume.

[24] Grammatische Angaben innerhalb von Verweisen finden sich z.B. in den LEXER-Artikeln **arn** *stv. red.*, **â-stiure** ‚ohne leitung, unbesetzt', **biuzen** ‚stossen' oder **bûzen-wendic** ‚auswendig, auswärts', innerhalb von fremdsprachigen Wortformen in den Artikeln **hôleht** ‚herniosus', **kôl** ‚kol, kolkopf', **mar** ‚quälendes nachtgespenst' oder **messe** ‚weibl. kalb von 1–2 jahren, das noch nicht gerindert hat'; die oben zitierten Beispiele stammen aus den Artikeln zu **bellunge** *stf.* und **margarîte** ‚perle'.

In einem letzten Punkt besteht Bedarf, die Richtlinien der TEI für unsere Zwecke zu modifizieren, um eine von der TEI-DTD nicht vorgesehene Einbettung eines Elements in ein anderes zu erlauben. Es ist nämlich verboten, das <def>-Element innerhalb von Passagen zu verwenden, die durch <gramGrp> ausgezeichnet worden sind. Doch können eigentliche Bedeutungserklärungen nicht selten mit den Angaben zur Morphologie durchmischt sein. Das ist häufig der Fall in den unter 3.1 erwähnten Wörterbuchartikeln, in denen insgesamt eine morphologisch-grammatische Perspektive dominiert, sich einzelne Wortbedeutungen aus den Formen ergeben, so daß eine Bedeutungsbeschreibung nicht von der Formenbeschreibung getrennt werden kann.[25] Abhilfe ist hier leicht zu schaffen, wenn <def> im *content model* von <gramGrp> in angemessener Weise berücksichtigt wird.

3.3 Probleme der (globalen) Attribuierung

Das Design der TEI-DTD ist nachhaltig geprägt durch das Prinzip, möglichst wenige Elemente zu definieren und stattdessen Attribute zu verwenden, um Elemente zu markieren, die sich nur geringfügig voneinander unterscheiden. Nach diesem Prinzip wurden vier sogenannte globale Attribute definiert, worunter solche Attribute zu verstehen sind, die zu jedem Element verwendet werden dürfen. Für die Auszeichnung vieler Positionen im Wörterbuchartikel, die kodiert werden müssen, auch ohne daß die von der TEI definierte Liste ‚passende' Attribute bereithält, liegt der Rückgriff auf die globalen Attribute daher immer nahe. Beim Markup der mittelhochdeutschen Wörterbücher wurde insbesondere das n-Attribut, das gewissermaßen zur Kommentierung von Elementen verwendet wurde, häufig herangezogen. Dieses Attribut dient z.B. im <entry>-Element dazu, auf Seite, Spalte und Zeile der Druckwörterbücher zu referieren, im <form>-Element dazu, um die sog. Sternchen-Lemmata oder fragliche Lemmaansätze zu kennzeichnen, im <def>-Element dazu, lateinische von neuhochdeutschen partiellen Synonymen zu unterscheiden.

Im BMZ waren zunächst auch die bei Verbartikeln zitierten Stammformen mit einem Attribut n=„Stammform" versehen. Einzelne Stammformen werden im zugrundeliegenden Druck allerdings gelegentlich auch mit kommentierenden Hinweisen versehen. Da diese Kommentare keine Stammformen sind, sondern die Stammformen lediglich näher erläutern, haben wir die Kommentare zunächst ebenfalls in n-Attribute ‚verpackt'.[26] Auf diese Weise kann sichergestellt werden, daß einerseits in einem Datenbankfeld nur wirkliche Stammformen ohne zugehörigen Kommentar aufgenommen werden, andererseits der Kommentar beim Wiederherstellen des Wörterbuchtextes nicht verloren geht. Bei diesem Vorgehen ist es allerdings möglich, daß in einem Element zwei n-Attribute zugleich auftreten. Eine derartige Auszeichnung erkennt der Parser als Verstoß gegen die TEI-DTD; das Dokument wird nicht validiert. Aus diesem Grund wurden alle n=„Stammform"-Attribute ersetzt durch rend-Attribute.

```
<form type="lemma" rend="Stammform">BRIUWEN</form>
<form type="lemma" rend="Stammform" n="und">BROUWEN</form>

<form type="lemma" rend="Stammform">GENESEN</form>
```

---

[25] Vgl. Anm. 19.

[26] Auf diese Weise wandert ein Teil des Wörterbuchtextes in das Markup. Ein solches Vorgehen widerspricht den Empfehlungen der TEI zur recoverability. Zu den Gründen dieses Verfahrens s.u. Abschnitt 4.1.

```
<form type="lemma" rend="Stammform" n="selten">GENEREN</form>

<form type="lemma">RUOFE <gram type="stv"></form>
<form type="lemma" rend="Stammform" n="prät.">RIEF</form>
```

Das ebenfalls global definierte rend-Attribut sollte allerdings eigentlich verwendet werden, um Hinweise auf Layout, Typographie und Format einzelner Elemente zu geben, die für unsere Kommentierung der Stammformen starker Verben jedoch irrelevant sind.

Von der ‚Normvorstellung', die nach den Empfehlungen der TEI mit der Verwendung des rend-Attributes verknüpft ist, weichen wir auch dann ab, wenn dieses Attribut zum Element <sense> tritt, um die in den Wörterbüchern belegten, verschiedenen syntaktischen Konstruktionsweisen von Verben (*transitiv*, *intransitiv*, *reflexiv mit Dativ* usw.) zu kennzeichnen und über das rend-Attribut gezielt auf Belege für diese Konstruktionen zugreifen zu können.[27] Mit der – hier – kursiven Wiedergabe der entsprechenden Konstruktionsangaben in den Wörterbüchern hat die Attribuierung jedoch nicht das geringste zu tun.

Die häufig erforderliche Verwendung unterschiedlichster Attribute kann leicht zu einer verwirrenden Vielfalt führen, wenn der eigentliche Wörterbuchtext fast vollständig hinter recht explizitem Markup verschwindet,[28] was zwar den Benutzer des fertigen Produkts nicht stört, für die Entwickler oder Bearbeiter jedoch leicht zum undurchschaubaren Auszeichnungs-Dschungel führen kann.

Beispielhaft sei hier dargestellt, wie sich die Verknüpfung des elektronischen Quellenverzeichnisses mit den Wörterbüchern auf eine Vielzahl von Attributen niederschlägt: Das elektronische Quellenverzeichnis folgt einer eigens entwickelten DTD mit nur wenigen definierten Elementen. In diesem Verzeichnis wurde jede Quelle als eigener <entry> definiert und mit dem globalen id-Attribut versehen, das ein eindeutiges Ansprechen der Quelle ermöglicht. Zur Klassifikation der Quelle nach Symptomwerten sind u.U. drei weitere Attribute für eine räumliche, zeitliche und textsortenspezifische Zuordnung vonnöten. Des weiteren müssen Siglen danach unterschieden werden, ob sie im LEXER, im BMZ oder nur im FINDEBUCH zitiert werden.

In den TEI-konform markierten Wörterbüchern selbst werden Belege über das globale type-Attribut als Quelle, Kommentar zur Quelle oder Forschungsliteratur gekennzeichnet (s.o.). Mit einem n-Attribut wird auf die Sprungadresse im elektronischen Quellenverzeichnis, das neben der bibliographischen Auflösung der Sigle die Informationen über sämtliche Symptomwerte enthält, hingewiesen. Nach den Konventionen unseres Projekts müssen die im Wörterbuch zitierten Quellentexte zudem nicht allein durch ein <title>-Element, sondern auch durch <bibl> und <author>-Elemente ausgezeichnet sein,[29] <ref> sowie <date>-Elemente können hinzutreten. Infolgedessen entstehen u.U. wahre Markup-Monstren. Gerade diese umständliche, oft als ‚geschwätzig' bezeichnete Auszeichnung ermöglicht allerdings die korrekte und effiziente Nutzung der Wörterbuchdaten.[30]

---

[27] Vgl. z.B. die LEXER-Artikel zum transitiven, intransitiven oder reflexiven Gebrauch der Verben **âsen, balden, baneken, bangen, erkoberen** oder **erkomen**.

[28] Ein Arbeiten nur mit Abkürzungen für die zu definierenden Elemente empfiehlt sich hier u.E. schon nicht mehr, weil zu viele Kürzel vom Bearbeiter ständig intellektuell aufgelöst werden müßten.

[29] <author> und <bibl> Elemente müssen verwendet werden, damit mehrgliedrige LEXER-Siglen wie z.B. DIEF. *n. gl.*, FRANKF. *richterb.*, PASS. *K.* oder TRIST. *U.* korrekt markiert werden können.

[30] Das Attribut *type* bezieht sich auf die Textsorte, das Attribut *place* auf die Zuweisung eines Schreibdialekts; *GL* steht für ‚Glossare und Wörterbücher', *GR 2* für ‚Heldenepik aus der Tradition der *chanson de geste*'.

```
<title n="QS0026" type="sigle"><bibl><author>Schm.</author>
Fr.</bibl> <ref>1,332 (<date n="a.">1399</date>).</ref></title>

<entry id="QS0026" type="GL">
<sigle type="Lexer">SCHM. Fr.</sigle>
<aufl>J.A.Schmeller: Bayerisches Wörterbuch. 2., mit des Verfassers
Nachträgen verm. Ausgabe, bearb. von G.K.Frommann. Bd. 1.2. München
1872-1877 [Neudr. (mit Vorworten von 1939 und 1961) Bd. 1.2. Aalen
1973].</aufl>
</entry>

<title n="QT0033" type="sigle"><bibl><author>Ulr.</author>
Wh.</bibl> <ref>165a</ref></title>

<entry id="QT0033" type="GR 2" place="südbairisch">
<sigle type="Lexer">TÜRH. Wh.</sigle>
<sigle type="Lexer">ULR. Wh.</sigle>
<sigle type="BMZ">Türh. Wh.</sigle>
<aufl>Ulrich von Türheim: Rennewart (früher 'Willehalm' genannt),
nach Lachmanns Abschrift der Heidelberger Hs. (H bei Hübner)</aufl>
<komm n="1">[Hübner vermerkt Blattzahl und Spalte der Hs. H jeweils
in runden Klammern; a und b bei Hübner = recto (bei Lexer: a), c und
d bei Hübner =
verso (bei Lexer: b)]. </komm>
<list type="date">
<item n="1243-1250">wohl zw. 1243 und 1250
</item>
</list>
<ausg>Ulrich von Türheim: Rennewart. Aus der Berliner und
Heidelberger Hs. [B und H] hg. von A. Hübner. Berlin 1938 (DTM 39)
[Neudruck Berlin 1964].</ausg>
<komm n="1">[Wortverzeichnis: S.559-614]</komm>
</entry>
```

Die DTDs der TEI zur Auszeichnung von Wörterbüchern wurden im wesentlichen an Wörterbüchern zum gegenwartssprachlichen Englisch, Französisch und Spanisch entwickelt.[31] Schreibsprachliche Varianten, mit deren Hilfe z.B. Texte klassifiziert und lokalisiert werden können, spielen hier nicht die wichtige Rolle, die ihnen in den Wörterbüchern zur mittelhochdeutschen Sprache zukommt. In deren Formteilen werden immer wieder dialektale Schreibformen angeführt, die auch ausgezeichnet werden sollten. Ein in der TEI-DTD definiertes <orth>-Element, mit dem die Schreibung der Stichwörter charakterisiert werden kann, darf allerdings nur innerhalb von <form>, nicht aber innerhalb von <gramGrp> verwendet werden. Unter den zu <gramGrp> definierten Elementen kommt einzig <gram> als Markup für Schreibsprachvarianten in Frage. Da <gram> jedoch bereits zur Auszeichnung der grammatischen Angaben verwendet wird, müssen <gram>-Elemente, mit denen Schreibsprachvarianten markiert werden, wiederum mit Hilfe eines Attributes von solchen <gram> geschieden werden, die sich auf das Markup grammatischer Angaben beziehen. Für differenziertere Möglichkeiten zur Auszeichnung schreibsprachlicher Varianten wäre es allerdings wünschenswert, im Rahmen der TEI-DTD ein zusätzliches Element zu definieren.

[31] Vgl. Anm. 13.

4 Zur Auswertung der TEI-konform markierten Dateien

4.1 Zur *recoverability* und maschinellen Wiederverwertung

Nach den Richtlinien der TEI sollten Texte möglichst so ausgezeichnet werden, daß das Markup ausschließlich Zusatzinformation bietet und der zu kodierende Text selbst nicht angetastet wird. Hintergrund dieser Empfehlung ist der Wunsch nach einem unproblematischen, leichten Austausch von Texten: Ist ein Wissenschaftler nur an einem Dokument, nicht aber an dem zu ihm gehörigen Markup interessiert, kann er alle Tags eliminieren und erhält so den ‚reinen' Text.

Bei der Retrodigitalisierung der mittelhochdeutschen Wörterbücher wird diese Empfehlung allerdings mit gutem Grund immer wieder mißachtet. Es gibt nämlich innerhalb vieler Artikel Informationen, die für einen gezielten Zugriff ausgezeichnet werden sollten, ohne daß die TEI-DTD eigene Elemente für diese Positionen definiert hätte. Deshalb werden häufig bestimmte Passagen des Wörterbuchtextes lediglich als Attribuierung einzelner Elemente ausgezeichnet. Erst bei der Ausgabe der Wörterbücher auf den Bildschirm wird der Text dieser Attribute wieder an die passende Stelle eingefügt. Z.B. ‚verschwinden' die Positionsmarken zur Artikelgliederung des BMZ in n-Attributen des <sense>-Elements; Asterisken, Fragezeichen oder eckige Klammern, die neue, fragliche oder falsch angesetzte Stichwörter kennzeichnen, werden als n-Attribute des Elements <form> kodiert; erläuternde Hinweise zu einzelnen Lemmavarianten – z.B. Hinweise auf ihre Häufigkeit durch ‚gelegentlich' oder ‚oder' – erscheinen ebenfalls als n-Attribute zu <form>.

In dem Bemühen, den ausgezeichneten Text der späteren Datenbankstruktur schon möglichst weit anzunähern, wird ein weiterer ‚Verstoß' gegen die Empfehlungen zur *recoverability* in Kauf genommen. Gelegentlich ändert sich nämlich auch die Reihenfolge der im Wörterbuch vorkommenden Elemente durch die oben beschriebene Art der Attribuierung. Dies gilt im wesentlichen für Fragezeichen, die im Druck unsichere Lemmaansätze oder unsichere grammatische Angaben andeuten. In diesen Fällen nämlich ist es sinnvoller – selbst unter Mißachtung der TEI-Empfehlungen –, gezielt auf die fraglichen Vorkommen zugreifen zu können. Aus diesem Grunde erscheinen die Fragezeichen als n-Attribute zum <form> oder zum <gram> Element.[32]

```
<gram type="stv. red. III" n="?">
<gram type="swv"> <gram type="st" norm="v" n="?">
<gram type="stf"> auch <gram type="swm" n="?">
```

So gesehen ergeben sich die hier erörterten ‚Verstöße' gegen Empfehlungen der TEI aus der Entscheidung, der Datenbankperspektive Priorität einzuräumen gegenüber einer ‚historischen', rein textuellen Perspektive auf die Wörterbücher; oberstes Kriterium ist und bleibt jedoch die Funktionalität der Auszeichnung für die spätere Verwendung der elektronischen Wörterbücher.

4.2 Über- und Unterauszeichnung

Ein großer Vorzug für die einfache digitale Umsetzung von Printwörterbüchern scheint uns die Tatsache zu sein, daß die TEI-DTD eine nur geringe Explizitheit der Auszeichnung erlaubt. Nicht jeder Wörterbuchartikel muß bis in seine feinsten Verästelungen hinein kodiert sein, bevor eine elektronische Publikation vorgenommen werden kann. Ein digitales

Wörterbuch kann schon dann umgesetzt werden, wenn z.B. noch nicht jede Wortform eines Formteils grammatisch genau klassifiziert worden ist. Da das *content model* von <gramGrp> nicht weiter ausgezeichnete Rohdaten (PCDATA) erlaubt, kann sich die Klassifikation der Elemente im Formteil auch in weiteren Schritten z.B. darauf beschränken, nur die grammatischen Hinweise zu markieren, die der eigentliche Wörterbuchtext selbst explizit benennt. Das Verfahren hätte den Vorteil großer Zeitersparnis, weil die expliziten Bestimmungen mit Hilfe automatisierter Prozeduren relativ leicht ausgezeichnet werden können, während alle nur implizit gegebenen Informationen eine aufwendige Nachmarkierung per Hand erfordern. Allerdings ist die Kehrseite dieses Verfahrens die dann stark eingeschränkte Möglichkeit zur maschinellen Wiederverwertung der elektronischen Wörterbücher: Eine solche setzte gerade die exakte grammatische Analyse voraus und dürfte ein Auslassen wesentlicher Informationspositionen gar nicht erlauben.[33]

Bestimmte Informationen, die nicht immer von höchster Relevanz sind, werden andererseits schon deshalb ausgezeichnet, weil sie z.B. leicht durch automatisierte Markup-Prozeduren erfaßt werden können. Nicht in allen diesen Fällen erkennt ein Benutzer (schon jetzt), daß hier bestimmte Sachverhalte markiert worden sind, auf die leicht zugegriffen werden könnte. Das gilt z.B. für die im vorigen Abschnitt erwähnten Sternchen-Lemmata oder fragliche grammatische Angaben. Auch sind im LEXER lateinische von neuhochdeutschen Bedeutungserläuterungen per Attribut unterschieden; unterschiedliche syntaktische Verwendungsweisen von Verben als intransitiv, reflexiv oder transitiv sind ebenfalls in Attributen angemerkt, ohne daß der Benutzer derzeit eine Möglichkeit hätte, diese Information abzurufen.

In anderen Fällen sind Elemente markiert und bereits abfragbar, ohne daß garantiert werden kann, daß wirklich alle relevanten Passagen ausgezeichnet worden sind. Denn bei allen nur implizit gegebenen Informationen – das gilt z.B. häufig für in fremden Sprachen zitierte Wortformen – kann nur die mühselige und meist langwierige manuelle Nachmarkierung gewährleisten, daß alle für elektronische Abfragen relevanten Passagen erfaßt werden.

5 Resümee

Läßt man die in den vorangehenden Abschnitten erörterten Erfahrungen im Umgang mit den TEI-Richtlinien noch einmal Revue passieren, zeigt sich dieses Bild: Probleme mit der Anwendung der Richtlinien auf die mittelhochdeutschen Wörterbücher ergeben sich nur zu einem geringen Teil durch die Architektur von SGML oder die DTD der TEI, sondern im wesentlichen aus dem Bemühen, die nicht streng standardisierten Wörterbuchartikel des BMZ und des LEXER mit Hilfe computergestützter Verfahren auszuzeichnen. Durch eine – wenngleich aufwendige und zeitraubende – nachträgliche Markierung von Hand ist es allerdings in den meisten Fällen möglich, eine Auszeichnung vorzunehmen, die sowohl der Struktur der Wörterbuchartikel als auch der Logik der TEI entspricht. Damit zahlreiche Inhalte eines Wörterbuchartikels, für die die Richtlinien keine eigenen Elemente definiert haben, als Attribute markiert werden können, wurden die TEI-Empfehlungen zur *recoverability* nicht in jeder Hinsicht befolgt. Dieses Verfahren, nach dem Wörterbuchinhalt nur

[32] Die drei folgenden Beispiele sind den LEXER-Artikeln zu **drouwen**, **rîden** ‚zittern' und **schricke** entnommen.

[33] Zu diesem Punkt s. die Zusammenfassung bei Heyn (1992, 187–192).

noch im Markup präsent ist, hat insbesondere Vorzüge für den effizienten Aufbau einer Wörterbuchdatenbank und entspricht damit dem Ansatz unseres Projekts, das die ‚Datenbank-Perspektive' weit stärker betont als die ‚historische Perspektive' auf die Wörterbücher. Aus den SGML-Dokumenten kann aber durch automatische Transformation ein *recoverable document* erzeugt werden (vgl. die korrekte Darstellung der mittelhochdeutschen Wörterbücher auf dem Bildschirm in Abbildung 4), welches jedoch nur auf Basis einer modifizierten TEI-DTD kodiert sein könnte. Die eingangs formulierte Erwartung, die auf ein zügiges Voranschreiten des Projekts durch den Einsatz der bereits in vielfältigen Anwendungen erprobten TEI-DTDs gerichtet war, hat sich voll und ganz bestätigt.

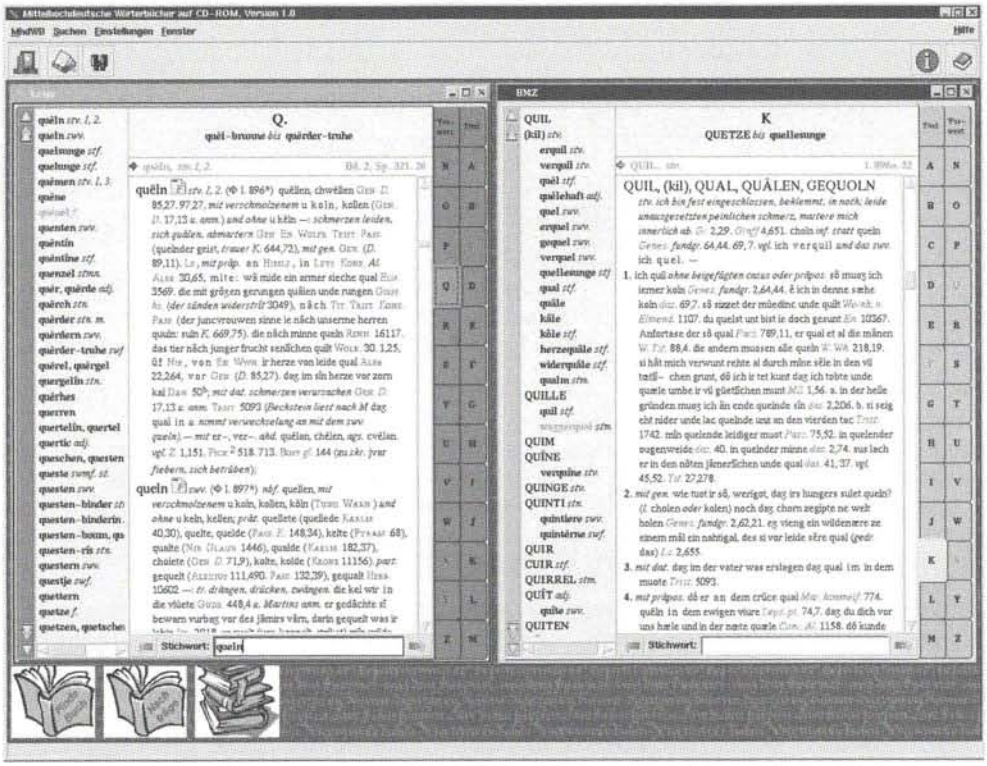


Abb. 4: Artikel *quëln* in der Darstellung der CD-ROM

6 Literatur

a) Wörterbücher

BMZ = Georg Friedrich Benecke/Wilhelm Müller/Friedrich Zarncke: Mittelhochdeutsches Wörterbuch. Leipzig 1854–1866. [Nachdruck mit einem Vorwort und einem zusammengefaßten Quel-

lenverzeichnis von Eberhard Nellmann sowie einem Alphabetischen Index von Erwin Koller, Werner Wegstein und Norbert Richard Wolf. Stuttgart 1990.]
FINDEBUCH = Kurt Gärtner/Christoph Gerhardt/Jürgen Jaehrling/Ralf Plate/Walter Röll/Erika Timm und Gerhardt Hanrieder (Datenverarbeitung): FINDEBUCH ZUM MITTELHOCHDEUTSCHEN WORTSCHATZ. Mit einem rückläufigen Index. Stuttgart 1992.
LEXER = Matthias Lexer: MITTELHOCHDEUTSCHES HANDWÖRTERBUCH. Leipzig 1872–1878. Nachdruck mit einer Einleitung von Kurt Gärtner. Leipzig 1992.
Nellmann, Eberhard (1997): Quellenverzeichnis zu den mittelhochdeutschen Wörterbüchern. Ein kommentiertes Register zum ‚Benecke/Müller/Zarncke' und zum ‚Lexer'. Stuttgart/Leipzig.
OED (1993): The Oxford English Dictionary Electronic Edition. Windows-CD-ROM mit 3,5"- und 5,25"-Diskette.

b) Sekundärliteratur

Alschuler, Liora (1995): ABCD... SGML. A User's Guide To Structured Information. London u.a.
Bachofer, Wolfgang (1988): Mittelhochdeutsches Wörterbuch in der Diskussion. Symposion zur mittelhochdeutschen Lexikographie. Hamburg, Oktober 1985. Tübingen.
Begegnung mit dem „Fremden" (1991): Grenzen – Tradtionen – Vergleiche. Akten des VIII. Internationalen Germanisten-Kongresses, Tokyo 1990. Hrsg. von Eijiro Iwasaki. Bd. 4. Sektion 4: Kontrastive Syntax; Sektion 5: Kontrastive Semantik, Lexikologie, Lexikographie; Sektion 6: Kontrastive Pragmatik. Hrsg. von Yoshinori Schichiji. München.
Burch, Thomas/Johannes Fournier/Kurt Gärtner (1998): Mittelhochdeutsche Wörterbücher auf CD-ROM und im Internet. Der Einsatz von SGML in der Retrodigitalisierung lexikographischer Standardwerke. In: Akademie-Journal 2. Mitteilungsblatt der Konferenz der deutschen Akademien der Wissenschaften, 17–24.
CHum 29 (1995): Computers And The Humanities. Volume 29. Special Issue on The Text Encoding Intitiative. Background and Contexts. Guest Editors: Nancy Ide and Jean Véronis.
Fournier, Johannes (1998): Mittelhochdeutsche Wörterbücher digital: Konzepte – Methoden – Entwicklung; s. u. http://193.174.98.10/scan1/MDZ/kolloquium/ref/fournier/vortrag.htm .
– (2000): Digitale Dialektik: Chancen und Probleme mittelhochdeutscher Wörterbücher in elektronischer Form. In: Wörterbücher in der Diskussion IV. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. Hrsg. Von Herbert Ernst Wiegand. (Lexicographica, Series Maior 100) Tübingen 85–108.
Gärtner, Kurt (1993): Das Handexemplar von Matthias Lexers ‚Mittelhochdeutschem Handwörterbuch'. In: Matthias von Lexer. Beiträge zu seinem Leben und Schaffen. Hrsg. von Horst Brunner. (Zeitschrift für Dialektologie und Linguistik; Beiheft 80). Stuttgart, 109–131.
– (1991): Ausgabenglossare und Wortverzeichnisse als Quellen eines neuen mittelhochdeutschen Wörterbuchs. In: Begegnung mit dem Fremden, 272–276.
– und Grubmüller, Klaus (Hrsg.; im Druck): Zur Konzeption des neuen mittelhochdeutschen Wörterbuchs. Abhandlungen der Göttinger Akademie der Wissenschaften.
Grubmüller, Klaus (1991): Elf Sätze zur Konzeption eines mittelhochdeutschen Wörterbuchs. In: Begegnung mit dem Fremden, 247–253.
Guidelines for Electronic Text Encoding and Interchange (1990–1994). Edited by C.M. Sperberg-McQueen and L. Burnard. Chicago, Oxford.
Hausmann, Franz Josef/Herbert Ernst Wiegand (1989): Component Parts and Structures of General Monolingual Dictionaries: A Survey. In: Wörterbücher I, 328–360.
Ide, Nancy/C.M. Sperberg-McQueen (1995): The TEI: History, Goals, and Future. In: CHum 29, 5–15.
Ide, Nancy/Jean Véronis (1995): Encoding Dictionaries. In: CHum 29, 167–179.
Jannidis, Fotis (1997): Wider das Altern elektronischer Texte: philologische Textauszeichnung mit TEI. In: edition. Internationales Jahrbuch für Editionswissenschaft 11, 152–177.
Jucker, Andreas H. (1994): New Dimensions in Vocabulary Studies: Review article of the *Oxford English Dictionary (2nd edition)* on CD-ROM. In: Literary and Linguistic Computing 9, 149–154.

Nellmann, Eberhard (1991): Die mittelhochdeutschen Wörterbücher. Ihre Qualitäten, ihre Grenzen, ihre mögliche Erneuerung. In: Begegnung mit dem Fremden, 254–263.

Plate, Ralf/Ute Recker (im Druck): EDV für Wörterbuchzwecke und neue lexikographische Arbeitsweisen. Erfahrungen beim Aufbau des elektronischen Text- und Belegarchivs für das mittelhochdeutsche Wörterbuch. In: Akten der 5. Internationalen Tagung „Maschinelle Verarbeitung altdeutscher Texte" (Würzburg 1997).

Rieger, Wolfgang (1995): SGML für die Praxis. Ansatz und Einsatz von ISO 8879. Mit einer Einführung in HTML. Berlin/Heidelberg.

Rösler, Uta (1998): Ein mittelhochdeutsches Wörterbuch auf CD-ROM. Strukturbeschreibung der Wörterbuchartikel in Matthias Lexers ‚Mittelhochdeutschem Handwörterbuch' für die Herstellung einer elektronischen Version auf CD-ROM. MA-Arbeit (masch.) Trier.

Schmidt, Frieder (1997): Neuland für die Buchgeschichte – Quellenaufbereitung im Zeitalter des WWW. Hypertext Markup Language (HTML), Standard Generalized Markup Language (SGML) und die Guidelines for Electronic Text Encoding and Interchange der Text Encoding Initiative (TEI). In: Leipziger Jahrbuch zur Buchgeschichte 7, 343–365.

Sperberg-McQueen, C.M./Lou Burnard (1995): The Design of the TEI Encoding Scheme. In: CHum 29, 17–39.

Storrer, Angelika (1998): Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In: Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. Hrsg. von Herbert Ernst Wiegand. (Lexicographica, Series Maior 84) Tübingen, 106–131.

Szillat, Horst (1995): SGML. Eine praktische Einführung. (Scientific Computing) Bonn.

Wiegand, Herbert Ernst (1989a): Aspekte der Makrostruktur im allgemeinen einsprachigen Wörterbuch: alphabetische Anordnungsformen und ihre Probleme. In: Wörterbücher I, 371–409.

– (1989b): Formen von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. In: Wörterbücher I, 462–501.

Wörterbücher. Ein internationales Handbuch zur Lexikographie. Hrsg. von Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand und Ladislav Zgusta. (Handbücher zur Sprach- und Kommunikationswissenschaft 5.1) Erster Teilband. Berlin/New York, 956–967.

Thomas Burch, Trier
Johannes Fournier, Trier

Ralf Plate, Ute Recker

Elektronische Materialgrundlage und computergestützte Ausarbeitung eines historischen Belegwörterbuchs. Erfahrungen und Perspektiven am Beispiel des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS

1 Das Vorhaben und sein EDV-Konzept

Das Vorhaben eines neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS ist im Gesamtzusammenhang der historischen Beleglexikographie des Deutschen zu sehen. Für das Althochdeutsche wie für das Frühneuhochdeutsche sind umfangreiche Belegwörterbücher im Erscheinen begriffen, die das Grimmsche Wörterbuch und seine ebenfalls in Bearbeitung befindliche Teilerneuerung (der Strecke A–F) als Epochenwörterbücher für die genannten älteren Sprachstufen ergänzen.[1] Für das Mittelhochdeutsche in den zeitlichen Grenzen von etwa 1050 bis 1350 ist man dagegen bislang noch auf die Pionierarbeiten des 19. Jahrhunderts angewiesen, deren historische Verdienste ebenso unbestritten sind wie ihre Unzulänglichkeit angesichts dessen, was heute möglich ist. Dies sind das von Georg Friedrich Benecke (1762–1844) konzipierte und von Wilhelm Müller und Friedrich Zarncke in den Jahren 1854 bis 1866 in vier Bänden ausgearbeitete MITTELHOCHDEUTSCHE WÖRTERBUCH (= BMZ) und das – auf dem BMZ beruhende und ihn vor allem für das späte 14. und das 15. Jahrhundert reichlich ergänzende – dreibändige MITTELHOCHDEUTSCHE HANDWÖRTERBUCH von Matthias Lexer (= LEXER), das 1872 bis 1878 erschien. Im vollen Wortsinne Belegwörterbuch zu nennen ist der BMZ, seine Belegbasis ist aber im Verhältnis zum heutigen Stand der Quellenerschließung verhältnismäßig schmal, und seine Anlage als Stamm- bzw. Wortfamilienwörterbuch legte eine starke Konzentration der Belegsammlung und -darbietung auf die Stammwörter nahe. Lexer hat die Belegbasis gegenüber dem BMZ

[1] Vgl. im Literaturverzeichnis unter AWB, [1]DWB, [2]DWB und FWB.

zwar erheblich erweitert, doch beschränkt er sich in großem Umfang auf Belegstellennachweise und verzichtet auf Belegzitate.

Die „zellen", meinte Lexer 1878 im Vorwort zum dritten Band seines Werks (S. IV), „sind durch die beiden mhd. wörterbücher gebaut und das weitere eintragen in dieselben wird eine verhältnismässig leichte arbeit der nachsammelnden sein". Das auf Lexers „Zellen" bezogene Nachsammeln geschah vor allem in Glossaren, Wortverzeichnissen und Registern zu Editionen und lexikographischen Untersuchungen mittelhochdeutscher Texte. Über 100 dieser Nachsammlungen sind kompiliert zu einem Gesamtverzeichnis, dem in Trier ausgearbeiteten ‚Findebuch zum mittelhochdeutschen Wortschatz' (1992). Das FINDEBUCH verweist in rund 35.000 Artikeln auf knapp 140.000 Glossarstichwörter, die ihrerseits in der Regel jeweils zu einer Vielzahl von Textbelegen führen.

Mit dem FINDEBUCH waren die Grenzen dessen, was auch dem kundigsten Benutzer an Eigenarbeit bei der Kombination verschiedener Wörterbücher und lexikographischer Hilfsmittel für das Mittelhochdeutsche zuzumuten ist, deutlich erreicht. Den gerade auch angesichts der sonst zu beobachtenden Fortschritte der historischen Beleglexikographie des Deutschen zunehmend als „unerträglich" empfundenen „Zustände[n] auf dem Gebiet der mittelhochdeutschen Lexikographie" (Stackmann 1990: V) war mit weiteren Anlagerungen an die alten Wörterbücher – sei es in der Form von Supplementen wie dem FINDEBUCH, sei es als Satellitenkranz von Spezialwörterbüchern – nicht mehr beizukommen, sie verlangten vielmehr zwingend einen Neuansatz in der Erfassung des mittelhochdeutschen Wortschatzes und der Beschreibung seines Gebrauchs, ein Wörterbuch also, das selbständig aus den Quellen zu erarbeiten ist.

Das neue Wörterbuch wird seit 1994 in zwei Arbeitsstellen an den Universitäten Göttingen und Trier mit Unterstützung der Deutschen Forschungsgemeinschaft und der Göttinger und Mainzer Akademien der Wissenschaften vorbereitet; zum 1.1.2000 sind die beiden Arbeitsstellen in das Akademieprogramm überführt worden. Die Vorbereitungsphase, die kurz vor dem Abschluß steht, gilt im wesentlichen der Materialbereitstellung, ferner der Diskussion und Erprobung konzeptioneller Grundsätze. Die Ausarbeitung des Werks, das auf insgesamt vier Bände zu je etwa 1000 Seiten berechnet ist, wird voraussichtlich 2001 beginnen und soll in einem Zeitraum von 20 Jahren abgeschlossen werden.

Als das jüngste der größeren Vorhaben zur historischen Beleglexikographie des Deutschen konnte das neue MITTELHOCHDEUTSCHE WÖRTERBUCH von vornherein ganz auf EDV-Basis gestellt werden. Zugute kamen den auf die EDV-Seite bezogenen Planungen die langjährigen Erfahrungen von Kurt Gärtner, dem Trierer Projektleiter, auf dem Gebiet des philologischen EDV-Einsatzes, die bis in die Mitte der 80er Jahre zurückreichenden computergestützten lexikographischen Vorarbeiten der Trierer Arbeitsstelle und nicht zuletzt der Beistand von Paul Sappler (Tübingen), der das für sein Vorhaben eines Wörterbuchs zu Gottfrieds ‚Tristan' entwickelte Programmsystem[2] zur Verfügung stellte und mit Rat und Tat bei der Anpassung und dem Ausbau für die Zwecke des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS behilflich war;[3] damit war gewährleistet, daß im Falle des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS die Sprachlexikographie von den avancierten Methoden und Techniken der Textlexikographie profitieren kann.[4]

---

[2] Vgl. Sappler/Schneider-Lastin 1991.

[3] Vgl. Recker/Sappler 1998.

[4] Wie „Methoden und Techniken der Textlexikographie auf die Sprachlexikographie zu übertragen" wären, diskutiert Sappler 1991 (Zitat S. 277).

Das EDV-Konzept des Vorhabens ist bereits mehrfach dargestellt worden unter Einschluß von Details der Arbeitsabläufe und ihrer technischen Realisierung.[5] Hier genügt daher eine kurze einleitende Rekapitulation der Grundzüge, die zum Verständnis des Zusammenspiels der einzelnen Programmkomponenten und Materialzustände nötig ist; in Bezug auf diese sollen dann anschließend dem Thema des Bandes entsprechend die „Chancen und Perspektiven", wie sie sich uns nach fünfjähriger Arbeit darstellen, erörtert werden.

Materialbereitstellung für ein historisches Belegwörterbuch heißt traditionell in erster Linie: Exzerption von Belegstellen aus den gedruckten Quellentexten, Verzettelung der Exzerpte, Ordnung der Belegzettel nach Artikelstichwörtern (Lemmatisierung).[6] Die elektronische Materialbereitstellung für das mittelhochdeutsche Wörterbuch lehnt sich an das traditionelle Verfahren an, geht aber in einigen Punkten anders vor; sie erfordert zusätzliche Arbeitsgänge, spart dafür andere ein und modifiziert insgesamt die Abfolge.

Zusätzliche Arbeitsgänge entstehen daraus, daß die Materialgrundlage in elektronischer Form vorliegen und für die weitere elektronische Verarbeitung eingerichtet werden muß. Materialgrundlage sind im Falle des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS überwiegend Volltexte, nur ein kleinerer Teil des Belegarchivs wird im Zuge der Ausarbeitung der Artikel durch Einzelerfassung von Belegen aus weiteren Texten erhoben werden. Das elektronische Textarchiv ist die erste Komponente der elektronischen Wörterbuchbasis.

Die weiteren Arbeitsgänge, die der herkömmlichen Verzettelung und anschließenden Lemmatisierung der Belege entsprechen, können bei entsprechender Vorbereitung der elektronischen Quellentexte automatisch bzw. halbautomatisch erledigt werden. Dafür wird – neben einer Reihe von geeigneten Programmen[7] – eine zweite Komponente benötigt, die die Lemmatisierungsinformation enthält, eine Art Thesaurus mit den Artikelstichwörtern und den ihnen zuzuordnenden Wortformen; diese Komponente heißt projektintern das Gerüst. Das Gerüst enthält eine auf der Grundlage der alten Wörterbücher und des FINDEBUCHS erarbeitete Stichwortliste mit insgesamt rund 80.000 Lemmata und die ihnen aus den bereits für das Belegarchiv bearbeiteten Texten zugewiesenen Wortformen.

Aus den Texten kann mit der Hilfe des Gerüsts eine lemmatisierte KWIC-Konkordanz (mit frei wählbarer Kontextlänge) erzeugt werden, der im herkömmlichen Verfahren der lemmatisierte Zettelkasten, also das eigentliche Belegarchiv entspricht. Wenn im folgenden in Anlehnung an die traditionelle Arbeitsweise vom Belegarchiv des mittelhochdeutschen Wörterbuchs die Rede ist, so gilt es stets, sich einen grundlegenden Unterschied vor Augen zu halten: Es handelt sich beim elektronischen Analogon eigentlich nicht um eine selbständige Komponente, sondern um einen Materialzustand, der auf der Grundlage von Texten und Gerüst bei Bedarf erzeugt wird und durch Änderungen in diesen beiden Komponenten beliebig beeinflußt werden kann. In der Phase der Materialbereitstellung dient die Ausgabe von Konkordanzen vor allem der Kontrolle der halbautomatischen Lemmatisierung. Erkannte Lemmatisierungsfehler werden durch Änderungen in den Texten (z.B. durch Einfügung von Homographenmarkierungen) und/oder im Gerüst (z.B. durch Umstellen von falsch zugeordneten Wortformen) behoben, ein neuer Konkordanzlauf stellt dann die berichtigte Lemmatisierung her.

---

[5] Vgl. zuletzt Plate/Recker 2000 und Sappler 2000b.

[6] Zur Exzerption für das FWB und DRW vgl. Lemberg 1996.

[7] Die verwendeten Programme sind auf der Basis des leistungsstarken Tübinger Systems von Textverarbeitungsprogrammen (TUSTEP) entwickelt, das erweiterbare Module für alle Anforderungen philologischer Datenverarbeitung zur Verfügung stellt.

Mit der Erzeugung des lemmatisierten Belegarchivs, das als elektronisches Analogon des herkömmlichen Zettelkastens gelten kann, ist die Arbeit der Materialbereitstellung geleistet. Es wäre nun möglich, hier den elektronischen Kreislauf zu unterbrechen, das heißt die Belege zu einem Stichwort für die Weiterarbeit in herkömmlicher Weise auf Zetteln auszugeben. Damit würden aber Vorteile verschenkt, welche die elektronische Materialgrundlage bietet. Für die elektronisch gestützte Artikelarbeit wird daher ein weiterer Materialzustand erzeugt, die Artikelarbeitsdatei, und es werden in einem Artikelredigierprogramm Funktionen zur Verfügung gestellt, die es dem Bearbeiter erlauben, die herkömmlichen Arbeitsgänge der Artikelarbeit bis hin zum Satz des Artikels elektronisch durchzuführen.

Weil der gesamte Arbeitsablauf computergestützt ist und alle Materialzustände in elektronischer Form vorliegen, ist auch die elektronische Publikation der Materialien wie des Wörterbuchs selbst ohne großen technischen Mehraufwand möglich. Die in diesem Zusammenhang zu erörternden Fragen sind daher im Falle des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS in erster Linie konzeptionell lexikographischer Natur.

Nach dieser kurzen Rekapitulation des EDV-Konzepts des Vorhabens sollen nun die einzelnen Komponenten und Materialzustände im Hinblick auf ihre „Chancen und Perspektiven" einer näheren Betrachtung unterzogen werden.

2 Das Textarchiv

Das elektronische Textarchiv besteht gegenwärtig aus den Volltexten des rund 75 Nummern (Einzeltexte und Textsammlungen unterschiedlichen Umfangs) umfassenden Grundkorpus der Wörterbuchquellen,[8] die Einwerbung von Mitteln für eine massive systematische Aufstockung zunächst um maschinenlesbare Versionen der rund 75 weiteren FINDEBUCH-Quellen, die noch nicht im Grundkorpus berücksichtigt sind,[9] wird zur Zeit vorbereitet.[10] Eine Reihe weiterer Texte, die zum Quellenbereich des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS zählen, sind aus dem Fundus der Projektbeteiligten und durch Deposita von Fachkollegen in maschinenlesbarer Fassung für das Wörterbuch zugänglich.

Angesichts der großen Zahl der Quellentexte wie des langen Zeitraums, in dem sie stets zugriffsbereit verfügbar gehalten werden müssen, ist eine sorgfältige einheitliche Einrichtung der Texte und ein softwareunabhängiges Auszeichnungs- und Archivierungssystem nötig.

Bei der Einrichtung der elektronischen Texte wird ‚ausgabendiplomatisch' verfahren. Daß diese Entscheidung nicht trivial ist, wird deutlich, wenn man sich vor Augen hält, daß das Quellenspektrum des Wörterbuchs die ganze Breite von verhältnismäßig stark normalisierten Editionstexten bis hin zu auch graphisch überlieferungsgetreuen Handschriftenabdrucken umfaßt. Die Alternative zum ausgabendiplomatischen Prinzip hätte darin bestanden, mindestens leichte Normalisierungen bei den graphisch überlieferungsnahen Editionen durchzuführen, um die Abbildung ungewöhnlicher und schwierig darzustellender, lexikographisch aber meist unerheblicher graphischer Differenzierungen zu ersparen. Genauere Prüfung dieser Alternative zeigte jedoch, daß es erstens immer zahlreiche

---

[8] Bibliographie im Anhang zu Gärtner (2000) und auf der Homepage des Projekts, dort auch mit einem alphabetischen Index.

[9] Bibliographie auf der Homepage des Projekts unter ‚Findebuchquellen'.

Einzelfälle gibt, in denen Unklarheit über die lexikographische Relevanz von Schreibvarianten besteht; die Aufstellung von Normalisierungsregeln und die Kontrolle ihrer Durchführung hätte also einen nicht unbeträchtlichen Aufwand erfordert. Zweitens aber, und dies ist der wichtigere Gesichtspunkt, wäre diese Normalisierung über den gesamten Text einer Quelle für Wörterbuchzwecke zum größten Teil verlorene Arbeit, denn nur ein Bruchteil der Belege eines Quellentextes wird später im Wörterbuch tatsächlich zitiert.

Neben diesem materiellen Aspekt hat die ausgabendiplomatische Einrichtung der elektronischen Wörterbuchquellen auch einen formalen: Er besteht zum einen in der Abbildung der Referenz (für die automatische Erzeugung der Fundstellenangabe bei den Belegtexten), zum anderen in der Auszeichnung von editorischem Beiwerk (wie Zwischenüberschriften des Herausgebers usw.), das nicht für das Belegarchiv ausgewertet werden soll. Beides zusammen läuft auf eine genaue Analyse und entsprechende sachliche Auszeichnung der formalen Struktur der Quellentexte hinaus.

Die über diese obligatorische Grundform der elektronischen Quellen des Textarchivs hinausgehenden weiteren Bearbeitungszustände enthalten zusätzliche Eintragungen, die der Steuerung der Lemmatisierung, der Auswahl oder dem Ausschluß von Belegen, ihrer Kommentierung usw. dienen. Sie brauchen hier im einzelnen nicht weiter beschrieben zu werden, dies ist bereits an anderer Stelle ausführlich geschehen.[11]

Der beträchtliche Aufwand, der für die Einrichtung und Bereithaltung des Textarchivs getrieben wird, scheint auf den ersten Blick kein Gegenstück im herkömmlichen Verfahren der Materialbereitstellung zu haben und lädt daher besonders zu kritischen Nachfragen ein. Was läßt sich zu seiner Begründung anführen?

Zunächst ist darauf aufmerksam zu machen, daß es sich in Wirklichkeit entweder um die ökonomische Vorverlegung und konzentrierte Durchführung von Arbeiten handelt, die im herkömmlichen Verfahren in späteren Phasen der Wörterbucharbeit ausgeführt werden müssen, oder um die Vorbereitung der programmgesteuerten Erledigung solcher Arbeiten. Erspart wird zum Beispiel

- das mühsame und fehleranfällige Abschreiben von einzelnen Belegen, sei es bereits bei der Exzerption selbst, sei es bei der Abfassung bzw. Erfassung der Artikelentwürfe
- die zusätzliche Konsultation der gedruckten Quellenwerke während der Artikelarbeit, die herkömmlich häufig dennoch nötig war, zum Beispiel weil die Kontextmenge auf dem Belegzettel nicht genügte oder weil die Transkription Zweifel erweckte
- die nicht weniger fehleranfällige Notierung der Quellensigle und genauen Fundstelle
- die in der Regel als letzter Arbeitsgang nötige Belegrevision (Kontrolle der Belegzitate des fertigen Artikels an den Quellen)
- schließlich ein großer Teil des Aufwands, der vor oder zu Beginn der Artikelarbeit für die Sortierung der Belegzettel unter verschiedenen Gesichtspunkten getrieben werden mußte (Lemmatisierung; Gruppierung lemmatisierter Belege innerhalb eines Stichworts nach verschiedenen Gesichtspunkten der Quellentypologie oder nach Wortformen usw.).

---

[10] Es handelt sich um ein deutsch-amerikanisches Kooperationsvorhaben, das an der Universität Trier in Zusammenarbeit mit dem Electronic Text Center der University of Virginia, Charlottesville, durchgeführt werden soll und die Einrichtung eines großen Digitalen mittelhochdeutschen Textarchivs für die Publikation im Internet und auf CD-ROM zum Ziel hat. Vorbereitende Arbeiten, die durch eine Anschubfinanzierung aus dem Forschungsfonds der Universität Trier unterstützt werden, wurden im April 2000 aufgenommen. Ein Antrag auf Drittmittelförderung ist eingereicht worden. Vgl. die Homepage des Vorhabens unter dem URL: http://gaer27.uni-trier.de/MhdTA/welcome.htm .

Ferner bietet die elektronische Auswertung der Wörterbuchquellen auch einen substantiell lexikographischen Vorteil, die Möglichkeit nämlich, Exzerption und Belegauswahl voneinander zu trennen und die letztere zu einem späteren Zeitpunkt vorzunehmen, wo sie mit größerer lexikographischer Einsicht durchgeführt werden kann. Herkömmlich muß schon wegen des Aufwands, der mit dem Abschreiben oder Kopieren von einzelnen Belegstellen und mit der weiteren Bearbeitung der Belegzettel (Sortierung usw.) verbunden ist, gleich bei der Exzerption, also in der Regel – wenn nicht nach vorhandenen Wortverzeichnissen exzerpiert wird – bei der fortlaufenden Lektüre einer Quelle, eine Auswahl getroffen werden: „Bei traditioneller Arbeit ist Auswahl ein praktischer Notbehelf und findet immer zu früh statt [...]" (Sappler 1991:281). Die Ergebnisse dieser frühzeitigen Auswahl können bei der späteren Artikelarbeit, wenn die Sichtweise stichwortbezogen ist und die Belegauswahl besser begründet vorgenommen werden könnte, nur noch ganz begrenzt durch Nachexzerption korrigiert werden.[12] Bei elektronischer ‚Verzettelung' der Quellen in der vorgestellten Weise können dem Artikelbearbeiter grundsätzlich alle Belege einer Quelle für ein Stichwort zugänglich gemacht werden. Daß dann im Bereich des hochfrequenten Wortschatzes Maßnahmen gegen nicht mehr überblickbare Überfülle von Belegen getroffen werden müssen, ist klar; näheres dazu unten unter 4. zum Belegarchiv.

Der Aufwand, der für die Bereitstellung der elektronischen Quellenbasis des Wörterbuchs getrieben werden muß und auf den ersten Blick unverhältnismäßig hoch erscheinen mag, ist also nicht allein durch die effektive Organisation der computergestützten Arbeitsabläufe, sondern vor allem lexikographisch gut begründet. Er entlastet die eigentliche Ausarbeitung des Wörterbuchs, weil ein großer Teil der ‚niederen', bei der Ausarbeitung der Artikel störenden und ablenkenden Arbeit bereits bei der vorbereitenden Materialbereitstellung erledigt wird; und er erlaubt es, die Belegbasis des Wörterbuchs lexikographisch zuverlässiger gesteuert einzurichten.

Wenngleich das elektronische Textarchiv genuin lexikographisch begründet, als Materialbasis für den Zweck der Ausarbeitung des Wörterbuchs eingerichtet ist, so ist nicht zu übersehen, daß es einen über die Wörterbuchzwecke hinausgehenden eigenen Wert für andere, auch wörterbuchferne linguistische oder literaturgeschichtliche Untersuchungsanliegen besitzt. Geeignet dafür macht es gerade die gewählte ‚ausgabendiplomatische' Einrichtung, weil sie einerseits die zugrundeliegenden Editionstexte ohne spezielle, an den Wörterbuchzweck gebundene Änderungen abbildet, andererseits aber bereits eine reiche sachliche Auszeichnung bietet. Das elektronische Quellenarchiv des Wörterbuchvorhabens dürfte sich daher gut eignen als Kern einer zukünftigen digitalen mittelhochdeutschen Bibliothek im WWW.

3 Die Lemmatisierungskomponente

Der Kern der Lemmatisierungskomponente ist das ‚Gerüst'. Es enthält die allgemeinen, nicht einzelstellenbezogenen Lemmatisierungsinformationen, das heißt Lemmata und ihnen zuzuordnende Wortformen; homographe Wortformen, die mehreren Lemmata gemeinsam

---

[11] Plate/Recker (2000).

[12] Ausführlich zu diesem Problem und zu den im herkömmlichen, nicht-computergestützten Verfahren möglichen Gegenmaßnahmen Reichmann (1986), hier Abschnitt „6.2. Die Quellenbearbeitung", S. 48–51 und Reichmann (1990).

sind, werden durch angehängte Markierungen unterschieden (und ihre Vorkommen in den Texten entsprechend ausgezeichnet). Die formale Struktur des Gerüsts ist an anderer Stelle bereits beschrieben und durch Abbildung veranschaulicht,[13] hier soll es nur um seine lexikographische Funktion gehen.

Das Gerüst ist vor allem Lemmatisierungsinstrument, es dient also zunächst – im Zusammenspiel mit den Texten und gesteuert von Programmen, die mit Texten und dem Gerüst als Eingabe arbeiten – der Bereitstellung des lemmatisierten Belegarchivs. Als solches ist es dynamisch, d.h. es wächst mit jedem weiteren Text vor allem um die darin neu auftretenden Wortformen, in geringem Maße auch um neue Lemmata; letzteres dann, wenn Lemmata belegt werden, die noch nicht in der aus den Vorgängerwörterbüchern und dem FINDEBUCH kompilierten und vor der Bearbeitung der ersten Quellentexte in das Gerüst eingetragenen Lemmaliste verzeichnet sind. Grundsätzlich wäre es möglich gewesen, auch auf Lemmaebene mit einem leeren Gerüst zu beginnen und die Stichwörter sukzessive mit ihrer erstmaligen Belegung in den bearbeiteten Texten anzusetzen. Der Vorteil des gewählten Verfahrens, das sich im Falle des Mittelhochdeutschen die Existenz umfangreicher Vorgängerwörterbücher zunutze machen konnte, dürfte aber auf der Hand liegen: Die bei der Bearbeitung der Texte für das Belegarchiv sonst in jedem einzelnen Fall eines neu belegten Lemmas durchzuführende Arbeit der Festlegung der Gestalt des Lemmazeichens und der häufig nötigen zusätzlichen unterscheidenden Angaben (grammatische Angaben, Homographenindizes) wird erspart. Die bereits vorhandene Lemmaliste ist die Grundlage für die Erzeugung automatischer Lemmatisierungsvorschläge für neue, noch nicht im Gerüst vorhandene Wortformen. Außerdem bietet die Gesamtlemmaliste aus den Vorgängerwörterbüchern eine ungefähre Zielvorgabe und damit einen Anhaltspunkt für die Qualifizierung des Arbeitsstandes der Materialbereitstellung, jedenfalls insoweit, wie die Lemmaebene betroffen ist; für weitergehende Fragen ist der jeweilige Stand des Belegarchivs einzubeziehen, s. dazu unten unter 4.

Weil die allgemeine Lemmatisierungsinformation in einer besonderen Komponente gespeichert wird, ist die Lemmatisierung leicht veränderbar. Dies betrifft sowohl Änderungen bei einzelnen Wortformen wie auch systematische Abwandlungen der äußeren Zugriffsstruktur überhaupt. Der erste Fall tritt regelhaft auf bei der Korrektur der halbautomatischen Lemmatisierung jedes Textes, der erstmals für die Aufnahme seiner Belegstellen in das Belegarchiv bearbeitet wird. Hier handelt es sich unter anderem darum, die automatische Einordnung neuer Wortformen ins Gerüst zu kontrollieren und gegebenenfalls falsch eingeordnete Wortformen umzustellen zu einem anderen Stichwort. Würde diese allgemeine Lemmatisierungsinformation statt im Gerüst (Zuordnung einer Wortform zu einem Stichwort) in den Texten (Zuordnung bei jedem Vorkommen dieser Wortform) aufbewahrt, dann müßte die Korrektur bei allen Vorkommen einer Wortform in den ausgewerteten Texten ausgeführt werden; im Gerüst dagegen ist die Korrektur nur an einer einzigen Stelle nötig.

Ebenso leicht sind Änderungen auf Lemmaebene möglich. Das Bedürfnis kann aus verschiedenen Gründen auftreten, die Durchführung ist aber im herkömmlichen Zettelkastenverfahren mit erheblichem Aufwand verbunden, der die Revision einmal getroffener Entscheidungen über die Lemmatisierungsprinzipien erschwert. Im Falle des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS waren bzw. sind zum Beispiel verschiedene Lemmatisierungsgrundsätze, die der aus den Vorgängerwörterbüchern kompilierten Lemmaliste zugrundeliegen, nicht unumstritten; die Entscheidung darüber, ob sie letztlich beibehalten oder verworfen werden, kann sich allein an dem sachlich begründeten Dafürhalten

[13] Plate/Recker (2000).

orientieren, weil der Arbeitsaufwand für die gegebenenfalls erwünschten Änderungen im Gerüst verhältnismäßig gering wäre. Erwogen wurde zum Beispiel, ob die trennbaren Partikelverben, ein Massenphänomen bereits im Mittelhochdeutschen, die in der Lemmaliste als artikelwertige Stichwörter an alphabetischer Stelle erscheinen, wie im BMZ unter den Simplexverben mitbehandelt oder wie im LEXER in Reihenartikeln zusammengefaßt werden sollten.[14] Weitere Beispiele, bei denen die Änderung in Richtung auf Zusammenlegung zu größeren Einheiten diskutiert worden ist, betreffen die in der Lemmaliste bislang ebenfalls als eigene artikelwertige Stichwörter behandelten Ansätze der Adjektivadverbien[15] und der *ge*-präfigierten Verben.[16] Alle erwogenen Änderungen könnten mit verhältnismäßig geringem Aufwand jederzeit bis zum Beginn der Ausarbeitung durchgeführt werden, ohne daß Arbeiten in den bis dahin ausgewerteten Texten selbst nötig würden. Selbstverständlich sind Umstellungen aller Art ferner auch, unabhängig von der letztlich im Wörterbuch durchzuführenden Zugriffsstruktur, für beliebige andere Zwecke jederzeit möglich.

Eine weitere Benutzungsmöglichkeit des Gerüsts sei hier angedeutet (vgl. dazu ausführlicher unten unter 5.2.1): Es kann durch eine einfache automatische Umordnung in ein alphabetisches Register aller im Belegarchiv bezeugten Schreibformen verwandelt werden, in dem von jeder Schreibform der bearbeiteten Quellentexte auf das bzw. die Lemmata verwiesen wird, dem oder denen sie zuzuordnen sind.

Das Gerüst ist zusammen mit den zugehörigen Programmen ein wirkungsvolles Instrument für die halbautomatische Lemmatisierung mittelhochdeutscher Texte und kann auch außerhalb des Wörterbuchprojekts benutzt werden. Bedarf dazu dürfte vor allem auf Seiten der Herausgeber mittelhochdeutscher Texte bestehen, die im Zuge ihrer Editionsarbeiten Konkordanzen benötigen und/oder ihren Ausgaben Glossare oder Wörterbücher beigeben wollen. Lemmatisierte Konkordanzen sollten eigentlich zum selbstverständlichen Rüstzeug eines jeden Herausgebers gehören, ihre Anfertigung erschien aber noch vor nicht allzu langer Zeit auch bei computergestützter Arbeit so aufwendig, daß man häufig darauf verzichtete oder sich mit einfachen Wortformenindizes behalf. Selbst die publizierten Produkte bleiben noch bis in die jüngste Zeit hinein beträchtlich hinter dem zurück, was angesichts der heutigen Möglichkeiten als Standard für derartige Hilfsmittel gelten müßte.[17]

4 Das Belegarchiv

Das Belegarchiv wird aus den dafür vorbereiteten Quellentexten mit der Hilfe des Gerüsts automatisch erzeugt. Es handelt sich um eine dynamische Größe, die mit jedem weiteren in geeigneter Weise präparierten Text wächst. Die Ausgabemöglichkeiten sind vielfältig, es kann in Konkordanzform mit frei wählbarer Kontextlänge vollständig oder für beliebige Teilstrecken, auf der Grundlage des gesamten Textarchivs oder von Teilcorpora ausgegeben und entweder auf Papier ausgedruckt oder mit verschiedenen Recherchemöglichkeiten elektronisch z.B. in einem Internetbrowser über die Homepage des Projekts zur Verfügung gestellt werden.[18] Zur Zeit dient die Ausgabe des Belegarchivs vor allem der

[14] Ausführliche Diskussion dieser Frage bei Plate (2000a).
[15] Vgl. Baumgarte (2000).
[16] Vgl. Tao (2000) und Wawer (2000).
[17] Vgl. die Übersicht bei Gärtner/Kühn (1998).

Kontrolle der halbautomatischen Lemmatisierung, daneben wurde es auch bereits für die Anfertigung von Probeartikeln benutzt (vgl. dazu unten unter 5) und ferner für verschiedene Recherchen im Zusammenhang mit der Diskussion konzeptioneller Fragen.

Umfang und Zusammensetzung des Belegarchivs hängen ab vom Umfang der bearbeiteten Texte und vom Grad ihrer Auswertung für das Belegarchiv. Oben unter 2 wurde bereits darauf hingewiesen, daß bei der Eingabe von Volltexten grundsätzlich jeder Beleg eines Textes ins Belegarchiv aufgenommen werden kann. Dies ist ein Vorteil gegenüber dem herkömmlichen Verfahren, bei dem Exzerption und Belegauswahl weitestgehend zusammenfallen und der späteren Artikelbearbeitung stark vorgreifen, indem sie ihr in teils beträchtlichem Umfang Belege vorenthalten, obwohl in der Phase der Exzerption die Entscheidung über die Belegauswahl wesentlich schlechter begründet durchzuführen ist als bei der eigentlichen Artikelarbeit, wo das Material erstmals in stichwortbezogener Zusammenstellung in den Blick kommt. Andererseits würde eine unbeschränkte Aufnahme aller Belege auch des höchstfrequenten Wortschatzes (vor allem der Synsemantika) eine für den Artikelbearbeiter kaum noch überschaubare Belegfülle ergeben. Für das Belegarchiv des mittelhochdeutschen Wörterbuchs werden die Quellen daher in unterschiedlicher Dichte ausgewertet. Das Grundkorpus des Belegarchivs gliedert sich entsprechend der Exzerptionsdichte in drei Gruppen. Die erste Quellengruppe aus insgesamt 18 Texten und Textsammlungen[19] wird vollständig für das Belegarchiv ausgewertet, das heißt jeder Beleg dieser Texte, einschließlich aller Belege für den höchstfrequenten Wortschatz, erscheint im Belegarchiv. Bei der zweiten, insgesamt 21 Texte umfassenden Quellengruppe[20] wird der höchstfrequente Wortschatz für das Belegarchiv ausgeblendet, die Belege für den übrigen Wortschatz gehen aber noch vollständig ins Belegarchiv ein. Nur die in vorhandenen Ausgabenglossaren, Wortverzeichnissen und Registern verzeichneten Belege werden schließlich in der dritten, 34 Texte zählenden Quellengruppe erhoben.[21] Das gezielte Ausblenden von Belegen für hochfrequenten Wortschatz (zweite Quellengruppe) bzw. die gezielte Aufnahme von Glossarbelegen (dritte Quellengruppe) wird technisch durch spezielle Markierungen an den betreffenden Wortformenvorkommen bewirkt.

Das Grundkorpus wird nach unserer Erwartung und den ersten Erfahrungen an Probeartikeln die Forderung nach einem „die Homogenität sichernden, möglichst breiten zentralen Teil" des gesamten Quellenkorpus (Reichmann 1990:1603, hier T3) erfüllen können. Für die weitere „Erfassung peripherer Wortschatzteile" (ebd.; ergänze: und peripherer Gebrauchsweisen des zentralen Wortschatzes) wird eine größere Zahl zusätzlicher Quellen herangezogen werden müssen. Systematisch vorbereitet werden kann das zum einen für jene durch das FINDEBUCH erschlossenen Quellen, die nicht bereits im Grundkorpus vertreten sind[22] und durch geeignete computergestützte Verfahren der Auswertung der beiden Vorgängerwörterbücher, die durch eine der Trierer Arbeitsstelle angeschlossene Arbeitsgruppe maschinenlesbar gemacht wurden und miteinander (und mit dem FINDEBUCH) verknüpft im WWW für elektronische Benutzung publiziert werden.[23]

Einige Zahlen zum Stand des Belegarchivs im Juli 1999 mögen eine Vorstellung vom Umfang der Sammlung vermitteln. Insgesamt sind nach der Bearbeitung aller Texte der

---

[18] Die Nutzungsmöglichkeit über die Projekthomepage bleibt zur Zeit noch den Arbeitsstellen vorbehalten, soll jedoch bald frei zugänglich gemacht werden.

[19] Vgl. in der Bibliographie im Anhang zu Gärtner (2000) und auf der Homepage des mittelhochdeutschen Wörterbuchs die Nummern P1–P9 und A1–A9.

[20] Vgl. die Bibliographie [wie Fußnote 18] unter B1–B21.

[21] Vgl. die Bibliographie [wie Fußnote 18] unter C1–C34.

[22] Vgl. die Bibliographie auf der Homepage des Projekts unter ‚Findebuchquellen'.

ersten, von acht Texten der zweiten und von neun Texten der dritten Quellengruppe des Grundkorpus 18.341 Lemmata mit zusammen rund 960.000 Stellen im Belegarchiv vertreten; die acht Texte der zweiten Quellengruppe enthalten für den höchstfrequenten Wortschatz weitere rund 350.000 Belege, die aber für das Belegarchiv ausgeblendet worden sind. Die weitere Bearbeitung der zweiten und dritten Quellengruppe wird die Zahl der im Belegarchiv ausgeblendeten Stellen (aus der zweiten Quellengruppe) stark erhöhen und insgesamt die Relation Lemma/Belege zugunsten der Lemmata verändern, d.h. es werden, vor allem durch die dritte Gruppe der gezielt nach Ausgabenglossaren bearbeiteten Quellen, verstärkt Belege für niederfrequenten Wortschatz hinzukommen.[24] Der niederfrequente Wortschatz macht in allen lexikalischen Corpora den weitaus größten Teil der Stichwörter und den geringsten Teil der Belege aus: Zur Zeit (September 1999) sind im Belegarchiv 7.156 Lemmata, das sind knapp 40 Prozent aller belegten Stichwörter, mit nur einem einzigen Beleg vertreten. Bei herkömmlicher Exzerption wird dieser periphere Wortschatz verhältnismäßig stark bevorzugt berücksichtigt. Die Feststellung Oskar Reichmanns, daß in der Belegsammlung des FWB der „allgemeinsprachliche Wortbestand" und seine „eigentlichen, allgemeinsprachlichen und geschichtlich relativ konstanten" Gebrauchsweisen verhältnismäßig zu gering belegt seien,[25] legt diese Schwäche der herkömmlichen Belegexzerption und Belegauswahl offen. Die breite Exzerption des häufiger belegten Kernwortschatzes, die bei elektronischer Auswertung der Quellen möglich ist, vergrößert für den Artikelbearbeiter in diesem Bereich die Menge des Überblickbaren und eröffnet daher die Möglichkeit, den Gebrauch des Kernwortschatzes im Wörterbuch angemessener als bislang zu beschreiben. Dies scheint uns entschieden zu den Chancen der computergestützten Lexikographie zu gehören.

Wie das Textarchiv, so ist auch das Belegarchiv über die Zwecke des Wörterbuchs hinaus für andere Anliegen benutzbar. Seine Bereitstellung im Internet für elektronische Nutzung mit verschiedenen Recherchemöglichkeiten wird zur Zeit vorbereitet, für die beiden Arbeitsstellen ist der Zugriff über einen internen Teil der Homepage des Projekts bereits realisiert. Bei herkömmlicher Arbeit waren Wörterbuch-Belegarchive für den weiteren Kreis der potentiell an solchen Sprachmaterialien Interessierten nur sehr schwer zugänglich; man mußte sich eben in die Arbeitsstelle selbst an die Zettelkästen begeben oder riskieren, mit umständlichen Anfragen an die Wörterbuchbearbeiter den Betrieb aufzuhalten. Die ungehinderte und bequeme Zugänglichkeit des Belegarchivs für jedermann ist ein Vorzug der elektronischen Materialbereitstellung.

5 Das Programmsystem für die Artikelarbeit

Die Erfahrungen mit der Artikelarbeit beschränken sich in der jetzigen Phase der Materialbereitstellung auf wenige gößere Artikel, die zur Erprobung konzeptioneller Überlegungen verfaßt wurden.[26] Aus ihnen ergibt sich jedoch schon eine Reihe von Gesichtspunkten hinsichtlich der Chancen und Perspektiven dieser Phase der computergestützten Wörterbucharbeit, die im folgenden benannt seien.

---

[23] Vgl. dazu Burch/Fournier/Gärtner 1998. – Selbstverständlich dienen die Vorgängerwörterbücher dabei nur als Wegweiser zu Belegmaterial, dieses muß dann aus den Quellen erhoben werden.

[24] Die Relation Lemma : Belege im Belegarchiv beträgt bei den Texten der ersten Quellengruppe 1:43, in der zweiten Quellengruppe 1:21, in der dritten 1:10.

5.1 Leistungsmerkmale des Artikelredigierprogramms

Es dürfte grundsätzlich Einigkeit in der Erwartung bestehen, daß die eingangs erwähnte Option, mit Beginn der Phase der Ausarbeitung des Wörterbuchs zum traditionellen Verfahren der Artikelarbeit überzugehen, die eigentlichen Chancen der elektronischen Materialbereitstellung verspielen würde. Andererseits ist es offensichtlich, daß es mit der einfachen Ausgabe der Belege in ein Textverarbeitungsprogramm nicht getan ist, sondern daß ein eigenes System von Programmen für die Artikelarbeit benötigt wird, für dessen Realisierung philologische Einsicht in die erforderlichen Leistungsmerkmale ebenso wie anspruchsvolle Programmierkenntnisse benötigt werden. Die Ermutigung, diese Herausforderung anzunehmen, kam vor allem aus der überzeugenden Demonstration eines wiederum von Paul Sappler (Tübingen) entwickelten Prototypen, der bereits 1988 verfügbar war.[27] Er wurde anläßlich der Abfassung von Probeartikeln in den Arbeitsstellen des mittelhochdeutschen Wörterbuchs vom Programmautor für die Bedürfnisse des Projekts eingerichtet und bietet das Vorbild für die Entwicklung eines umfassenderen Programmsystems für die Artikelarbeit.

Die Leistungsmerkmale des Artikelredigierprogramms lassen sich in drei Hauptgruppen einteilen. Die erste Gruppe unterstützt vorbereitende Sortierläufe, denen bei herkömmlicher Arbeit das Ordnen der Belegzettel für ein Stichwort nach bestimmten äußeren Kriterien entspricht, z.B. alphabetisch nach Wortformen, nach Quellen in chronologischer Reihenfolge oder einem beliebigen anderen quellentypologischen Merkmal, innerhalb der Belege einer Quelle aufsteigend nach Vorkommen usw. Ferner gehört dazu die Unterstützung einer inhaltlichen Sortierung nach ausdrucksseitig abfragbaren Merkmalen der Textumgebung; wie bei der Artikelarbeit Kontextabfragen im elektronischen Belegmaterial zu einem Stichwort sinnvoll eingesetzt werden können, zeigt an einem Beispiel Sappler (1999:280).

Die zweite Gruppe betrifft die eigentlichen Editierfunktionen. Dabei handelt es sich um Unterstützungen

- für das Umordnen der Belege nach Gebrauchstypen
- für die Einfügung von Gliederungsmarken und entsprechenden Gliederungskommentaren (Bedeutungserläuterung usw.)
- für die Verdopplung von Belegen, die mehr als einmal im Artikel zitiert werden sollen
- für das Ausblenden von Belegen, die nicht im Artikel erscheinen sollen, oder des Belegzitats, wenn nur die Stelle angegeben werden soll
- für die mit dem Belegzitieren im engeren Sinne zusammenhängenden Arbeiten wie Festlegung des Belegschnitts, (gekennzeichnete) Kürzung oder Ergänzung, einzelstellenbezogene Kommentierung einschließlich der Möglichkeit, alle diese Eingriffe wieder aufzuheben und den originalen Wortlaut automatisch zu restituieren
- schließlich für das Verweisen innerhalb des Artikels und auf andere Artikel.

Bei der dritten Gruppe von Funktionen des Artikelredigierprogramms handelt es sich um durch Prüfprogramme unterstützte Sicherungsfunktionen, die z.B. Datenteile gegen unerlaubte Aktionen des Artikelautors schützen (vor allem die Fundstellenangabe, die unantastbar ist, aber auch die Belege selbst, die zwar in jeder beliebigen Weise geändert oder

---

[25] Lexikographische Einleitung (1986), S. 160 (zu den Gründen vgl. dort S. 48–51).

[26] Sie sind zum Teil publiziert in den Beitragen des von Gärtner/Grubmüller (2000) herausgegebenen Tagungsbandes.

[27] Vgl. die anschauliche Darstellung bei Sappler/Schneider-Lastin 1991.

auch mit einer Markierung für Ausblendung versehen, nie aber aus der Artikelarbeitsdatei völlig gelöscht werden dürfen) und verschiedene weitere formale Konsistenzmerkmale geltend machen.

Die Erfahrungen mit dem bereits realisierten Leistungsumfang des Artikelredigierprogramms sind so positiv, daß eine Rückkehr zur herkömmlichen Arbeit mit Belegzetteln oder dem Ausformulieren der Artikel in einem Textverarbeitungsprogramm für uns nicht mehr in Frage kommt. Als Hauptvorteil erscheint uns, daß die Aufmerksamkeit in höherem Maße der eigentlichen philologischen Arbeit im Belegmaterial zugewendet werden kann, während die lästigen Schreib-, Kontroll- und Sortierarbeiten weitgehend der Maschine übertragen sind. Ein nicht zu unterschätzender Vorteil ist es auch, daß mit der Hilfe eines anschließenden Satzprogramms jederzeit eine Vorschau auf den gesetzten Artikel möglich ist.

Abschließend seien zwei Eigenschaften der elektronischen Artikeltexte hervorgehoben, die für die Funktionsweise eines Programmsystems mit dem beschriebenen Leistungsumfang unerläßlich sind, daneben aber auch der späteren Benutzung des elektronischen Wörterbuchs (vgl. dazu unten unter 6) zugute kommen: Zum einen die Tatsache, daß die Belegzitate mit dem elektronischen Belegarchiv und damit den ihm zugrundeliegenden elektronischen Quellentexten verknüpft sind, zum anderen die Strukturierung des Artikeltexts in Felder (dies ist vor allem für die Prüfprogramme nötig). Ersteres ermöglicht es, direkt aus der elektronischen Fassung des Wörterbuchs heraus die für ein Belegzitat angegebene Fundstelle in einem Quellentext anzusprechen (und so das Belegzitat in seinem ursprünglichen Gebrauchszusammenhang zu prüfen, vgl. unter 6.1), die zweite Eigenschaft erlaubt das auf Felder beschränkte Suchen im elektronischen Wörterbuchtext (vgl. unter 6.2.2).

5.2 Mehrwert der Artikelarbeitsdateien gegenüber dem Artikel im publizierten Wörterbuch

Vor allem bei hochfrequentem Wortschatz und großen Artikeln haben die Artikelarbeitsdateien einen gewissen Mehrwert gegenüber dem fertiggestellten Artikel, der im publizierten Wörterbuch erscheint. Dies ergibt sich daraus, daß die Belegauswahl zu einem großen Teil erst während der Artikelarbeit vorgenommen wird und nicht bereits bei der Auswertung (Exzerption) der Quellen für das Belegarchiv. Bei der Artikelarbeit kann sich die Auswahl bereits auf eine Ordnung des Belegmaterials nach Gebrauchstypen (‚Bedeutungen' usw.) stützen. Im Ergebnis wird daher bei höherfrequentem Wortschatz für bestimmte Artikelpositionen immer eine größere Menge des Belegmaterials gesichtet und grob geordnet worden sein als im publizierten Wörterbuchartikel zitiert wird. Für Spezialuntersuchungen, feinere Durcharbeitung usw. steht in den Artikeldateien dann also bereits weiteres vorsortiertes Material zur Verfügung. Ähnliches gilt für bestimmte Materialordnungen, die auf dem Weg zur endgültigen Artikelgliederung erprobt, bei der Weiterarbeit dann aber als im Artikelzusammenhang nicht aussagekräftig verworfen worden sind; für andere Untersuchungszwecke können sie jedoch durchaus ihren Wert haben.

6 Konzeptionelle Fragen der Publikation des Wörterbuchs für elektronische Nutzung

Als erstes historisches Belegwörterbuch des Deutschen wird das neue MITTELHOCHDEUTSCHE WÖRTERBUCH vollständig computergestützt erarbeitet. Quellen, Belegarchiv und Artikelarbeit bis hin zum Satz der Artikel sind Zustände und Prozesse eines geschlossenen elektronischen Kreislaufs. Daher besteht selbstverständlich auch die Möglichkeit der Publikation des Wörterbuchs selbst in den ‚neuen Medien' WWW und CD-ROM; ja man könnte die elektronische Publikation sogar als das Gegebene, die Ausgabe im Druck dagegen als auffälligen sekundären Medienwechsel und als zusätzlichen Schritt des lexikographischen Prozesses auffassen.

Die Möglichkeiten der neuen elektronischen Publikationsformen werden aktuell auch unter Wörterbuchmachern engagiert diskutiert, wobei eine begrüßenswerte Verlagerung der Aufmerksamkeit von eher unspezifischen, gelegentlich geradezu trivialen Gesichtspunkten hin zu substantiell lexikographischen Fragen zu beobachten ist. Die ‚neuen Medien' verstärken also in erfreulicher Weise die Forderung nach konzeptioneller Reflexion und Explizitheit, der die Wörterbuchmacher bereits durch die seit den 70er Jahren auch im deutschen Sprachbereich blühende theoretische Wörterbuchforschung und Wörterbuchkritik ausgesetzt sind und deren erste reiche Ernte in Oskar Reichmanns ‚Lexikographischer Einleitung' (1986) zum FRÜHNEUHOCHDEUTSCHEN WÖRTERBUCH gehalten wurde.

Zwei konzeptionelle Probleme der mit den neuen Medien verbundenen „Chancen und Perspektiven" historischer Belegwörterbücher seien hier diskutiert. Sie betreffen das Verhältnis des ausgearbeiteten Wörterbuchs zu seiner Basis (Quellen und Belegarchiv) und das Verhältnis der Benutzung des Wörterbuchs über Artikelstichwörter zur Benutzung über andere, registerartige Zugriffshilfen.

6.1 Das Verhältnis von Wörterbuch und Wörterbuchbasis

Die Frage nach dem Verhältnis von Wörterbuch und Wörterbuchbasis ergibt sich bei elektronischer Publikation dadurch, daß die Möglichkeit der Verknüpfung sowohl mit dem lemmatisierten Belegarchiv wie auch mit den Quellentexten selbst besteht; beide können dann direkt aus einem Wortartikel heraus angesprochen werden. Im Falle des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS ist diese Verknüpfung in der Datenbasis durch die Art der Materialbereitstellung und die Anlage des Programmsystems für die Artikelarbeit (vgl. dazu oben unter 5.1) von vornherein gegeben, sie muß also nicht nachträglich hergestellt werden.

Das Belegarchiv enthält in der Regel mehr Belege als im Wörterbuch-Artikel zitiert werden, es bietet also für Spezialuntersuchungen zusätzliches Belegmaterial und es erlaubt – vor allem bei Unterstützung durch entsprechende Editorfunktionen – die Arbeit mit ihm. Das ist für die Wörterbuchbenutzung unter bestimmten speziellen Fragestellungen unbedingt ein Zugewinn der elektronischen Publikation, und insgesamt möglicherweise ein Anreiz zu selbständigerem, kritischerem Umgang mit dem Wörterbuch. Um falschen Erwartungen vorzubeugen, ist aber daran zu erinnern, daß das elektronische Belegarchiv die Form einer lemmatisierten KWIC-Konkordanz hat, die noch keine weiteren lexikographischen Strukturierungen aufweist. „Das Lesen in einer Konkordanz kann für den Philologen durchaus anregend sein; man nimmt in Kauf, daß es kein Gespräch mit einem dar-

stellenden Lexikographen ist und man nicht auf Beobachtungen und Wissen anderer aufbaut, sondern selber ganz von unten anfängt" (Sappler 1991:278). Der substantielle Unterschied zwischen Rohmaterialien und Wörterbuch darf im Interesse der Benutzer also nicht verwischt werden. Der Regelfall der Konsultation verlangt einen ausgearbeiteten Artikel, der für den kundigen Benutzer des betreffenden Werks aus sich heraus verständlich ist; die Bereitstellung mehr oder weniger unbearbeiteter Materialien kann hier nur ein zusätzliches Angebot an diejenigen sein, die bereit und in der Lage sind, ihre Fragestellung ohne lexikographische Hilfe zu verfolgen.

Die Verknüpfung des Wörterbuchs mit den Quellentexten kann theoretisch dem Zugriff in beiderlei Richtung dienen: Zum einen von den im Wörterbuch zitierten Belegen zu den jeweiligen Stellen in den Quellentexten; zum anderen von einer Textstelle (Wortform) zum betreffenden Wörterbuchartikel. Einleuchtend ist vor allem der Weg vom Belegzitat zum Quellentext, der die Prüfung einer Stelle in ihrem ursprünglichen Gebrauchszusammenhang erlaubt.

Auch die umgekehrte Richtung sieht auf den ersten Blick vielversprechend aus, bei näherem Hinsehen erweist sie sich jedoch als problematisch. Reizvoll ist aus der Sicht der Wörterbuchmacher die Vorstellung, daß man auf diese Weise die Schwelle zur Benutzung des großen Belegwörterbuchs absenken und bereits während der Ausarbeitung des Wörterbuchs eine größere Benutzergemeinde aufbauen könnte; dies wäre besonders im Falle des Mittelhochdeutschen zu begrüßen, weil leider immer wieder festgestellt werden muß, daß nur wenige der potentiellen Benutzer den Umstieg vom TASCHENLEXER auf die großen Wörterbücher bewältigen.[28] Doch könnte eine solche Verknüpfung von Quellentexten mit dem Wörterbuch bei den Benutzern leicht Erwartungen wecken, die in der Regel enttäuscht werden würden. Denn da nur ein geringer Teil der Stellen jeder Quelle im Sprachwörterbuch zitiert wird, und da ferner die lexikographische Erläuterung dort in aller Regel nicht einzelnen Textstellen, sondern Typen des Gebrauchs einer Wortschatzeinheit gilt, leistete diese Verknüpfungsrichtung meist nicht mehr als die Hinführung zu dem einschlägigen Artikel des großen Wörterbuchs. Diese Hilfe kann erwünscht sein von Benutzern, denen die Lemmatisierungs- und Normalisierungsprinzipien des betreffenden Wörterbuchs nicht vertraut sind, doch ist sie gerade bei elektronischer Publikation des Wörterbuchs rationeller durch ein Schreibformenregister und gegebenenfalls durch weitere Hilfs-Zugriffsstrukturen des Wörterbuchs zu realisieren (vgl. dazu unten unter 6.2.1). Wer darüber hinaus einzelstellenbezogene Verständnishilfen sucht, der konsultiert auch in Zukunft besser die dafür gedachten textbezogenen Hilfsmittel wie Stellenkommentare, Ausgabenglossare oder Textwörterbücher.

6.2 Benutzung des Wörterbuchs über Artikelstichwörter, über Register und über registerähnliche Hilfen

Wenn es um die Chancen und Perspektiven der neuen elektronischen Publikationsformen geht, findet unter Lexikographen der Gesichtspunkt erweiterter Zugriffsmöglichkeiten auf das Wörterbuch gegenwärtig besondere Aufmerksamkeit. Das gedruckte Werk, so die Überlegung, erlaubt in der Regel nur eine einzige Zugriffsmöglichkeit, jene über die Arti-

[28] Zum Zusammenhang von BMZ, LEXER und TASCHENLEXER vgl. Plate (1997). Das neuerdings mit dem TASCHENLEXER konkurrierende ‚Kleine mittelhochdeutsche Wörterbuch' von Beate Hennig ist von vornherein nicht auf die Hinführung zur Benutzung der großen Wörterbücher angelegt und erschwert sie sogar; vgl. dazu Plate (2000b).

kelstichwörter. Die elektronische Fassung gewährt dagegen grundsätzlich freien Zugang über beliebig definierbare Merkmale des Wörterbuchtexts (Volltextsuche) bzw. in bestimmten Textteilen der sachlich strukturierten Wörterbuchdaten (Suche in Feldern, z.B. in bestimmten Artikelpositionen oder sachlich markierten Textteilen wie Belegtext, Erläuterungstext, Quellensiglen usw.).

6.2.1 Register

Die systematische Beschäftigung mit der Einrichtung zusätzlicher Zugriffsstrukturen für Wörterbücher ist verhältnismäßig jungen Datums, sie stammt aber noch aus der Zeit vor dem Siegeszug der neuen elektronischen Publikationsformen und bedurfte nicht der Inspiration durch deren technische Möglichkeiten, denn sie konnte sich an dem herkömmlichen Hilfsmittel orientieren, das im gedruckten (und lange zuvor schon im geschriebenen) Buch zusätzlich zur Hauptzugriffsstruktur (wie sie z.B. ein Inhaltsverzeichnis darstellt) weitere Zugriffsmöglichkeiten anbietet, dem Register also.

Im Bereich der historischen Lexikographie des Deutschen hat wohl erstmals Oskar Reichmann in der ‚Lexikographischen Einleitung' (1986) des FWB nachdrücklich auf die Registermöglichkeiten hingewiesen;[29] nicht weniger als 15 verschiedene Register zur Strecke *a* bis *ausgang* wurden 1995 vollständig oder in Ausschnitten vorgelegt und ausführlich diskutiert in einer von Reichmann gemeinsam mit Ulrich Goebel und Ingrid Lemberg verfaßten Monographie, deren einprägsamer (Haupt-)Titel „Versteckte lexikographische Information" inzwischen zum Schlagwort für eine bislang unerschlossene Wörterbuch-Dimension geworden ist. Er bezeichnet das „bekannte Faktum, daß Wörterbücher in aller Regel nur über eine einzige Zugriffsstruktur verfügen und daß damit alle Daten, die sich dieser Struktur entziehen, versteckt bleiben, so wichtig sie für andere lexikographische, linguistische, philologische, wortgeschichtliche sowie für kulturhistorische usw. Fragestellungen und außerdem zur Kontrolle der eigenen lexikographischen Tätigkeit auch sein mögen" (Goebel/Lemberg/Reichmann 1995: VII). Geeignete Register können solche „versteckte Information" zugänglich machen. Welche der zahlreichen möglichen Register realisiert werden sollten und können, hängt von den antizipierten Benutzungsanliegen, von der Ergiebigkeit des zugrundeliegenden Wörterbuchs hinsichtlich bestimmter Informationstypen und nicht zuletzt auch von dem je nach Registertyp ganz unterschiedlichen Aufwand ab, der für die Gewinnung der Registerdaten und ihre weitere Bearbeitung zu einem tauglichen Hilfsmittel für die jeweilige Fragestellung nötig ist.

Nicht bei allen Registern handelt es sich tatsächlich um eigenständige weitere Zugriffsstrukturen. Das von Reichmann in der ‚Lexikographischen Einleitung' zum FWB (1986:160) erwogene Schreibformenregister z.B., das alle im zitierten Belegmaterial erscheinenden Schreibformen enthält und zu jeder das betreffende Artikelstichwort bzw. (bei homographen Schreibformen) die möglichen Lemmatisierungen angibt, ist eine bloße Hilfsstruktur für den Zugriff über Artikelstichwörter, die sich von der herkömmlichen Realisierung in Form von Verweisartikeln nur durch ihren Umfang und die Zusammenfassung in einem eigenen Alphabet unterscheidet. Im Falle des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS läßt sich ein solches Register automatisch aus dem Gerüst erzeugen (vgl. oben unter 3).

---

[29] Im Abschnitt „21. Register zur Erschließung von Datentypen des Wörterbuchs", S. 158–163.

Eine weitere mögliche Hilfsstruktur, die bei Reichmann (1986) und bei Goebel/ Lemberg/Reichmann (1995) nicht besprochen wird, könnte dazu dienen, Abweichungen von den aus gegenwartssprachlichen Wörterbüchern gewohnten Lemmatisierungsprinzipien aufzufangen. In größerem Umfang treten sie in der historischen Lexikographie des Deutschen besonders im Bereich der Wortbildungen auf, speziell in der Tradition der mittelhochdeutschen Lexikographie bietet z.B. die Behandlung der trennbaren Partikelverben Anstoß.[30] Die Diskussion solcher strittigen Lemmatisierungsprinzipien kann sich ganz auf das sachlich Gebotene konzentrieren, wenn gewährleistet ist, daß abweichende Nachschlagegewohnheiten durch entsprechende Zugriffshilfen unterstützt werden. Dies gilt nicht ausschließlich, aber in besonderer Weise für die elektronische Wörterbuchfassung, weil bei ihrer Benutzung der Aufwand für das ‚Nachschlagen' generell stark reduziert ist.

Nicht um solche Hilfsstrukturen für den Zugriff über Artikelstichwörter, sondern um wirklich eigenständige andere Zugriffsmöglichkeiten geht es, wenn bei Goebel/Lemberg/ Reichmann (1995) und in ihrer Nachfolge von der Erschließung „versteckter lexikographischer Information" die Rede ist. Diese Idee, deren Realisierung, wie schon gesagt, nicht an die technischen Möglichkeiten der neuen Medien gebunden ist, wird gleichwohl in der gegenwärtigen Diskussion über die elektronische Benutzung von Wörterbüchern stark betont. Der Grund dafür sind bestimmte Vorteile, die elektronische Suchsysteme gegenüber dem herkömmlichen gedruckten Register bieten. Sie werden von Sappler (2000a) prägnant wie folgt charakterisiert: „Wo bisher ein – vom Schöpfer des Registers mit Überlegung formulierter – Registereintrag im Register aufgesucht wurde und wo anschließend den Referenzen in die so erschlossenen Grunddaten nachgegangen wurde, kann jetzt unmittelbar in den Grunddaten sichtbar gemacht werden, wo bestimmte, auch komplexe Abfragebedingungen erfüllt sind. Der Benutzer dieses „Registers" hat mehr zur Verfügung (nämlich die gesamten Grunddaten), kann genauer zielen (durch Kombination der Bedingungen) und kommt schneller ans Ziel (mit elektronischer Geschwindigkeit)." Andererseits ist nicht zu übersehen: „Meist aber, wenn auch nicht notwendig, bekommt er weniger Führung und Vorarbeit als durch ein traditionelles systematisches Register. Trotz guten Ansätzen scheint diese zweite Art des Umgangs mit Registern [d.h. der Umgang mit elektronischen Suchsystemen] methodisch erst am Anfang zu stehen" (ebd.).

Es bedarf heute keiner prophetischen Gaben mehr, um zu erkennen, daß in sehr absehbarer Zeit die Publikation von Nachschlagewerken in elektronischer Form, mindestens zusätzlich zum Druck, der Regelfall sein wird. Daher müssen sich die lexikographischen Langzeitprojekte bereits jetzt der Frage stellen, ob und gegebenenfalls wie den im Vorstehenden angedeuteten neuen Möglichkeiten der Wörterbuchbenutzung konzeptionell, also in der Anlage des Wörterbuchs selbst, Rechnung zu tragen ist. Zwei Bemerkungen allgemeinerer Art sollen dazu hier abschließend gemacht werden.

Zunächst gilt es festzuhalten, daß die Zugriffsstruktur alphabetisch-semasiologischer Wörterbücher, die herkömmlich in aller Regel als einzige realisiert ist, nicht eine beliebige unter mehreren und prinzipiell mit den Register-Zugriffsstrukturen gleichrangig ist, sondern diejenige, die dem usuellen Gebrauch des Wörterbuchs gemäß seinem genuinen Zweck (im Sinne von Wiegand 1998: 303ff.) dient; die von Goebel/Lemberg/Reichmann 1995 diskutierten Registerstrukturen dagegen bedienen, wie es in der oben einleitend zitierten Charakterisierung zutreffend heißt, „andere [...] Fragestellungen" als jene, auf die die sprachwissenschaftliche Darstellungsform ‚(alphabetisch-semasiologisches) Wörterbuch' zielt. Genaugenommen ist es daher auch unzutreffend, hinsichtlich des Registerziels von der Er-

[30] Vgl. Plate (2000).

schließung „versteckter lexikographischer Information" zu sprechen: Insofern die gemeinte Information eine lexikographische ist, steht sie im jeweiligen Artikelzusammenhang an der richtigen Stelle; wenn Register bestimmte Informationstypen aus dem Artikelzusammenhang isolieren und einen anderen Zusammenhang herstellen, ist in aller Regel keine lexikographische Information intendiert.[31] Ein Beispiel mag dies illustrieren. Die Angabe, daß ein bestimmtes Verb der Objektsprache in einer bestimmten Gebrauchsweise mit Genitiv oder Akkusativ oder Präpositionalobjekt konstruiert werden kann, ist im Zusammenhang des Wörterbuch-Artikels zu diesem Verb eine genuin lexikographische Information. Der entsprechende Eintrag in einem Register der Verbkonstruktionen unter einer Rubrik, die sämtliche im Wörterbuch dokumentierten Fälle der Konstruktion mit Genitiv als Objektskasus verzeichnete, wäre dagegen eine grammatische Information (die z. B. im Zusammenhang einer Untersuchung der Kasusfunktionen wertvoll sein könnte). Das gedachte Register der Verbkonstruktionen wäre ebensowenig als Mittel der Erschließung „versteckter lexikographischer Information" zu bezeichnen, wie man umgekehrt das Wortregister zum Syntaxteil einer Grammatik ein Hilfsmittel für die Erschließung „versteckter syntaktischer Information" nennen würde. Der beträchtliche Wert, den Wörterbuch-Register haben können, liegt also gerade darin, daß sie es erlauben, das Wörterbuch nicht nur als Wörterbuch zu benutzen, sondern auch zu anderen Zwecken, z.B. als Materialsammlung für grammatische Untersuchungen. Es ist vor allem eine Frage der für die Ausarbeitung des Wörterbuchs zur Verfügung stehenden Ressourcen, ob eine Wörterbuch-Arbeitsstelle es sich leisten kann, auch solche ‚uneigentlichen' Benutzungsweisen des Wörterbuchs zu unterstützen. Wenn sich die Wörterbuchmacher – wie Goebel/Lemberg/Reichmann (1995: 260) im Falle des FWB – dagegen entscheiden müssen, können sie es mit gutem lexikographischen Gewissen tun; ihrem Wörterbuch wird deswegen als Wörterbuch nichts Wesentliches fehlen.[32]

6.2.2 Volltextsuche und Suche in Feldern

Das eben Ausgeführte gilt unabhängig davon, ob das Wörterbuch im Druck oder elektronisch benutzt wird. Die folgende Bemerkung betrifft die speziellen Möglichkeiten elektronischer Publikation. Hier trifft die Idee erweiterter Zugriffsmöglichkeiten bei Lexikographen auf besonders fruchtbaren Boden, weil sie zum Teil ohne den abschrek-

---

[31] Eine Ausnahme bildet das Register des Erläuterungswortschatzes, wenn es zu einem onomasiologischen Umkehrwörterbuch ausgearbeitet wird. Dieser Typ von Wörterbuchregistern ist methodisch besonders schwierig, wie die Forschungsdiskussion zeigt (zuletzt Plate 1992 und Goebel/Lemberg/Reichmann 1995, 1–127). Entsprechend selten ist es bislang realisiert worden, im Bereich der historischen Lexikographie des Deutschen vgl. vor allem den ‚Neuhochdeutschen Index zum mittelhochdeutschen Wortschatz' von Koller/Wegstein/Wolf (1990, dazu Plate 1992) und die umfangreiche Probe eines onomasiologischen Umkehrwörterbuchs zum FWB bei Goebel/Lemberg/Reichmann (1995:53–127). Diese Probe, der anspruchsvollste Versuch in dieser Gattung überhaupt, bezieht sich auf die Strecke **a** bis **ausgang** des FWB und bietet für diese Strecke des Ausgangswörterbuchs nur die Artikel **a,A** bis **Ausgang** des Umkehrwörterbuchs. Welchen Umfang das vollständige Umkehrwörterbuch zum vollständigen FWB hätte, läßt sich daran ermessen. Eine größere Strecke hätte vermutlich manche methodischen Probleme noch deutlicher gemacht, zum Beispiel das der onomasiologischen Verweise (vgl. dazu Goebel/Lemberg/Reichmann 1995:35).

[32] Der Gesichtspunkt des Wertes der Register als Mittel der lexikographischen Selbstkontrolle ist dabei unberührt.

kenden Aufwand eigener registrographischer Anstrengungen der Wörterbuchmacher realisiert werden kann, nämlich in der Form der Volltextsuche oder der Suche in Feldern. Für die Beurteilung dieser beiden Zugriffsweisen ist es nützlich, sich den Unterschied zu einer durch Register unterstützten Wörterbuchabfrage vor Augen zu halten. Volltextsuche ist in einem elektronischen Wörterbuchtext immer möglich, sie liefert jedoch erstens in der Regel zu einem großen Teil unspezifische Daten (während beim Registermachen bereits die Rohdaten einen spezifischen Filter passiert haben, den der Textauszeichnung fürs Register), zweitens sind diese Daten unstrukturiert (für die Strukturierung sorgt der zweite Schritt des Registermachens, die Formulierung der Registereinträge, ihre Hierarchisierung, die Einfügung von Verweisen usw.). Entsprechend gering ist der Nutzen einer Volltextsuche für anspruchsvollere Fragestellungen. Auch die Suche in Feldern liefert unstrukturierte Daten, doch sind in diesem Fall die Rohdaten bereits in gewisser Weise gefiltert, weil sie nur in bestimmten Teilen des Textes, den Feldern eben, erhoben wurden. Wenn das elektronische Wörterbuch mehr bieten soll als die Möglichkeit einer bloßen Volltextsuche, andererseits aber die – an sich wünschenswerte – Ausarbeitung von eigenen Zugriffsstrukturen (Registern) für bestimmte Datentypen nicht geleistet werden kann, dann sehen sich die Wörterbuchmacher also auf die Unterstützung von gezielten Abfragen durch die Strukturierung der Artikeltexte in *Felder* verwiesen.

Verhältnismäßig aufwendig kann selbst eine einfache Feldstrukturierung (z.B. Stichwort, Formteil, grammatische Angabe, Erläuterungstext, Belegtext) sein, wenn ein im Druck erschienenes Wörterbuch nachträglich digitalisiert und für die elektronische Publikation aufbereitet wird, weil automatisch nur das ausgezeichnet werden kann, was im zugrundeliegenden gedruckten Werk durch explizite Angaben oder durch typographisch eindeutige Merkmale abgesetzt ist (generell dazu vgl. Dummer/Michaelis/Schlaefer [1998]). Beispiele für Retrodigitalisierungsvorhaben mit unterschiedlich großem Ehrgeiz – und entsprechend geringerem oder höheren Zeit- und Kostenaufwand – hinsichtlich der inhaltlichen Strukturierung sind der elektronische Verbund der vorhandenen mittelhochdeutschen Wörterbücher (BMZ, LEXER, FINDEBUCH)[33] und das Projekt der Retrodigitalisierung der älteren Bände des DRW.[34]

Weniger aufwendig ist die inhaltliche Strukturierung der Daten und damit die Vorbereitung für gezielte Recherchen in der elektronischen Fassung bei Wörterbüchern, die sich eines speziellen Programmsystems für die Unterstützung der Artikelarbeit bedienen, wie es das DRW seit Band IX (erschienen 1992–1996) tut[35] und zukünftig das neue Mittelhochdeutsche Wörterbuch (vgl. dazu oben unter 5). Denn schon die Programme, die verschiedene Aspekte der Artikelarbeit unterstützen, verlangen eine formal eindeutige Absetzung von bestimmten Elementen des Artikeltextes (z.B. des Stichworts, der Belegzitate, der Fundstellenangaben, des lexikographischen Kommentartextes), und die Absetzung beliebiger weiterer Datentypen (z.B. der verschiedenen Typen lexikographischen Kommentars wie: Angabe von Übersetzungsäquivalenten, von Bedeutungsverwandten oder Gegensatzwörtern, von räumlichen, zeitlichen oder textsortenspezifischen Gebrauchsbeschränkungen, von Konstruktionsangaben usw.) kann durch entsprechende formale Vorgaben in der Artikelarbeitsdatei unterstützt oder gefordert werden. Die Grenzen solcher vorbereitenden

---

[33] Zu den Problemen der strukturellen Auszeichnung für gezielte Recherchen vgl. Fournier (2000) und den Beitrag von Burch/Fournier in diesem Band.

[34] Vgl. dazu Speer (1998:15) mit ausdrücklicher Absetzung von dem bei den mittelhochdeutschen Wörterbüchern gewählten Digitalisierungsverfahren.

[35] Vgl. das Vorwort von Heino Speer zu Bd. IX.

Kennzeichnung bei der Artikelarbeit für spätere gezielte Feldabfragen im Wörterbuchtext wären dort überschritten, wo sie ein nachgeordnetes Anliegen zur Hauptsache machte und die Artikelarbeit selbst behinderte, indem sie z.B. die nötige Formulierungsfreiheit der Artikelverfasser durch schematische Vorgaben beschränkte oder ihre Aufmerksamkeit vom Artikelzusammenhang abzöge und stattdessen auf mögliche Registerzusammenhänge ablenkte.

7 Resümee

Der Beitrag erörtert Chancen und Perspektiven der computergestützten Erarbeitung eines historischen Belegwörterbuchs, wie sie sich den Verfassern nach fünfjähriger Vorarbeit für das neue MITTELHOCHDEUTSCHE WÖRTERBUCH darstellen. Die Ausarbeitung des Wörterbuchs wird voraussichtlich 2001 beginnen, innerhalb von 20 Jahren soll das in der Druckfassung auf vier Bände zu je ca. 1000 Seiten berechnete Werk abgeschlossen sein und dann an die Stelle der Vorgängerwörterbücher des 19. Jahrhunderts für den Quellenzeitraum 1050 bis 1350 treten können. Als jüngstes der größeren Vorhaben der historischen Beleglexikographie des Deutschen konnte das neue MITTELHOCHDEUTSCHE WÖRTERBUCH von vornherein vollständig auf EDV-Basis gestellt werden. Die Materialsammlung beruht auf einem für das Wörterbuch eingerichteten umfangreichen elektronischen Textarchiv mit Volltexten aller Quellen des Grundkorpus und auf einer aus den Vorgängerwörterbüchern kompilierten Lemmakandidatenliste, die rund 80.000 Artikelstichwörter umfaßt. Mit einem Programmsystem für die halbautomatische Lemmatisierung wird nach entsprechender Vorbereitung der Quellentexte aus diesen und einer eigenständigen Lemmatisierungskomponente, die aus der um die Wortformen der bearbeiteten Texte erweiterten Lemmaliste besteht, das lemmatisierte Belegarchiv erzeugt. Auf seiner Grundlage und mit der Hilfe eines weiteren Systems von Programmen, die die Artikelarbeit von einer ersten vorbereitenden Sortierung der Belege bis hin zum fertig gesetzten Artikel unterstützen, wird das Wörterbuch ausgearbeitet werden. Die vollständig elektronisch gestützte Erarbeitung des Wörterbuchs erlaubt es, neben der als Lieferungswerk erscheinenden Druckfassung das Wörterbuch und seine Materialbasis mit überschaubarem zusätzlichem Aufwand auch für elektronische Nutzung mit weiteren Recherchemöglichkeiten zu publizieren.

Die im vorliegenden Beitrag am Beispiel des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS erörterten Chancen der computergestützten Lexikographie liegen in der Möglichkeit, mit Rechnerhilfe die herkömmlichen Arbeitsgänge und ihre Abfolge in lexikographisch vorteilhafter Weise zu reorganisieren und zu rationalisieren. Ein großer Teil des herkömmlich mit dem Abschreiben, Korrigieren und Lemmatisieren einzelner Belegstellen verbundenen Arbeitsaufwands kann eingespart werden bei einmaliger Bereitstellung sorgfältig korrigierter und für die weitere Verarbeitung präparierter elektronischer Volltexte der Wörterbuchquellen und dem Einsatz eines halbautomatischen Lemmatisierungsverfahrens. Dies gestattet es insbesondere, die Belegauswahl, die herkömmlich bereits bei der Exzerption vorgenommen werden muß, zum größten Teil dem Artikelverfasser zu überlassen, der sie lexikographisch begründet vornehmen kann. Die Ausarbeitung der Artikel wird unterstützt von einem Programmsystem, das die mechanischen Sortier-, Schreib- und Kontrollarbeiten zum größten Teil automatisch erledigt, so daß der Artikelverfasser sich ungestörter der Interpretation des Belegmaterials und dem Formulieren des lexikographi-

schen Befundes widmen kann. Zugleich sind in der Artikeldatei durch ihre obligatorische formale Anlage bestimmte Mindestinformationen über die Struktur des Wörterbuchtextes mitkodiert, die später gezielte Recherchen in der elektronischen Fassung des Wörterbuchs erlauben. Im weiteren Sinne zu den Chancen der computergestützten Lexikographie zu zählen ist die Tatsache, daß im Zuge der Vorbereitung und Ausarbeitung des Wörterbuchs Materialien und Hilfsmittel entstehen, die zusätzlich zum Wörterbuch oder sogar unabhängig von ihm für andere linguistische und philologische Anliegen von Interesse sein können. Im Falle des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS sind dies vor allem das elektronische Textarchiv, das elektronische Belegarchiv und schließlich ein wirkungsvolles Instrument für die halbautomatische Herstellung lemmatisierter Indices und Konkordanzen zu beliebigen mittelhochdeutschen Texten.

Die diskutierten Perspektiven sind solche, die sich für die Wörterbuchbenutzung aus der Möglichkeit der elektronischen Publikation des Wörterbuchs ergeben und Fragen der Wörterbuchkonzeption aufwerfen. Diese betreffen zum einen das Verhältnis zwischen dem ausgearbeitetem Wörterbuch und den ihm zugrundeliegenden Materialien (vor allem Text- und Belegarchiv), wenn diese gemeinsam mit dem Wörterbuch elektronisch publiziert werden und die Möglichkeit besteht, sie direkt aus einem Wörterbuchartikel heraus anzusprechen. Zum anderen stellt sich die Frage, in welchem Umfang das Wörterbuch neben seiner eigentlichen Aufgabe auch nicht-usuellen Benutzungsanliegen, die Register oder registerähnliche Zugriffsmöglichkeiten verlangen, entgegenkommen soll.

8 Literatur

(a) Wörterbücher (gedruckte)

AWB = ALTHOCHDEUTSCHES WÖRTERBUCH. Auf Grund der von Elias von Steinmeyer hinterlassenen Sammlungen im Auftrage der sächsischen Akademie der Wissenschaften zu Leipzig. Hgg. Elisabeth Karg-Gasterstädt, Theodor Frings. Bisher 4 Bde. Berlin: Akademie-Verlag 1968ff.

BMZ = MITTELHOCHDEUTSCHES WÖRTERBUCH. Mit Benutzung des Nachlasses von Georg Friedrich Benecke ausgearbeitet von Wilhelm Müller und Friedrich Zarncke. Nachdruck der Ausg. Leipzig 1854–1866 mit einem Vorwort und einem zusammengefaßten Quellenverzeichnis von Eberhard Nellmann sowie einem alphabetischen Index von Erwin Koller, Werner Wegstein und Norbert Richard Wolf. 5 Bde. Stuttgart: Hirzel 1990.

DRW = DEUTSCHES RECHTSWÖRTERBUCH. Wörterbuch der älteren deutschen Rechtssprache. Hg. Heidelberger Akademie der Wissenschaften. Bisher 10 Bde. Weimar: 1930ff.

[1]DWB = DEUTSCHES WÖRTERBUCH von Jacob Grimm und Wilhelm Grimm. 16 Bd. in 32 Bdn. und Quellenverzeichnis. Leipzig 1854–1971.

[2]DWB = DEUTSCHES WÖRTERBUCH von Jacob Grimm und Wilhelm Grimm. Neubearbeitung hg. von der Akademie der Wissenschaften der DDR in Zusammenarbeit mit der Akademie der Wissenschaften zu Göttingen. Bisher Bd. 1, 2 (Lfg. 1–3), 6, 7 und 8 (Lfg. 1–7). Leipzig: Hirzel 1983ff.

FINDEBUCH = FINDEBUCH ZUM MITTELHOCHDEUTSCHEN WORTSCHATZ. Mit einem rückläufigen Index. Hg. Kurt Gärtner, Christoph Gerhardt, Jürgen Jaehrling, Ralf Plate, Walter Röll, Erika Timm (Datenverarbeitung: Gerhard Hanrieder). Stuttgart: Hirzel 1992.

FWB = FRÜHNEUHOCHDEUTSCHES WÖRTERBUCH. Hg. Ulrich Goebel, Oskar Reichmann. Begründet von Robert R. Anderson, Ulrich Goebel und Oskar Reichmann. Bisher Bd. 1, Bd. 2, Bd. 3 (Lfg. 1–3), Bd. 4 (Lfg. 1+2) und Bd. 8 (Lfg. 1). Berlin: de Gruyter 1986ff.

Koller, Erwin/Wegstein, Werner/Wolf, Norbert Richard (1990): Neuhochdeutscher Index zum mittelhochdeutschen Wortschatz. Stuttgart: Hirzel.

LEXER = Mittelhochdeutsches Handwörterbuch von Matthias Lexer. Zugleich als Supplement und alphabetischer Index zum Mittelhochdeutschen Wörterbuche von Benecke-Müller-Zarncke. Nachdruck der Ausg. Leipzig 1872–1878 mit einer Einleitung von Kurt Gärtner. 3 Bd. Stuttgart: Hirzel 1992.

Stackmann, Karl (1990), Wörterbuch zur Göttinger Frauenlob-Ausgabe. Unter Mitarbeit von Jens Haustein. Göttingen: Vandenhoeck & Ruprecht (=Abhandlungen der Akademie der Wissenschaften in Göttingen, Philologisch-Historische Klasse; Folge 3, Nr. 186).

TASCHENLEXER = Matthias Lexer (1992): MITTELHOCHDEUTSCHES TASCHENWÖRTERBUCH. Mit den Nachträgen von Ulrich Pretzel. 38., unveränderte Aufl. Stuttgart: Hirzel. – Vgl. auch: Matthias Lexer: Mittelhochdeutsches Taschenwörterbuch in der Ausgabe letzter Hand. 2. Nachdruck der 3. Aufl. von 1885 mit einem Vorwort von Erwin Koller, Werner Wegstein und Norbert Richard Wolf und einem biographischen Abriß von Horst Brunner. Stuttgart: Hirzel 1992.

(b) Forschungsliteratur (gedruckte)

Baumgarte, Susanne (2000): Vorstellung des Probeartikels *schœne*. – In: Gärtner, Kurt/Grubmüller, Klaus (2000).

Burch, Thomas/Fournier, Johannes/Gärtner, Kurt (1998): Mittelhochdeutsche Wörterbücher auf CD-ROM und im Internet. Der Einsatz von SGML in der Retrodigitalisierung lexikographischer Standardwerke. – In: Akademie-Journal. Mitteilungsblatt der Konferenz der deutschen Akademien der Wissenschaften 1998/2, S. 17–24.

Burch, Thomas/Fournier, Johannes (2000): Zur Anwendung der TEI-Richtlinien auf die Retrodigitalisierung mittelhochdeutscher Wörterbücher (in diesem Band)

Dummer, Sven/Michaelis, Frank/Schlaefer, Michael (1988): Zur Digitalisierung historischer Wörterbücher. – In: Lexikos 8 (Afrilex-Reeks/Series 8:1998), 194–222.

Fournier, Johannes (2000): Digitale Dialektik. Chancen und Probleme mittelhochdeutscher Wörterbücher in elektronischer Form. – In: Wörterbücher in der Diskussion IV. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. 85–108. Hg. v. Herbert Ernst Wiegand. – Tübingen: Niemeyer.

Gärtner, Kurt (2000): Quellenauswahl, Arbeitsverfahren, Zeitplanung und Kooperation. – In: Gärtner, Kurt/Grubmüller, Klaus (2000).

– und Grubmüller, Klaus (Hgg.) (2000): Ein neues Mittelhochdeutsched Wörterbuch. Prinzipien, Probeartikel, Diskussion. (= Nachrichten der Akademie der Wissenschaften in Göttingen. I Philologisch-historische Klasse. Jahrgang 2000, Nr. 1).

Gärtner, Kurt/Kühn, Peter (1998): Indices und Konkordanzen zu historischen Texten des Deutschen. Bestandsaufnahme, Typen, Herstellungsprobleme, Benutzungsmöglichkeiten. – In: Werner Besch u.a. (Hgg.) Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. 1. Halbband. (Handbücher zur Sprach- und Kommunikationswissenschaft 2). 2.,vollst. neubearb. u. erw. Aufl. Berlin/New York, 715–742.

Goebel, Ulrich/Lemberg, Ingrid/Reichmann, Oskar (1995): Versteckte lexikographische Information. Möglichkeiten ihrer Erschließung dargestellt am Beispiel des Frühneuhochdeutschen Wörterbuchs. – Tübingen: Niemeyer (= Lexicographica, Series Maior 65).

Lemberg, Ingrid (1996): Die Belegexzerption zu historischen Wörterbüchern am Beispiel des Frühneuhochdeutschen Wörterbuchs und des Deutschen Rechtswörterbuchs. – In: Herbert Ernst Wiegand (Hg.): Wörterbücher in der Diskussion II. Vorträge aus dem Heidelberger Lexikographischen Kolloquium, 83–102. Tübingen: Niemeyer.

Plate, Ralf (1992): Onomasiologische Umkehrlexikographie auf dem Prüfstand. Anläßlich des Erscheinens von: Erwin Koller/Werner Wegstein/Norbert Richard Wolf: Neuhochdeutscher Index zum mittelhochdeutschen Wortschatz. – In: Zeitschrift für Dialektologie und Linguistik 59, 312–329.

– (1997): Votum für ein kleines Belegwörterbuch zur mittelhochdeutschen Klassik. Zugleich ein Beitrag zur Kritik der beiden großen mittelhochdeutschen Wörterbücher und des TASCHENLEXER. In: Energeia 23 (Tokyo), 57–97.

– (2000a): Zum Lemmastatus und Buchungsort der trennbaren Partikelverben im neuen Mittelhochdeutschen Wörterbuch. – In: Gärtner, Kurt/Grubmüller, Klaus (Hgg.): Ein neues MITTELHOCHDEUTSCHES WÖRTERBUCH. Prinzipien, Probeartikel, Diskussion.
– (2000b): „Erster schneller Zugriff" oder Fehlgriff? Zum alten und zu einem neuen mittelhochdeutschen Wörterbuch für Studierende. – In: Zeitschrift für Dialektologie und Linguistik 67 [im Druck].
– und Recker, Ute (2000): EDV für Wörterbuchzwecke und neue lexikographische Arbeitsweisen. Erfahrungen beim Aufbau des elektronischen Text- und Belegarchivs für das MITTELHOCHDEUTSCHE WÖRTERBUCH [im Druck).
Recker, Ute/Sappler, Paul (1998): Aufbau des maschinenlesbaren Text- und Belegarchivs für das MITTELHOCHDEUTSCHE WÖRTERBUCH. – In: Rudolf Große (Hg.): Bedeutungserfassung und Bedeutungsbeschreibung in historischen und dialektologischen Wörterbüchern. Beiträge zu einer Arbeitstagung der deutschsprachigen Wörterbücher, Projekte an Akademien und Universitäten vom 7. bis 9. März 1996 anläßlich des 150jährigen Jubiläums der Sächsischen Akademie der Wissenschaften zu Leipzig. (=Abhandlungen der Sächsischen Akademie der Wissenschaften zu Leipzig. Philologisch-historische Klasse. Bd. 75, H. 1). 249–253. Stuttgart/Leipzig.
Reichmann, Oskar (1986): Lexikographische Einleitung. – In: FWB [s. oben unter (a)]. Bd. 1, 10–164.
– (1990): Formen und Probleme der Datenerhebung I: Synchronische und diachronische historische Wörterbücher. – In: Franz Josef Hausmann u.a. (Hgg.): Wörterbücher: Ein internationales Handbuch zur Lexikographie. 1588–1611. Berlin/New York.
Sappler, Paul (1991): Strukturierungs- und Auswahlhilfen bei Autorwörterbuch und Sprachwörterbuch. – In: Eijirō Iwasaki (Hg.): Begegnung mit dem „Fremden": Grenzen – Traditionen – Vergleiche. Akten des VII. Internationalen Germanisten-Kongresses, Tokyo 1990. Bd. 4, 277–281. München.
– (2000a): Probleme literarhistorischer und inhaltlicher Erschließung durch Register. (Im Druck.)
– (2000b): Prinzipien des EDV-Konzepts. – In: Gärtner, Kurt/Grubmüller, Klaus (2000).
Sappler, Paul/Schneider-Lastin, Wolfram (1991): Ein Wörterbuch zu Gottfrieds ‚Tristan'. – In: Kurt Gärtner, Paul Sappler, Michael Trauth (Hgg.): Maschinelle Verarbeitung altdeutscher Texte IV. Beiträge zum Vierten Internationalen Symposion, Trier 28. Februar bis 2. März 1988. 19–28. Tübingen.
Speer, Heino (1996): Vorwort. – In: DRW (s.o. unter [a]). Bd. IX, III–VII.
– (1998): Ein Wörterbuch, die elektronische Datenverarbeitung und ihre Folgen. – In: Akademiejournal 2/98. Mitteilungsblatt der Konferenz der Deutschen Akademien der Wissenschaften. Mainz. 11–16.
Tao, Jingning (2000): Vorstellung des Probeartikels *nëmen* (B). – In: Gärtner, Kurt/Grubmüller, Klaus (2000).
Wawer, Annne (2000): Vorstellung des Probeartikels *nëmen* (A). – In: Gärtner, Kurt/Grubmüller, Klaus (2000).
Wiegand, Herbert Ernst (1998): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. Berlin/New York: de Gruyter.

(c) Wörterbücher und Literatur im WWW

Homepage des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS im WWW. April 2000. http://gaer27.uni-trier.de/MhdWB .
Burch, Thomas/Fournier, Johannes/Gärtner, Kurt (1999): Mittelhochdeutsche Wörterbücher im Verbund [miteineinander verknüpfte digitale Versionen von BMZ, LEXER und Findebuch; diese vgl. oben unter (a)]. April 2000. http://gaer27.uni-trier.de/MWV-online/MWV-online.html .
Lemberg, Ingrid (1998): DRW-digital. Neue Wege zur versteckten lexikographischen Information. Vortrag gehalten auf dem 7th International Symposium on Lexicography, Kopenhagen, April 1998. April 2000. http://www.rzuser.uni-heidelberg.de/~q63/kopenhag.html .

Plate, Ralf (1997): [wie oben unter (b)]. April 2000. http://gaer27.uni-trier.de/MhdWB/plate97.pdf.
Speer, Heino (1999): Digitale Version des DRW [dieses vgl. oben unter (a)]. April 2000. http://www.uni-heidelberg.de/institute/sonst/adw/drw.
– (1998): [wie oben unter (b)]. April 2000. http://www.rzuser.uni-heidelberg.de/~cd2/drw/publikat/speer_akadjourn.htm.

Ralf Plate, Trier
Ute Recker, Trier

Gerd Richter

Das elektronische Flurnamenbuch – Innovationen in der Flurnamenforschung durch den Einsatz neuer Medien[1]

1 Einleitung

1988 stellte Halfer fest: „Die Flurnamenforschung im deutschsprachigen Raum ist heute nicht geprägt von spektakulären Neuansätzen, sondern von der ständigen Verfeinerung der Untersuchungsmethoden." (1988:12)

Erwähnenswerte Ausnahmen von dieser Entwicklung stellen lediglich die Arbeiten und Forschungsmethoden des Anfang der 80er Jahre u.a. von Hans Ramge wiederbelebten „Hessischen Flurnamenarchivs in Gießen" dar. Zu nennen sind hier exemplarisch

- der im Vergleich zur restlichen Philologie (ganz zu schweigen von der Flurnamenforschung) frühe und konsequente Einsatz der Elektronischen Datenverarbeitung Mitte der 80er Jahre,
- der „Hessische Flurnamenatlas" (Ramge 1987a), als nicht nur erstem Flurnamenatlas überhaupt, sondern als bis dato einzigem Namenatlas, der vollständig (von der Datenerfassung bis zur Kartenerstellung) computativ bearbeitet wurde, und
- das „Südhessische Flurnamenbuch" (SHFLNB), das als erstes Flurnamenbuch eine größere Region mit umfangreichem Gesamtdatenbestand (ca. 150.000 Belege) mit z.T. neuen konzeptionellen Methoden bearbeiten und präsentieren will.

Diese kurze Aufzählung macht bereits deutlich, daß die 80er Jahre in der Flurnamenforschung durchaus nicht ohne methodische Innovationen waren – entgegen der oben angeführten Sichtweise Halfers. So wie der grundsätzliche Einsatz der EDV der Flurnamenforschung – wie auch anderen philologischen Disziplinen – bisher zu neuen methodischen Möglichkeiten verholfen hat, kann die Entwicklung eines Flurnamenbuchs als Hypertext

---

[1] Dieser Aufsatz ist Teil meiner Magisterarbeit (Richter 1999) und ist im Rahmen des DFG-Projekts „Mittelhessisches Flurnamenbuch als Hypertext" entstanden. Dem Leiter dieses Projekts, Herrn Prof. Dr. Hans Ramge, sowie Dr. Herbert Schmidt und Dr. Thomas Gloning danke ich für hilfreiche Hinweise und Anregungen, meiner Frau für die sorgfältige Korrektur des Aufsatzes.

diese ebenso mit neuartigen Aspekten bereichern. Die diesem Aufsatz zu Grunde liegende These lautet daher: Die Darstellung von (Flur)Namenbüchern in elektronischer Form – z.B. als Hypertext – schafft neuartige Präsentations-, Zugriffs- und Auswertungsmöglichkeiten.

2 Gründe für die Umsetzung eines Flurnamenbuchs in einen Hypertext

Die Gründe, die für ein Flurnamenbuch[2] als Hypertext sprechen, haben sich zum einen aus den Erfahrungen ergeben, die das Hessische Flurnamenarchiv mit dem DFG-Projekt SHFLNB gesammelt hat, und zum anderen aus Überlegungen, wie man neuartige, flexiblere und benutzerfreundlichere Zugangsformen schaffen kann, die die Benutzung eines Nachschlagewerkes (z.B. eines Flurnamenbuchs) grundlegend verbessern könnten.

Ein Flurnamenbuch als Hypertext führt zu einer Entschärfung

- der Belegauswahl, da Platzgründe im Distributionsmedium CD-ROM oder World Wide Web (WWW) eine sekundäre bzw. so gut wie keine Rolle spielen.
- der lemmaübergreifenden Verweistechnik, die – aus der Perspektive des Benutzers – die Rezeption erheblich vereinfacht und – aus der Perspektive der Produktion und der Arbeitsökonomie – beträchtliche Entlastung für die Autoren bedeutet.

Zur Entschärfung der Belegauswahl: In der Printversion des SHFLNB mußte aus Gründen des zur Verfügung stehenden Platzes und der Benutzerfreundlichkeit der Belegteil eines Namenartikels in seiner Mikrostruktur nicht nur stark verdichtet werden (z.B. durch Ortskürzel, Abkürzungen, Quellensiglen etc.), sondern anhand von zuvor festgelegten Regeln die Gesamtmenge der potentiellen Belege mal mehr, mal weniger drastisch reduziert werden.[3] Eine der Aufnahmeregeln lautete, daß abgeleitete, sekundäre Flurnamen nicht aufgenommen werden, wenn ein Simplex mit dem gleichen Namen im selben Ort existiert (Bsp.: Die Belege „Am Brühlweg" bzw. „Am Brühlgraben" werden dann nicht aufgenommen, wenn im selben Ort ein Simplex-Beleg „Brühl" vorhanden ist.). Im Hypertext sind Regeln, die allein das Ziel der quantitativen Reduktion verfolgen, von untergeordneter Bedeutung; sie können wegfallen, müßten überdacht oder durch neue qualitative Konventionen ersetzt werden.

Zur Entschärfung der lemmaübergreifenden Verweistechnik: Zur Optimierung der Übersichtlichkeit, der Auffindbarkeit und des Zugriffs auf die Namenartikel wurde für die Printversion das „Erstgliedprinzip" beim Lemmaansatz gewählt (vgl. Ramge 1996:166). Daraus folgt: (1) Namenkomposita werden nur einmal im gesamten Flurnamenbuch nachgewiesen. (2) Ausschlaggebend für die Einsortierung von Belegen mit Namenkomposita ist das lemmatisierte Erstglied des Namens, also der Bestimmungsteil (BT). Der Beleg „Am Brühlgraben" steht im Belegteil des Namenartikels „Brühl", im Namenartikel „Graben" – das Lemma des Grundteils (GT) – findet sich ein Verweis auf den Artikel „Brühl". Durch neue Darstellungsmöglichkeiten (Filter, Stretchtext usw.), die Hypertext bietet, können Belege mit Komposita in allen Namenartikeln erscheinen, die als Lemma Bestandteil von diesen sind: Ein „Brühlgraben" wäre somit für den Benutzer sowohl unter „Brühl" als auch unter „Graben" zu finden. Die Vorteile liegen auf der Hand: Die Verweisteile, die in der Print-

---

[2] Zum Namenbuch vgl. u. a. Greule 1984 und 1990 und Reichardt 1984 und 1995, zum Flurnamenbuch vgl. Reichardt 1995:307f., v.a. Ramge 1990 und 1996.

[3] Zum Problem der Reduktion vgl. Ramge (1990:103–105), auch (1996:163ff.).

version durch das beschriebene Verfahren einen z.T. beträchtlichen Umfang erfuhren, werden stark entlastet, bei konsequenter Anwendung sind sie obsolet. Das bedeutet für den Benutzer eine navigationsärmere Benutzung, da der durch die Verweisteile teilweise indirekt entstandene Zwang zum Verfolgen von Verweisen wegfällt.[4] Für die Flurnamenbuch-Produktion hat dies ebenfalls einen entlastenden Effekt: Die extrem zeitaufwendige und fehleranfällige Erstellung und Wartung von Verweisen entfällt. Verweise im elektronischen Medium können wesentlich einfacher – weil automatisch oder halbautomatisch – und kontrollierter gepflegt werden.

3 Vorteile und Mehrwerte eines Hypertext-Flurnamenbuchs

Die folgenden sechs Aspekte, die m. E. als Vorteile und Mehrwerte im wesentlichen dafür sprechen, weshalb ein Flurnamenbuch wie das SHFLNB als Hypertext umgesetzt werden sollte, ergeben sich teilweise aus den im vorherigen Kapitel genannten Gründen.

3.1 Umfangreichere Datenpräsentation

Flurnamenbelege, die in der Printversion aus Gründen des zur Verfügung stehenden Platzes wegfallen mußten, können im elektronischen Medium (Off-Line oder On-Line) aufgenommen werden. Während im Buch ‚nur' rund 50% des Gesamtkorpus im Durchschnitt dokumentiert werden, könnten in einem Hypertext im Prinzip alle Belege abgebildet werden. Meines Erachtens sollte die optimale quantitative Dokumentation eine der zentralen Funktionen von elektronischen Flurnamenbüchern sein: Das elektronische Medium schafft durch nahezu ‚unbegrenzten' Platz die technischen Voraussetzungen, Hypertext die neuen Methoden, die es ermöglichen, die gewachsenen Belegmengen zu verwalten und benutzergerecht zu organisieren und aufzubereiten.

Die Maxime der optimalen quantitativen Korpusdokumentation zieht zwei qualitative Forderungen nach sich:

1. Der Anspruch eines Namenbuchs, auch selbst wieder als (zitierfähige) Quelle für weitergehende Forschungen dienen zu können, muß auch bei gestiegener quantitativer Dokumentation aufrechterhalten werden: der qualitative Aspekt darf nicht dem quantitativen geopfert werden. Das heißt, Belege, die in ihrer Korrektheit als zweifelhafte Belege eingestuft werden[5] und nicht mit vertretbarem Aufwand[6] überprüft werden

---

[4] Als ‚Gegengewicht' zum gestiegenen Belegaufkommen müßten natürlich neue Mittel einer Belegreduktion nach Bedarf des Benutzers geschaffen werden. Steigende quantitative Datendokumentation erfordert zusätzliche Hilfsmittel zur Erhaltung der Orientierung bei der Rezeption.

[5] Fehlerhafte Belege können auf mehreren Ebenen entstanden sein, z.B. durch Fehler beim Lesen und Auswerten der Quelle, Schreibfehler, Lesefehler bei der Eingabe in die EDV (bedingt durch schlecht leserliche Vorlagen oder einfach durch menschliches Versagen), technische Restriktionen seitens der EDV (begrenzte und/oder starre Zeilenlänge usw.), technische Übertragungsfehler durch Portierung (z.B. vom Großrechner auf den PC) und Konvertierung der Daten (von einem Datenbanksystem in ein anderes).

[6] Was unter ‚vertretbarem Aufwand' zu verstehen ist, kann prinzipiell festgelegt werden: Bspw. könnte vereinbart werden, daß historische Belege zwecks Korrektur nur dann bis zur Quelle

können, sollten entweder trotzdem dokumentiert werden (müssen dann aber als solche speziell und unübersehbar als zweifelhaft markiert werden) oder aber ignoriert werden.

2. Durch eine umfangreichere Datenpräsentation steigt aber zugleich auch der Bedarf an einer stärkeren Strukturierung und Qualifizierung der Dokumentationstiefe, um Rezeptions-, Orientierungs- und Navigationsproblemen vorzubeugen. Denkbar wäre, dem Prinzip der vertiefenden Detaillierung folgend, für den Belegteil sog. Dokumentationstiefen festzulegen und jeden Beleg mit einer von diesen zu attribuieren. Die Kriterien, die diesen Dokumentationstiefen zugrunde liegen, müßten selbstverständlich dem Benutzer offengelegt werden. Eine Möglichkeit wäre ein Vier-Ebenen-Modell. Auf der ersten Ebene werden nur ausgewiesene rezente und historische Belege präsentiert, die die folgenden Bedingungen erfüllen: Für rezente Belege gilt, daß – wenn vorhanden – je ein Namen-Simplex und ein -Kompositum aufzuführen sind; für historische Belege gilt, daß der (ggf. die) älteste(n) Beleg(e) genannt werden, evtl. noch interessante Varianten. Die zweite Ebene führt alle noch nicht durch Ebene 1 präsentierten rezenten Belege auf, Ebene 3 alle weiteren historischen. Mit der Ebene 4 werden all jene Belege qualifiziert (und angezeigt), in denen das Lemma im Grundteil positioniert ist. Um das Vier-Ebenen-Modell auf ein Drei-Ebenen-Modell zu reduzieren, könnte man die Ebenen zwei und drei ggf. zusammenfassen.

Für den Deutungsteil schlägt Ramge ein Zwei-Ebenen-Modell vor: An erster Position (Deutungsstufe 1) wird

> die Globaldeutung des Lemmas angegeben mit Angabe des Bezugsappellativums bzw. des Namentyps (z.B. Familienname), dem Wahrscheinlichkeitsgrad der Deutungssicherheit und der Literatur, auf die sich die Globaldeutung stützt. [...] An zweiter Position [Deutungsstufe 2] werden, soweit notwendig und möglich, Erläuterungen zu den Deutungsproblemen gegeben, sprachhistorische Aspekte aufgeführt, vor allem auch der Bezug zu den realen Vorkommen verdeutlicht, verifiziert durch Realproben, ggf. in Auseinandersetzung mit der Lokalliteratur. Hier wird dann auch ggf. ergänzende Literatur genannt. [1997]

Diese ersten Überlegungen machen deutlich, daß die Erarbeitung trennscharfer und plausibler Qualifizierungskriterien ein schwieriges Unterfangen ist. Das haben auch die Erfahrungen bei der Erarbeitung und – in verstärktem Maße – der Anwendung der Kriterien zur repräsentativen Belegauswahl bei der Produktion des SHFLNB in der Printversion gezeigt:

> Die theoretisch einsichtigen und folgerichtigen Prinzipien der Datenauswahl führten in der prinzipienkonformen Anwendungspraxis häufig zu intuitiv, aber auch sachlich begründbar unbefriedigenden Ergebnissen, sei es, dass zu viele Belege ‚geopfert' werden mussten, sei es, dass das Auswahlprinzip zu einer stereotypen Wiederholung des immer gleichen oder fast gleichen Belegtyps führte. Wir haben dieses Problem pragmatisch durch Subregeln gelöst. Es hat sich aber das Bewusstsein gefestigt, dass eine weniger rigide Reduktion des Materials wünschenswert wäre – bei gleichzeitiger Vermeidung von Beleg-Redundanz. [Ebd.]

3.2 Einfachere und sichere Verweistechnik – direkter ‚Verweiszugriff'

Die im Kapitel zuvor ausgeführten Aspekte wie Reduktion bis Wegfall des Verweissystems und ‚elektronische' Unterstützung beim Erstellen von Verweisen und Warten des Verweis-

---

zurückverfolgt werden, wenn diese vor Ort vorhanden ist: Belege, die anhand von Quellen in Archiven zu überprüfen wären, werden nicht korrigiert. Natürlich kann und muß man von diesem Prinzip Ausnahmen zulassen können, etwa wenn sehr alte und daher namenkundlich interessante Belege fehlerhaft sind.

systems sind klare Vorteile, die einen Hypertext gegenüber einem konventionellen Text auszeichnen. Hinzu kommt die eigentlich nicht erwähnenswerte, weil selbstverständliche Eigenschaft von Hypertext, über Verweise direkt mit einem einfachen Mausklick zugreifen zu können: Dies ermöglicht eine direkte und schnellere und somit letztlich eine benutzerfreundlichere Rezeption, da langes Suchen und Blättern entfällt.

3.3 Visuelle und akustische Dokumentierbarkeit

Gänzlich neue Möglichkeiten für die Namenforschung im allgemeinen und die Flurnamenforschung im besonderen eröffnet das elektronische Medium bzw. das Hypertextkonzept:

- Der akustische Bereich – als im Vergleich zum Printmedium vollkommen neue Präsentationsform – bietet z.B. die Möglichkeit, mündliche Belege (die in phonetischer Umschrift repräsentiert sind) in einem dialektalen Kontext zu präsentieren.
- Im visuellen Bereich könnten nicht nur die auch in der Printversion etablierten Verbreitungskarten mit neuen Funktionen in den Hypertext integriert werden, sondern auch Realaufnahmen in Form von statischen Fotos und bewegten Bildern[7], Karten (Flurplatten im Maßstab 1:5.000, Gemarkungskarten bzw. topographische Karten in verschiedenen Maßstäben, historische Karten u.a.), Abbildungen von Quellen (Urkunden u.ä.) usw. dargestellt werden.

3.4 ‚Interaktion' zwischen Benutzer und Hypertext

Die ‚Interaktion' zwischen Benutzer und Hypertext bezieht sich zum einen auf sog. ‚on-the-fly'-Karten und zum anderen auf im Vergleich zum Printbuch neue, erweiterte und flexiblere Auswertungs- und Zugriffsmöglichkeiten.

Unter On-the-Fly-Karten sind Vorkommens- und Distributionskarten zu verstehen, die

- interaktiv zwischen Benutzer und System,
- on-demand, d.h. bei Bedarf und
- nach benutzerspezifischen Vorgaben erstellt werden.

Der entscheidende Vorteil des elektronischen Mediums besteht demnach nicht nur darin, theoretisch alle möglichen Verbreitungskarten darstellen zu können (in einem Print-Flurnamenbuch können aus Platzgründen meist nur einige wenige Karten abgebildet werden, die zudem durch den Benutzer nicht modifizierbar sind), sondern auch darin, Karten nach individuellen Angaben (Symbolwahl usw.) anfertigen zu lassen. So könnte ein Benutzer folgende einfache Anweisung zum Anfertigen einer Verbreitungskarte an das System geben: „Zeige mir alle historischen Belege zum Lemma X mit jenem Symbol und alle rezenten mit diesem Symbol an!" Verbreitungskarten haben u.a. die Funktion, die geographische Verteilung von Namen in einem abgegrenzten Sprachraum zu präsentieren,[8] so daß

---

[7] Damit sind bspw. Videoaufnahmen, z.B. der Flug oder Gang über eine Gemarkung oder Simulationen gemeint. Der Technik sind diesbezüglich heutzutage (fast) keine Grenzen mehr gesetzt. Die Grenzen sind durch sinnvolle funktionalen Bezug zu den präsentierten Namen und Namentypen gegeben.

[8] Zu Namengeographie, Namenkarten und -atlanten vgl. den sehr guten Überblick von Ramge (1995). Zur Methodik der Flurnamengeographie, Aufbau und namenkundlicher Bearbeitung der

nicht nur die sprachwissenschaftlichen Ergebnisse durch die kartographische Visualisierung (im optimalen Falle) schneller und besser zu erfassen sind als durch verbale Beschreibung,[9] sondern auch (strukturelle) Zusammenhänge deutlich werden, die ansonsten im Verborgenen geblieben wären. Aus diesem Grund ist die interaktive Verbreitungskarte auch dafür prädestiniert, Ergebnisse, die mit Hilfe von Selektionsabfragen (als sog. ‚Kreuzklassifikationen'; s.u.) erzielt wurden, zu visualisieren, um dem Benutzer einen ersten Überblick über die Verteilung bzw. die Vorkommenshäufigkeit u.a. zu geben.

3.5 Neue, erweiterte und flexiblere Auswertungs- und Zugriffsmöglichkeiten

Anhand von zwei Aspekten möchte ich darstellen, was hierunter zu verstehen ist:

1. Selektion von Teilmengen: Hypertext bietet anhand verschiedener Verfahren die Möglichkeit, dem Benutzer Teilmengen nach zuvor definierten Kriterien anzeigen zu lassen. Dies ist zum einen durch die benutzergesteuerte Ausführung von Selektionsabfragen möglich, die mit vordefinierten Kriterien eine Teilmenge aus dem Gesamtkorpus ‚extrahiert', zum anderen durch Filter (s.u.). Zu definierende Kriterien könnten sich z.B. (a) auf einen bestimmten geographischen Raum (lokale Teilsammlung, Orts-/Gemarkungsliste, Kreisliste etc.), (b) auf einen bestimmten zeitlichen Raum, (c) auf eine bestimmte Sprachform (alle Namen, die ein „ei" enthalten) oder (d) bei historischen Belegen auf eine oder mehrere Quellen beziehen. Zudem hätte der Benutzer die Möglichkeit, die durch ein einzelnes Kriterium erhaltene Teilmenge durch Kombination mit anderen Kriterien weiter einzuschränken. Als Beispiel einer solchen Kreuzklassifikation könnte vom Benutzer folgende Selektionsanfrage an das System gestellt werden: „Liste alle (historischen) Belege auf, die aus dem Kreis Bergstraße stammen und vor 1600 datiert sind!"
2. ‚Views' durch Filter: Ein anderes technisches Verfahren für die Selektion von Daten sind Filter. Mit ihnen können sog. ‚views' erzeugt werden, also auf bestimmte Teilstrukturen reduzierte Sichten.[10] Durch ‚Vorschalten' eines entsprechenden Filters könnten in einem Flurnamenbuch als Hypertext z.B. alle Namenartikel auf die mikrostrukturellen Bauteile „Lemmaangabe" und „Deutungsteil" reduziert werden. Aber auch die oben erwähnten Dokumentationstiefen und Deutungsstufen wären durch Filter zu realisieren. Eine andere Anwendungsmöglichkeit von Filtern wäre z.B., die Makrostruktur eines Flurnamenbuchs von der alphabetischen in die onomasiologische Anordnung umzuwandeln.[11]

Flurnamenkarten im „Hessischen Flurnamenatlas" s. Ramge (1987b). Händler (1987) stellt die computative Bearbeitung des Hessischen Flurnamenatlasses dar.

[9] Vgl. hierzu Ramge: „Das Kartenthema ersetzt eine ausführliche verbale Beschreibung, ergänzt, illustriert und entlastet Artikel, vor allem Artikel in Namenbüchern." (1995:315)

[10] Das OXFORD ENGLISH DICTIONARY (OED) auf CD-ROM arbeitet mit solchen Views (vgl. Raymond und Tompa 1988).

[11] Dazu müßten zuvor die einzelnen Namenartikel bestimmten Referenzbereichen zugeordnet und entsprechend codiert werden. Nicht zu erwähnen brauche ich, daß das *theoretische* Problem der Zuordnung einzelner Namen zu Referenzbereichen mit Hypertext bzw. Filtern nicht gelöst werden kann, da „sich das Spannungsverhältnis zwischen benennbarer Wirklichkeit und tatsächlich vorkommender Namengebung nicht ohne Gewaltsamkeit und nicht restlos befriedigend auflösen läßt." (Ramge 1987b: 13) Ohne jedoch an dieser Stelle auf die theoretischen Probleme der onoma-

Abfragen und Filter sind, neben anderen Mitteln und Verfahren, dazu geeignet, unterschiedlichen Benutzergruppen und -typen gerecht zu werden. Filter sind hinsichtlich ihrer Flexibilität und Individualisierbarkeit Abfragen deutlich unterlegen. Abfragen hingegen setzen andererseits (a) voraus, daß der Benutzer weiß, wonach er sucht, und (b) er die Suche mit den vom System zur Verfügung stehenden Mitteln operationalisieren kann.

3.6 Erhöhte Reichweite durch Publikation auf unterschiedlichen Medien

Daß eine erhöhte Reichweite erzielt wird, indem in unterschiedlichen Medien publiziert wird, braucht nicht weiter ausgeführt werden. Besonders interessant ist eine erhöhte Reichweite v.a. für die auflagenschwachen Publikationen, die den Wissenstransfer zwischen Wissenschaft und Forschung auf der einen und dem wissenschaftlichen ‚Laien' auf der anderen Seite leisten wollen und sollen.[12] Die Publikation eines Hypertexts als linear organisiertes Printbuch widerspricht den konstitutiven Eigenschaften eines nonlinear organisierten Hypertexts. Die Konzeption eines Flurnamenbuchs als Hypertext muß daher so geplant werden, daß ohne größere Hindernisse aus einer Datenquelle heraus in verschiedene Medien publiziert werden kann: konventionelles Printbuch, off-line auf CD-ROM und on-line im WWW.

4 Die hypertextuellen Potentiale eines Flurnamenbuchs: Hypertextualisierung der Mediostruktur und neue Darstellungsformen am Beispiel des *Südhessischen Flurnamenbuchs*

Die folgenden Analysen sollen zeigen, daß sich durch die Hypertextualisierung der konkreten Mediostruktur des SHFLNB gänzlich neue Verknüpfungs- und Darstellungsmöglichkeiten ergeben.[13]

---

siologischen Anordnung von Flurnamenbüchern eingehen zu wollen, müßte für die Fälle, in denen eine eindeutige Zuordnung zu einem Referenzbereich nicht möglich ist, die ‚Hauptzielrichtung' das ausschlaggebende Kriterium sein. In den Fällen, in denen eine Zuordnung nach diesem Verfahren nicht entschieden werden kann, bietet Hypertext die neue – wenngleich theoretisch nicht zufriedenstellende – Chance, das Zuordnungsproblem zumindest praktisch zu lösen, z.B. dadurch, daß ein Namenartikel in verschiedenen Referenzbereichen aufgeführt und entsprechend gekennzeichnet wird.

[12] Ramge spricht von einer Bringschuld der Universität gegenüber der finanzierenden Öffentlichkeit: „Das Gesamtarchiv der rezenten hessischen Flurnamen und die Teilarchive der rezenten und historischen Flurnamen Süd- und Mittelhessens, mit öffentlichen Mitteln aufgebaut, schulden der interessierten Öffentlichkeit, nicht nur den Experten, auch den Heimatforschern, den interessierten Laien, die Präsentation nutzbarer, handhabbarer Ergebnisse." (Ramge 1997)

[13] Die folgenden Ausführungen orientieren sich bezüglich der Vorgehens- und Darstellungsweise an Kammerer (1998). Er untersucht meines Wissens als erster die Frage der möglichen Hypertextualisierung von Verweisstrukturen durch Hyperlinks. Zur metalexikographischen Behandlung von Mediostrukturen in Wörterbüchern vgl. Wiegand (1996). Mit Verweisen beschäftigen sich u. a. Blumenthal et al. (1988), Wiegand (1996) und Kammerer (1998). Die im folgenden verwendete Terminologie orientiert sich an den beiden zuletzt genannten Autoren.

Die Mediostruktur (auch: Verweisstruktur) beschreibt die Verweisbeziehungen in einem Wörterbuch oder Namenbuch (vgl. Kammerer 1998:145). Sie ist eine von vielen anderen Strukturen – bspw. der Makro- und Mikrostruktur[14] –, die gedruckte sprachlexikographische Nachschlagewerke aufweisen (können).

Eine erste Betrachtung der Mediostruktur von Wörter- und Namenbüchern läßt den Schluß zu, daß diese zu den „idealen Textsortenträgern gehören, die in einem Hypertext modelliert werden können." (vgl. ebd.: 154). Vor allem die für diese Textsorten charakteristische (und konstitutive) Eigenschaft einer ausgeprägten und komplexen Mediostruktur fordert zu einer Übertragung in eine Hyperlink-Struktur geradezu auf – nicht zuletzt aufgrund der Einsicht, daß das lineare Medium Buch die Verweisstruktur nur über Umwege und letztlich unbefriedigend wiedergibt.

Die folgenden Ausführungen sollen daher am konkreten Beispiel „Südhessisches Flurnamenbuch" zeigen, ob es sinnvoll ist, jeden in der Mediostruktur des SHFLNB existierenden Verweis – sei er expliziter, impliziter oder auch potentieller Art – in einen Hyperlink zu transformieren oder ob es in bestimmten Fällen besser ist, die Mikrostruktur den Bedingungen des Hypertexts anzupassen, um somit überkommene Formen der Verweisung zu vermeiden.

Abb. 1 stellt die zwei Namenartikel „Dribsach" und „Abtei" aus der Printversion des SHFLNB dar.[15] In den abgebildeten Namenartikeln sind alle Verweisadressenangaben von expliziten und impliziten Verweisen einfach unterstrichen, jene von potentiellen Verweisen unterpunktet. Sind Verweisadressenangaben ‚polyfunktional' in bezug auf ihr Verweispotential, sind sie doppelt unterstrichen.

Ein erster Blick auf die beiden abgebildeten und markierten Namenartikel zeigt eine Fülle von möglichen und notwendigen Verweisungen. Fangen wir mit dem Artikel „Dribsach" an: Hierbei handelt es sich um einen reinen Verweisartikel, einen Namenartikel mit rudimentärer Mikrostruktur, der neben dem Lemma (auch: Lemmaangabe) nur aus einer Verweisangabe, im SHFLNB aus der Verweisbeziehungsangabe „→" und der Verweisadressenangabe – hier: „Trieb" –, besteht. Es handelt sich hierbei nach der Teilklassifizierung von Wiegand (1996) also um einen wörterbuch- bzw. namenbuchinternen, expliziten Artikelverweis. Des weiteren kann man sagen, daß dieser Verweis

- obligatorisch und nicht fakultativ in bezug auf die Notwendigkeit seiner Befolgung ist;
- ein adressierter Verweis ist, da er eine Verweisadressenangabe aufweist (im Gegensatz z.B. zur Verweisangabe „s.o.", die keine Verweisadressenangabe besitzt);
- ein adkurrenter Verweis ist, d.h. der Verweis führt den Benutzer zum Lemma eines bestimmten Namenartikels (hier: „Trieb") und nicht wie beim inkurrenten in diesen hinein;[16]

---

[14] Die textuelle Namenbuchstruktur, Makro- und Mikrostruktur von Flurnamenbüchern im allgemeinen und ihre konkreten Ausprägungen im SHFLNB im besonderen werden in Richter (1999:35ff.) detaillierter untersucht.

[15] Aus Gründen der besseren Lesbarkeit ist der Durchschuß etwas erweitert und die Schriftgröße aller Artikelsegmente etwas erhöht worden, ansonsten entsprechen sie weitgehend dem Layout des Endprodukts.

Dribsach → Trieb

Abtei

• BAB: ° *Abteischneise* (abdai:ʃna:s) / 1356 W *der geheiszen ist die aptie* (Reimer Bd. 3, 1891–7, S. 183), 1730 *An der Aptey* (Rühl 1953, S. 46). • EPP: ° *Die Abtei* / ° *Bei der Abtei* (Da, O 61, Buxb, 1) / 1689 *auf die Abtey* (Da, C 2, 30/1 , f. 141') / ° *Abteischneise*.

→ Alt, Breidert, Schwarz.

Abtei ‚Kloster', ahd. *abbateia*, eine Entlehnung aus mlat. *abbâtia*, und mhd. *aptei*. Als FlN ist *Abtei* in Südhessen stets auf das Kloster Seligenstadt bezogen. Zu unterscheiden sind dabei Besitztümer des Klosters, so der Wald *der geheiszen ist die Aptie* in BAB (der rezente Beleg in Epp bezeichnet wohl ursprünglich dasselbe Waldstück, das der Abt von Seligenstadt 1356 verkauft hat[1]), und Flurstücke, die an Wege angrenzen, die in Richtung des Klosters führen.

Karg-Gasterstädt/Frings 1,12, LEXER 1,1, FWB 1,431f.; DWB Neub. 1,1123f.; Hiersche 20; Vielsmeier (1995), S. 29; Rühl (1953), S. 46. [1]Müller (1937), S. 38.

Vgl. auch Kirche, Stift.

Abb. 1: Namenartikel „Dribsach" und „Abtei" aus dem SHFLNB

- ein lexikologischer Verweis ist, da dies per definitionem als Konstitutionsregel für Verweisartikel festgelegt ist (diese Regel muß dem Benutzer auch in einem Außentext mitgeteilt werden).

[16] Im SHFLNB können keine inkurrenten Verweise vorkommen, da es sich um eine glattalphabetische Anordnungsform (bzw. eine striktalphabetische Makrostruktur *ohne* Gruppierung) handelt. Inkurrente Verweise sind in nischen- und nestalphabetisch angeordneten Wörter- und Namenbüchern aufgrund ihrer Anordnungsstruktur notwendig.

Geht man davon aus, daß jeder Namenartikel einen eigenen Hypertextknoten bildet und die Gesamtheit der Namenartikelknoten die Hypertextbasis „Namenverzeichnis" repräsentiert, muß dieser Verweis als intertextueller node-to-node-Link[17] hypertextualisiert werden.

Die Modellierung eines Flurnamenbuchs als Hypertext eröffnet auch gänzlich neue Möglichkeiten, wie z.B. die, die **Makrostruktur des Namenverzeichnisses** nicht nur in alphabetischer Anordnung zu präsentieren, sondern parallel hierzu dem Benutzer eine onomasiologische Anordnung anzubieten. Die onomasiologische Makrostruktur von Namenbüchern, also die Einordnung von Namenartikeln in ihre Referenzbereiche, wird seit geraumer Zeit immer wieder gefordert,[18] in nur wenigen Fällen[19] ist sie tatsächlich realisiert worden. Drosdowski (1985) merkt an: „Bei Datenbanken, die einen Zugang unter den verschiedensten Aspekten von allen Seiten erlauben, kann... die Frage ‚alphabetische oder begriffliche Anordnung' gegenstandslos werden!" (66f.) Daher sollten m.E. die neuen Potentiale von Hypertext genutzt werden, um – in Analogie zum vielfach geforderten „integrierten Wörterbuch"[20] – ein *integriertes Namenbuch* als ein Ziel zu formulieren. Kirkness resümiert schon 1985: „Auf jeden Fall sollten sich die gegenwärtigen Lexikographen diese Zukunftschance [des integrierten Wörterbuchs] nicht entgehen lassen, zumal sie mit Hilfe des Rechners genutzt werden kann. Die notwendige Software existiert heute schon." (1985:49)

Der **Belegteil** in unserem Beispiel weist neben impliziten eine Reihe von potentiellen Verweisen auf. Beginnen wir mit dem Ortskürzel „BAB": Alle Abkürzungen können, besonders wenn sie – wie in diesem Fall – wörter- bzw. namenbuchspezifisch sind, potentielle

---

[17] Zur Typologie der Hyperlinks vgl. z.B. Kuhlen (1991).

[18] So z.B. von Boesch: „Unerläßlich ist endlich die Gliederung des Namengutes nach sachlichen Gesichtspunkten. Eine alphabetische Liste der Namen ist noch keine Darstellung: der Name gehört in den Kreis der Sachen, die er bezeichnet." (Boesch 1959/60:7)

[19] Eine Ausnahme stellt Halfer (1988) dar, dem es „gelungen ist, den gesamten Namenbestand ziemlich befriedigend nach Sachgruppen zu sortieren". (Ramge 1989:257) Er stellt richtig fest: „Über die geltenden Einteilungskriterien für die Gliederung des Namenmaterials nach Sachgruppen herrscht im großen und ganzen Einvernehmen." (1988:20) Zu den Schwierigkeiten der sachlichen Anordnung vgl. Ramge (1987b: 12f.) und (1989:257).

[20] Kirkness stellt zunächst fest, daß in der aktuellen metalexikographischen Diskussion die „Forderung nach einer expliziten Abbildung der systematischen Bedeutungsverwandtschaft *und* der systematischen Mehrdeutigkeit des Wortschatzes, d.h. die Forderung nach der Integration von Onomasiologie und Semasiologie, insbesondere im alphabetischen Wörterbuch" (1985:48; Hervorhebung im Original) ein beliebtes Thema ist. Henne formuliert dies als Aufgabe: „Textphilologisches Interesse mag sich der alphabetischen Ordnung begnügen; aber dieses Interesse ist nicht das ganze Interesse an Sprache. Wörter werden im Zusammenhang erlernt, was bedeutet, daß das Erlernen von Wörtern in subjektiv orientierten und intersubjektiv korrigierten Wortfeldern erfolgt. Die alphabetischen Wörterbücher versuchen zum Teil, ihren Mangel durch Verweissysteme auszugleichen [vgl. hierzu die Funktion des Referenzteils bzw. die der darin enthaltenen Verweise im SHFLNB; GR]; die synonymischen Wörterbücher durch Hinweise auf ausgesparte Teilbedeutungen. Die Aufgabe für die Gegenwart liegt darin, integrierte, zumindest aufeinander bezogene Wörterbücher zu schaffen, die der bedeutungsstrukturellen Erfahrung der Sprachbenutzer Rechnung tragen." (1977:47) Wiegand stellt als These auf: „Alphabetische Wörterbücher müssen schrittweise zu integrierten Wörterbüchern umgestaltet werden, so daß sie in Situationen der Textlektüre und Textproduktion gleichermaßen benutzbar sind. Die totale Herrschaft des Alphabets, die die Wortschatzstrukturen zertrümmert, muß durch Kodifikationsverfahren überwunden werden, die die onomasiologische Blindheit der alphabetischen Wörterbücher beseitigt." (Wiegand 1977:102) Auch Drosdowski ist der Meinung, daß trotz aller Schwierigkeiten die Entwicklung hin zum integrierten Wörterbuch gehe (vgl. 1977:130).

Verweise sein. Ein Benutzer-in-actu, der wissen will, was die Abkürzung „BAB" bedeutet, muß (a) diese Angabe als Ortskürzel identifizieren (durch die Position im Belegteil, die Struktur der Ortsangabe (‚Ortskürzel' < „:")[21] und seine mikrotypographische Markierung in ‚Kapitälchen') und (b) im Außentext liegenden Ortsverzeichnis nachschlagen. Eine Hypertextualisierung liegt nahe: Intertextuelle Hyperlinks verknüpfen alle Ortskürzel mit ihren jeweiligen Einträgen bzw. Knoten im Ortsverzeichnis.[22] Zwei mögliche Darstellungsformen sind denkbar: (1) Der Zielknoten (hier: Ortsverzeichnis) ersetzt den Ausgangspunkt (hier: Namenverzeichnis bzw. Namenartikel). (2) Ein Pop-up-Fenster, nur aus der Auflösung des betreffenden Akronyms bestehend, überlagert nur zum Teil den Ausgangsknoten. Die zweite Möglichkeit wäre auch in der Art eines sensitiven Hyperlinks, wie sie von den QuickInfos bekannt sind, vorstellbar: Nur durch Berühren eines solchen Links öffnet sich nach einer festgelegten Zeit automatisch ein kleines Pop-up-Fenster mit der Auflösung der Abkürzung.

Gesagtes trifft sowohl auf die Abkürzungen der Kulturart wie „W" im Belegteil des Artikels „Abtei" (s. Abb. 1), als auch auf die wörterbuch- bzw. namenbuchspezifischen Abkürzungen „ahd.", „mlat." und „FlN" zu. Für die Abkürzung „Vgl." gilt dies in ihrer Funktion als Verweisbeziehungsangabe[23] zwar ebenso, sie unterscheidet sich jedoch von den anderen Abkürzungen durch ihre Bifunktionalität (vgl. Kammerer 1998:161): Zum einen repäsentiert sie die Verweisbeziehungsangabe, zum anderen typisiert sie die Verweisbeziehung, indem sie als „Erkennungsmarke für die Angaben der semantischen Merkmalsteilkongruenz" (ebd.) fungiert. Richtig ist zunächst Kammerers Feststellung, daß bei Tilgung der Verweisbeziehungsangabe nicht nur der Verweis vom Typ „explizit" zum Typ „implizit" transformiert würde, sondern mit dieser Tilgung auch die zweite Funktion mit getilgt würde (vgl. ebd.). Dies gilt jedoch nicht für die Verweisangaben im Referenzteil des SHFLNB: Hier wäre es durchaus möglich, den Referenzteil dadurch weiter zu verdichten, indem die Verweisbeziehungsangabe „vgl. auch" getilgt wird. Es müßte weiterhin die Bifunktionalität der Verweise gewährleistet sein, einerseits durch die Position (dem fakultativen Literaturteil bzw. dem obligatorischen Deutungsteil folgend), andererseits durch die (ggf. zu modifizierende) mikrotypographische Markierung. Bei allgemeinen einsprachigen Wörterbüchern kann aufgrund höherer Strukturkomplexität – u.a. verursacht durch ausgeprägte Textverdichtungsmaßnahmen – die Tilgung zu einer unzureichenden Interpretation der Verweisadressenangabe seitens des Benutzers führen.

Die Belegstellenangabe besteht aus zwei Textelementen: (a) der Quellenangabe (Bsp.: „Da, C 2, 30/1") und (b) der Seitenangabe („f. 141'"). Die Belegstellenangabe (hier: „Da, C 2, 30/1, f. 141'") als Ganze ist Teil der quellenbezogenen Mediostruktur und verweist auf die entsprechende Textstelle in dieser Quelle. Die Quellenangabe hingegen alleine ist ein Außentextverweis, also Element der wörterbuchinternen Mediostruktur. Die Quellenangabe ist in diesem Fall identisch mit der Verweisadressenangabe und verknüpft die (gekürzte) Quellenangabe im Belegteil mit der ausführlichen Quellenangabe im Quellenverzeichnis, das sich entweder im Vor- oder Nachspann befindet.[24] Nimmt man an, das Namenverzeich-

---

[21] „<" ist ein Anordnungszeichen mit der Bedeutung ‚geht voraus' und fungiert in der Darstellung einer konkreten Mikrostruktur als Sequenzkante (vgl. Wiegand 1989:412).

[22] Zur Frage, ob in einem Hypertext nicht generell soweit wie irgend möglich Abkürzungen vermieden werden sollten, um die Verweisdichte (und damit auch die kognitive Belastung) zu verringern, s. weiter unten.

[23] Es handelt sich eigentlich nur um einen Teil der Verweisbeziehungsangabe: Die vollständige ist durch „vgl. auch" repräsentiert.

nis und der Vorspann, u.a. bestehend aus dem Bauteil Quellenverzeichnis, seien als eigenständige Hypertextbasen konzipiert, so müßte man bei der Umsetzung in einen Hypertext Quellenverweise als extrahypertextuelle point-to-point-Verknüpfungen einrichten.

Problematischer scheint mir zu sein, daß es sich – worauf Kammerer zu Recht hinweist – streng genommen um zwei sich z.T. überdeckende Verweisadressenangaben handelt (vgl. 1998:162). Wie kann man dieses ‚Problem' in einem Hypertext lösen? Man könnte zunächst daran denken, daß nicht die Belegstellenangabe als Ausgangspunkt des Verweises auf die Textstelle (an der der Beleg im Kontext rezipiert werden kann), sondern der Beleg (hier: „auf die Abtey") selber als Verknüpfungsausgangspunkt fungiert. Favorisiert man jedoch die erste Variante (Belegstellenangabe = Verweis auf Quelle), handelt es sich um einen 1:n-Link mit der spezifischen 1:2-Relation. Diese in einem Hypertext zu realisieren, ist nicht weiter schwierig.

Die im Literaturteil aufgeführten Literatur-, Namenbuch- und Wörterbuchangaben sind Bestandteil ihrer jeweiligen literatur- bzw. namenbuch- und wörterbuchvernetzenden Mediostruktur.[25] Die Eigenschaften, die diese Angaben in der Funktion von Verweisen zeitigen, und die Probleme, die bei einer Hypertextualisierung auftreten, sind dieselben (1:2-Relation) wie bei den Quellenangaben; für sie gilt analog das bereits für die Quellenangaben Gesagte.

Die historischen Belege selber sollten m.E. – wie eben erwähnt – als Verweisausgangspunkt dienen, die bei Klick auf diese zu der Textstelle in der Quelle führen, an der der Beleg in seinem Kontext dargestellt ist. Bei den amtlichen Belegen, z.B. „Abteischneise", die in der Abbildung als unterpunktet und damit als potentielle Verweise markiert sind, böte es sich bei der Hypertextualisierung an, eine Verknüpfung zur aktuellen topographischen Karte (1:25.000 oder 1:10.000) der jeweiligen Gemarkung herzustellen, um den Beleg in dieser verorten zu können. Der historische Beleg wird über eine Verknüpfung *textuell*, der amtliche *geographisch* lokalisiert. Der mundartliche Beleg („abdai:ʃna:s" im Namenartikel „Abtei") ist ähnlich gelagert wie die Quellenangabe: Es handelt sich – auch wenn es die Abbildung nicht richtig wiedergeben kann – um einen zweifachen potentiellen Verweis. Zum einen könnte es ein wörterbuchinterner Verweis auf den Bauteil im Vorspann sein, in dem die verwendete Lautschrift in ihrer Aussprache erklärt wird, zum anderen könnte man den Mundart-Beleg auch als Ausgangspunkt für eine Verknüpfung zu einer Audiodatei

---

[24] Die Wiegandsche Grobgliederung der Systematik der Mediostrukturen bei gedruckten Wörterbüchern (vgl. Abb. 2 in 1996:14) macht deutlich, daß die Angaben im Quellen- bzw. Literaturverzeichnis selbst wiederum Verweiszieladressenangaben darstellen und dadurch die quellen- bzw. literaturbezogene Mediostruktur generieren. Die textuellen Namenbuchstrukturen von Print-Flurnamenbüchern zeigen mit wenigen Ausnahmen, daß sich als ‚heimlicher' Standard herausgebildet hat, das Quellenverzeichnis im Nachspann aufzuführen. Im Hypertext ist aufgrund der nonlinearen Organisation die Unterscheidung in Vor- und Nachspann ohnehin obsolet.

[25] Besonders die wörter- bzw. namenbuchvernetzenden Mediostrukturen, die dem Benutzer die Möglichkeit offerieren, direkt zur entsprechenden Textstelle im referenzierten Wörterbuch oder Namenbuch (auch zu sonstiger Sekundärliteratur oder Quellen) zu springen, sind keine Visionen mehr, sondern durch das WWW in greifbare Nähe gerückt. Einige Wörterbuch-Projekte, z.B. „LEXER-Online", das im Rahmen des DFG-Projekts „Mittelhochdeutsche Wörterbücher auf CD-ROM und im Internet" (s. Plate et al. und Burch & Fournier in diesem Band) erarbeitet wird, oder das DRW (vgl. Lemberg et al. 1998; erste Beispiele unter http://www.uni-heidelberg.de/institute/sonst/adw/drw/demo/frameset.htm), arbeiten verstärkt daran, Online-Versionen zu erarbeiten und anzubieten. Eine direkte Vernetzung der Wörterbücher untereinander wäre damit realisierbar.

betrachten. Klickt der Benutzer auf einen solchen Beleg, wird eine Aufnahme abgespielt, in der der Beleg von einem einheimischen Dialektsprecher gesprochen wird.[26]

Der **Verweisteil** hat beim Print-SHFLNB die genuine Funktion, auf den bzw. die Namenartikel zu verweisen, deren Belege das Lemma des Ausgangs-Namenartikels im Grundteil ‚tragen'. In obigem Beispiel wird im Verweisteil des Artikels „Abtei" auf die Namenartikel mit dem Lemma „Alt", „Breidert" und „Schwarz" verwiesen. Nach erwähnter Regel müßte sich also im Namenartikel „Alt" ein Beleg finden lassen, der der Struktur „Alt" im Bestimmungs- und „Abtei" im Grundteil entspricht (z.B. „An der alten Abtei" o.ä.). Der einzige Grund für die Existenz des Verweisteils ist darin zu sehen, daß Belege im Printbuch nach dem lemmatisierten Bestimmungsteil einsortiert werden („Brühlweg" erscheint im Namenartikel „Brühl"). Führt man hingegen das System der Dokumentationstiefen ein, fällt der Verweisteil weg, ansonsten handelt es sich hier um wörterbuchinterne Artikelverweise, die als intra- oder interhypertextueller point-to-node- bzw. point-to-point-Link zu realisieren sind.

Der Belegteil bietet weitere vier Verweisarten, auf die ich noch eingehen will. In Flurnamenbüchern kommen – im Gegensatz zu den allgemeinen einsprachigen Wörterbüchern – sog. ‚lexical links', als eine Art von potentiellen Links, die selbstreferentiell auf das eigene (Namenbuch-)Material verweisen, wesentlich weniger vor.[27] Daher ist es auch nicht vonnöten, einen Algorithmus zu entwickeln (wie dies Kammerer und Lehr (1996) versuchen), der entscheidet, ob ein Wort als potentieller Verweis ‚aktiviert', d.h. vom System als Verweis freigegeben werden soll. In Flurnamenbüchern müssen, so denke ich, die wenigen vorkommenden potentiellen Verweise dieser Art manuell ‚verlinkt' werden.

Als potentiellen Verweis habe ich „Kloster Seligenstadt" markiert. In einem Hypertext könnte dies ein extrahypertextueller Link sein, der auf einen Knoten verweist, der für die Deutung des Flurnamens von außergewöhnlicher Bedeutung ist bzw. die Deutung transparenter macht. Die Notwendigkeit von solchen Sachinformationen stellt Ramge heraus:

> Die Deutung besteht aus einem Satz kontrollierter Entscheidungsverfahren, die die philologisch-dialektologische Analyse des Namenzeichens ebenso umfaßt wie die außersprachlich-referentiellen-pragmatischen Bezüge des Zeichens (den Bezug auf „Sachen" im weitesten Sinne und in verschiedenartigen Relationen). (1990:107)

Hypertext bietet mit seiner nonlinearen Organisationsform kombiniert mit dem elektronischen Medium die neuartige Chance, die für die Deutung relevanten Sachinformationen dem Benutzer visuell oder auditiv aufbereitet zu präsentieren. Drei Beispiele sollen den Mehrwert, den Hypertext damit erzielt, unterstreichen: Ramge weist darauf hin, daß die

---

[26] Das Hessische Flurnamenarchiv in Gießen hat in den Sammelaktionen der 80er Jahre viele Gemarkungen erhoben, die entweder noch gar nicht oder bereits erhoben worden waren, aber eine unzureichende Materialbasis darstellten. Im Rahmen dieser Sammeltätigkeiten wurden nicht nur die amtlichen, sondern auch vor Ort mit Hilfe einer Kontaktperson die mundartlichen Belege erfaßt. Diese wurden nicht nur – je nach Kompetenz des Sammlers bzw. des Gewährsmannes – in Lautschrift (Laienumschrift oder IPA-ähnlich) transkribiert, sondern zumeist auch auf Kassetten oder Tonbändern mitgeschnitten. Auch wenn im Einzelfall kritisch die Qualität der Aufnahme geprüft werden muß, so verfügt doch das Hessische Flurnamenarchiv damit über ein sehr wertvolles Korpus.

[27] Zu den potentiellen Verweisen vgl. Kammerer (1998:147f.) und Kammerer und Lehr (1996). Raymond und Tompa sehen es in bezug auf die Entwicklung des elektronischen Oxford English Dictionary (OED) als höchst wünschenswert an, sog. „lexical links" als potentielle Verweise zu realisieren – auch wenn bei der automatischen Linkgenerierung einige Probleme zu lösen sind (vgl. 1988:875, 877).

Herkunft des Flurnamens „Steinrutsche" in der Gemarkung Kaichen (Wetteraukreis) nicht besser und einfacher zu verdeutlichen sei, als durch die Wiedergabe der farbigen Luftaufnahme, die im Feld den Grundriß einer römischen Villa erkennen läßt, über deren Steine der Pflug früher ‚gerutscht' ist (vgl. Ramge 1997). Mit Luftbildaufnahmen sind z.T. auch nicht mehr existierende Flurformen aufgrund von Bewuchsmerkmalen zu erkennen, die in den Flurnamen noch weiterleben und damit evtl. auf das vergangene ursprünglich namengebende Motiv hinweisen. Mit Hilfe einer Luftbildaufnahme der Gemarkung Hungen, Landkreis Gießen, konnten die Grundrisse der Kirche (zu erkennen anhand negativer Bewuchsmerkmale) visualisiert und rekonstruiert werden, die einst zur Wüstung Meßfelden gehörte. Noch heute verwendete Flurnamen weisen auf die untergegangene Siedlung hin.

Gebräuchliche Fachtermini, die im **Deutungsteil** verwendet werden, sind ebenso potentielle Verweise. Im Printbuch sind die Fachtermini nicht als solche markiert: Dem Benutzer bleibt nichts anderes übrig, als bei Begriffen, die er nicht versteht, *auf Verdacht* im Glossar – sofern es vorhanden ist – nachzuschlagen. Der Hypertext kann dem Benutzer die Frustration, keinen Eintrag im Glossar gefunden zu haben, insoweit abnehmen, indem Fachbegriffe, die im Glossar erklärt werden, als extrahypertextuelle Verknüpfungen (z.B. farblich hervorgehoben) angezeigt bzw. anderweitig signalisiert werden (etwa dadurch, daß sich der Mauszeiger beim ‚Überfahren' in eine zeigende Hand verwandelt).

Zu sehen ist im Deutungsteil ein Fußnotenzeichen in der Funktion der Verweisadressenangabe, das auf die dem Literaturteil angegliederte Fußnote („Müller (1937), S. 38.") verweist. Fußnoten werden wie Glossareinträge oder Abkürzungsauflösungen als Pop-up-Fenster oder in Form der QuickInfo gestaltet. Der Unterschied zwischen Fußnoten einerseits und Glossareinträgen bzw. Abkürzungsauflösungen andererseits besteht in der Relationalität ihrer Verknüpfung: Fußnotenverknüpfungen haben das 1:1-Beziehungsverhältnis, Glossareinträge und Abkürzungsauflösungen (auch Literaturverzeichnis u.a.) hingegen sind n:1-Relationen. Aus diesem Grund sollte man – v.a. wenn die Fußnote nicht als überlagerndes (bspw. als Pop-up-Fenster) konzipiert ist – Fußnotenverweise im Hypertext immer als bidirektionale Verknüpfungen anlegen.

Ich denke, es ist deutlich geworden, daß die 1:1-Konvertierung, bei der die Darstellungsformen und -strukturen aus dem Printmedium übernommen werden, zu unnötig vielen Hyperlinks führt. Vor allem die Abkürzungen, die im Printmedium als potentielle Verweise vorliegen, müßten im Hypertext expliziert werden, d.h. als Hyperlinks angelegt werden. Die Folgen sind unmittelbar sichtbar (s. Abb. 1): Es entstünde ein wahrer ‚Flickenteppich', bestehend aus einer Vielzahl von Verknüpfungen, die nur dafür da sind, Abkürzungen aufzulösen. Es muß daher überlegt werden, ob Abkürzungen wie die Ortskürzel nicht generell durch die vollen Ortsnamen ersetzt werden, um damit die Verweisdichte zu reduzieren. Die dann noch immer verbleibende hohe Anzahl von Verknüpfungen müßte mittels geeigneter Präsentationsformen (Stichwort „invisible link" oder „link on demand") so aufbereitet werden, daß der Benutzer-in-actu nicht mehr als nötig kognitiv beansprucht wird.

Neue Präsentationsformen sind m.E. eine Gratwanderung: Auf der einen Seite gilt es, den Hypertext so benutzerfreundlich wie möglich zu gestalten (z.B. durch weitgehenden Verzicht auf nicht-typographische Strukturanzeiger). Auf der anderen Seite nimmt der Benutzer aufgrund bestehender Rezeptionsmuster, die durch individuelle Rezeptionserfahrungen mit Nachschlagewerken aller Art gebildet worden sind (und sich daher stark von Benutzer zu Benutzer unterscheiden können), eine bestimmte Erwartungshaltung in bezug auf die (Makro- und Mikro-)Struktur ein. Hypertext kann – und das ist ein entscheidender Vorteil gegenüber dem Printmedium – unterschiedliche Nutzungsformen bedienen. Welche Nutzungsformen existieren und welche Erwartungen überhaupt an ein Flurnamenbuch ge-

stellt werden, kann letztlich nur eine empirische Namenbuch-Benutzungsuntersuchung herausfinden.

Neben der Hypertextualisierung von ‚Printverweisen' treten neue Möglichkeiten hinzu, bspw. durch die graphische oder auditive Aufbereitung von deutungsrelevanten Sachinformationen oder durch integrierende Verknüpfungen von semasiologischer und onomasiologischer Anordnung des Namenverzeichnisses. Der dadurch entstandene Mehrwert rechtfertigt in jedem Fall den höheren Aufwand in der Herstellung eines Flurnamen-Hypertexts.

5 Resümee

Im folgenden möchte ich anhand von drei Aspekten die wesentlichen Ergebnisse zusammenfassen:

1. Verhältnis von Sprach- und Sachlexikographie in der Flurnamenforschung
2. Die stark heterogene Menge der potentiellen Benutzer(typen) und Benutzungsformen von Flurnamenbüchern
3. Aspekt der Dokumentation: Repräsentative vs. vollständige Dokumentation des Belegmaterials

Zweifelsohne ist einerseits die Erarbeitung eines Flurnamenbuchs mittels linguistischer Methoden ein in erster Linie sprachlexikographisches Produkt, andererseits kommt man bei der Deutung der Flurnamen ohne sachlexikographische Informationen und ggf. Hinweise auf diese nicht aus. Ganz im Gegenteil: Sachlexikographische Informationen (also z.B. Informationen zur Geschichte und Geographie der bearbeiteten Region) sind zwingend notwendig für eine – auch philologisch – exakte Deutung.

Was in einem Print-Flurnamenbuch z.B. aus Platzgründen oder aus mediumbedingten Gründen nicht möglich ist, kann in einem elektronischen Medium bspw. als Hypertext realisiert werden: Die sachlexikographischen Informationen, die für eine Deutung der einzelnen Flurnamen notwendig und hilfreich sind, können

- visualisiert werden – z.B. mittels historischen und/oder topographischen Karten, Luftbildern, Realabbildungen usw. und
- mit den sprachlexikographischen Informationen mittels elektronischer Verweise verknüpft werden.

Die Flurnamenforschung ist nicht nur für die Sprachwissenschaft im allgemeinen und die Dialekt- und die Sprachgeschichtsforschung im speziellen, sondern auch für andere wissenschaftliche Disziplinen wie die Geschichtsforschung (v.a. die Siedlungsgeschichte) oder die Geographie interessant und von Nutzen. Aber auch für Beschäftigungen, die nicht streng wissenschaftlich fundiert sind, wie die sog. Heimatforschung, stellt die Flurnamenforschung ein wichtiges Hilfsmittel dar. Die Erarbeitung eines Flurnamenbuchs als Hypertext bedeutet also nicht nur für die sprachwissenschaftlichen Disziplinen einen Zugewinn hinsichtlich der Auswertung und Benutzung des Materials, auch die anderen fachlichen Forschungszweige profitieren von den neuen und zusätzlichen sachlexikographischen Informationen.

Der zweite Hauptaspekt, der für eine Produktion eines Flurnamenbuchs als Hypertext spricht, ist die Tatsache, daß ein Flurnamenbuch als eine Form des Extrakts der Flurnamenforschung unterschiedlichsten Benutzertypen, Benutzergruppen und Benutzungsmöglichkeiten gerecht werden sollte. Die heterogene Menge der Benutzer umfaßt sowohl den Heimatforscher, den in erster Linie lokal begrenzte Informationen interessieren, als auch den Sprachhistoriker, der überregionale Sprachphänomene untersucht. Ein Print-Flurnamenbuch kann die Erwartungen, die diese in mehrfacher Hinsicht unterschiedlichen Benutzer an es stellen, in vielen Fällen nicht erfüllen.

Durch die Publikation eines Flurnamenbuchs als Hypertext kann man sich von Zwängen befreien, die dem Printmedium eigen sind, und zum anderen neue Möglichkeiten des elektronischen Mediums und des Konzepts ‚Hypertext' nutzen:

- Neue und flexible Präsentationsformen und Zugriffsverfahren, die den unterschiedlichen Benutzergruppen, -typen und -zwecken stärker angepaßt sind.
- Der zum Druck zur Verfügung stehende Platz, die Anzahl der Bände, die Anzahl der Seiten je Band usw., spielen bei der Veröffentlichung auf elektronischen Datenträgern (CD-ROM, Internet usw.) eine untergeordnete Rolle.

Kurzum: Die Potentiale, die den Mehrwert des elektronischen Medium ausmachen, müssen auch und v.a. für eine benutzerorientierte Gestaltung genutzt werden. Dazu ist es jedoch unerläßlich, daß eine empirische Studie zu den genannten Aspekten der Benutzung von Flurnamenbüchern durchgeführt wird. Ihre Ergebnisse dienen zum einen als Grundlage für die Konzeptionierung der neuen Präsentationsformen und Zugriffsverfahren und zum anderen für das Design der Bedieneroberfläche der Hypertextanwendung „Flurnamenbuch als Hypertext".

Vielfach erreichen Flurnamenbücher ihre potentiellen Benutzer nicht. Wichtig scheint mir aber zu sein, die Reichweite und die Erreichbarkeit von Flurnamenbüchern zu erhöhen – nicht zuletzt deswegen, da mit dem fortschreitenden Aussterben von Flurnamen auch das Interesse an diesen zu verschwinden scheint. Um eine signifikant höhere Reichweite erreichen zu können, ist es m.E. unbedingt notwendig, „Hypertext-Flurnamenbücher" auch im Internet zu publizieren, nicht nur, um den wissenschaftlichen Austausch und Diskurs zu erleichtern, sondern auch, um das Interesse von Personen, die mit einem traditionellen Print-Flurnamenbuch nicht erreicht werden können, für einen Teil der sprachlichen Wirklichkeit zu gewinnen. Meist sind das Angehörige der jüngeren Generation, für die der Umgang mit dem Internet selbstverständlich ist.

Ein zentraler Zweck von Flurnamenbüchern ist die Dokumentation des Belegmaterials. Im SHFLNB mußte – wie in jedem Flurnamenbuch, das ein größeres regionales Untersuchungsgebiet bearbeitet und darstellt – eine repräsentative Auswahl der Belege getroffen werden, einerseits um aus linguistischer Sicht unbedeutende Belege auszuschließen und andererseits, um einigermaßen handlich, benutzerfreundlich und im Kaufpreis erschwinglich zu bleiben.

Unsere Erfahrungen mit der Erstellung des SHFLNB haben gezeigt, daß die theoretisch einsichtigen und folgerichtigen Prinzipien der Datenauswahl in der Anwendungspraxis häufig zu unbefriedigenden Ergebnissen führten. Eine deutlich weniger rigide Datenauswahl vornehmen zu müssen, scheint wünschenswert zu sein. Durch eine umfangreiche Datendokumentation steigt aber zugleich auch der Bedarf an einer stärkeren Strukturierung und Qualifizierung des Materials, um

(a) nicht nutzlose, weil nicht benutzbare Datenmüllhalden zu produzieren und
(b) mögliche Rezeptions- und Navigationsprobleme zu vermeiden.

Zu den neuen und veränderten Möglichkeiten der quantitativen Dokumentation kommt ein gänzlich neuer qualitativer Aspekt: die visuelle und akustische Dokumentierbarkeit (Tonaufnahmen, interaktive ‚On-the-Fly-Karten' usw.).

Dieser Aufsatz ist als Versuch zu verstehen, die methodischen Innovationen, die neuen Perspektiven aber auch Probleme, die sich für ein Hypertext-Flurnamenbuch ergeben, darzustellen. Es scheint so, als könnte mit der Entwicklung eines Flurnamenbuchs als Hypertext ein – nach dem „Hessischen Flurnamenatlas" und dem Print-SHFLNB – methodisch weiterer Schritt vollzogen werden.

6 Literatur

Blumenthal, Andreas/Lemnitzer, Lothar & Storrer, Angelika (1988): Was ist eigentlich ein Verweis? Konzeptionelle Datenmodellierung als Voraussetzung computergestützter Verweisbehandlung. In: Harras, Gisela (Hg.): Das Wörterbuch. Artikel und Verweisstrukturen. – Düsseldorf. (Jahrbuch des Instituts für deutsche Sprache 1987. Sprache der Gegenwart; 74). S. 351–373.

Boesch, Bruno (1959/60): Die Auswertung der Flurnamen. – *Mitteilungen für Namenkunde*, H. 7, S. 1–9.

Drosdowski, Günther (1977): Nachdenken über Wörterbücher: Theorie und Praxis. In: Drosdowski et al. 1977. S. 103–143.

– (1985): Einige Anmerkungen zur heutigen Lexikographie. In: Stötzel 1985. S. 63–68.

– /Henne, Helmut & Wiegand, Herbert Ernst (1977): Nachdenken über Wörterbücher. – Mannheim; Wien; Zürich.

Eichler, Ernst/Hilty, Gerold/Löffler, Heinrich/Steger, Hugo & Zgusta, Ladislav (Hgg.) (1995): Namenforschung. Ein internationales Handbuch zur Onomastik. 1. Teilband. – Berlin; New York. (Handbücher zur Sprach- und Kommunikationswissenschaft; 11.1).

Greule, Albrecht (1984): Die Lexikographie der deutschen Ortsnamen. In: Wiegand, Herbert Ernst (Hg.): Studien zur neuhochdeutschen Lexikographie V. – Hildesheim; Zürich; New York. S. 135–157. (Germanistische Linguistik; 3–6/1984).

– (1990): Ortsnamenwörterbücher. In: Hausmann, Franz Josef/Reichmann, Oskar/Wiegand, Herbert Ernst & Zgusta, Ladislav (Hgg.): Wörterbücher. Ein internationales Handbuch zur Lexikographie. Zweiter Teilband. – Berlin; New York. (Handbücher zur Sprach- und Kommunikationswissenschaft; 5.2). S. 1276–1284.

– und Prinz, Michael (1999): Auf dem Weg zum digitalen Namenbuch. Multimedia in Namenforschung und Namendidaktik. In: Franz, Kurt & Greule, Albrecht (Hgg.): Namenforschung und Namendidaktik. Gerhard Koß zum 65. Geburstag. – Baltmannsweiler. S. 10–25.

– und Prinz, Michael und Korten, Heinz (1998): Multimedia in der Namenforschung. In: Lehner, Franz/Braungart, Georg & Hitzenberger, Ludwig (Hgg.): Multimedia in Lehre und Forschung. Systeme – Erfahrungen – Perspektiven. – Wiesbaden. S. 157–178.

Halfer, Manfred (1988): Die Flurnamen des Oberen Rheinengtals. Ein Beitrag zur Sprachgeschichte des Westmitteldeutschen. – Stuttgart. (Mainzer Studien zur Sprach- und Volksforschung; 12).

Händler, Harald (1987): Die computative Bearbeitung. In: Ramge 1987a. S. 25–27.

Henne, Helmut (1977): Nachdenken über Wörterbücher: Historische Erfahrungen. In: Drosdowski et al. 1977. S. 7–49.

Kammerer, Matthias (1998): Hypertextualisierung gedruckter Wörterbuchtexte: Verweisstrukturen und Hyperlinks. Eine Analyse anhand des *Frühneuhochdeutschen Wörterbuches*. In: Storrer, Angelika & Harriehausen, Bettina (Hgg.): Hypermedia für Lexikon und Grammatik. – Tübingen. (Studien zur deutschen Sprache; 12). S. 145–171.

– und Lehr, Andrea (1996): Potentielle Verweise und die Wahrscheinlichkeit ihrer Konstituierung. In: Wiegand, Herbert Ernst (Hg.): Wörterbücher in der Diskussion II. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. – Tübingen. (Lexikographica: Series Maior; 70). S. 311–354.

Kirkness, Alan (1985): Deutsche Wörterbücher – ihre Geschichte und Zukunft. Vorschläge zur lexikographischen Praxis. In: Stötzel 1985. S. 44–54.

Kuhlen, Rainer (1991): Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank. – Berlin; Heidelberg; New York.

Lemberg, Ingrid/Petzold, Sybille & Speer, Heino (1998): Der Weg des Deutschen Wörterbuchs in das Internet. In: Wiegand 1998. S. 262–284.

Ramge, Hans (unter Mitarbeit von Sonja Hassel-Schürg, Ulrich Reuling, Gerda Weigel, Bernd Vielsmeier – computativ bearbeitet von Harald Händler und Wolfgang Putschke) (Hg.) (1987a): Hessischer Flurnamenatlas. – Darmstadt. (Arbeiten der Hessischen Historischen Kommission NF; 3).

– (1987b): Einleitung: Der Aufbau und die namenkundliche Bearbeitung. In: Ramge 1987a. S. 11–24.

– (1989): Besprechung von Halfer 1988. – *Beiträge zur Namenforschung*, NF 20. Jg., H. 1/2, S. 254–257.

– (1990): Zur Konzeption regionaler Flurnamenbücher. Am Beispiel des künftigen „Südhessischen Flurnamenbuchs". In: Schützeichel, Rudolf & Seidensticker, Peter (Hgg.): Wörter und Namen. Aktuelle Lexikographie. Symposium Schloß Rauischholzhausen, 25.–27. September 1987. – Marburg. S. 97–121.

– (1995): Arbeits- und Darstellungstechniken der Namenforschung: Atlanten. In: Eichler et al. 1995. S. 312–317.

– (1996): Datenpräsentation, Artikelstruktur und Namenkontinuität im Südhessischen Flurnamenbuch. In: Tiefenbach, Heinrich (Hg.): Historisch-philologische Ortsnamenbücher. Regensburger Symposion, 4. und 5. Oktober 1994. – Heidelberg. (Beiträge zur Namenforschung: Beiheft; N.F. 46). S. 161–183.

– (1997): Projektantrag „Mittelhessisches Flurnamenbuch in multimedialer Form". Unveröff. Mskr. – Gießen.

Raymond, Darrell R. & Tompa, Frank Wm. (1988): Hypertext and the Oxford English Dictionary. – *Communications of the ACM*, 31. Jg., H. 7, S. 871–879.

Reichardt, Lutz (1984): Zur Anlage und Herstellung landschaftlicher Namenbücher. – *Beiträge zur Namenforschung*, NF 19. Jg., S. 184–200.

– (1995): Arbeits- und Darstellungstechniken der Namenforschung: Namenbücher. In: Eichler et al. 1995. S. 304–312.

Richter, Gerd (1999): Möglichkeiten und Probleme der linguistischen Dokumentation in Hypertextumgebungen. Studien zur elektronischen Umsetzung des Südhessischen Flurnamenbuchs. – Universität Gießen, Fachbereich Germanistik, Magisterarbeit.

Storrer, Angelika (1998): Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In: Wiegand 1998. S. 106–131.

Stötzel, Georg (Hg.) (1985): Germanistik – Forschungsstand und Perspektiven. Vorträge des Deutsche Germanistentages 1984. 1. Teil: Germanistische Sprachwissenschaft, Didaktik der Deutschen Sprache und Literatur. – Berlin; New York.

Wiegand, Herbert Ernst (1977): Nachdenken über Wörterbücher: Aktuelle Probleme. In: Drosdowski et al. 1977. S. 52–102.

– (1989): Der Begriff der Mikrostruktur: Geschichte, Probleme, Perspektiven. In: Hausmann, Franz Josef/Reichmann, Oskar/Wiegand, Herbert Ernst & Zgusta, Ladislav (Hgg.): Wörterbücher. Ein internationales Handbuch zur Lexikographie. Erster Teilband. – Berlin; New York. (Handbücher zur Sprach- und Kommunikationswissenschaft; 5.1). S. 409–462.

– (1996): Über die Mediostrukturen bei gedruckten Wörterbüchern. In: Zettersten, Arne & Pedersen, Viggo Hjørnager (Hgg.): Symposium on Lexicography VII. Proceedings of the Seventh Symposium on Lexicography May 5–6, 1994 at the University of Copenhagen. – Tübingen. S. 11–43.

– (Hg.) (1998): Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Kolloquium. – Tübingen. (Lexikographica. Series Maior; 84).

Gerd Richter, Gießen

Krzysztof Petelenz

Das Informationsdesign auf der Speicherungsebene eines zweisprachigen Online-Wörterbuchs Polnisch-Deutsch

1 Vorbemerkung

Seitdem Nachschlagewerke auf CD-ROMs erscheinen, vergrößert sich die praktische und kommerzielle Bedeutung des Mediums Hypertext für die Lexikographie.[1] Das Aufkommen des WWW und der Hypertext-Sprache HTML brachte neue Qualitäten der Informationsverknüpfung mit sich: offene Systeme, aktuelle Inhalte, erleichterte Kommunikation zwischen den Autoren und den Rezipienten und – dank der Interaktivität – bessere Berücksichtigung von Benutzerbedürfnissen. Deshalb stellt sich die Frage, wie diese im WWW entwickelten Möglichkeiten auch für bessere Online-Wörterbücher nutzbar gemacht werden können.

Der folgende Beitrag sieht sich als Versuch, die Grundlagenarbeit für die Präsentation eines zweisprachigen Wörterbuchs in der Form eines hochwertigen Hypertextes zu leisten. Insbesondere möchte ich im folgenden die Strategie der Strukturierung und Speicherung von Wörterbuchdaten erörtern und am Beispiel eines polnisch-deutschen Wörterbuchteils vorstellen. Zuerst werde ich jedoch kurz die Stellung der zweisprachigen Computerwörterbücher innerhalb von Hypermediasystemen diskutieren und auf einige Besonderheiten der

[1] Zum Begriff Hypertext siehe J. Conklin (1987), R. E. Horn (1989), R. Kuhlen (1991), S. Freisler (1994); in Bezug auf die Lexikographie A. Storrer in diesem Band.

zweisprachigen Lexikographie im allgemeinen hinweisen, die für die Strukturierung von zweisprachigen Online-Wörterbüchern von entscheidender Bedeutung sind. Besonderes Augenmerk gilt auch der Verwendung und der Kategorisierung von Bildern, denen in dem neuen Medium endlich genügend Platz eingeräumt werden kann. Abschließend will ich noch auf einige „Mehrwerte" eines Online-Wörterbuchs eingehen- u.a. die Vernetzungsmöglichkeiten mit anderen Web-Projekten sowie auf die Möglichkeit kontinuierlicher Erweiterung der virtuellen Wörterbücher im Internet.

2 Das Computerwörterbuch als Hypermediasystem

2.1 Die Tendenz zur Kumulation der lexikographischen Inhalte

Die logische Zusammenführung bestehender Informationssysteme – und jedes zweisprachige Wörterbuch kann man als ein Informationssystem verstehen – hat sich in den letzten Jahren im Bereich der Informationswissenschaften zu einem wichtigen Untersuchungsgegenstand entwickelt. Seitdem hypertextbasierte Netzdienste wie das World Wide Web Zugriff auf verschiedene, heterogene Informationsquellen ermöglichen, begannen die Anstrengungen um transparente föderative Oberflächen, die Informationen aus unterschiedlichen Online-Wörterbüchern integrieren. Gleichzeitig bemüht man sich um die Aufbereitung der Informationen aus bereits bestehenden Quellen, so daß bei Neuentwicklungen auf das schon gesammelte lexikalische Wissen zurückgegriffen werden kann.

Im Bereich der lexikographischen Online-Angebote sind Aktivitäten beider Art anzutreffen.[2] Es wurde zum Beispiel ein WWW-Projekt angekündigt, in dem eine wörterbuchübergreifende Präsentationsform für Inhalte verschiedener zweisprachiger Wörterbücher mit den Sprachen Englisch, Spanisch und Baskisch gefunden werden soll (vgl. Patric; Zubigulla 1994). Atkins (1996) schlägt eine Verbindung von zwei einsprachigen Wörterbüchern mit einem polyfunktionalen, bidirektionalen zweisprachigen Wörterbuch für das Sprachenpaar Englisch-Französisch vor – die Daten dieses virtuellen Nachschlagewerkes sollen in einer Frame-basierten Datenbank gespeichert und online zugreifbar sein.[3] Das einsprachige „Hypertext Webster Gateway" ermöglicht schließlich unter einer einheitlichen Bedienungsoberfläche den Zugang zu mehreren einsprachigen Englischwörterbüchern; durch Anklicken eines in der Definition auftretenden Wortes wird ein Hyperlink zu demjenigen Backend-Wörterbuch aktiviert, aus dem die jeweilige Definition stammt, und ein neuer Nachschlagevorgang für dieses Wort ausgelöst.[4]

---

[2] Um den Aufbau wiederverwendbarer, in mehreren Kontexten nutzbarer lexikalischer Online-Ressourcen bemühen sich z.B. die Projekte *Deutscher Wortschatz* (30.08.1999, http://wortschatz.uni-leipzig.de/inhalt/) und *LEKSIS* (30.08.1999, http://www.ids-mannheim.de/wiw/).

[3] Frames sind komplexe Datenstrukturen, die stereotypische Situationen als Module des syntaktisch-semantischen Wissens über bestimmte Sprachwirklichkeitsausschnitte repräsentieren (siehe Wegner 1985, Lowe, Baker, Fillmore 1997). Atkins (1996:532ff.) schlägt vor, das Frame-Konzept – als ein linguistisches Prinzip der Modellierung eines Datenbankschemas – zur Überwindung der kognitiven Unzulänglichkeit (vgl. Kap. 3.1) des zweisprachigen Wörterbuchs zu benutzen.

[4] Vgl. (30.08.1999) <http://work.ucsd.edu:5141/cgi-bin/http_webster>.

Während derartige Modelle im Internet noch am Anfang ihrer Entwicklung stehen, lassen sich unter lexikographischen Produkten, die für die Offline-Nutzung bestimmt sind, etliche kumulative Ansätze finden (vgl. dazu auch Petelenz 1998). In den neunziger Jahren sind mehrere Wörterbücher – meist Deutsch-Englisch/Englisch-Deutsch sowie deutsche Fremd- und Synonymwörterbücher – auf CD-ROM in Verbindung mit anderen Nachschlagewerken wie Lexikon, Enzyklopädie, Landkartenatlas u.ä. erschienen. Da sie in sich Texte, Bilder und Ton vereinigen, lassen sie sich als Hypermediasysteme bezeichnen.[5] Untereinander verbundene lexikographische Kompendien wie DUDENs LexiRom, Bertelsmanns INFOROM oder Infopedia von Tewi gestatten dem Benutzer den Zugriff auf Informationen aller Werke gleichzeitig, da die Daten zentral in einer Datenbank abgelegt wurden. In diesen lexikographischen Megawerken der ersten Stunde geht man oftmals davon aus, daß das Übermaß an Information nicht schaden kann. Schlägt man in der Infopedia 2.0 im deutsch-englischen Wörterbuch das Lemma *Auge* nach, weil man z.B. die Übersetzung für die deutsche Redewendung *unter vier Augen* sucht, so bekommt man ungefragt – zusammen mit dem Eintrag des Oxford-Taschenwörterbuchs Dt.-Engl. – eine Bildanimation zu sehen und einen kurzen Vortrag (auf Deutsch) über die Funktion und Aufbau dieses menschlichen Sinnesorganes zu hören. Die englische Sprachausgabe oder IPA-Transkription für *eye* oder *in privacy* wird dem Benutzer aber nicht angeboten. Infopedia 2.0. erlaubt es nicht, die Suche im Datenbestand auf bestimmte Nachschlagewerke einzuschränken. Das deutet auf die Unterschätzung der Hypertextfunktionalität hin; statt dem Benutzer die Möglichkeit zu geben, gezielt weiterführende Informationen vom System anzufordern, wird er in diesem Fall mit enzyklopädischen Angaben versorgt, die bei dieser Nachschlageoperation vollkommen redundant und von den meisten Benutzern sicherlich auch nicht erwünscht sind. Ein derartiges Überangebot an Informationen läßt sich nach Conklin (1997:40) als Beispiel für eine Ablenkung vom eigentlichen Ziel des Nachschlagens („disorientation") und für die kognitive Überlast („cognitive overhead") bezeichnen, die allerdings in diesem Fall nicht aus der verwirrenden Menge der einzuholenden Auskünfte resultiert, sondern aus einer Unterschätzung der Fähigkeiten des Benutzers, sich die im gegebenen Kontext interessierenden Informationen selbst aussuchen zu können.

Das Konzept der PC-Bibliothek von DUDEN in Deutschland und des LEKSYKONIA-Systems in Polen setzt ebenfalls auf die Versorgung der Benutzer mit Informationen aus verschiedenen Nachschlagewerken auf einmal, d.h. bei einem Nachschlagevorgang werden – falls der Benutzer es wünscht – mehrere Datenbestände durchsucht. Während Infopedia nicht immer die Informationen aus unterschiedlichen Werken auf eine gelungene benutzerfreundliche Art miteinander kombiniert, funktioniert die PC-Bibliothek in dieser Hinsicht wesentlich besser (vgl. dazu auch Neth; Swanson 1999:106). Auch das reichlich bebilderte dt.-poln./poln.-dt. LEKSYKONIA-Wörterbuch läßt sich um ein externes dt.-poln./poln.-dt. Fachwörterbuch der Wirtschaftssprache oder ein einsprachiges polnisches illustriertes Lexikon der technischen Termini erweitern. So besteht die Möglichkeit, den Bestand des allgemeines Wörterbuchs gleichsam um Lemmata und zusätzliche Angaben weiterer Werke zu vergrößern. Auf diese Weise kann man je nach Bedarf in einer, in zwei oder in mehreren „Informationsbasen", wie sie von den Autoren des LEKSYKONIA-Systems benannt wurden, gleichzeitig nachschlagen.

Die Tendenz zur Angabe heterogener Informationstypen in einem zweisprachigen Wörterbuch scheint sich auch auf die Printlexikographie auszuweiten: das Power-Englischwörterbuch von Langenscheidt (mit dem Untertitel „Zum Nachschlagen und Lernen") ver-

[5] Zur Definition der Hypermedia in Bezug auf Lexikographie siehe Storrer (1998:107).

einigt in sich – außer den „reinen" Angaben eines zweisprachigen Wörterbuchs – bisher selten verwendete Elemente. So sind da etwa farbige Bilder zu finden, die jeweils entweder einen einzelnen Begriff oder einen größeren Ausschnitt der Wirklichkeit illustrieren und das entsprechende, themengebundene Vokabular samt Beispielsätzen dazu liefern. Kleine Kästchen, die immer wieder die herkömmliche Makrostruktur unterbrechen, enthalten zusätzliche „Sprachglossen" – wie sie im Wörterbuchvorwort genannt werden – mit Informationen zum richtigen Sprachgebrauch, zu grammatischen, pragmatischen und phonetischen Schwierigkeiten, „falschen Freunden" und kulturbezogenen Hinweisen zur Landeskunde. Der Benutzer kann folglich zwischen den Wörterbuchartikeln und den ergänzenden Informationstafeln „hin- und herspringen", ähnlich wie zwischen den Fenstern einer Computeranwendung.

2.2 Drei Generationen von zweisprachigen Hypermediawörterbüchern

Artents und Bogaerts (1991) unterscheiden zwischen drei Generationen von Hypermediasystemen. Die erste Generation arbeitet mit festen Standardattributen, die mit Knoten, Ankern und Links verbunden sind.[6] Die Knoten beinhalten hauptsächlich Text, eventuell integrierte Medieninhalte dienen als ergänzende Illustration, die Benutzer können weder Knoten noch Links verändern, die Kommunikation mit anderen PC-Anwendungen, wie z.B. Textverarbeitungsprogrammen, ist nicht vorgesehen. Zu dieser Gruppe zählen viele der heute verfügbaren zweisprachigen PC-Wörterbücher auf CD-ROM aus der niedrigeren Preisklasse (vgl. Heth, Swanson 1999).

Die zweite Generation eröffnet die Möglichkeit, dem System benutzerdefinierbare Knoten und Links, Attribute und Schlüsselwörter hinzuzufügen. Benutzer können auf diese Weise – wie das beispielsweise in Langenscheidts Handwörterbuch Englisch aus der PC-Bibliothek 2.0. der Fall ist – neue Einträge, Querverweise und Schlagwörter definieren, Artikeltexte der einzelnen Knoten mit Anmerkungen versehen und mit „Leuchtstiften" kennzeichnen. Über eine DDE-Schnittstelle lassen sich Wörter aus einer Textverarbeitung heraus nachschlagen und die Knoteninhalte über die Zwischenablage ganz oder teilweise in andere Anwendungen übernehmen.[7] Gleichzeitig ist es möglich, ohne das System zu verlassen, auf weitere gleich strukturierte Datenbankinhalte – d.h. andere Wörterbücher und Lexika desselben Herausgebers – zuzugreifen. Das LEKSYKONIA-Wörterbuch, das „Umlaut"-Wörterbuch und der „Grundwortschatz Polnisch" im GlobeDisc-System können z.T. mit einer Funktionalität aufwarten, die es erlaubt, diese Produkte zu der zweiten Generation zu rechnen.[8]

---

[6] Knoten sind die im Hypertext mit Verknüpfungen (Links) verwobenen Inhalte (multimediale Informationseinheiten: Text, Bild, Ton), Anker stellen Ausgangspunkte der Links dar, mit den Links wird aus einer Informationseinheit auf Informationen in anderen Knoten verwiesen. Die Abbildung der Verweisbeziehungen zwischen den Informationseinheiten (lexikographischen Textsegmenten) eines Printwörterbuchs mithilfe von Links beschreibt ausführlich – am Beispiel der einsprachigen Lexikographie – Kammerer (1998). Zu den Verweisbeziehungen in Wörterbüchern allgemein vgl. Wiegand 1996a.

[7] DDE steht für *D*ynamic *D*ata *E*xchange.

[8] Augenfällig ist jedoch die Diskrepanz zwischen dem Einsatz fortgeschrittener Softwaretechnik beim Erstellen des Wörterbuchbrowsers und der extrem schlechten lexikographischen Qualität der Daten im Fall des GlobeDisc-Wörterbuchs sowie der ziemlich konzeptionslosen lexikographischen Ausrichtung des LEKSYKONIA-Werkes.

Die dritte Generation von Hypertextsystemen sieht eine komplexe Systemunterstützung für die Handhabung der Knoteninhalte und für die benutzergerechte Navigation vor; sie setzt zu diesem Zweck sog. „link abstract structures" an. Diese Linkstrukturen stellen die Voraussetzungen für die Entwicklung einer wissensbasierten Systemarchitektur dar, die die Kluft zwischen dem „reader's mental model" und dem „system's internal model" beseitigen würde (vgl. Arents, Bogaerts 1991:133). Das Konzept beruht auf der Integration eines semantischen und pragmatischen Rahmens für Inhalt, Kontext und Struktur der Informationen auf den Ebenen der Links, der Knoten und des gesamten Systems. Die Idee der abstrakten Linkstrukturen läuft auf die Behandlung von Knoten und Links als gleichrangige Objekte hinaus – die Links seien keine einfachen Attribute der Knoten, sondern völlig eigenständige Objekte, die je nach Verwendungskontext zur Restrukturierung des Hypertextes dienen können. Diese Restrukturierung setzt eine saubere funktionale Trennung zwischen verschiedenen Typen der Semantik, eine statische Struktur der Informationen sowie eine klar definierte Navigationssemantik voraus. Die von Arents und Bogarts vorgeschlagene MMVP-Architektur (Model-Map-View-Praxis), die in einem Prototyp IKON (Intelligent Knowledge Objects Navigator) realisiert wurde, basiert auf zwei Prinzipien:[9]

- „Links do not express meaning by themselves, but express meaning through their navigation",
- „Link navigation through message passing" – eine Anfrage des Benutzers an das System wird durch den „message-passing"-Browsingmechanismus bearbeitet (vgl. Arents, Bogaerts 1991:137, 138).

Die Ideen von Arents und Bogaerts können auch als Anregung für die Computerlexikographie dienen. Ein elektronisches Wörterbuch, das den Ansprüchen der dritten Generation der Hypermediasysteme genügen sollte, müßte folglich über semantisch und pragmatisch typisierte Links zur Steuerung und Kontrolle der Systemunterstützung für unterschiedliche Benutzergruppen und Benutzungssituationen verfügen (siehe Kap. 4. 1); die Knoten der Hypertextbasis müssen zugleich mit referentiellen Links verbunden sein, die die gegenseitigen Beziehungen der Knoten untereinander ausdrücken. Im derart konstruierten Wörterbuchsystem soll auf der einen Seite ein Formalismus zur wissensbasierten Strukturierung der Informationsinhalte, der Knoten und der Links implementiert sein, auf der anderen Seite soll eine erfolgreiche Suche und eine freie Navigation (individuelles Browsing) gewährleistet sein.

Darüber hinaus sollte ein innovatives Computerwörterbuch ein offenes, erweiterbares System darstellen und – anders als eine Offline-Anwendung auf einem lokal verfügbaren Speichermedium – eine breite Kommunikation zwischen den Wörterbuchautoren und -benutzern ermöglichen (siehe den Beitrag von I. Lemberg in diesem Band). Dank dieser Kommunikation lassen sich die Knoteninhalte korrigieren, ergänzen und erweitern.

---

[9] IKON speichert Informationen in Form von „information units"; sie sind klein, exakt definiert und dienen der Dekontextualisierung von Informationfragmenten, so daß die korrespondierenden Informationseinheiten sinnvoll miteinander verbunden und manipuliert (wiederverwendet und rekombiniert) werden können. Die „information units" werden im semantischen Hyperindex charakterisiert. Zur semantischen Hyperindexing-Technik siehe Arents, Bogaerts (1993).

2.3 Neue Perspektiven für zweisprachige Computerwörterbücher

Völlig neu entwickelte elektronische zweisprachige Wörterbücher, sowohl auf CD-ROM als auch im WWW, fehlen noch (vgl. dazu auch Feldweg 1997). Der Grund dafür ist auf der einen Seite die Zurückhaltung der Verlage, wenn es darum geht, viel Geld in die neuen Medien zu investieren, und sie ist auch verständlich, solange die elektronischen Ressourcen auf Speichermedien dem zügellosen Raubkopieren ausgesetzt sind und solange keine Möglichkeiten der Abrechnung für virtuelle WWW-Angebote bestehen. Auf der anderen Seite fehlt wohl einigen Verlagen noch das Know-how, um die Informationen medienneutral aufzubereiten und in einer zentralen Datenbank aufzubewahren. Es ist vor allem nicht möglich, die linear abgefaßten Wörterbuchartikel der Printlexikographie maschinell in eine mehrdimensionale, hierarchisch organisierte Hypertextstruktur zu überführen, da die paradigmatischen Querverbindungen auf der Ebene der Semantik nur aufgrund intellektueller Analyse abbildbar sind. Ebenso lassen sich die Textverdichtungen in den Wörterbuchartikeln nicht völlig automatisch auflösen. Die CD-ROM-Publikationen auf der Basis herkömmlicher Printwörterbücher – auch wenn sie recht gelungen sind, wie z.B. die elektronische Version des OED oder des Petit Robert – orientieren sich immer noch am linearen Textparadigma. Das 1996 von Heidecke skizzierte Redaktionssystem bringt hinsichtlich der Hypertextualisierung von zweisprachigen Wörterbüchern wenige Innovationen, da es haupsächlich auf Erstellung von Druckwerken ausgerichtet wurde.

Eine endgültige Technologie der Erstellung und Wartung von großen, komplexen elektronischen Dokumenten und zur Erschließung der Papierwelt für Computer hat sich noch nicht etabliert. Man kann die Bemühungen um die innovativen Publikations- und Kommunikationsmethoden mit dem vagen Begriff Wissensmanagement (Knowledge Management) bezeichnen, der in der Informationswissenschaft u.a. als Oberbegriff für Document Engineering und intelligentes Information Retrieval funktioniert. Document Engineering verbindet die Methoden der angewandten Text- und Computerlinguistik und der objektorientierten Softwareentwicklung; alle Arbeitsschritte werden zu einem ingenieurmäßigen, linguistisch fundierten Vorgehen gebündelt (vgl. Kesseler; Freisler 1995, Williams 1997). Die eigentlichen Probleme des Document Engineerings liegen nicht nur auf der technischen Ebene der Textproduktion und -präsentation, sondern vorwiegend auf der semantischen Ebene: Die bestehenden Informationen müssen analysiert und in Bezug auf ihre Beschaffenheit, Bedeutung und Funktion charakterisiert werden. Für eine Computeranwendung, die die Informationen handhaben soll (d.h. dynamisch zueinander in Verbindung setzen soll), muß außer der Information selbst auch ihre Semantik repräsentiert werden. Document Engineering entwickelt Techniken, mit denen der semantische Gehalt von Informationen und die Semantik der Relationen zwischen einzelnen Informationen explizit und deklarativ modelliert und mittels geeigneter Schlußfolgerungsverfahren interpretiert werden. Benutzermodelle sollen die Erstellung von „maßgeschneiderten" Artikeln möglich machen. Im folgenden wird eine mögliche Organisation der Informationen eines zweisprachigen Wörterbuchs in einer Hypertextbasis skizziert, die – um eine Navigationskomponente ergänzt – den Benutzern die lexikographischen Daten auf eine innovative Art und Weise präsentieren könnte. Auf die Aspekte des Retrievals von Wörterbuchdaten wird aus Platzgründen nicht eingegangen.

Zunächst möchte ich jedoch die spezifischen Probleme des zweisprachigen Wörterbuchs überblickartig beleuchten, wohl wissend, daß sie den mit der zweisprachigen Metalexikographie vertrauten Leserinnen und Lesern hinlänglich bekannt sein dürften. Diese Probleme spielen aber eine entscheidende Rolle für das lexikographische Informationsdesign –

und umgekehrt: ein gelungenes Informationsdesign kann viele dieser Probleme lösen –, sie können also im vorliegenden Beitrag nicht unerwähnt bleiben.

3 Zu den Besonderheiten der zweisprachigen Lexikographie

Um sich der Anordnung der lexikographischen Daten in einer Hypertextbasis zu widmen, muß man zuvor kurz auf die Grundbausteine der zweisprachigen Printwörterbücher zu sprechen kommen. Die Werke lassen sich nach Skopus, Funktion, Direktion und dem primären linguistischen Ordnungsprinzip des Datenmaterials charakterisieren (vgl. Hausmann, Werner 1991).

Nach dem Skopus sind ein- und zweiteilige Werke zu unterscheiden, z.B. ein dt.-poln. und poln.-dt. Wörterbuch (biskopal) vs. ein poln.-dt. Wörterbuch ohne Pendant für die Gegenrichtung (monoskopal). Ein anderes Kriterium ist das viel diskutierte Aktiv-Passiv-Prinzip, d.h. die Ausrichtung des Werkes auf die „aktive" Hinübersetzung und Textproduktion in der ZS[10] bzw. „passive" Herübersetzung und Textrezeption in der AS (folglich jeweils monofunktional) oder für beide diese Funktionen gleichzeitig (bifunktional).[11] Als bidirektional läßt sich ein Wörterbuch bezeichnen, wenn es die Bedürfnisse beider Benutzergruppen, d.h. der Sprecher der AS und der Sprecher der ZS, gleichermaßen gut befriedigt. Monodirektional ist hingegen ein solches, das grundsätzlich nur eine der beiden Benutzergruppen als Zielpublikum ins Auge faßt.

Weitere Kriterien stellen die Paradigmatik und die Syntagmatik dar (vgl. Hausmann 1991). Sie determinieren die Materialanordnung in der Makro- und Mikrostruktur des Wörterbuchs. Während sich die paradigmatischen zweisprachigen Werke (wie Bildwörterbuch, thematischer Thesaurus oder Wörterbuch der kontrastiven Synonymik) der Onomasiologie und der lexikalischen Austauschbarkeit widmen, beschreiben syntagmatische Wörterbücher (Idiomatik-, Valenz-, Kontextwörterbuch etc.) die kontextuellen und grammatischen Abhängigkeiten der Lexik, ihre Gebundenheit in phraseologischen Einheiten. Im Zentrum der Aufmerksamkeit stehen dabei die Translate, d.h. die syntagmatischen Übersetzungseinheiten. Diese müssen nicht identisch sein mit den aus einer innersprachlichen Analyse hervorgegangenen Phrasemen, das betrifft Kollokationen, Routineformeln und „stehende" Vergleiche. Ein allgemeines zweisprachiges Wörterbuch versucht sowohl die paradigmatische als auch die syntagmatische Betrachtungsweise in sich zu vereinen, mit dem Resultat, daß an vielen Stellen große Abstriche in der lexikographischen Beschreibung auf allen Ebenen gemacht werden müssen. Die mangelhafte Behandlung der Phraseme im zweisprachigen Wörterbuch wird in Kap. 4. 4 näher besprochen.

Pädagogische Überlegungen legen den Gedanken nahe, zwischen den Bedürfnissen der Laienbenutzer (die weder mit der Zielsprache gut vertraut sind noch sich mit den lexikographischen Konventionen auskennen) und denen der kundigen Benutzer (die sowohl die Zielsprache gut beherrschen als auch über die linguistische Vorbildung und die nötige Übung in der Wörterbuchbenutzung verfügen) zu unterscheiden. Diese Differenzierung soll jedoch an dieser Stelle lediglich signalisiert und nicht weiter diskutiert werden.

---

[10] ZS = Zielsprache, AS = Augangssprache.

[11] Die Zusammenfassung der Diskussion um das Aktiv-Passiv-Prinzip ist bei Tarp (1995) zu finden.

4 Der typenübergreifende Wörterbuchserver: Zur allgemeinen Vorgehensweise

4.1 Das Schema des Wörterbuchsystems

Der erste Schritt auf dem Weg zu einem typenübergreifenden Wörterbuchserver ist die Analyse der lexikographischen Textsegmente verschiedener Typen von zweisprachigen Wörterbüchern für Deutsch-Polnisch. [12] Die dabei isolierten mikrostrukturellen Textsegmente werden anschließend im Hinblick auf die Bifunktionalisierung und/oder Bidirektionalisierung um entsprechende Angabetypen ergänzt. Die Bifunktionalisierung und Bidirektionalisierung führt z.T. zur Vereinheitlichung, Restrukturierung und Expandierung der in den Printwörterbüchern vorgesehenen Informationstypen.

Die eingescannten Artikel verschiedener Wörterbücher für Deutsch-Polnisch bilden ein Korpus mit den lexikographischen Daten. Diese Daten werden (korrigiert und ergänzt) in die Lexikoneinträge der Hypertextbasis übernommen. In die so entstandene Gesamtstruktur fließen noch zusätzliche Informationstypen mit ein, die in der zweisprachigen Printlexikographie entweder ansatzweise (umfangreiche Textbeispiele und Bilder) oder überhaupt nicht (Töne) vertreten sind. Die ursprüngliche Erfassung erfolgt im Textverarbeitungsprogramm (MS Word), den einzelnen Datentypen werden Absatz- und Zeichenformatvorlagen zugeordnet. Anschließend ist die Konvertierung der Daten in eine XML-Datei vorgesehen – die formbezogenen Informationen aus den Word-Stylesheets sollen durch semantische Tags ersetzt werden[13]. Schließlich findet die Übernahme der lexikographischen Daten in die wörterbuchtypen-übergreifende Struktur statt. Diese Struktur, die sich als „Export-Standard" im Sinne von Bläsi et al. (1994) bezeichnen läßt, muß auf der Speicherungsebene des Wörterbuchservers mithilfe spezieller Software – SchemaText – modelliert werden (vgl. Abbildung 3).[14] Zuvor wird jedoch der Entwurf einer Hypertextbasis in Angriff genommen: Das konzeptuelle, implementierungsunabhängige Schema der Hypertextbasis setzt sich auf der einen Seite aus den typisierten Knoten zusammen, die die ‚kohäsiv geschlossenen informationellen Einheiten' hierarchisch ordnen[15], und auf der anderen Seite aus den typisierten Links, die die Semantik der Relationen zwischen den Knotentypen festlegen. Die typisierten Links, die die einzelnen Knotentypen verbinden, stellen ‚informationelle Funktionen' (vgl. Kuhlen 1991:89) dar, sie drücken entweder die hierarchischen Relationen auf der syntagmatischen und paradigmatischen Ebene des Lexikons aus – die hierarchischen Links – oder sie bezeichnen die Beziehungen der Einheiten zueinander im bestimmten situativen Kontext – die pragmatischen Links.

Die höchste Hierarchiestufe der Hypertextbasis unterscheidet zwischen sprachlichen und metasprachlichen Einheiten. Die sprachlichen Einheiten bilden die Lemmata des Wörterbuchsystems, alle an das jeweilige Lemma adressierten Angaben stellen metasprachliche Einheiten dar. Alle Einheiten zusammen sind in eine semantisch und pragmatisch vordefinierte Struktur organisiert, die modular aufgebaut ist, wie die Abbildung 1 zeigt. Die Unterscheidung auf der Ebene der sprachlichen Einheiten führt zu dem Mehrwortlexem-Modul

---

[12] Analysiert wurden folgende Wörterbücher: DBWDP, GWPD/DP, HWBPD/DP, ISNP, SWBPD, STJN und TWBP.

[13] XML steht für E*X*tensible *M*arkup *L*anguage.

[14] (30.8.1999) http://www.schema.de/sitehtml/site-d/schemat2.htm .

[15] Die zusammenhängenden informationellen Einheiten (‚information units') können größere Einheiten (sog. ‚information chunks' oder ‚informations blocks', vgl. dazu Horn 1989:40ff.) bilden.

(aus Idiomen, Kollokationen und Sprichwörtern) und Einwortlexem-Modul (aus Verben, Nomen, Adjektiven/Adverbien und einer „Restkategorie").

Die Unterscheidung auf der Ebene der metasprachlichen Einheiten führt zu dem Monosem-Modul und Polysem-Modul, wobei die Templates beider Module (in der Abb. 1 der Hauptknotentyp ‚Bedeutung') gleich strukturiert sind; das Schrift- und Lautform-Modul befindet sich auf der selben Hierarchiestufe. Auf der nächstniedrigeren Ebene befinden sich einzelne Beschreibungssegmente der ZS-Äquivalente für die jeweilige Lemmabedeutung, die je nach Wortart unterschiedliche Slots mit Angaben zur Herkunft, Grammatik, Semantik und Pragmatik sowie Beispielen und Bildern aufweisen können.[16] Die Knoteninhalte mehrerer Slots können identisch sein (z.B. eine Paraphrase verschiedener Redewendungen mit gleicher Bedeutung), was bedeutet, daß derselbe Knoten in mehreren Kontexten verwendbar sein kann.

Die herkömmlichen zweisprachigen Printwörterbücher ermöglichen den Zugriff auf die Lexik in der Regel nur „auf der Form-, nicht auf der Bedeutungsebene" (Martin 1994:17.; vgl. auch Storrer 1998:116). Die hierarchischen Links stellen also Verknüpfungen hauptsächlich nach semantischen Kriterien her und eröffnen dadurch neue Zugriffswege zum lexikalischen Material. Die pragmatischen Links haben hingegen die Aufgabe, die Generierung der benutzergerechten Artikel zu steuern.

Auf der nächstniedrigeren Ebene kommt der Parameter Adressatenkreis richtig ins Spiel – die metasprachlichen Angaben sind nämlich in der Hypertextbasis grundsätzlich doppelt vorhanden: der eine Angabetyp ist für die deutschen, der andere für die polnischen Benutzer bestimmt.

4.2 Die Typisierung als Grundlage der wissensbasierten Relationierung von Hypertexteinheiten

Die Anpassung der präsentierten Informationen an die Bedürfnisse der Benutzer geschieht in einem Interaktionsprozeß. Am Anfang einer Sitzung mit dem Wörterbuchsystem (d.h. auf der ersten HTML-Seite), wird der Benutzer gefragt, welche Muttersprache er spricht und an dieser Stelle somit die Entscheidung gefällt, welcher Teil des Schemas ausmaskiert bleibt und welche Expander für die Generierung der HTML-Seiten aus der Datenbasis zum Einsatz kommen. (Das Layout für die Zielformate wird in SchemaText in den sog. Expandern festgelegt.[17] Expander sind kleine Scheme-Programme, die das Layout für das Textschema, für Knotentypen, für Knoten, für Linktypen und für Links generieren. Sie können rekursiv weitere Expander aufrufen).[18]

Der „SchemaEditor" dient in SchemaText zur Modellierung von Informationstypen und deren Relationen. Dies erfolgt auf einer graphischen Oberfläche (vgl. Abb. 1). In einem Schema werden Typen von Textobjekten und Verweisbeziehungen bestimmt (Knotentypen: Idiom, Kollokation, Nomen; Äquivalent, Genusangabe usw. und Linktypen ‚Bezieht-sich-auf', ‚Ist-Teil-von', ‚Synonym-zu' etc.) und damit Operationen auf globalen und lokalen Textstrukturen ermöglicht. Komplexe heterogene Vernetzungsstrukturen können – zwecks

---

[16] Die Knoteninhalte mit Bildern und Tönen enthalten in der Tat lediglich einen Verweis auf die entsprechende Datei im Ressourcenpool mit binären Daten. Dieser Pool stellt ein separates Modul der Hypermediabasis dar (s. Kap. 4.4.2).

[17] Wie das Navigieren auf der Oberfläche des Wörterbuchsystems im WWW-Browser erfolgt, wird im Petelenz (2000:216ff.) skizziert.

[18] Scheme ist ein weitverbreiteter, international standardisierter LISP-Dialekt.

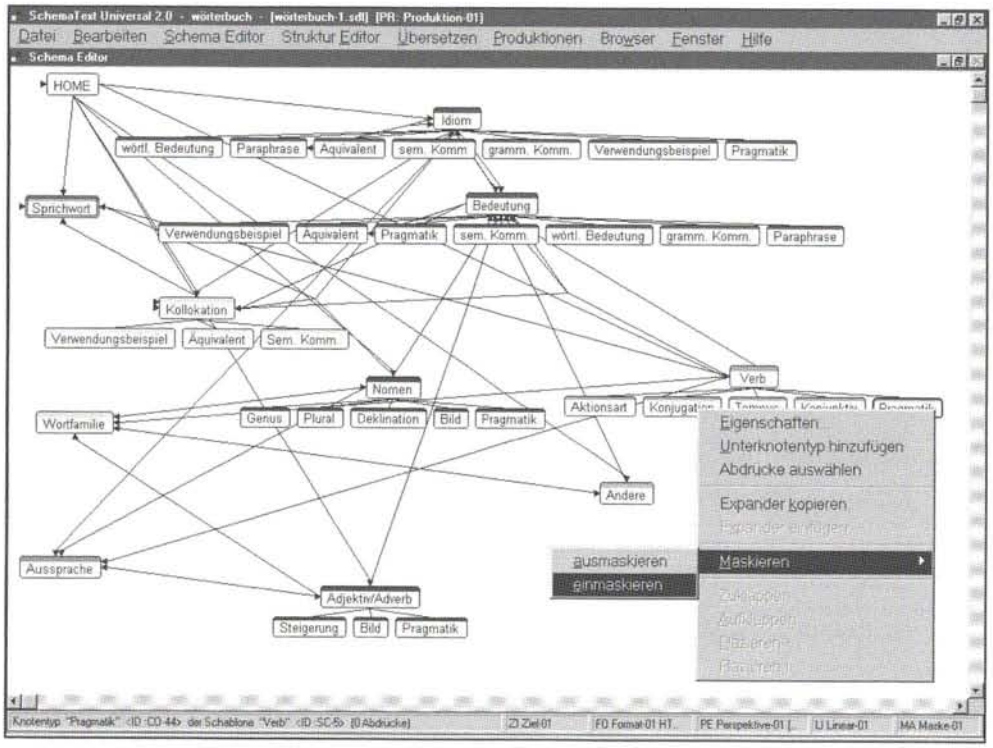


Abb. 1

Mehrfachverwendung und Rekombination der Inhalte – über sog. Multilinks verwaltet werden. Ein Nomen (als Knotentyp) kann z.B. gleichzeitig einen Teil von einer Kollokation, einem Idiom oder einem Sprichwort bilden.

Schema der Hypertextbasis			
Schema 1 WB Poln.-Dt.	Schema 2 WB Dt.-Poln.	Schema 3 Grammatikhilfe	Ressourcen-Pool Bild- und Tondateien

Abb. 2

Die Hypertextbasis setzt sich im SchemaText auf der abstrakten Strukturebene aus Schemata zusammen, diese wiederum aus den hierarchisch organisierten Templates (Schablonen). Jede Schablone hat in der Datenbasis eine physikalische Entsprechung: einen Abdruck, der aus Knoten und Links besteht. Ein Abdruck muß mit dem definierten Template konform sein. (Die Templates lassen sich allerdings auch nachträglich problemlos verändern). Die konkreten Textstrukturen des Wörterbuchs kann man im sog. „Struktur-Editor" festlegen. Die im Schema gespeicherte Struktur ist von den Instanzen selbst unabhängig, d.h. die Layouteigenschaften der Knoten und Links werden getrennt vom Schema

festgelegt. Die Vererbung bedeutet dann z.B. die Fähigkeit, Änderungen, die am Knotentyp durchgeführt wurden, auf alle Knoten dieses Typs zu übertragen.

Der Skopusbezug wird durch die Teilung der Hypertextbasis in zwei Wörterbuchschemata bestimmt: ein Schema enthält die deutschen Lemmata mit polnischen Äquivalenten, das andere die polnischen Lemmata mit deutschen Äquivalenten. Die Benutzerprofile (1–4 im Kap. 4.3) sind für jedes Schema einzeln zu implimentieren. Das dritte Schema enthält die Grammatikhilfe (vgl. Abbildung 5), ein für beide Benutzergruppen bestimmtes Handbuch, auf deren Einträge (Kapitel und Unterkapitel) kontextsensitiv aus dem Wörterbuch (Schema 1 und Schema 2 in der Abbildung 1) verwiesen wird. Das Handbuch läßt sich jedoch auch als ein inhärentes Hypertextdokument lesen, ähnlich wie etwa eine Onlinehilfe von üblichen Computerprogrammen. Die einmal erstellte Struktur des Textschemas für die Grammatikhilfe ist in beiden Handbüchern, d.h. in dem für die polnischen und in dem für die deutschen Benutzer, gleich. Der benutzergerechte Zugang wird über die sog. Perspektive gesteuert, also eine Dimension des Schemas, die in diesem Fall zwei Textstrukturkopien der aktuellen Textstruktur enthält, und zwar für beide Sprachen. In diesen zwei Sprachperspektiven sind die Knoteninhalte des Schemas verschieden, die Textstruktur und das verwendete Expanderset jedoch gleich.

4.3 Die expandierten Informationstypen für die Wörterbuchinstanzen

Das System hat Sorge dafür zu tragen, daß der Benutzer nicht mit für ihn redundanten Informationen vom Wörterbuchsystem überversorgt wird. Die Verknüpfung der informationellen Einheiten mit semantisch typisierten Links soll in der Interaktion mit dem Benutzer die dynamische Anpassung des Systems an die jeweilige Wörterbuchbenutzungssituation bewirken. Demnach sind Benutzerprofile zu definieren, sie basieren auf dem in der Künstlichen Intelligenz benutzen Konzept des Benutzermodells. Grundsätzlich sind vier Benutzerprofile denkbar:

1. Benutzerprofil: AS Deutsch ‚aktiv' (Hinübersetzung)
2. Benutzerprofil: AS Deutsch ‚passiv' (Herübersetzung)
3. Benutzerprofil: AS Polnisch ‚aktiv' (Hinübersetzung)
4. Benutzerprofil: AS Polnisch ‚passiv' (Herübersetzung)

Kuhlen (1991:330, 331) bezeichnet das Benutzermodell als ein wissensbasiertes Verfahren mit dem Ziel, „Informationssysteme mit dem Wissen über individuell oder stereotypisch definierte Nutzer oder Nutzerklassen zu versorgen, damit Systeme selektiv auf die Bedürfnisse der Nutzer beim Antwortverhalten oder beim Problemlösen eingehen können. [...] Benutzermodelle sind bei Hypertext notwendig, um die mit dem dialogischen Prinzip verknüpften Erwartungen einlösen zu können". „Dialogisches Prinzip" sei weiterhin eine „Konzeption von Hypertext, nach der System und Nutzer wechselnde Initiative im Dialog haben sollten, d.h. weder soll die Nutzung über vordefinierte Pfade vom System gesteuert werden, noch soll dem Nutzer, gemäß dem Prinzip der direkten Manipulation, die Zuständigkeit für das Navigieren alleine überlassen werden. Vielmehr soll das System in der Lage sein, dem Benutzer situationsgemäße Nutzungsangebote, z.B. über Pfade, zu machen, denen er folgen kann (oder auch nicht). Voraussetzung für die Realisierung dialogischer Prinzipien sind reiche, wissensbasierte Strukturierung der Hypertexteinheiten und differenzierte Verknüpfungsstrukturen [...]. Das Modellieren der Benutzermodelle kann folglich als

„realistische Zwischenstufe" zwischen „‚dummen' Volltextsystemen und ‚intelligenten' Wissensbanken" angesehen werden (vgl. Kuhlen 1991:61).

Die Modellierung von Benutzerprofilen soll mittels des Maskierungsmechamismus geschehen: Mit Masken ist es möglich, durch Eingabe von Regeln oder durch Selektieren auf der graphischen Oberfläche (vgl. das kontextsensitive Menü in der Abb. 1) Varianten eines Dokumentes als Untermenge des Gesamtdokumentes zu definieren. Die Variantenbildung erstreckt sich nicht nur auf die grobe Strukturebene, sondern auch auf die feinkörnige Inhaltsebene.[19] Das Aus- und Einmaskieren von Link- und Knotentypen in bestimmten Verwendungssitationen soll gemäß erstellten Benutzerprofilen erfolgen.

4.4 Die Behandlung der Phraseme

Auf der Ebene der Mikrostruktur in den Printwörterbüchern fällt auf, daß Phraseme –abgesehen von Ausnahmen – als infralemmatische Adressen in verdichteter Form in den Artikeln für Einwortlexeme auftreten. Dieses Verfahren wird den Mehrwortlexemen (MWL) nicht gerecht, da wichtige Angaben (die an MWL adressiert sein müßten) meist im allgemeinen zweisprachigen Wörterbuch fehlen, – zum einen, weil der verfügbare Platz nicht ausreicht, zum anderen, weil das grammatische Verhalten oder die stilistische Markierung des MWLs mit denen des Eingangslemma nicht übereinstimmen und die Wörterbuchautoren die Verwirrung der Benutzer vermeiden wollen. Die Phraseme und andere MWL sind deshalb als eigenständige Lemmata – und nicht als Sublemmata bzw. infralemmatische Adressen der Einwortlexem-Lemmata – in die Datenbasis des neuen Wörterbuchsystems aufzunehmen (vgl. Abbildung 1).[20] Das Phrasem-Modul eröffnet die Möglichkeit, die Mehrwortlexeme als eigenständige Lemmata ausführlich zu bearbeiten (vgl. Abbildung 3).

Das Nachschlagen der Phraseme soll von der Retrieval-Komponente des Wörterbuchsystems erleichtert werden. Hierzu könnte auf Elemente der bereits bestehenden computerlexikographischen Tools, wie etwa des von Xerox patentierten experimentellen LOCOLEX-Nachschlagesystems, zurückgegriffen werden. LOCOLEX leistet nicht nur morphologische Analyse, d.h. die Rückführung auf die Grundform und die Kompositasegmentierung, sondern löst darüber hinaus das Problem der Erkennung von „MWLs ausgehend von einzelnen Textwörtern, die Bestandteil des MWL sind" sowie der „Erkennung von getrennten Prefixverben und anderen komplexen Ausdrücken" (vgl. Thielen, Breidt, Feldweg 1998). Dafür wurde die sog. IDAREX-Grammatik (Breidt, Segond, Valetto 1996) benutzt[21]:

---

[19] Textteile können auch für bestimmte Medien bzw. Zielformate ein- oder ausgeschaltet werden.

[20] „Die in Nischen und Nestern gruppierten Lemmata heißen Sublemmata. Die an sie adressierten Angaben sind sublemmatisch adressiert. Sublemmata gehören zur Makrostruktur. Das erste Lemma einer Gruppe von Sublemmata heißt Eingangslemma (Nischeneingangslemma/Nesteingangslemma). Es ist also streng zwischen Sublemma und infralemmatischen Adressen zu unterscheiden. Während die Sublemmata der Makrostruktur angehören – sie können über einen alphabetischen Suchpfad gefunden werden (der freilich im Falle der Nester recht kurvenreich verlaufen kann) –, kommt infralemmatischen Adressen nur mikrostruktureller Rang zu" (Hausmann, Werner 1991:2747).

[21] IDAREX steht für *ID*ioms *A*s *R*egular *Ex*ressions. Über die Einschränkungen bei der Verwendung der IDAREX-Grammatik für die Beschreibung von MWL in einem englisch-polnischen Wörterbuch berichtet Piotrowski (1999). Diese Einschränkungen resultieren hauptsächlich aus der starken Orientierung an den MWL-Ansatzformen des „Oxford-Hachette English and French Dictionary" (vgl. Piotrowski 1999:118–119).

„Zweisprachige Wörterbücher enthalten [...] Mehrwortlexeme (MWL, z.B. *sein Teil abbekommen, in Erfahrung bringen*), feste Ausdrücke (z.B. *Eile mit Weile*) und grammatische Kollokationen (z.B. *sich in etw* or *zu etw verwandeln*), die für die Rezeption durchaus relevant sind und deshalb in ein maschinell verarbeitbares Format überführt werden müssen. [...] Zur Formalisierung von Mehrwortlexemen, festen Ausdrücken und grammatischen Kollokationen werden diese zuerst in ihre kanonische Form umgeformt, wofür wir die in den herkömmlichen Wörterbüchern gebrauchte Notation etwas erweitert haben. Die kanonische Form eines Mehrwortausdrucks enthält alle lexikalisch fixierten Komponenten in der Form, in der sie vorkommen müssen, bzw. bei morphologisch flexiblen Bestandteilen in einer ‚neutralen' Form (Verb im Infinitiv Präsens, Nomen im Singular etc.). Verbargumente werden mit Metavariablen (etw, jmd, NPakk) ausgedruckt. Außerdem wird mit ‚°' markiert, ob ein Wort morphologisch variiert werden kann. Lexikalische Varianten, die im Wörterbuch durch ‚/' getrennt sind, werden mit ‚^' geklammert, wenn sie mehr als ein Wort enthalten, um den Skopus deutlich zu machen". (Thielen, Breidt, Feldweg 1998:187–188; weiterführende Details zu IDAREX kann man auch bei Breidt, Segond, Valetto [1996:21ff.] nachlesen.)

Im Rahmen des EU-Projektes STEEL (Developing Specialized Translation/Foreign Language Understanding Tools for Eastern European Languages) wurden die ersten Erfahrungen mit der Anwendung von IDAREX-Formalismen zur Beschreibung der Phraseme in einem Wörterbuch mit der ZS-Sprache Polnisch gemacht (vgl. Piotrowski 1999). Inwieweit sich diese Technik für die Beschreibung von polnischen MWL eignet, muß in Zukunft noch untersucht werden.

5 Zum Problem der Standardisierung

Um lexikalische Daten in standardisierter Form als Hypertext zu organisieren, würde es ideal sein, auf Standards der Printlexikographie zurückgreifen zu können. Für Wörterbücher gibt es jedoch keine typen- und länderübergreifenden Standards, sondern spezifische lexikographische Konventionen, die einerseits aus der Tradition der Wörterbucherstellung in einem gegebenen Land resultieren und andererseits von den spezifischen Erfordernissen eines Wörterbuchtyps abhängen. Die meisten Konventionen betreffen die Textverdichtungsmechanismen, die für das Gebot der Platzersparnis in Printwörterbüchern von entscheidender Bedeutung sind. Sie sollen Wörterbücher handhabbar und benutzerfreundlich machen. Aber tun sie das immer? Mit Sicherheit ja – weil ohne Textverdichtungskonventionen überhaupt kein Wörterbuch möglich ist – aber zugleich nur sehr eingeschränkt. Viele Benutzer haben erhebliche Schwierigkeiten mit den Abkürzungen, der Nischentechnik und der Tildierung zurecht zu kommen. Das Alphabet als primäres Ordnungsprinzip im Printwörterbuch ist ebenfalls problematisch (vgl. auch Kap. 4.2.). Die Hypertextualisierung der zweisprachigen Wörterbücher wurde u.a. durch die Medienpädagogik und die Sprachdidaktik angeregt, aus der Überzeugung heraus, daß paradigmatische Navigationsstrukturen viel besser als die alphabetische Anordnung der Lemmata der Lernfunktion der zweisprachigen Wörterbücher gerecht werden. Die Konvention der alphabetischen Materialanordnung kann ein Beispiel für die psycholinguistische Unzulänglichkeit des zweisprachigen Wörterbuchs sein. [22] Die nach linguistischen Prinzipien gebildeten Nester wichen in

[22] Das alphabetische zweisprachige Wörterbuch ist deswegen psycholinguistisch unzulänglich, da es das Bild von der sprachlichen Wirklichkeit verfälscht: Die zielsprachlichen Äquivalente eines Lemma können vom Benutzer erst dann richtig erfaßt und behalten werden, wenn sie nicht nur im

zweisprachigen Wörterbüchern zunehmend den nach stupider Alphabetordnung gereihten Nischen, da nur diese Methode der Materialanordnung sich in der Printlexikographie als praxistauglich erwiesen hat. Innerhalb von „graphischen Nischen" (Hausmann, Werner 1991:2747) sind die Lemmata striktalphabetisch sortiert, ohne Rücksicht auf Semantik und Wortbildung. Für die Angabe der Antonyme, Hypo- bzw. Hyperonyme oder der Wortfamilie des Lemma ist in der Regel kein Platz. Ebenfalls gibt es keine Standards für die Behandlung von Phrasemen – unter welchem Stichwort ein Mehrwortlexem lemmatisiert wird, dies wird meist der individuellen Entscheidung des Autors überlassen.

Bei Experimenten mit der Nutzbarmachung von lexikographischen Daten aus traditionellen Wörterbüchern für sprachverarbeitende Systeme stellte man dennoch Überlegungen über abstrakte lexikographische Standards an. Solche Standards sind für die automatische Erkennung von groben syntaktischen Strukturen in Wörterbuchtexten sehr nützlich. Bläsi et al. (1994) unterscheiden zwischen internen Standards und Export-Standards für einzelne Artikel und ganze Wörterbücher. Die internen Standards stellen die abstrakte Artikelhierarchie (die je nach Wortart unterschiedlich ausgeprägt ist) dar, die Export-Standards bestimmen hingegen die Datenbankstruktur, in die die Wörterbuchinhalte überführt werden sollen. Die „Rohdaten" eines Wörterbuchs müssen folglich anhand von internen Standards validiert und anschließend an die Export-Standards angepaßt werden – dann lassen sie sich in einer Wörterbuchdatenbank[23] ablegen (vgl. Bläsi et al. 1994). In unserem Fall geht es jedoch nicht um das Parsen von einzelnen Wörterbüchern, die man komplett in eine Datenbankanwendung übernehmen möchte. Vielmehr geht es um eine exemplarische Analyse der Informationstypen der einzelnen Werke im Hinblick auf Skopus, Funktion und Direktion. Um eine Vorstellung davon zu bekommen, welche potentiellen Textsegmente ein biskopales, bifunktionales und bidirektionales Wörterbuch haben kann und haben soll, gilt es die Intention der Lexikographen nachzuvollziehen, mit der sie bestimmte funktionale Textsegmente in die Artikel ihrer Wörterbücher aufnahmen. Bei dieser Analyse leistete die metalexikographische Methode der exhaustiven funktional-positionalen Segmentierung von E.H. Wiegand (vgl. Wiegand 1986 und Wiegand 1996b) eine große Hilfe.

Da alle analysierten dt.-poln./poln.-dt. zweisprachigen Wörterbücher für menschliche Benutzer geschrieben wurden, genügen sie nur sehr bedingt einem internen metalexikographischen Standard, da sie in ihre Artikel oft von der schablonenhaften „Idealhierarchie" mehr oder minder abweichen. Das spielt aber keine große Rolle, da sie, wie schon gesagt – aus unterschiedlichen Gründen – nicht für das Parsen in Frage kommen.[24]

Bedeutungsfeld, sondern auch im Bezeichnungsfeld gesehen werden. Die Formebene des Alphabets läßt nur eine – nämlich die semasiologische – Betrachtungsweise des sprachlichen Zeichens zu, sie geht vom Wortkörper aus und gibt Äquivalente für die verschiedenen homonymen Bedeutungen an. Die Onomasiologie geht hingegen nicht von der Form, sondern vom Begriff aus; sie könnte folglich im zweisprachigen Wörterbuch die Verbindungen von einem Begriff zu unterschiedlichen Wortkörpern herstellen und diese innerhalb eines Bedeutungs-, Sach- oder Situationsfeldes je nach Bedarf verdeutlichen. Nur das gleich berichtigte Nebeneinander beider Betrachtungsweisen der Lexik wird den kognitiven und kommunikativen Aspekten des Fremdsprachenerwerbs und -gebrauchs gerecht.

[23] Als Wörterbuchdatenbank ist in diesem Kontext ein geparstes Wörterbuch zu verstehen – vgl. Bläsi, Koch (1991); Hauser, Storrer (1996).

[24] Die wichtigsten Gründe sind: 1. Das Fehlen maschinenlesbarer Vorlagen. 2. Die komplizierte urheberrechtliche Lage. 3. Die Notwendigkeit, die Wörterbuchdaten nicht nur strukturell, sondern vor allem sachlich in Hinblick auf Korrektheit, Aktualität und Vollständigkeit zu überprüfen. Daher würde sich der Aufwand, eine Mikrostrukturgrammatik für die Wörterbücher zu schreiben, nicht lohnen.

6 Die expandierten lexikographischen Datentypen in der Hypertextbasis

6.1 Verschiedene Angabetypen für verschiedene Wörterbuchbenutzungssituationen

Die Analyse der Wörterbuchtexte erlaubt Rückschlüsse auf die lexikographischen Datentypen, die in der Datenbasis des Online-Wörterbuchs ihren Platz finden sollen, um Bedürfnisse beider Benutzergruppen bei allen potentiellen Nachschlageoperationen zu befriedigen. Im polnisch-deutschen Teil sind somit folgende Angaben zu jedem Lemma vorgesehen:

Für die polnischen Benutzer, sowohl im Wort- als auch im Phrasem-Modul:

1. eine Paraphrase der jeweiligen Bedeutung des AS-Lexems (auf Polnisch)[25]
2. ein oder mehrere ZS-Äquivalente als Vorschläge für die Hinübersetzung
3. ein Beispiel der Verwendung des AS-Lexems im Kontext
4. eine Übersetzung des Beispiels
5. Informationen über Lautform (Rechtschreibung) und Aussprache (als Sprachausgabe und IPA-Angabe) zu jedem ZS-Äquivalent
6. ggf. eine Herkunftsangabe (auf Polnisch)
7. Bei sog. Nulläquivalenz eine ZS-Paraphrase der AS-Bedeutung (deutlich als solche gekennzeichnet) – wenn möglich – als ein einsetzbares Translat
8. ggf. semantisch-pragmatische Kommentare zur konnotativen Verwendung des jeweiligen ZS-Äquivalentes (auf Polnisch)
9. grammatische Konstruktionshinweise bezüglich der Valenz und Rektion des AS-Lemma (auf Polnisch)
10. ggf. ein kontextsensitiver Verweis auf ein entsprechendes Kapitel der polnischen Grammatikhilfe (d.h. zur kleinen Grammatik des Deutschen auf Polnisch)
11. grammatische Konstruktionshinweise bezüglich der Valenz und Rektion zu jedem ZS-Übersetzungsvorschlag (auf Polnisch)
12. paradigmatische Verweise auf sinnverwandte Lexeme, Synonyme, Antonyme, Hypo- und Hyperonyme des jeweiligen ZS-Äquivalenten (zur Abgrenzung von ZS-Äquivalenten deutlich als solche gekennzeichnet)
13. ggf. Verweis(e) auf graphische Illustration(en)
14. ggf. Verweis auf eine verdichtete Artikelübersicht[26]

Zusätzlich im Wort-Modul (d.h. wenn ein Einwortlexem als Äquivalent in Frage kommt) auf der Ebene des jeweiligen Semems:

1. ggf. lexikalische Informationen über die bevorzugten Kontextpartner des jeweiligen ZS-Äquivalentes (Verweise auf Kollokationen)
2. ggf. Verweise auf Redewendungen mit dem jeweiligen ZS-Äquivalent (Idiome, „stehende" Vergleiche, Routineformeln etc.)
3. ggf. Verweise auf Sprichwörter in denen das ZS-Äquivalent vorkommt
4. ggf. paradigmatische Verweise auf Mitglieder der Wortfamilie zur Veranschaulichung des Wortbildungspotentials des ZS-Äquivalentes

---

[25] Zur Angabe der jeweiligen Lemmabedeutung soll man anmerken, daß sie in vielen Fällen – besonders bei Verben und Adjektiven – mit der Angabe der Kollokation des Lemma gleichzusetzen ist.

[26] Das impliziert die Notwendigkeit, einige metasprachlichen Angaben zweifach, d.h. einmal in verdichteter und einmal in expliziter Form, in der Datenbasis abzulegen. Die verdichteten Angaben lassen sich auch in einer Printversion des Wörterbuchs einsetzen, falls solche generiert werden soll.

Für die deutschen Benutzer, sowohl im Wort- als auch im Phrasem-Modul:

1. Informationen über Lautform (Rechtschreibung) und Aussprache (als Sprachausgabe und IPA-Angabe) des AS-Lexems
2. ggf. eine Herkunftsangabe (auf Deutsch)
3. ggf. eine Paraphrase der jeweiligen Bedeutung des AS-Lexems (auf Deutsch)
4. ein Beispiel der Verwendung des AS-Lexems im Kontext
5. eine Übersetzung des Beispiels
6. ggf. semantisch-pragmatische Kommentare zur denotativen Rezeption des AS-Lexems (auf Deutsch)
7. ein oder mehrere ZS-Äquivalente als Vorschläge für die Herübersetzung
8. grammatische Konstruktionshinweise bezgl. Valenz und Rektion des AS-Lemma (auf Deutsch)
9. ggf. ein kontextsensitiver Verweis auf ein entsprechendes Kapitel der deutschen Grammatikhilfe (d.h. zur kleinen Grammatik des Polnischen auf Deutsch)
10. grammatische Konstruktionshinweise bezüglich der Valenz und Rektion zu jedem Übersetzungsvorschlag (auf Deutsch)
11. paradigmatische Verweise auf sinnverwandte Lexeme, Synonyme, Antonyme, Hypono- und Hyperonyme des jeweiligen AS-Semems (deutlich als solche gekennzeichnet)
12. bei sog. Nulläquivalenz eine deutsche Paraphrase der Bedeutung (deutlich als solche gekennzeichnet) als ein einsetzbares Translat
13. ggf. Verweis(e) auf graphische Illustration(en)
14. ggf. Verweis auf eine verdichtete Artikelübersicht

Zusätzlich im Wort-Modul (d.h. wenn das ZS-Lemma ein Einwortlexem ist) auf der Ebene des jeweiligen Semems:

1. ggf. lexikalische Informationen über die bevorzugten Kontextpartner des AS-Lexems (Verweise auf Kollokationen)
2. ggf. Verweise auf Redewendungen mit dem AS-Lemma (Idiome, „stehende" Vergleiche, Routineformeln etc.)
3. ggf. Verweise auf Sprichwörter in denen AS-Lemma vorkommt
4. ggf. paradigmatische Verweise auf Mitglieder der Wortfamilie zur Veranschaulichung des Wortbildungspotentials des AS-Lexems

Es ist nicht möglich, alle diese Informationstypen in einem Zug mit den lexikalischen Daten zu füllen, das kann nur nach und nach geschehen. Man muß also auch hier das Prinzip der Modularität walten lassen. Dennoch ist für den Anfang ein Grundgerüst zu erstellen, in dem alle vorgesehenen Angabetypen tatsächlich vollständig existieren. Auf dieses Fundament kann man dann weiter bauen, d.h. die für weitere Phasen vorgesehenen Instanzen (konkrete Knoteninhalte) hinzufügen. In der Hypertextbasis müssen jedoch für diese Daten entsprechende Informationstypen, die auf dem Grundgerüst aufbauen, im voraus vorgesehen sein.

Die im folgenden Kapitel dargestellten Abbildungen veranschaulichen die Beziehungen zwischen den lexikographischen Datentypen am Beispiel des Lemma *piwo* im polnisch-deutschen Teil des Wörterbuchs sowie der mit ihm assoziierten lexikalischen Daten: Es handelt sich um Lemmata (Ein- und Mehrwortlexeme) sowie die auf sie adressierten Angaben. Die Pfeile (in der Abb. 3 und Abb. 4) drücken lediglich Querverbindungen zwischen den einzelnen Knoten (Kästchen) aus. Diese Verbindungen sollen als Hyperlinks umgesetzt werden, die Pfeile sagen jedoch nichts über die Linktypen aus. Ob die jeweiligen Links ein- oder bidirektional, bzw. als sog. Multilinks realisiert werden sollen, muß im Einzelfall entschieden werden. Die Entscheidung über die Linktypen wird zwar generell auf der Ebene des Schemas durch das im ‚SchemaEditor' erstellte abstrakte Modell getroffen. Auf

der Ebene der Wörterbuchinstanzen ist es aber möglich – und wünschenswert – die Default-Attribute der Links gegebenenfalls (im ‚StrukturEditor') zu ändern.

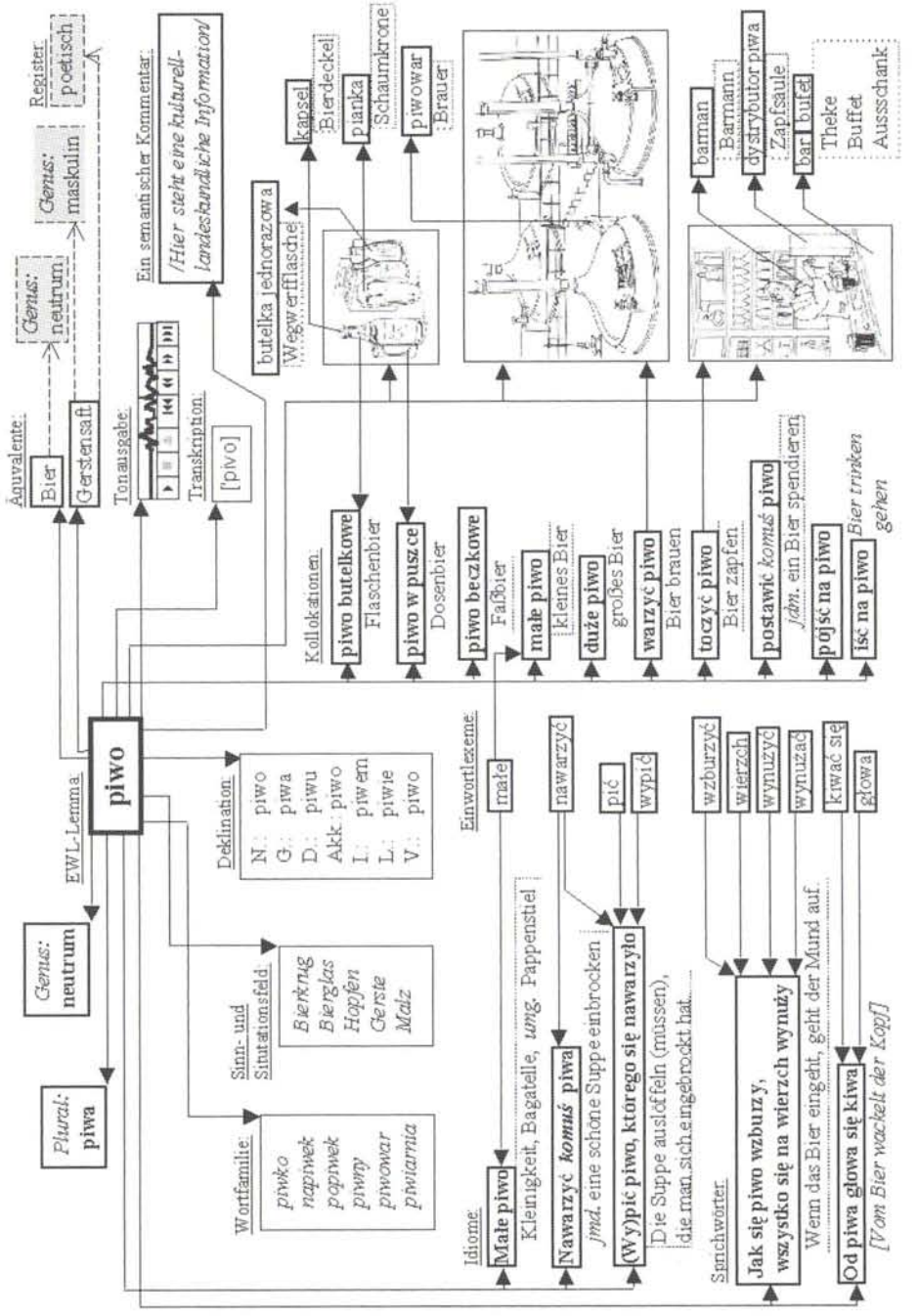
6.2 Die Datenmodellierung

Die Abbildung 3 zeigt ein abstraktes Schema mit den lexikographischen Daten zum Lemma *piwo* (δτ. ‚Bier'). Das Einwortlexem (EWL) *piwo* ist auf der einen Seite mit Angaben zur Grammatik (Genus- und Pluralform, Deklinationsmuster) und zur Aussprache (Ton und IPA-transkription) sowie mit morphologisch- und sinnverwandten Lexemen verknüpft, auf der anderen Seite ist das Einwortlexem ‚piwo' mit den Lemmata verbunden, deren Bestandteil es bildet. So wird z.B. auf das Mehrwortlexem (MWL) *małe piwo* (dt. Kleinigkeit, Bagatelle) aus dem Knoten ‚piwo' (Bier) und aus dem Knoten ‚małe' (klein) verwiesen. Sowohl das EWL ‚piwo' als auch alle Idiome, Sprichwörter und Kollokationen, die das Wort ‚piwo' beinhalten, sind mit den in Frage kommenden deutschen Äquivalenten verlinkt. (In der Abbildung 3 sind der Übersichtlichkeit halber nur Verweise auf ‚Bier' und ‚Gerstensaft' aus ‚piwo' dargestellt). Kollokationen bzw. semantische Zusammensetzungen *piwo butelkowe* (Flaschenbier), *warzyć piwo* (Bier brauen) und *toczyć piwo* (Bier zapfen) sind mit entsprechenden graphischen Illustrationen verbunden. Aus den Bildern sollen auch Verweise auf Lexeme im Sach- und Situationsfeld möglich sein: *kapsel* (Bierdeckel), *pianka* (Schaumkrone), *piwowar* (Brauer) usw.

Die Auskünfte über Genus, Plural und Aussprache zu ‚piwo' sind natürlich für die polnischen Benutzer redundant, die in der Abbildung 3 grau schattierten Knoten mit den Angaben zu den deutschen Äquivalenten – ‚Bier' und ‚Gerstensaft' – (hier nur mit einem Beispiel angedeutet) sind wiederum für die deutschen Benutzer irrelevant.

Wenn der deutsche Benutzer das poln. Lexem ‚piwo' nachschlägt, werden die Informationen über Genus, Plural aus den dazugehörigen Knoten inkludiert. Der Benutzer bekommt – indem er einen entsprechenden Button anklickt – die Möglichkeit, sich alle Idiome (vgl. Abbildung 5), Sprichwörter bzw. Kollokationen mit ‚piwo' sowie einen kulturell-landeskundlichen Kommentar über die Stellung des Biers in Polen (auf Deutsch) oder die Wortfamilie sowie Wörter im semantischen Umfeld (auch über die Verweise auf Bilder) von ‚piwo' einblenden lassen. Er kann sich auf die gleiche Art und Weise über die Aussprache und Deklination von ‚piwo' informieren. Wenn der deutsche Benutzer ein MWL als solches identifiziert, muß er nicht unter einem der Bestandteile nachschlagen, sondern kann direkt z.B. *nawarzyć piwa* (jmd. ‚eine schöne Suppe einbrocken') als Suchbegriff eingeben. Daraufhin werden die Knoten mit dem grammatischen Kommentar, der Paraphrase, dem Äquivalenten, dem Beispielsatz (oder -sätzen) samt dt. Übersetzung inkludiert und Verweise auf synonyme bzw. sinnverwandte Phraseme angezeigt. Einige Angaben beinhalten weitere Informationen: z.B. die Information über das pragmatische Register bezieht sich auf die Äquivalentangabe zu ‚nawarzyć (komuś) piwa' (dt. ‚jmd. eine schöne Suppe einbrocken') – siehe Abbildung 4. Für diese Benutzungssituation wird in den Äquivalent-Knoten eine auf deutsch verfaßte Entität „umgangssprachlich" inkludiert.

Manche Unterknoten – etwa mit der Kasus- und Valenzangabe oder mit dem Aktionsart-Formativ des poln. Verbs – sind nur in der AS-Sprache verfaßt, manche sind in der Datenbasis doppelt vorhanden: Die einen beinhalten lexikographische Daten auf Deutsch, die anderen auf Polnisch. Je nach der Benutzungssituation wird ein entsprechender Unterknoten eingeschlossen.


piwo
EWL-Lemma
Genus: neutrum
Plural: piwa
Äquivalente:
Bier
Gerstensaft
Genus: neutrum
Genus: maskulin
Register: poetisch
Tonausgabe:
Transkription:
[pivo]
Ein semantischer Kommentar:
/Hier steht eine kulturell-landeskundliche Information/
butelka jednorazowa
Wegwerfflasche
kapsel
Bierdeckel
pianka
Schaumkrone
piwowar
Brauer
Kollokationen:
piwo butelkowe
Flaschenbier
piwo w puszce
Dosenbier
piwo beczkowe
Faßbier
małe piwo
kleines Bier
duże piwo
großes Bier
warzyć piwo
Bier brauen
toczyć piwo
Bier zapfen
postawić komuś piwo
jdm. ein Bier spendieren
pójść na piwo
iść na piwo
Bier trinken gehen
barman
Barmann
dystrybutor piwa
Zapfsäule
bar/bufet
Theke
Buffet
Ausschank
Deklination:
N.: piwo
G.: piwa
D.: piwu
Akk.: piwo
I.: piwem
L.: piwie
V.: piwo
Sinn- und Situationsfeld:
Bierkrug
Bierglas
Hopfen
Gerste
Malz
Wortfamilie:
piwko
napiwek
popiwek
piwny
piwowar
piwiarnia
Einwortlexeme:
małe
nawarzyć
pić
wypić
wzburzyć
wierzch
wynurzyć
wynurzać
kwiać się
głowa
Idiome:
Małe piwo
Kleinigkeit, Bagatelle, umg. Pappenstiel
Nawarzyć komuś piwa
jmd. eine schöne Suppe einbrocken
(Wy)pić piwo, którego się nawarzyło
Die Suppe auslöffeln (müssen), die man sich eingebrockt hat
Sprichwörter:
Jak się piwo wzburzy, wszystko się na wierzch wynuży
Wenn das Bier eingeht, geht der Mund auf.
Od piwa głowa się kiwa
[Vom Bier wackelt der Kopf]

Wenn der polnische Benutzer ‚piwo' nachschlägt, so bleiben die Verweise auf die für Deutsche bestimmten Informationen ausmaskiert. Dem polnischen Benutzer werden Angaben zu den dt. Äquivalenten von ‚piwo' und Verweise auf die polnischen sinnverwandten Lexeme (*kufel, szkalnka, chmiel, słód, jeczmień, podchmielony, browar, dystrybutor*, usw.) angeboten. Die metasprachlichen Informationen (wie Paraphrasen, explizite Kommentare und Registerangaben), die sich auf das Lemma und das gegebene Äquivalent beziehen, erfolgen auf Polnisch. So bekommt etwa beim Nachschlagen des Lemma *małe piwo* (siehe Abb. 4) der Benutzer, neben dem Hinweis auf die wörtliche Bedeutung ‚kleines Bier', zwei idiomatische Bedeutungen zur Auswahl: ‚łatwe do załatwienia' (mühelos zu Erledigendes) und ‚coś niewiele znaczącego, przeciętnego' (etwas Unbedeutendes, Mittelmäßiges). Wählt er im weiteren Verlauf des Nachschlageprozesses die erste Bedeutung der Redewendung aus, so bekommt er den Inhalt des Knoten mit dem dt. Äquivalent, in den ein Unterknoten mit der Registerangabe auf Polnisch inkludiert wurde: ‚Kleinigkeit, Bagatelle; *potocznie* Kinkerlitzchen, Pappenstiel'.

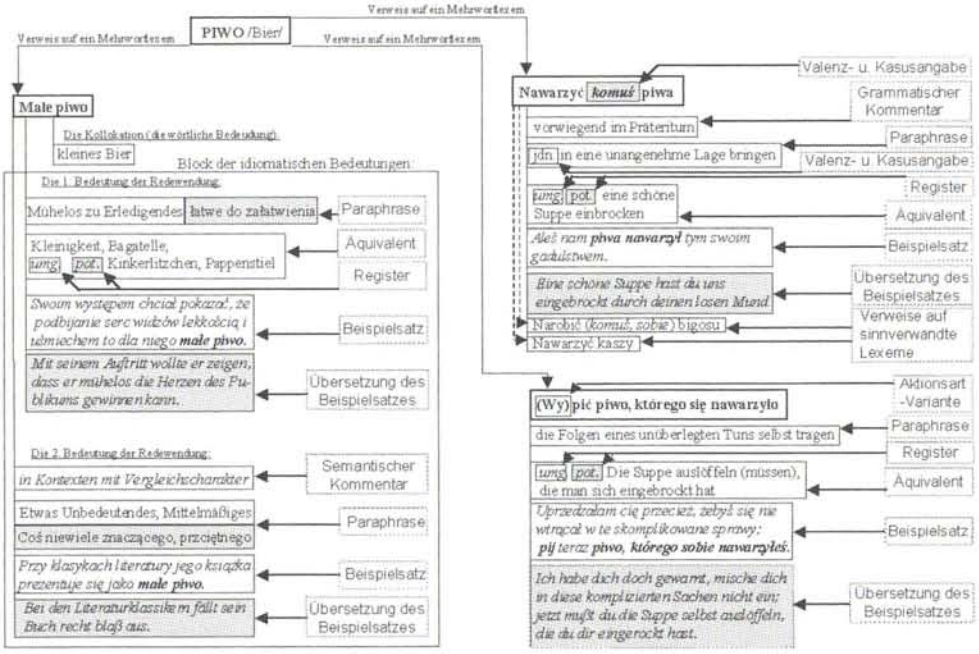


Abb. 4: Mehrwortlemmata

In der Abbildung 4 sind Angaben zu drei poln. MWL dargestellt: ‚małe piwo', ‚nawarzyć piwa' und ‚pić/wypić piwo, którego się nawarzyło'. Da das Lemma ‚małe piwo' polysem ist, sind die Paraphrasen – als bedeutungsdifferenzierende Glossen – für beide Benutzergruppen nützlich und deshalb in der Datenbasis doppelt vorhanden: auf Polnisch und auf Deutsch. (Im Falle der monosemen Phraseme ist es anders, die Bedeutung, zumindest in den meisten Fällen, kann für den polnischen Benutzer als bekannt vorausgesetzt werden.) Das Lemma *nawarzyć piwa* ist mit zwei sinnverwandten Lemmata verknüpft: *narobić*

bigosu i *nawarzyć kaszy*. In Abbildung 4 stellen die Angaben in grau schattierten Kästchen beispielhaft diejenigen Informationen dar, die im Zuge der Bidirektionalisierung des Materials den lexikalischen Daten aus den analysierten Werken hinzugefügt wurden.

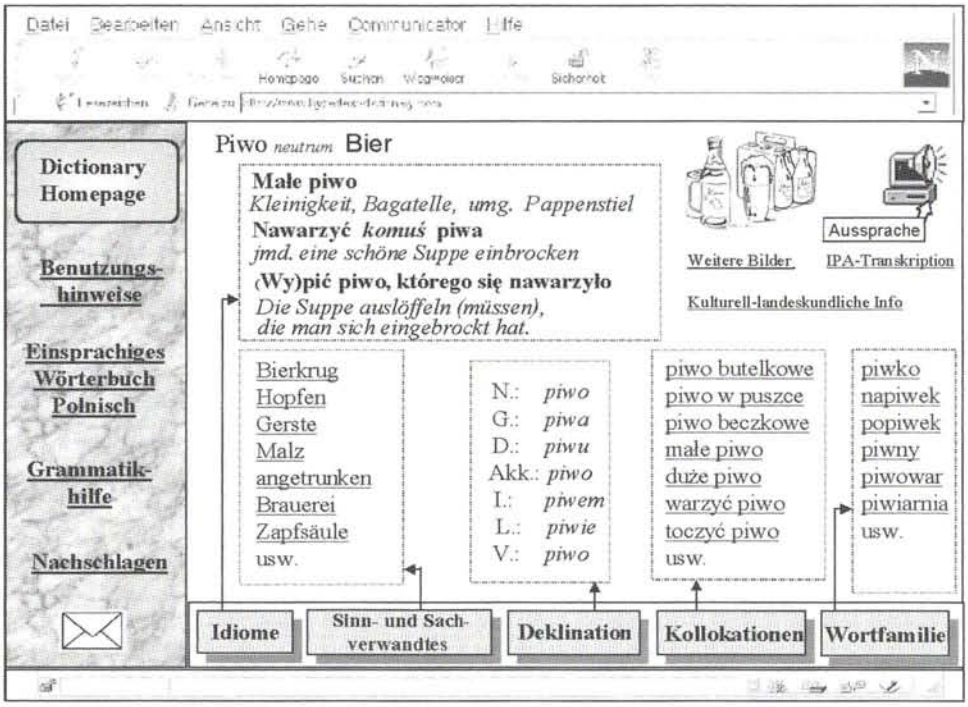


Abb. 5: Benutzeroberfläche

Auf der Benutzeroberfläche (siehe Abb. 5) sollen die deutschen Benutzer des Wörterbuchsystems durch das Anklicken der entsprechenden Buttons z.B. die Mehrwortlexem-Lemmata abfragen können, die das Einwortlexem ‚piwo' als Bestandteil enthalten. Würde der Benutzer hingegen den Button „Deklination" mit der Maus betätigen, so könnte er das Deklinationsparadigma für das Nomen ‚piwo' auf dem Bildschirm sehen. In der Abb. 5 stellen sowohl alle Mehrwortlexeme als auch die unterstrichenen Elemente anklickbare Linkanker dar. Die Pfeile und die gepunkteten Kästchen mit lexikographischen Daten haben auf dem Bild folgende Bedeutung: wenn der Benutzer einen Button drückt, bekommt er die gewünschten (in dem mit entsprechenden Pfeil angedeuteten Kästchen dargestellten) Informationen eingeblendet. Das HTML-Layout wird – wie im Kap. 4. 2 besprochen – über die mit den einzelnen Datenobjekten assoziierten Expander in Verbindung mit dem Maskierungsmechanismus generiert.

7 Übersichtsknoten

Viele Autoren plädieren – mit Recht – für die Auflösung der verdichteten Angaben, wie sie in Büchern zum Einsatz kommen, wenn die Wörterbuchdaten in elektronische Nachschlagewerke konvertiert werden oder Neuentwicklungen entstehen. Es gibt jedoch Situationen, in denen die Beibehaltung der Textverdichtung auch in einem elektronischen Wörterbuch sinnvoll zu sein scheint.

Der wesentliche Nachteil des Mediums Hypertext ist bekanntlich die Zersplitterung der Informationen in feinkörnige Einheiten, deren gegenseitige Beziehungen oft nicht klar genug dargestellt werden. Die Benutzer laufen dann Gefahr, aufgrund fehlerhaft konstruierter Navigationsmechanismen den Überblick zu verlieren. Deshalb wären Überblickstafeln mit verdichteten Informationen (die den Artikeln eines Printwörterbuchs gleichen und im Hypertext als ‚information chunks' organisiert sind) besonders nützlich: Einerseits leisten sie Orientierungshilfe für alle Benutzer, andererseits bieten sie den kundigen Benutzern komprimierte Informationen. Ein mit den Konventionen der Printlexikographie vertrauter Benutzer kann die Abkürzungen sofort entschlüsseln, den Inhalt des Übersichtsfensters wie im Buch überfliegen und ganz gezielt die Verweise zu den ihn in der konkreten Benutzungssituation interessierenden Informationen verfolgen. Es ist allerdings wichtig, daß die Abkürzungen und Symbole auf einen Mausklick des Benutzers hin angezeigt werden können, sei es, indem sie durchgängig in ganzem Knoten mit voll ausgeschriebenen Entitäten ersetzt werden, sei es, indem die Auflösung kontextsensitiv im Pop-up-Fenster erscheint. Denjenigen Benutzern, die im Umgang mit herkömmlichen Wörterbüchern ungeübt sind, helfen hoffentlich die leicht auflösbaren Textverdichtungselemente, die Bedeutung der in der Printlexikographie gängigen Abkürzungen für grammatische und pragmatische Kategorien zu lernen. So kann der Hypertext vielleicht auch dazu beitragen, daß die Konventionen der Printlexikographie, die für Nichtlinguisten nicht immer leicht zu verstehen sind, auf dem „Umweg" über ein elektronisches Medium besser kennengelernt werden.

Natürlich passiert es oft, daß die Benutzer nicht nur die Abkürzungen des Wörterbuchs nicht verstehen, sondern auch mit den aufgelösten Begriffen und explizit formulierten Kommentaren zur grammatischen, semantischen oder pragmatischen Verwendung des Lemmas wenig anfangen können. Deswegen scheint es wichtig, daß sie kontextsensitiv in der „Grammatikhilfe" nachschlagen können. Die Grammatikhilfe ist als ein separates Schema mit zwei Perspektiven angelegt, d.h. die deutsche Grammatik (auf Polnisch) und die polnische Grammatik (auf Deutsch) sind in Kapiteln mit gleicher Struktur gegliedert und lassen sich deshalb in einem Schema unterbringen.

8 Das objektorientierte Datenmodell

Die Implementierung der geeigneten Datenmodelle, die die oben umrissene Funktionalität unterstützen, ist mit dem Programm SchemaText dank komplexer Merkmale der Objektorientiertheit möglich.[27] Zur genauen Veranschaulichung all dieser Merkmale müßte man jeweils ein geeignetes Beispiel anführen, was hier aus Platzgründen nicht möglich ist. Trotzdem sollen diese Prinzipien nicht unerwähnt bleiben, da ihnen in Zukunft bei der Entwicklung von Computerwörterbüchern eine immer größere Rolle zukommen wird:

- Prinzip der Modularisierung (Vgl. Kap. 4. 1, 4. 2 und Kap. 10).
- Prinzip der Maskierung (Vgl. Kap. 4. 2).
- Prinzip der Typisierung von Knoten und Links (Vgl. Kap. 4. 1).
- Prinzip der Vererbung (Es regelt u.a. die Vererbung von Layouteigenschaften, die in den Expandern implementiert wurden. Dadurch, daß alle Informationstypen gleichen Strukturvorgaben genügen, können Eigenschaften von denen auf höherer Ebenen auf solche unterer Ebenen vererbt werden).
- Prinzip der Mehrfachverwendung (Manche Knoten können in mehreren Kontexten verwendet werden, etwa Beispielsätze und ihre Übersetzungen, Bilder oder Paraphasen, die sinngemäß mehreren Lexemen entsprechen).

9 Die Rolle der Bilder

Die zweisprachigen Printwörterbücher setzen Bilder entweder gar nicht oder nur sehr sparsam ein (vgl. Hupka 1989:39–43), Illustrationen würden zusätzlichen Platz erfordern und die Kosten in die Höhe treiben.[28] Außerdem stellt sich zunächst die Frage, ob die Bilder im zweisprachigen Wörterbuch überhaupt Sinn machen und, wenn ja, welche Arten von Illustrationen zu welchem Zweck an welcher Stelle am besten geeignet sind. Fest steht, daß die Bilder in einigen PC-Wörterbüchern bis jetzt nur der werbewirksamen „Multimedialität" wegen eingesetzt werden und durch den unüberlegten und konzeptionslosen Einsatz von Illustrationen kaum zur Verbesserung des Wörterbuchs beitragen. Das LEKSYKONIA-Wörterbuch präsentiert etwa beim Lemma *Egipt* ein Bild mit der Flagge Ägyptens (das Verfahren gilt für alle Länder) und beim Lemma *Glühbirne* ein Photo einer Glühbirne. Dieser konsequente Illustrationsansatz, substantivische Lemmata mit mehr oder weniger aussagekräftigen Photographien zu illustrieren, kann den Benutzern wohl wenige Vorteile bieten. Mehr Sinn machen schon einige unikale Illustrationen, z.B. für verschiedene Gemüsesorten (Lemmata *Kohl*, *Erbse*, *Gurke* usw.), wobei jedes dieser Lemmata einen Verweis auf den hyperonymischen Eintrag *Gemüse* enthält, der wiederum Verweise auf illustrierte Lemmata der einzelnen Gemüsesorten enthält.[29]

Für die Lernfunktion ist es wichtig, verschiedene Bildtypen zu differenzieren und zu überlegen, mit welchen lexikographischen Texten welche Typen von Illustrationen verbunden werden. Hupka (1989, 200–202 und 235–242) unterscheidet zwischen neun Typen von lexikographischen Illustrationen in gedruckten Nachschlagewerken. Alle diese Typen

---

[27] Das ganze Wörterbuch stellt als ein elektronisches Dokument – in der Sprache der Informatik gesagt – ein Objekt dar. Auf den niedrigeren Ebenen läßt sich dieses Objekt in weitere Objekte „zerlegen". Jedes dieser Objekte verfügt potentiell über gewisse Eigenschaften, die von einem Laserdrucker, einem WWW-Browser oder einer Belichtungsmaschine anders interpretiert werden. Jedes Objekt trägt gleichsam die „Rezepte" für seine Verarbeitung mit sich. Das Objekt Wörterbuch zeigt somit ein anders „Gesicht" gegenüber verschiedenen Ausgabemedien und verschiedenen Benutzern.

[28] Eine Ausnahme stellt nur das zweisprachige Bildwörterbuch dar.

[29] Eine nomenklatorische Illustration, die alle Gemüsesorten auf einmal präsentiert, und somit dem Benutzer die Möglichkeit gibt, die ihm unbekannte ZS-Bezeichnung für eine bestimmte Gemüsesorte (z.B. durch das Anklicken des entsprechenden Bildes) kennenzulernen, gibt es im LEKSYKONIA-Wörterbuch nicht.

können auch für ein elektronisches zweisprachiges Wörterbuch sehr vorteilhaft sein, wenn sie zweckmäßig angewandt werden:

1. Unikale Illustrationen helfen die referenzierten Denotate zu identifizieren. Sie sind in der Herübersetzungssituation bei Lemmata etwa aus dem technischen und naturwissenschaftlichen Bereich einzusetzen, da, wo der Benutzer auch in seiner Muttersprache über keine oder nur vage Vorstellungen von dem mit dem Lemma bezeichneten Begriff verfügt. Eine unikale Illustration kann zugleich sequentiell sein (Typ 3). Unikale Illustrationen dienen der Absicherung und der besseren Memorierung der Information, die durch die Äquivalentangabe vermittelt wird.
2. Aufzählende Illustrationen sind für Lexeme vorgesehen, die eine Klasse von Gegenständen unter einer gemeinsamen Bezeichnung vereinen. Das Aussehen der Referenten kann stark voneinander abweichen. (Hupka nennt als Beispiel das Lemma *Säge*. Zu illustrieren wären dann *Stichsäge, Bandsäge, Metallsäge, Kreissäge, Motorsäge* etc.).
3. Sequentielle Illustrationen veranschaulichen Bewegungsabläufe und sind im zweisprachigen Wörterbuch nur in Ausnahmefällen einzusetzen, etwa wenn Verben (*tröpfeln, schaukeln, rutschen, auskippen, zwinkern* etc.), Substantive (*Purzelbaum, Liegestütz*), Kollokationen (*ein Rad schlagen*) oder kulturspezifische Lemmata (bayer. *Schuhplattler*) illustriert werden sollen. Als technisches Mittel eigenen sich hierzu besonders Videosequenzen.
4. Strukturelle Illustrationen verdeutlichen die Teil-Ganzes-Beziehungen und helfen z.B. Teile eines Fahrrads zu benennen.
5. Funktionale Illustrationen zeigen die referierten Objekte im Kontext der angrenzenden Umgebung, die zum Verständnis des Wortes notwendig ist, da sie einen integralen, nicht abtrennbaren Teil einer größeren Einheit darstellen (z.B. das Lemma *Pupille*). Sie sind auf ähnliche Weise wie die unikalen und enzyklopädischen Illustrationen einzusetzen.
6. Nomenklatorische Illustrationen geben Instanzen einer Klasse wieder, z.B. *Gemüse*: *Gurke, Tomate* usw.
7. Szenische Illustrationen bilden einen Realitätsausschnitt ab: z.B. *Campingplatz, Bahnhofshalle, Krankenhauszimmer, Büro*. Sie eignen sich auch zur Veranschaulichung der Verwendung von Präpositionen.
8. Funktionsschemata könnten z.B. – in Form von Diagrammen – die Frequenz einzelner Äquivalente graphisch präsentieren, vorausgesetzt, es gäbe ein repräsentatives Korpus, an dem die Frequenzmessungen und andere statistischen Auswertungen vornehmbar wären.[30]
9. Enzyklopädische Illustrationen ergänzen Definitionen von landes- und kulturspezifischen Lexemen und bilden typisch deutsche Gegenstände wie etwa *Adventskranz, Gartenzwerg* bzw. unikale Eigennamen-Denotate wie *Friedensengel* ab.

Einen für die zweisprachigen Wörterbücher prädestinierten Bildertyp stellt die kontrastierende Illustration dar, die in der Typologie von Hupka (1989) nicht erfaßt wurde.[31] Kontrastierende Illustrationen wurden z.B. im Umlaut-Wörterbuch (monodirektional, d.h. nur für poln. Benutzer) eingesetzt, und verdeutlichen (immer in Form von einer Photographie) die besonderen Merkmale, die bestimmte Denotate in der deutschen Realität charakterisieren: z.B. *Telefonzelle, Briefkasten, Fahrkarte, Aufenthaltserlaubnis, Polizist, Streifenwagen, Imbißbude*, diverse Schilder wie *Naturschutzgebiet, Bushaltestelle, U-Bahn*

---

[30] Ein großes, ausbalanciertes, projekteigenes Korpus, in dem Lexikographen und Benutzer recherchieren könnten, wäre natürlich sehr wünschenswert. Dieses Desiderat, ähnlich wie das Morphologieanalysesystem, muß jedoch als separates Unterfangen betrachtet und ggf. realisiert werden.

[31] Unerwähnt bleibt bei Hupka auch die komplementäre Illustration, z.B. *Steckdose* : *Stecker* (vgl. LGDaF:942) oder *aufblasen:platzen* und die polysemieaufdeckende Illustration, z.B. *Flügel*: Flügel eines Vogels, Konzertflügel, Gebäudeflügel usw.

usw.[32] Diese Illustrationen vermitteln dem Benutzer zusätzliche landeskundliche Kenntnisse, die ihm einerseits die Orientierung während seiner Deutschlandreise erleichtern, und andererseits zusätzliche Konnotationen für die Memorierung der Begriffe liefern. (Allerdings muß noch der Einfluß einer bildhaften Kontrastierung von mutter- und fremdsprachlichen Referenten auf den Lernprozeß besser untersucht werden).

Alle Typen von Bildern haben primär die Aufgabe, den Lernprozeß zu fördern, indem sie entweder zusätzliche enzyklopädische (onomasiologische und semasiologische) Informationen vermitteln oder assoziative und paradigmatische Beziehungen der Lexik veranschaulichen.[33] Sie sind mit den lexikographischen Textsegmenten gleichrangig, folglich gilt es auch für sie zu entscheiden, in welcher Benutzungssituation sie eingesetzt werden sollen. Ferner gilt es zu bestimmen

- von welchem Knoten auf die Illustration verwiesen wird, d.h. ob der Verweis auf der Ebene des Äquivalentes bzw. der Paraphrase erfolgt oder auf der Ebene des semantischen bzw. pragmatischen Kommentars,
- ob, und wenn ja, wann eine unikale Illustration einen Teil einer aufzählenden, strukturellen, szenischen oder nomenklatorischen Illustration bilden soll. Das würde bedeuten, daß Illustrationen ggf. untereinander verlinkt werden müssen.

10 Ausblick

Das Prinzip der Modularisierung ist auf der einen Seite notwendig, um das System inhaltlich zu strukturieren (siehe Kap. 4.2), und auf der anderen Seite, um es von anderen Webprojekten abzugrenzen. Es ist nämlich nicht leicht, der Versuchung zu widerstehen, auch andere als Äquivalenz-Wörterbuchtypen in das System zu integrieren, und dadurch Umfang und Funktionalität immer weiter auszubauen. Besonders im Hinblick auf die Diskussion über erklärende zweisprachige (Übersetzungs-)Wörterbücher (vgl. Duda 1986, Worbs 1997) könnte eine solche zu weite Öffnung bedeuten, daß einsprachige Bedeutungswörterbücher und enzyklopädische Lexika mit eingeschlossen wären, auf die ein Zugriff im Prinzip sehr wünschenswert wäre. Solch ein Ansatz, der für die Benutzer mit Sicherheit einen großen Vorteil bedeuten würde, muß jedoch aus praktischen Gründen modularisiert werden, d.h. es sind getrennte, von zweisprachigen Wörterbüchern unabhängige einsprachige Webprojekte vonnöten, auf deren Lemmata punktuell verwiesen werden kann. Das erlaubt den Autoren nach dem Motto multum non multa zu verfahren, birgt jedoch die Gefahr in sich, daß die referentielle Integrität u.U. nicht mehr gewährleistet ist. Hier ist man mit dem grundsätzlichen Problem des WWW konfrontiert:

> „Während in einem geschlossenen Hypertextsystem, z.B. einem auf CD-ROM vertriebenen Lernsystem, alle Links, denen der Leser nachgehen kann, vom Autor kontrolliert und mit einer klaren

---

[32] Das Beispiel ‚Telefonzelle' zeigt, daß auch Illustrationen schnell obsolet werden können. Die im Umlaut-Wörterbuch abgebildete gelbe Telefonzelle der Post wurde mittlerweile durch eine rosagraue Telekom-Zelle beinahe verdrängt. Wenn der Telefonmarkt in Zukunft weiter privatisiert wird, kann es sein, daß es bald kein dominierendes länderspezifisches Design für eine Telefonzelle mehr gibt. Dann soll auch die Illustration zu diesem Lemma wegfallen.

[33] Auskünfte über die Zusammenhänge zwischen dem Text, Bild und Ton in neuen Medien gibt aus der Sicht der kognitiven Psychologie Weidenmann (1997:108ff.).

> Intention angelegt und somit integraler Bestandteil des Textes sind, gilt das in einem WWW-Dokument nur für einen Teil der Links, nämlich die internen. Ansonsten ist jedoch das eigene Dokument ein Knoten in einem unkontrollierten, weil unkontrollierbaren, dynamischen, globalen Netzwerk. Bereits die von einem Autor selbst angelegten, von seinem eigenen Dokument nach außen führenden Links unterliegen nur bedingt seiner Kontrolle, da er auf die Kontinuität des referenzierten Materials keinen Einfluß hat. Noch schwieriger ist es mit den externen Verweisen auf den eigenen Text: Hier wird deutlich, dass die pragmatischen Rahmenbedingungen der Textrezeption in einem offenen Hypertextsystem zu einer qualitativen Nebenbestimmung des Textbegriffs führen. [...] Durch dieses (vom Autor selbst nicht kontrollierbare) Wegfallen oder Veralten der Referenzen und die potentiellen Veränderungen des referenzierten Materials verlieren Texte als WWW-Dokumente ihren statischen Charakter, sie verändern sich im Laufe der Zeit" (Frisch 1998:228–229).

Deswegen wäre es sinnvoll, eine Koordination mit anderen Webprojekten vorzunehmen. Ideale Bedingungen dafür sind geschaffen, wenn man dieselbe Software als Dokumentenverwaltungssystem anwendet.

Sogar ein sehr umfangreiches zweisprachiges Wörterbuch kann nicht alle möglichen Kontexte der Lemmaverwendung berücksichtigen – künftige Wörterbücher sollten also einen Anschluß an Korpusabfragesysteme bieten. Solche Systeme, heute noch selten und nur experimentell betrieben, können bald als ein wichtiges lexikographisches Arbeitsmittel dienen, nicht nur für die Wörterbuchautoren, sondern auch für die gewöhnlichen Benutzer.[34] Weitere Links, etwa auf Seiten mit Online-Sprachkursen, Vokabeltrainern, Informationen zur Landeskunde u.ä. sind denkbar und wünschenswert.

Ebenso nützlich wie ein ausgereifter Bildereinsatz kann die Verwendung der Töne sein. Heute noch, wegen beschränkter Bandbreite der Leitungen, eher die Domäne der Offline-Wörterbücher, bald jedoch wird die Tonübertragung im Internet wahrscheinlich kein technisches Problem mehr darstellen. Dann können ganze Redewendungen vertont werden.

Neue Lemmatypen wie Neologismen, Fachtermini, aber vor allem zahlreiche Varianten eines Lexems, sobald sich in diesen Varianten veränderte Semantik des Lemma manifestiert (sehr wichtig für den polnischen Aspekt und die Diminutiva), lassen sich endlich in das zweisprachige Wörterbuch aufnehmen. Das ist für die stark flektierenden Sprachen wie Polnisch und Deutsch von besonderer Bedeutung. Man findet z.B. bislang in keinem poln.-dt. Wörterbuch alle Verbindungen des Typs Prefix+Verb des poln. Verbs gotować (zagotować, przegotować, wygotować, nagotować, rozgotować, odgotować, dogotować). Ein Lemmatisierungsalgorithmus, der alle polnischen Formen ‚iść,, ‚pójść,, ‚chodzę,, ‚szedł' usw. auf die Grundformen ‚chodzić,/‚iść' (wo als Äquivalent ‚gehen' in Frage kommt) zurückführen würde, wäre der komplexen Problematik des Aspektes im Polnischen und der Aktionsarten im Deutschen nicht gerecht.

Die vom Autor durchgeführten Experimente, mit dem Ziel, ausgewählte Wörterbuchfragmente in die mit SchemaText erstellte Hypertextstruktur zu überführen, lassen sich noch nicht als völlig erfolgreich bezeichnen – die vollständige Implementierung und Evaluierung eines Prototyps konnte deswegen noch nicht in Angriff genommen werden. Ein inhärentes semantisch-pragmatisches Organisationsprinzip, das eine dynamische Generierung von Online-Wörterbucheinträgen gemäß modellierter Benutzerprofile unterstützt, stößt noch auf technische Hürden auf der Produktionsseite und auf urheberrechtliche Probleme bei der Redaktion der Inhalte. Nach wie vor gilt die Aussage von Drewek (1990:272): „Das Urheberrecht regelt nicht, wem die formale Spezifikation eines Wörterbuchdesigns gehört, obwohl sich hierin die individuelle Handschrift von LexikographInnen

[34] Vgl. Figge (1994), Martin (1995), Krishnamurthy (1996).

schöpferisch manifestiert. Hier öffnet sich eine rechtliche Grauzone, die vom Knowledge Engineering her bekannt ist".

Auf der technischen Seite ergeben sich außerdem schwerwiegende Probleme mit der Unterstützung der länderspezifischen Buchstaben und der Sonderzeichen. Die proprietären Datenformate können zwar mit diesen Zeichen in der Regel gut umgehen, aber wenn es darum geht, die Daten zwischen den Softwarekomponenten oder Anwendungen auszutauschen, kommt man um trickreiche Methoden nicht herum. Der lautstarken Werbung mancher Hersteller, in ihren Produkten den Unicode-Standard zu unterstützen, darf man nicht immer Glauben schenken.

Der im vorliegendem Beitrag skizzierte Ansatz würde ernste Konsequenzen sowohl für die Lexikographen als auch für die Wörterbuchbenutzer mit sich bringen, und zwar einen Paradigmenwechsel bei der Erstellung von lexikographischen Werken und eine Reorganisation der Wörterbucharbeit durch das Aufheben der strikten Grenze zwischen Makro- und Mikrostrukturen, durch Mehrfachverwendung der Inhalte sowie durch ein dichtes Verknüpfen von lexikographischen Daten mithilfe von Hyperlinks. In das Wörterbuch würden sich nicht nur neue Lemmata und Angaben jederzeit aufnehmen lassen, sondern auch veraltete – wenn nötig – entfernen. Dadurch könnte das Wörterbuch stets aktuell bleiben, es wäre aber nie endgültig fertiggestellt.

Ein ähnlicher Paradigmenwechsel wäre für das Benutzerverhalten notwendig, um das Wörterbuch als ein interaktives Medium erfolgreich zu gebrauchen und es eventuell mit- bzw. weitergestalten. Eine entscheidende Rolle bei weiterer Entwicklung von Online-Nachschlagewerken wird wahrscheinlich die Flexibilität der Systeme spielen: erst wenn Hypertextwörterbücher den individuellen Bedürfnissen verschiedener Benutzergruppen entsprechen werden, kann ihre Akzeptanz und damit auch praktische Relevanz steigen. Hypertexte bringen für die Benutzer bekanntlich nicht nur Vorteile mit sich, sondern bergen etliche Gefahren – Unübersichtlichkeit, Ablenkung und Angebotsfülle können bei der Rezeption der Informationen sehr hinderlich sein. Helfen können gut strukturierte Hypertextbasen und leistungsfähige Navigationsmechanismen. Zudem müssen wir noch mehr über die Beziehungen zwischen dem mentalen Lexikon und der formalen Repräsentation bzw. Rekonstruierung des lexikalischen Wissens für die Konstruktion von Hypertextwörterbüchern in Erfahrung bringen.

11 Literatur

11.1 Wörterbücher und Lexika

DBWDP = DUDEN BILDWÖRTERBUCH DEUTSCH UND POLNISCH. Bibliographisches Institut & FA Brockhaus AG & Wiedza Powschechna. Aktualisierte Neuauflage. Warszawa 1998.

GWPD/DP = GROSSWÖRTERBUCH DEUTSCH-POLNISCH/POLNISCH-DEUTSCH von Juliusz Ippoldt und Jan Piprek. Wiedza Powszechna & Verlag Enzyklopädie Leipzig. Warszawa 1969/1971. 4 Bd.

GlobeDisc 2.0 GRUNDWORTSCHATZ POLNISCH. (Deutsch-Polnisch/Polnisch-Deutsch). Rossipaul Verlag München & EPP Nürnberg 1996. 3 Disketten.

HWBPD/DP = HANDWÖRTERBUCH DEUTSCH-POLNISCH/POLNISCH-DEUTSCH von Andrzej Bzdęga, Jan Chodera und Stefan Kubica. Wiedza Powszechna. Warszawa 1971. 2 Bd.

Infopedia 2.0: COMPTON'S ENZYKLOPÄDIE, OXFORD ENGLISH DICTIONARY ENG.-GERM./GERM.-ENG., WELTLÄNDER-LEXIKON, WELTATLAS, RECHTSCHREIBWÖRTERBUCH, FALKEN'S BUCH DER ZITATE,

FREMDWÖRTERBUCH, SYNONYMENWÖRTERBUCH, ZWEIFELSFÄLLE DEUTSCH, Leonardo der Erfinder, Jahrgang 1995 der Zeitschrift „Bild der Wissenschaft". Tewi Verlag. München 1997. 3 CD-ROMs.

INFOROM '98/'99: BERTELSMANN UNIVERSALLEXIKON, WÖRTERBUCH ZUR ALTEN UND ZUR NEUEN RECHTSCHREIBUNG, WÖRTERBUCH ENGLISCH-DEUTSCH/DEUTSCH-ENGLISCH, WÖRTERBUCH FRANZÖSISCH-DEUTSCH/DEUTSCH-FRANZÖSISCH, WELTATLAS, AUTOATLAS DEUTSCHLAND, WAHRIG-FREMDWÖRTERLEXIKON, Rechtschreibprüfung für MS Winword. Bertelsmann. München 1998. 2 CD-ROMs.

ISNP = ILUSTROWANY SŁOWNIK NIEMIECKO-POLSKI von Aldona Brzeska und Alojzy Brzeski. Wiedza Powszechna. Warszawa 1975.

LANGENSCHEIDTS POWER DICTIONARY ENGLISCH (ENGLISCH-DEUTSCH/DEUTSCH-ENGLISCH) hrsg. von Vicent Docherty. Langenscheidt. München 1997.

LANGENSCHEIDTS HANDWÖRTERBUCH ENGLISCH (ENGLISCH-DEUTSCH/DEUTSCH-ENGLISCH) - DIE PROFILINE. In der PC-Biblithek 2.0. Langenscheidt. München 1997. 1 CD-ROM.

LE PETIT ROBERT. LA REFERENCE DE LA LANGUE FRANÇAISE. Liris Interactive. Paris 1997. 1 CD-ROM.

LGDaF = LANGENSCHEIDTS GROßWÖRTERBUCH DEUTSCH ALS FREMDSPRACHE hrsg. von Dieter Götz, Günther Haensch und Hans Wellmann. Langenscheidt. München 1993.

LEKSYKONIA 4.0 MULTIMEDIALNY SŁOWNIK NIEMIECKO-POLSKI I POLSKO-NIEMIECKI. WNT Warszawa & Lexland Knurów 1997. 1 CD-ROM.

LexiRom 3.0: MEYERS LEXIKON IN DREI BÄNDEN, DUDEN DEUTSCHE RECHTSCHREIBUNG, DUDEN FREMDWÖRTERBUCH, DUDEN WÖRTERBUCH DER SINN- UND SACHVERWANDTEN WÖRTER, LANGENSCHEITS TASCHENWÖRTERBUCH ENGLISCH, WELTATLAS. Bibliographisches Institut & FA Brockhaus AG. Mannheim 1998. 1 CD-ROM.

OED = OXFORD ENGLISH DICTIONARY ON CD-ROM prepared by John Simpson and Edmund Weiner. Oxford University Press & AND Software Inc. Second Edition. Oxford 1996. 1 CD-ROM.

OXFORD-HACHETTE ENGLISH AND FRENCH DICTIONARY hrsg. von Marie-Helene Corréard und V. Grundy. Oxford University Press. Oxford 1994.

PC-Bibliothek 2.0. Langenscheidts HANDWÖRTERBUCH ENGLISCH. München 1997. 1 CD-ROM.

PWP = PHRASEOLOGISCHES WÖRTERBUCH POLNISCH von Erika Ehegötz, Walter Duda, Maria Frenzel, Maria Gehrmann und Stanisław Skorupka. Verlag Enzyklopädie. Leipzig 1990.

SPRACHBROCKHAUS. DEUTSCHES BILDWÖRTERBUCH VON A-Z. Neubearbeitung. Wiesbaden 1984.

STJN = SŁOWNIK TEMATYCZNY JĘZYKA NIEMIECKIEGO von Grażyna Hatałal und Małgorzata Bielicka. Kanion. Zielona Góra 1996.

SWBPD = SPRICHWÖRTERBUCH DEUTSCH-POLNISCH/POLNISCH-DEUTSCH von Alina Wójcik und Horst Ziebart. Wiedza Powszechna. Warszawa 1997.

TWBP = LANGENSCHEIDTS TASCHENWÖRTERBUCH POLNISCH von Stanisław Walewski. Langenscheidt. München 1979.

UMLAUT. MULTIMEDIALNY SŁOWNIK EDUKACYJNY NIEMIECKO-POLSKI I POLSKO-NIEMIECKI. Wiedza Powszechna & Premiere Training Company Ltd. Warszawa 1996. 1 CD-ROM.

11.2 Sonstige Literatur

Arents, Hans C.; Bogaerts, Walter (1991): Towards an Architecture For Third-order Hypermedia Systems. In: Hypermedia 3 (2), 133–152.

– (1993): Concept-based Rertrieval of Hypermedia Information: From Term Indexing to Semantic Hyperindexing. In: Information Processing & Management 29 (3), 373–386.

Atkins, Sue (1996): Bilingual Dictionaries: Past, Present and Future. In: Gellerstam, Martin; Järborg, Jerker; Malmgren, Sven-Göran; Norén, Kerstin; Rogström, Lena; Papmehl, Catarina (Hg.): Euralex'96 Proceedings. 7th International Congress on Lexicography. – Göteborg. Sweden, 515–546.

Bläsi, Christoph; Koch Heinz-Detlev; Lehr, Andrea; Wiegand, Herbert Ernst (1994): Lexicographic Standards and Reusability – From Metalexicographic Description to a Parsing Procedure. In: Lexicographica 10/94, 221–248.

Breidt, Elisabeth; Segond, Frédérique; Valetto, Giuseppe (1996): Local Grammars for the Description of Multi-Word Lexemes and their Automatic Recognition in Texts. In: Kiefer, Ferenc; Kiss, Gábor; Pajzs Júlia (Hg.): Proceedings of the 4th International Conference on Computational Lexicography COMPLEX'96. – Budapest. Hungary, 9–28.

Conklin, Jeff (1987): Hypertext: An Introduction and Survey. In: IEEE Computer 20 (9), 17–41.

Drewek, Raimund (1990): Skizzen aus der real produzierenden Lexikographie. In: Schaeder, Burkhard; Rieger, Burghard (Hg.): Lexikon und Lexikographie. – Hildesheim: Olms (= Sprache und Computer 11, Vorträge aus der Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung e.V., Siegen, 26.–28.3.1990), 265–274.

Duda, Walter (1986): Ein „aktives" russisch-deutsches Wörterbuch für deutschsprachige Benutzer? In: Beiträge zur Lexikographie slawischer Sprachen. – Berlin (Ost): Akademie der Wissenschaften der DDR (= Linguistische Studien, Reihe A, 147), 9–15.

Feldweg, Helmut (1997): Wörterbücher und neue Medien: Alter Wein in neuen Schläuchen? In: Zeitschrift für Literaturwissenschaft und Linguistik 27/107, 110–123.

Figge, Udo L. (1994): Lexikographische Datenverarbeitung und maschinelle Lexikographie. In: Figge, Udo L. (Hg.): Portugiesische und portugiesisch-deutsche Lexikographie. – Tübingen: Niemeyer (= Lexicographica, Series Maior 56), 209–221.

Freisler, Stefan (1994): Hypertext. Eine Begriffsbestimmung. In: Deutsche Sprache 22/94, 19–50.

Frisch, Elisabeth (1998): Ausgewählte Aspekte des Publizierens im WWW. In: Storrer, Angelika; Harriehausen, Bettina (Hg.): Hypermedia für Lexikon und Grammatik. – Tübingen: Narr (= Studien zur deutschen Sprache 12), 217–232.

Hauser, Ralf; Storrer, Angelika (1996): Probleme und Lösungen beim Parsen von Wörterbüchern. In: Feldweg, Helmut; Hinrichs, Erhard W. (Hg.): Lexikon und Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen. – Tübingen: Niemeyer (= Lexicographica, Series Maior 73), 53–68.

Hausmann, Franz-Josef (1991): Typologie der zweisprachigen Spezialwörterbücher. In: Hausmann, Franz-Josef; Reichmann, Oskar; Wiegand, Herbert Ernst; Zgusta, Ladislav (Hg.): Wörterbücher. Ein internationales Handbuch der Lexikographie. 3. Teilband – HSK 5.3 (= HSK 5.1, 5.2, 5.3 Handbücher zur Sprach- und Kommunikationswissenschaft 1989, 1990, 1991). – Berlin: De Gruyter, 2877–2881.

– und Wiegand, Herbert Ernst (1989): Component Parts and Structures of General Monolingual Dictionaries: A Survey. In: Hausmann, Franz-Josef; Reichmann, Oskar; Wiegand, Herbert Ernst; Zgusta, Ladislav (Hg.): Wörterbücher. Ein internationales Handbuch der Lexikographie. 1. Teilband – HSK 5.1 (= HSK 5.1, 5.2, 5.3 Handbücher zur Sprach- und Kommunikationswissenschaft 1989, 1990, 1991). – Berlin: De Gruyter, 328–359.

– und Werner, Reinhold Otto (1991) Spezifische Bauteile und Strukturen zweisprachiger Wörterbücher. In: Hausmann, Franz-Josef; Reichmann, Oskar; Wiegand, Herbert Ernst; Zgusta, Ladislav (Hg.): Wörterbücher. Ein internationales Handbuch der Lexikographie. 3. Teilband – HSK 5.3 (= HSK 5.1, 5.2, 5.3 Handbücher zur Sprach- und Kommunikationswissenschaft 1989, 1990, 1991). – Berlin: De Gruyter, 2729–2769.

Heidecke, Stefanie (1996): SGML-Tools in the Dictionary-making Process – Experiences with a German-Polish/Polish-German Dictionary. In: Gellerstam, Martin; Järborg, Jerker; Malmgren, Sven-Göran; Norén, Kerstin; Rogström, Lena; Papmehl, Catarina (Hg.): Euralex'96 Proceedings. 7th International Congress on Lexicography. – Göteborg. Sweden, 395–404.

Horn, Robert E. (1989): Mapping Hypertext. The Analysis, Organisation and Display of Knowledge for the Text Generation of On-line-text and Graphics. – Waltham, MA.: Information Mapping Inc.

Hupka, Werner (1989): Wort und Bild. Illustrationen in Wörterbüchern und Enzyklopädien. – Tübingen: Niemeyer (= Lexicographica, Series Maior 22).

Kammerer, Matthias (1998): Hypertextualisierung gedruckter Wörterbuchtexte: Verweisstrukturen und Hyperlinks. In: Storrer, Angelika; Harriehausen, Bettina (Hg.): Hypermedia für Lexikon und Grammatik. – Tübingen: Narr (= Studien zur deutschen Sprache 12), 145–171.

Kesseler, Markus; Freisler, Stefan (1995): Document Engineering. In: Materialien aus der Herbsttagung der Tekom e.V. im November 1995 in Dortmund. (30.3.1999) http://www.schema.de/html-deu/schemapu/vortrag/document.htm.

Krishnamurthy, Ramesh (1996): The Data is The Dictionary: Corpus at the Cutting Edge of Lexicography. In: Kiefer, Ferenc; Kiss, Gábor; Pajzs Júlia (Hg.): Proceedings of the 4th International Conference on Computational Lexicography COMPLEX'96. – Budapest. Hungary, 117–144.

Kuhlen, Reiner (1991): Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank. – Springer, Edition SEL-Stiftung.

Lowe, John B.; Baker, Collin F.; Fillmore, Charles J. (1997): A Frame-Semantic Approach to Semantic Annotation. In: Proceedings of the SIGLEX Workshop „Tagging Text with Lexical Semantics: Why, What and How?" (in conjunction with ANLP-97), 4–5 April 1997. Washington, D.C. HTML-Datei: (24.09.1999). URL:
http://www.icsi.berkeley.edu/~framenet/docs/siglex.html, PostScript-Datei:
http://www.icsi.berkeley.edu/~framenet/docs/siglex.ps .

Martin, Willy (1995): Maschinelle Lexikographie. Ein Blick in die Zukunft. In: Hitzenberger, Ludwig (Hg.): Angewandte Computerlinguistik. – Hildesheim: Olms (= Sprache und Computer 15, Vorträge aus der Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung e. V., Regensburg, 30.–31.3.1995), 1–21.

Neth, Hansjörg; Swanson, Heidi (1999): Zwei Sprachen, eine Scheibe. Englischwörterbücher auf CD-ROM. In: c't. Magazin für Computertechnik. 2/99, 100–106.

Patric, Jon; Zhang, Jun; Zubillaga, Artola Xabier (1996): An Architecture for a Federation of Heterogeneous Lexical and Dictionary Databases. In: Proceedings of the 1996 Joint Conference of The Association for Literary and Linguistic Computing and The Association for Computers and the Humanities (ALLC/ACH). – Bergen. Norway. URL: http:\\fims-www.massey.ac.nz\~is (25.3.1999).

Petelenz, Krzysztof (1998): Lexicon ex machina. Wie können wir elektronische Maschinen für die multimediale Wörterbucharbeit nutzen? In: Orbis Linguarum Vol. 8, 147–158.

– (2000): Zur Hypertextualisierung von zweisprachigen Wörterbüchern. Einige Vorschläge am Beispiel des Sprachenpaares Deutsch-Polnisch. In: Wiegand, Herbert Ernst (Hg.): Wörterbücher in der Diskussion IV. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. – Tübingen: Niemeyer (= Lexicographica, Series Maior 100), 203–231.

Piotrowski, Tadeusz (1999): Tagging and Conversion of a Bilingual Dictionary for XeLDA, a Xerox Computer Assisted Translation System. In: Kiefer, Ferenc; Kiss, Gábor; Pajzs Júlia (Hg.): Proceedings of the 5th International Conference on Computational Lexicography COMPLEX'99. – Budapest. Hungary, 113–120.

Storrer, Angelika (1998): Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In: Wiegand, Herbert Ernst (Hg.): Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. – Tübingen: Niemeyer (= Lexicographica Series Maior 84), 107–134.

Tarp, Sven (1995): Wörterbuchfunktionen: Utopische und realistische Vorschläge für die bilinguale Lexikographie. In: Wiegand, Herbert Ernst (Hg.): Studien zur zweisprachigen Lexikographie mit Deutsch II. – Hildesheim: Olms (= Germanistische Linguistik 127–128), 17–62.

Thielen, Christine; Breidt, Elisabeth; Feldweg, Helmut (1998): COMPASS – Ein intelligentes Wörterbuchsystem für das Lesen fremdsprachiger Texte. In: Storrer, Angelika; Harriehausen, Bettina (Hg.): Hypermedia für Lexikon und Grammatik. – Tübingen: Narr (= Studien zur deutschen Sprache 12), 173–194.

Wegner, Immo (1985): Frame-Theoreie in der Lexikographie. Untersuchungen zur theoretischen Fundierung und computergestützter Anwendung kontextueller Rahmenstrukturen für die Lexikographische Repräsentation von Substantiven. – Tübingen: Niemeyer (= Lexicographica, Series Maior 10).

Weidenmann, Bernd (1997): Multicodierung und Multimodalität im Lernprozeß. In: Issing, Ludwig; Klisma, Paul (Hg.): Information und Lernen mit Multimedia. – Weinheim: Beltz, 65–81.

Wiegand, Herbert Ernst (1986): Shanghai bei Nacht. Auszüge aus einem metalexikographischen Tagebuch zur Arbeit beim Großen Deutsch-Chinesischen Wörterbuch. In: Wiegand, Herbert Ernst (Hg.) Studien zur neuhochdeutschen Lexikographie VI. – Hildesheim: Olms (= Germanistische Linguistik 87–90), 522–626.

– (1996a): Über die Mediostrukturen bei gedruckten Wörterbüchern. In: Zettersten, Arne; Pedersen, Vigo Hjørnager (Hg.): Proceedings of the Symposium on Lexicography VII, 5.-6. May 1994 at the University of Copenhagen. – Tübingen: Niemeyer (= Lexicographica, Series Maior 76), 11–43.
– (1996b): Das Konzept der semiintegrierten Mikrostrukturen. Ein Beitrag zur Theorie zweisprachiger Printwörterbücher. In: Wiegand, Herbert Ernst (Hg.): Wörterbücher in der Diskussion II. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. – Tübingen: Niemeyer (= Lexicographica, Series Maior 70), 1–82.
Williams, Melinda (1997): Document Engineering: A Unique Design and Modeling Discipline. (30.3.1999), http://www.goinside.com/97/2/doceng.html .
Worbs, Erika (1997): Plädoyer für das zweisprachige Wörterbuch als Hilfsmittel des Translators. In: Drescher, Horst W. (Hg.): Transfer. Übersetzen-Dolmetschen-Interkulturalität. – Frankfurt a. M.: Peter Lang, 497–510.

Krzysztof Petelenz, Stuttgart

Claudia Kunze, Andreas Wagner

Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche

1 GermaNet, eine lexikalische Ressource für die Informationserschließung

Elektronische Wörterbücher, welche über die Modellierung kleinerer Sprachdomänen hinausgehend den strukturierten Zugriff auf lexikalische Einheiten des Grundwortschatzes gestatten, sind ein zentrales Desiderat der Informationsgesellschaft. Lexikalische Wissensbasen stehen im Mittelpunkt des Interesses, da die Verfügbarkeit elektronischer Bedeutungswörterbücher für zahlreiche Anwendungen innerhalb der Maschinellen Sprachverarbeitung die unabdingbare Voraussetzung darstellt.

Semantische Netze im Stil des Princeton WordNet (vgl. Miller et al. 1990, Fellbaum 1998), die eine Vielzahl lexikalischer Einheiten in ihren grundlegenden semantischen Relationen abbilden, stellen geeignete Grundlagen-Ressourcen für effiziente computerlinguistische Verfahren zur Bedeutungsdisambiguierung bereit.

Das in solchen Ressourcen repräsentierte Wissen fungiert in computerlinguistischen Anwendungen als Referenzwissen und wird über statistische Verfahren mit den im zu verarbeitenden Text vorkommenden Wörtern abgeglichen (vgl. Yarowsky 1992; Harley 1994). WordNet kann so wie ein klassischer Thesaurus oder ein Wörterbuch sinn- und sachverwandter Wörter eingesetzt werden.

Folgende Anwendungen bedürfen der lexikalisch-semantischen Disambiguierung (vgl. Kapitel 4):

- die Maschinelle Übersetzung;
- die Informationserschließung;
- die semantische Annotierung von Korpora;

- die Entwicklung von Sprachlernwerkzeugen, Übersetzungswerkzeugen und Werkzeugen zum Informationserwerb;
- die Entwicklung automatischer Summarizer,
- die Realisierung von Sprachgenerierungswerkzeugen.

Mit der Entwicklung des deutschen Wortnetzes GermaNet ist die Lücke, die es in bezug auf deutschsprachige semantische Lexika zu verzeichnen gab, gefüllt worden.

In Kapitel 2 beschreiben wir die Aufbauprinzipien und die Merkmale des GermaNet. Kapitel 3 zeigt, wie GermaNet in einen multilingualen Kontext gestellt worden ist. Die Rolle, die GermaNet in computerlinguistischen Anwendungen spielen kann, insbesondere bei der Akquisition von Selektionspräferenzen, wird im 4. Kapitel erörtert. Das Schlußwort faßt zusammen, welche Perspektiven und Modifikationen sich für GermaNet ergeben.

2 Aufbau des GermaNet

Der festgestellte Mangel an computertechnisch verfügbaren lexikalisch-semantischen Ressourcen für das Deutsche hat die Entwicklung eines deutschen semantischen Wortnetzes motiviert, das im wesentlichen an den Strukturierungsprinzipien des Princeton WordNet 1.5 orientiert ist.[1]

Diese Anlehnung am Datenbankmodell und an den Aufbauprinzipien des WordNet bedeutet allerdings nicht, daß GermaNet aus einer Übersetzung der WordNet-Konzepte hervorgegangen ist. Vielmehr ist GermaNet aus verschiedenen lexikographischen Quellen (u.a. aus dem DEUTSCHEN WORTSCHATZ und dem DUDEN 8 der sinn- und sachverwandten Wörter) unter Berücksichtigung von Korpusfrequenzen von Hand aufgebaut worden. Darüber hinaus setzt GermaNet eigene Schwerpunkte sowohl auf der strukturellen als auch auf der konzeptuellen Ebene (zu den Unterschieden zwischen GermaNet und WordNet s. u.).

GermaNet modelliert den Grundwortschatz des Deutschen auf konzeptueller Ebene und verbindet Nomen, Verben und Adjektive durch elementare semantische Relationen und leistet somit einen wichtigen Beitrag zur Schaffung einer geeigneten Dateninfrastruktur für deutschsprachige computerlinguistische Anwendungen.

2.1 Abdeckung

Ausgangsziel des SLD-Projekts war, mit GermaNet einen on-line Thesaurus, der den deutschen Grundwortschatz abdeckt, zu erstellen (vgl. Hamp/Feldweg 1997). Zentrales Konzept der lexikalischen Kodierung sind die sogenannten *synsets*, die als abstrakte Bedeutungseinheiten zu gegebenen Konzepten eine Synonymenmenge bereitstellen. Es gibt semantische Relationen zwischen Konzepten (*synsets*) oder zwischen Wortbedeutungen

---

[1] GermaNet wurde unter der Leitung von Helmut Feldweg im Rahmen des SLD-Projektes („Ressourcen und Methoden zur semantisch-lexikalischen Disambiguierung") aufgebaut, das 1996 und 1997 vom Land Baden-Württemberg gefördert wurde. An der Realisierung des Projektes waren im weiteren Valérie Béchet-Tsarnos, Birgit Hamp, Michael Hipp, Claudia Kunze, Karin Naumann, Susanne Schüle, Rosmary Stegmann, Karen Steinicke, Christine Thielen und Andreas Wagner beteiligt.

(einzelnen Synonymen aus den *synsets*). Solche *synsets* werden gleichermaßen für Nomen, Verben und Adjektive implementiert.

Die zentrale Relation ist die Hyponymie-Beziehung, welche die Konzepte aller Wortarten (auch der Adjektive) hierarchisch gliedert.

Zur Zeit enthält die Datenbank, deren Abdeckung sich in kontinuierlicher Erweiterung befindet, ca. 25000 *synsets* und etwa 30000 Wortbedeutungen. Einträge der Datenbank werden mit Frequenzlisten, die aus Korpora extrahiert sind, abgeglichen, um fehlende Konzepte systematisch zu ergänzen.

Im Netz sind nur morphologische Vollformen kodiert und lediglich sehr geläufige Mehrwortlexeme wie *erste Hilfe* oder *gesprochene Sprache*. Eigennamen werden, sofern sie berücksichtigt sind (z.B. im Wortfeld der Geographie die Namen der Städte, Länder und Flüsse), speziell markiert. Einige wichtige Abkürzungen, etwa für die politischen Volksparteien, sind ebenfalls in GermaNet repräsentiert.

Der Datenbestand ist in fünfzehn semantische Felder unterteilt, die weitgehend von WordNet übernommen wurden und die zur Bearbeitung in den sog. ‚lexicographer files' hilfreich sind.

2.2 Relationen

GermaNet unterschiedet zwischen *lexikalischen* und *konzeptuellen Relationen*:

- Lexikalische Relationen wie Synonymie und Antonymie bestehen zwischen verschiedenen lexikalischen Realisierungen von Konzepten und sind bidirektionale Relationen, die für alle drei Wortklassen gelten.
- Konzeptuelle Relationen wie Hyponymie, Hyperonymie, Meronymie, Implikation und Kausation bestehen zwischen gegebenen Konzepten in all ihren Lexikalisierungen.

Ferner gibt es noch eine *semantische Derivationsrelation*, die kategorienübergreifend relevant ist für denominale Adjektive (**finanziell** zu **Finanzen**), deverbale Nominalisierungen (**Entdeckung** zu **entdecken**) und deadjektivische Nominalisierungen (**Müdigkeit** zu **müde**).

Das grundlegende Strukturierungsprinzip stellt die *Hyponymierelation*, wie sie z.B. zwischen **Rotkehlchen** und **Vogel** besteht, dar. Deszendentenketten für Nomen weisen oft eine beträchtliche Hierarchietiefe auf, aber auch im verbalen und adjektivischen Bereich ist die Taxonomie wesentliche Gliederungsrelation.

Die *Teil-Ganzes-Beziehung* (Meronymie) wird nur für Nomen spezifiziert. So ist ein **Arm** nur unzureichend als eine Art **Körper** klassifiziert, sondern zählt als Teil eines Körpers. Teil-Ganzes-Beziehungen liegen auch auf abstrakter Ebene vor, etwa in bezug auf Mitgliedschaft in einer Gruppe (**Vorsitzender** eines **Vereins**) oder als Material in einer Komposition (**Fensterscheibe** aus **Glas**).

Die *Implikationsbeziehung* ist anhand einiger weniger Beispiele kodiert. Hier sind Verbkonzepte in einem logischen Zusammenhang (‚backword presupposition') erfaßt, wie dieser z.B. zwischen **gelingen** und **versuchen** besteht.

Wichtiger und in größerem Ausmaß kodiert ist die klassenübergreifende *Kausationsrelation*, die lexikalische Resultative betrifft und z.B. **töten** und **sterben** oder **öffnen** und **offen** verknüpft.

Die folgenden Abbildungen zeigen einen Verbeintrag und einen nominalen Eintrag mit allen korrelierten Konzepten. Diejenigen Lesarten, die zu den Ausgangskonzepten keine

hyponymische bzw. hyperonymische Relation aufweisen, sind grau markiert. Wir haben die Lesartennummern der *synset*-Varianten aufgeführt, um zu verdeutlichen, daß in GermaNet Wortbedeutungen (‚senses') repräsentiert und semantisch miteinander verknüpft werden.

Das **synset** {**öffnen#3, aufmachen#2**} hat als Hyperonym {**wandeln#4, verändern#2, ändern#2**} sowie die vier Hyponyme {**aufschieben#1**}, {**aufstoßen#2**}, {**aufbrechen#1**} und {**aufsperren#1**}. Es gibt eine kausale Relation zum inchoativen **öffnen** (vgl. {**öffnen#1, aufgehen#1**}). Interessanterweise haben die Varianten im synset unterschiedliche Antonyme: **öffnen#3** hat das Antonym **schließen#7**, **aufmachen#2** das Antonym **zumachen#2**. Zur Verdeutlichung sind die bilateralen Antonym-Pfeile direkt auf die entsprechende Variante gerichtet und nicht auf den gesamten Konzeptknoten.

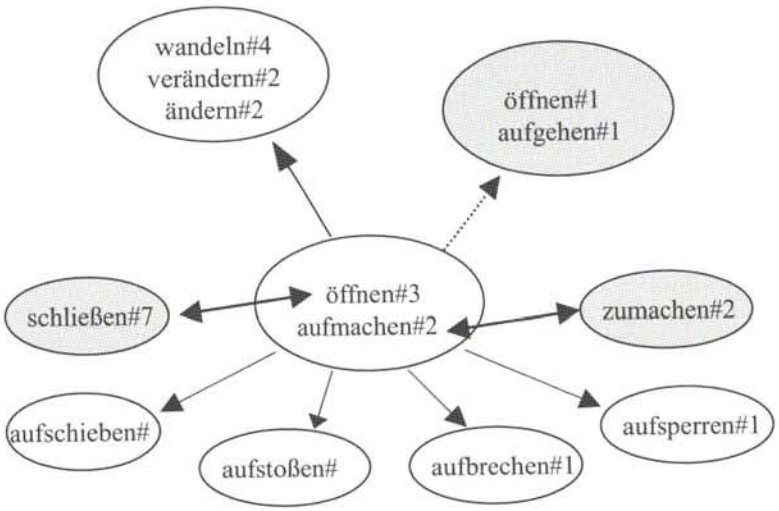


Abb. 1: Semantische Relationen des kausativen Verbs **öffnen**. Die einfachen Pfeile indizieren Überordnung (Pfeilspitze weist nach oben) und Unterordnung (mit der Pfeilspitze nach unten). Antonymie ist durch den Doppelpfeil gekennzeichnet, die kausative Relation mittels des gepunkteten Pfeils.

Abbildung 2 (folgende Seite) zeigt das Beispiel **Atmungsorgan** mit zwei Hyponymen (**Lunge** und **Kieme**), einem Hyperonym (**Organ**) und zwei Holonymen (**Oberkörper** und **Atemsystem**). Ein Meronym (**Luftröhre**) ist ebenfalls kodiert.

2.3 Kreuzklassifikation

In GermaNet werden Konzepte, die unterschiedlichen Hierarchien zugehören, kreuzklassifiziert. Das *Kaninchen* ist als *Haustier*, *Nutztier* und *Hasentier* klassifiziert, der *Hase* lediglich als *Hasentier*, der *Wellensittich* als *Haustier* und *Vogel*, der *Hund* lediglich als *Haustier* und die *Drossel* nur als *Vogel* (Abbildung 3).

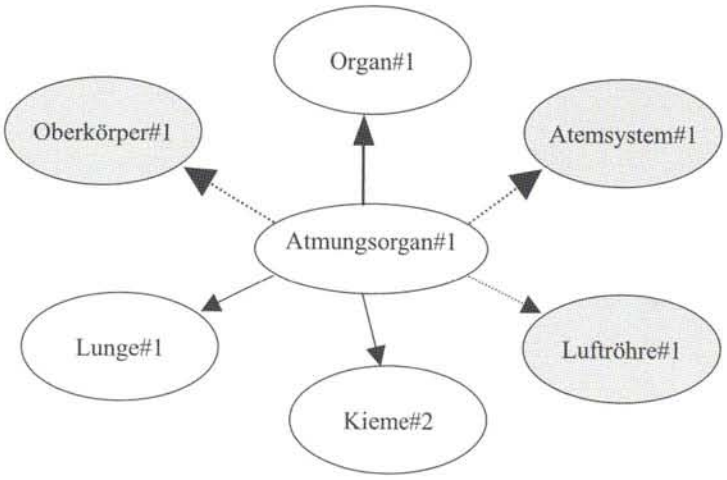


Abb. 2: Semantische Relationen des Nomens **Atmungsorgan**. Die Teil-Ganzes-Beziehung wird durch die gepunkteten Pfeile angezeigt, Meronyme mit nach unten, Holonyme mit nach oben weisender Pfeilspitze.

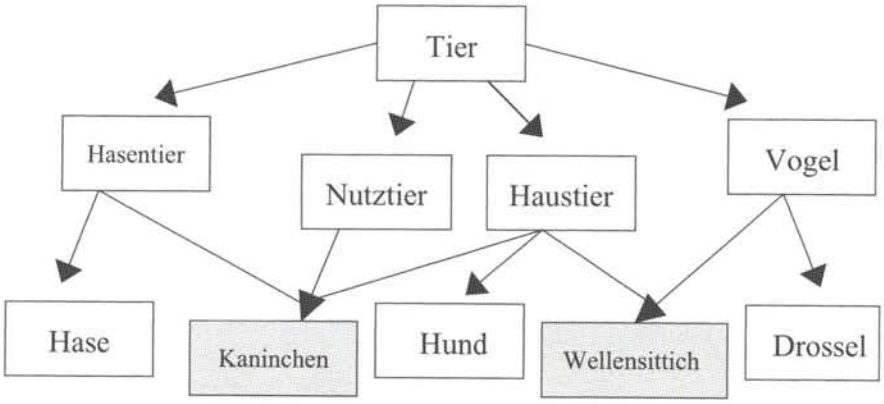


Abb. 3: Kreuzklassifikation im Tierreich: **Kaninchen** und **Wellensitich**

Nicht nur die Zugreifbarkeit der Lexeme unter verschiedenartigen Bedeutungsaspekten ist ein Vorteil des systematischen Kreuzklassifizierens.

Anhand durchgängiger Kreuzklassifikation können wir Muster regulärer Polysemie ausmachen, die für Restrukturierungen des Lexikons nützlich sind, vgl. aus der Klasse der Früchte diejenige Teilmenge, die wie *Banane* zugleich als *Nahrungsmittel* und als *Pflanze* klassifiziert ist. Andere Fälle regulärer Polysemie betreffen *Birke* als *Pflanze* und als *Holzart* oder *Tennis* als *Veranstaltung* und als *Sportart*. Die empirische Analyse deutet auf sehr viele produktive Muster (vgl. Buitelaars Analyse der CoreLex-Pattern 1998).

2.4 Artifizielle Konzepte

Auf der konzeptuellen Ebene tragen eigens eingeführte künstliche Knoten zu einer ausgewogeneren Taxonomie bei. Künstliche Konzepte können auf lexikalische Lücken in der Sprache bezogen sein (z.B. das fehlende Antonym zu *durstig*), aber auch rein konzeptuelle Konstrukte wie etwa **?Charakterbeschaffener** betreffen. Oftmals helfen künstliche Knoten, unmotivierte Kohyponymie zu vermeiden.

Das Beispiel in Abbildung 4 enthält mit **?Schullehrer** und **?hierarchischer Lehrer** zwei künstliche Konzepte, welche das Teilnetz im Wortfeld **Lehrer** symmetrischer strukturieren. Nach Cruse (1986:22) sollten Kohyponyme eines Mutterknotens möglichst inkompatibel zueinander sein. Diese Inkompatibilität operiert auf einer Ebene von Ähnlichkeit, die durch den gemeinsamen Oberbegriff gegeben ist, vgl. **Baby**, **Kleinkind**, **Vorschulkind**, **Schulkind** als Unterbegriffe zu **Kind**, die einander wechselseitig ausschließen.

Da ein Fachlehrer aber an einem Schultyp in einer hierarchischen Position unterrichtet, wären die sechs Endknoten des Beispielnetzes als direkte Deszendenten des **Lehrer**-Knotens nicht inkompatibel genug, so daß die nicht-lexikalisierten Konzepte, die in GermaNet durch ein initiales Fragezeichen gekennzeichnet werden, sinnvoll eingeführt sind.

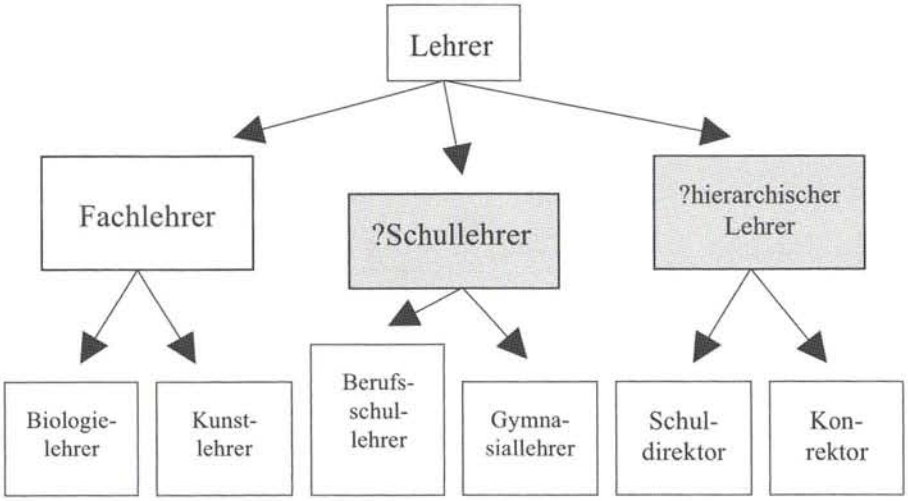


Abb. 4: Künstliche Konzepte im Wortfeld **Lehrer**

2.5 Subkategorisierungsrahmen

Alle Verbeinträge in GermaNet sind mit Subkategorisierungsrahmen und diesbezüglichen Beispielen versehen. Die kodierten Rahmen geben Aufschluß über das syntaktische Komplementierungsverhalten der GermaNet-Prädikate, leisten also einen Beitrag zur Syntax-Semantik-Schnittstelle.

Die Notation orientiert sich an den *Celex*-Frames, ist aber in bezug auf die Kodierung von Subjektphrasen und Reflexivphrasen leicht modifiziert worden. Unterschiedliche Verbrahmen zu einem Konzept helfen bei der Lesartendisambiguierung, vgl. das Beispiel *setzen*:

setzen1	NN.AN	*Er setzt die Fahnen.*
setzen2	NN.AR	*Sie setzt sich.*
setzen3	NN.AN.BL	*Er setzt den Schüler ans Fenster.*
setzen4	NE.AN	*Es setzt Prügel.*
setzen5	NN.AN.PP	*Er setzt seine Hoffnungen auf sie.*
setzen6	NN.An	*Die Häsin hat (Junge) gesetzt. (Jägersprache)*
setzen7	NN.BL	*Das Pferd setzt über die Hürde.*
setzen8	NN.AN.Dn	*Das Protokoll setzt (ihnen) Schranken.*

2.6 Unterschiede zwischen GermaNet und WordNet

Trotz der strukturellen Ähnlichkeit des GermaNet zum Princeton WordNet lassen sich folgende Unterschiede skizzieren:

- GermaNet orientiert sich im Gegensatz zu WordNet an linguistischen und nicht an psychologischen Strukturierungsprinzipien der Daten.
- Um eine ausgewogene Konzepthierarchie zu gestalten und unmotivierte Kohyponymie zu reduzieren, wird in GermaNet systematischer Gebrauch von artifiziellen Konzepten gemacht, die entsprechend markiert sind.
- GermaNet kodiert Partikelverben (vgl. Hamp 1997), die in WordNet nicht berücksichtigt werden.
- Die Kausationsrelation, die in WordNet lediglich als Relation zwischen Verbinstanzen vorgesehen ist, kann in GermaNet zwischen allen Wortarten kodiert werden.
- In GermaNet sind Adjektive taxonomisch strukturiert und unterliegen nicht dem Satelliten-Ansatz des WordNet, einem assoziativen Verbund von Adjektiven, der auch zu wenig intuitiven Konzepten wie *unschwanger* führen kann.

Zunehmend zeigt sich, daß die Großressource WordNet, welche ca. die dreifache Menge an Einträgen enthält, zu feinkörnige Lesartenunterscheidungen vornimmt, um effizient genug in computerlinguistischen Anwendungen zu sein. Für den Eintrag *go* gibt es 32 Lesarten. Der „richtige Polysemiegrad" ist gefragt, um erfolgreich Bedeutungsdisambiguierung leisten zu können (vgl. Buitelaar 1998 zur Modellierung regelgeleiteter Polysemie).

Eine Ressource mittlerer Größe wie GermaNet, die zudem noch von Restrukturierungsansätzen zur Lesartenreduktion mittels des sogenannten ‚Sense Clustering' (vgl. Peters et al. 1998) profitieren kann, kann durchaus leistungsfähiger in der lexikalisch-semantischen Disambiguierung sein als das behäbige WordNet.

3 GermaNet im multilingualen Kontext

Das Basisvokabular des GermaNet ist Bestandteil des multilingualen semantischen Netzes EuroWordNet, das im Rahmen eines Projektes der Europäischen Gemeinschaft für acht europäische Sprachen aufgebaut worden ist.[2] EuroWordNet ist eine wertvolle Ressource für

[2] Genau gesagt handelt es sich um ein zweiteiliges Projekt, EuroWordNet-1 und EuroWordNet-2. Tübingen ist Partner des EuroWordNet-2-Projektes LE4 8328, vgl. Vossen 1998: „Extending EuroWordNet with four languages".

die Sprachtechnologie in bezug auf multilinguale Anwendungen der Informationserschließung.

Durch die kontrastive Analyse der erstellten Daten im Projekt haben wir auch für die monolinguale Weiterarbeit an GermaNet profitieren können, z.B. in Hinblick auf die Datenabdeckung und durch die Verwendung der statistischen Methoden zur Evaluierung der Daten.

3.1 Interlingualer Index und Basiskonzepte

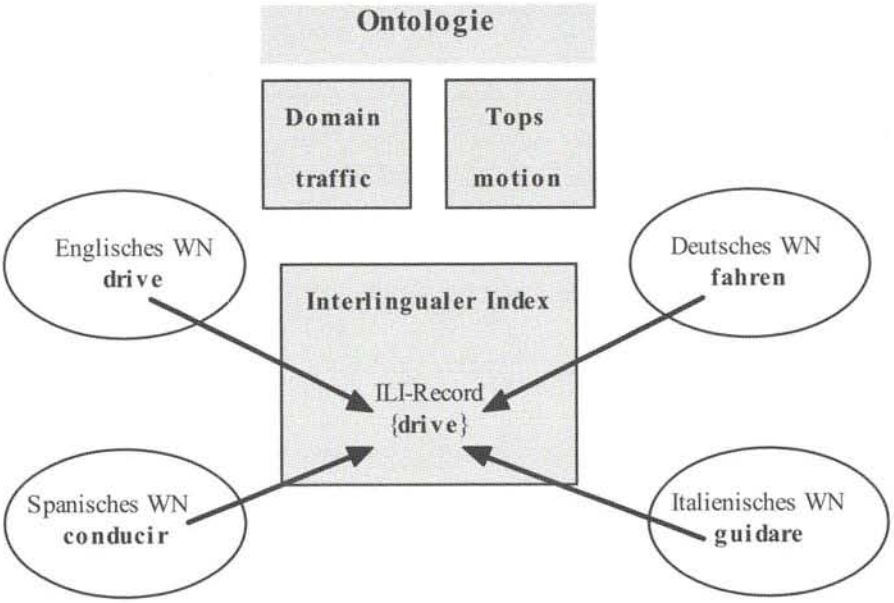


Abb. 5: Ausschnitt aus der EuroWordNet-Architektur. Die sprachunabhängigen Komponenten, zu denen neben dem ILI auch die merkmalsbasierten Ontologien gehören, sind grau markiert. Einzelsprachliche Konzepte werden über Äquivalenzrelationen an den ILI gelinkt.

Die EuroWordNet-Datenbank ist ein multilinguales Netz, das Basiskonzepte acht europäischer Sprachen (Englisch, Holländisch, Spanisch, Italienisch, Deutsch, Französisch, Estnisch und Tschechisch) in ihren semantischen Relationen modelliert. In der Datenbankarchitektur sind die einzelsprachlichen Komponenten über den sprachunabhängigen Interlingualen Index (ILI) korreliert. Trotz seiner Übersprachlichkeit ist der ILI, welcher eine unstrukturierte Liste sogenannter ILI-Records[3] enthält, durch die Dominanz und Vorreiterrolle des Princeton WordNet stark an den englischen Konzepten bzw. den WordNet-Einträgen orientiert. Die sprachspezifischen Konzepte der einzelnen Sprachen werden über eine Äquivalenzrelation an passende ILI-Records gelinkt. Einzelne Sprachpaare zu erfragten Konzepten werden also mittelbar (über den ILI) erzeugt, vgl. Abbildung 5.

---

[3] Ein ILI-Record ist durch einen eindeutigen Code, den *unique identifier*, gekennzeichnet.

Um die Abdeckung der sprachspezifischen Wortnetze kompatibel zu gestalten, müssen alle Sprachnetze anhand der sogenannten ‚Base Concepts' strukturiert sein. Die 1300 ‚Base Concepts' (ca. 1000 Nomen und 300 Verben), die durch einzelsprachliche Selektionen und statistische Evaluierungen dieser Selektionen ermittelt worden sind, müssen folgenden Kriterien genügen:

- ‚Base Concepts' sollten einen hohen Abstraktionsgrad aufweisen, der sich an der Menge der dominierten Unterbegriffe und an der Hierarchietiefe der dominierten Kette manifestiert. Base Concepts sollten spezifischer sein als semantische Merkmale der Top Ontology wie **Funktion, Eigenschaft, Dynamisch**, etc., aber auch abstrakter als Roschs ‚Basic Level Concepts'[4] (z.B. **Tisch** und **Hammer**). Den richtigen Abstraktionsgrad weisen deren semantische Oberbegriffe **Möbel** und **Werkzeug** auf.
- Außerdem sollen Konzepte, die in einer Sprache (und möglichst auch sprachübergreifend) sehr häufig vorkommen, als ‚Base Concepts' berücksichtigt werden, auch wenn sie nicht den gewünschten Abstraktionsgrad aufweisen, wie etwa **lieben** und **mögen**.

Dies Inventar gemeinsamer ‚Base Concepts' ist in einem ersten Schritt über Äquivalenzrelationen an den ILI zu binden, um dann sowohl die leicht erfaßbaren Topknoten als auch die Hyponyme erster Ordnung (meist ‚Basic Level Concepts') zu linken. Die einzelsprachlichen Netze können so unabhängig voneinander, jedoch mit einem Großmaß an Kompatibilität, integriert werden.[5] Durch die Vererbung der semantischen Merkmale der Top Ontology ist es weiterhin auch möglich, die Abdeckung der Netze in den einzelnen semantischen Feldern statistisch zu evaluieren.

3.2 Relationstypen

Nicht immer können äquivalente ILI-Records als Übersetzungen der einzelsprachlichen Konzepte ausgemacht werden. Neben den unterschiedlichen Lexikalisierungspattern, die auf sprachliche und kulturelle Unterschiede zurückgehen, sind dafür auch unterschiedliche Gewichtungen der Konzepte sowie Kodierungslücken verantwortlich. So gibt es im WordNet (das ja weitgehend den ILI prägt) kein Konzept, das dem deutschen Lexem **Lebensgefährte** (als unverheirateter Partner einer eheähnlichen Lebensbeziehung) entspricht. Im deutsch-englischen *COLLINS* hingegen konnten wir das Literal *companion through life* finden. Ein weiteres Beispiel betrifft den Wettbewerbstyp **championships** ‚Meisterschaft', der eine Lexikalisierung im Englischen hat, aber in EuroWordNet nicht als ILI-Record vorhanden ist. Neben der Synonymiebeziehung und der Quasi-Synonymiebeziehung stehen auch nicht-synonymische Äquivalenzlinks der Hyperonymie und Meronymie, ferner Rollenbeziehungen und Kausationsbeziehungen zur Verfügung.

Mitunter ist ein Konzept gut abzubilden, indem mehrere nicht-synonymische Verknüpfungen verwendet werden, vgl. Abbildung 6:

---

[4] Vgl. Rosch (1978).

[5] Mittels dieser Prozedur ist das erste Daten-Ensemble mit ca. 7500 Einträgen entstanden.

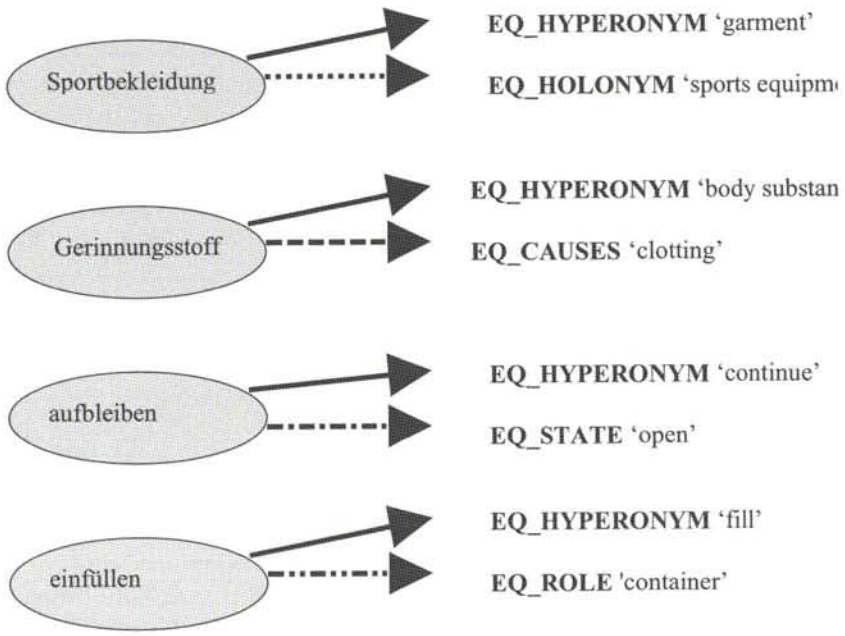


Abb. 6: Beispiele für nicht-synonymische Äquivalenzlinks

Der ebenfalls in der EuroWordNet-Spezifikation vorgesehene Near-Synonym-Link verleitet mitunter zu Ungenauigkeiten. Auf ihn wird oft rekurriert, wenn die Quasi-Äquivalente im Bedeutungsumfang nicht ganz deckungsgleich sind oder wenn zu einem Konzept mehr als ein Synonymlink zu unterschiedlichen ILI-Konzepten etabliert wird, die im WordNet nicht dem gleichen *synset* angehören.

3.3 Synergien für GermaNet

Die Integration des GermaNet in die EuroWordNet hat einige wesentliche Impulse für Optimierungsansätze unserer monolingualen Ressource beigesteuert:

- Im Projektkontext konnten wir ein lexikalisches Werkzeug unserer Amsterdamer Projektpartner für das Deutsche adaptieren, das sehr benutzerfreundlich zum Browsen und Editieren lexikalischer Datenbanken einsetzbar ist.[6]
- Mithilfe der ‚Base Concepts' konnte eine strukturierte, interlingual hinterfragte Überprüfung der Grundwortschatzabdeckung vorgenommen werden. So hatte es im GermaNet u. a. eine Abdeckungslücke im Bereich der Ereignisnominalisierungen gegeben, die nun ausgeglichen werden kann.

---

[6] Hierbei handelt es sich um das ALS (Amsterdam Lexical System), das von Boersma und Vossen zwischen 1992 und 1997 entwickelt worden ist. Die Adaption für das Deutsche von A. Wagner haben wir TALS (Tübinger ALS) genannt. Im Verlauf des Projektes haben wir TALS für die ILI-Anbindung verwendet und nicht zur Editierung des GermaNet.

- Aus der EuroWordNet Spezifikation wollen wir aus dem Inventar innersprachlicher Relationen die Rollentypen für GermaNet übernehmen. Diese sprachinternen Relationen (**ROLE_AGENT, role_patient, role_location**) und deren konverse Relationstypen (**involved_agent, involved_Patient, involved_location**) wollen wir zur Kodierung semantischer Rollen in GermaNet und zur Formulierung von geeigneten Selektionsrestriktionen für Verbeinträge implementieren.

4 Anwendungsperspektiven des GermaNet

In diesem Kapitel werden die Einsatzmöglichkeiten lexikalisch-semantischer Netze wie WordNet und GermaNet in der maschinellen Sprachverarbeitung anhand der Beispiele Lesartendisambiguierung, Informationserschließung und Textkategorisierung exemplarisch aufgezeigt. Insbesondere soll die Ermittlung von Selektionspräferenzen im Mittelpunkt stehen.

4.1 Lesartendisambiguierung

Lesartendisambiguierung (‚word sense disambiguation') ist von entscheidender Bedeutung für maschinelles Sprachverstehen. Gerade häufig verwendete Wörter sind mehrdeutig. Um einen Satz semantisch interpretieren zu können, müssen die mehrdeutigen Wörter in diesem Satz disambiguiert werden. Hierbei ist ein lexikalisch-semantisches Netz wie WordNet oder GermaNet in zweifacher Hinsicht nützlich: Zum einen liefert ein solches Netz (mit seinen Knoten) ein Inventar semantischer Konzepte, mit denen Wortbedeutungen repräsentiert werden können, die *synsets*. Ist ein Wort mehrdeutig, so gehört es zu mehreren *synsets*, z.B. **Ton** mit den *synsets* {**Ton**, **Laut**} und {**Ton**, **Tonerde**}, was den beiden Lesarten dieses Wortes entspricht. Zum anderen modelliert ein lexikalisch-semantisches Netz (mit seinen Kanten) Beziehungen zwischen Konzepten. Diese Beziehungen liefern wichtige Informationen für die Disambiguierung. So nutzt der nun skizzierte Ansatz die durch die Hyponymie-Relationen definierte Hierarchie, um die semantische Ähnlichkeit von Konzepten zu quantifizieren.

Stetina et al. (1998) entwickeln ein Verfahren zur semantischen Disambiguierung von Inhaltswörtern in geparsten Texten. Zur Bestimmung der Lesart eines Wortes werden die Lesarten derjenigen Wörter herangezogen, die mit diesem Wort in einer syntaktischen Relation (z.B. Subjekt–Verb) stehen. Verschiedene Kombinationen von Wortbedeutungen werden mit unterschiedlicher Wahrscheinlichkeit durch bestimmte syntaktische Relationen miteinander verbunden. Den Inhaltswörtern in einem Satz werden nun diejenigen (WordNet-)Lesarten zugewiesen, die gemäß den in diesem Satz vorhandenen syntaktischen Relationen die wahrscheinlichsten sind. Die zu Grunde liegenden Wahrscheinlichkeiten (z.B. die Wahrscheinlichkeit, daß ein Verb mit der Lesart y ein Subjekt mit der Lesart x hat) werden durch eine statistische Analyse des semantisch annotierten Korpus SemCor (Miller u.a. 1993) eingeschätzt. SemCor besteht aus etwa 200000 Wörtern, die jeweils mit ihrer WordNet-Lesart annotiert sind. Hierbei ergibt sich das Problem, daß viele Wörter, die später disambiguiert werden müssen, in diesem Korpus nicht vorkommen, so daß für sie keine Wahrscheinlichkeit geschätzt werden kann. Dies Problem wird dadurch gelöst, daß

für die Einschätzung der Lesarten solcher Wörter die Wahrscheinlichkeiten semantisch ähnlicher Lesarten herangezogen werden, die in SemCor vorkommen. Die semantische Ähnlichkeit zweier Lesarten wird über die Hyponymie-Hierarchie von WordNet ermittelt: Je näher beieinander die entsprechenden Konzepte in der Hierarchie angeordnet sind, desto größer ist die semantische Ähnlichkeit.

4.2 Informationserschließung

Bei der automatischen Informationserschließung (‚Information Retrieval') geht es darum, aus einem umfangreichen Inventar von Dokumenten diejenigen Texte zu finden, die bestimmte, durch eine Anfrage spezifizierte Informationen enthalten. Für diese Aufgabe kann ein semantisches Netz nützliche Hinweise liefern. Wenn sowohl die Anfrage als auch die zu durchsuchenden Dokumente semantisch disambiguiert sind, kann gezielt nach Begriffen in der intendierten Lesart gesucht werden. Wenn z.B. nach Informationen zum Stichwort **Bank** (im Sinne von **Geldinstitut**) gesucht wird, so werden keine Texte über Sitzmöbel geliefert. Dadurch wird die „Treffergenauigkeit" der ermittelten Dokumente (‚precision') erhöht. Ein zweiter Vorteil eines semantischen Netzes ist, daß mit seiner Hilfe die Anfrage um Konzepte erweitert werden kann, die mit den Suchbegriffen in einer semantischen Beziehung stehen. So können bei einer Anfrage nach **Bank** auch Texte gefunden werden, in denen der Begriff selbst nicht vorkommt, jedoch **Geldinstitut** oder **Sparkasse**. Dadurch wird die Anzahl der korrekt ermittelten Dokumente (‚recall') erhöht. Mit einem multilingualen semantischen Netz wie EuroWordNet ist so auch die Durchsuchung von Texten in unterschiedlichen Sprachen möglich, indem die Anfrage um Konzepte aus verschiedenen Sprachen erweitert wird, die zu den Suchbegriffen äquivalent sind.

Gonzalo et al. (1998) haben durch entsprechende Experimente herausgefunden, daß die Performanz von Information Retrieval signifikant erhöht wird, wenn die Anfrage und die zu durchsuchenden Texte mit WordNet-*synsets* indiziert sind (d.h. jedem Wort das entsprechende *synset* zugewiesen wird). Diese Indizierung liefert erstens Lesartendisambiguierung und zweitens die Erweiterung der Anfrage um Synonyme der Suchbegriffe.

4.3 Textkategorisierung

Textkategorisierung befaßt sich mit der Klassifikation von Texten im Hinblick auf eine (vorgegebene) Menge von Kategorien (z.B. Domänen oder Textsorten). Systeme zur automatischen Textkategorisierung werden zunächst mit Hilfe einer Kollektion von Texten trainiert, die manuell mit dem vorgesehenen Inventar von Kategorien klassifiziert wurden. Ein neu zu kategorisierender Text wird mit den Texten in dieser Kollektion bzgl. der Vorkommenshäufigkeit bestimmter Begriffe verglichen. Die ausgewählte Kategorie ergibt sich aus diesem Vergleich.

Buenaga Rodríguez et al. (1997) ziehen WordNet als zusätzliche Informationsquelle heran: In das Kategorisierungsverfahren gehen auch die Vorkommenshäufigkeiten der Kategoriebezeichnungen selbst sowie ihrer Synonyme im zu klassifizierenden Text ein. Die Synonyme werden aus WordNet extrahiert. Auch hier ist die Einbeziehung anderer Konzepte denkbar, die mit den Kategoriebezeichnungen durch semantische Relationen verbunden sind.

4.4 Eine lexikographische Anwendung: Selektionsbeschränkungen

In diesem Abschnitt wird exemplarisch eine lexikographische Anwendung des GermaNet ausführlicher beschrieben: die Akquisition von Selektionsbeschränkungen. Diese Aufgabe stellt gleichzeitig eine Perspektive für die qualitative Weiterentwicklung von GermaNet dar.

Selektionsbeschränkungen sind semantische Beschränkungen, die ein Prädikat (z.B. ein Verb oder Adjektiv) seinen Argumenten (z.B. einer Verbergänzung oder einem durch ein Adjektiv modifizierten Nomen) auferlegt. So fordert beispielsweise das Verb *essen* einen menschlichen oder tierischen Agens und einen Patiens, der ein Nahrungsmittel bezeichnet.

4.4.1 Motivation

Die Akquisition von Selektionsbeschränkungen ist aus mehreren Gründen sinnvoll. Zum einen können sie einen wichtigen Beitrag zur syntaktischen und lexikalischen Disambiguierung leisten. Im Satz

(1) Das Brot schneidet die Mutter.

folgt aus den Selektionsbeschränkungen von *schneiden*, daß *das Brot* die Akkusativ- und *die Mutter* die Nominativergänzung ist, nicht umgekehrt, wie es nach rein morphologischen und syntaktischen Kriterien möglich wäre. Im Beispiel

(2) Der Mann tritt gegen den Ball.

wird *Ball* aufgrund der Selektionsbeschränkungen von *treten* als Spielgerät (und nicht als Tanzveranstaltung) disambiguiert. Selektionsbeschränkungen sind also als eine Informationsquelle zur Disambiguierung für maschinelle Sprachverarbeitungssysteme interessant.

Daneben kann es jedoch auch zweckmäßig sein, Selektionsbeschränkungen in Lexika für menschliche Benutzer aufzunehmen. Vor allem für Fremdsprachenlerner können sie wichtige Hinweise für den Wortgebrauch liefern, die aus der Wortbedeutung nicht unbedingt und aus Verwendungsbeispielen höchstens indirekt hervorgehen. Z.B. verwendet man *tranchieren* nur im Zusammenhang mit Fleisch, nicht mit Fisch, Gemüse oder Holz, was aus der Wortbedeutung nicht zwingend folgt. In einigen deutschen Wörterbüchern sind Selektionsbeschränkungen explizit angegeben (z.B. in VERBEN IN FELDERN und im WÖRTERBUCH ZUR VALENZ UND DISTRIBUTION DEUTSCHER VERBEN), in anderen sind sie implizit in Bedeutungsdefinitionen und Beispielen enthalten.

4.4.2 Akquisition von Selektionsbeschränkungen

Es ist offensichtlich, daß die manuelle Akquisition von Selektionsbeschränkungen zeit- und arbeitsaufwendig ist, wenn sie für ein Lexikon breiteren Umfangs durchgeführt werden soll. Dies ergibt sich nicht zuletzt aus der Tatsache, daß prinzipiell jede semantische Eigenschaft eine Rolle bei Selektionsbeschränkungen spielen kann und es folglich empirisch nicht adäquat ist, Selektionsbeschränkungen mit einem relativ kleinen Inventar semantischer Merkmale wie **belebt** oder **abstrakt** zu modellieren. Manche Prädikate stellen sehr spezielle Selektionsanforderungen an ihre Argumente, die von einem solchen Inventar nicht erfaßt

werden können. So kann man nichts anderes *diagonalisieren* als eine *Matrix*, und nur eine *Geschwulst* kann als *gutartig* charakterisiert werden.

In den letzten Jahren sind Verfahren entwickelt worden, um Selektionsbeschränkungen durch statistische Analyse großer Textkorpora zu ermitteln (Resnik 1993, Ribas 1994, Abe/Li 1996). Diese Verfahren ermitteln WordNet-Konzepte, die von einem Prädikat präferiert werden (z.B. *food* als Objekt von *eat*). Hierbei weisen sie den Konzepten jeweils einen Präferenzwert zu, der die Stärke der Präferenz charakterisiert und auf der Grundlage relativer Häufigkeiten von Prädikat-Argument-Kookkurenzen im untersuchten Korpus berechnet wird. Tatsächlich haben Selektionsbeschränkungen eher den Charakter von Präferenzen als von scharfen Restriktionen. So sind kontextbedingte oder metaphorisch zu interpretierende Abweichungen von Selektionspräferenzen wie in *Angst essen Seele auf* durchaus gängig. Außerdem können auch innerhalb des durch Selektionsbeschränkungen sanktionierten „semantischen Raumes" unterschiedliche Präferenzgrade vorliegen. So lassen die Selektionsbeschränkungen von *lesen* für den Satz

(3) Der Student liest den Artikel.

die Interpretation von *Artikel* sowohl als ein Determinans als auch als Text zu. Jedoch wird die Text-Lesart stärker präferiert, sofern kein spezifischer Kontext die andere Interpretation nahelegt. Die statistischen Verfahren haben also neben der automatischen Akquisition auf breiter empirischer Basis den Vorteil, daß sie durch die Quantifizierung des Präferenzverhaltens eines Prädikats das Phänomen Selektionsbeschränkungen adäquater modellieren.

Mit Hilfe von GermaNet sollen mit diesen Verfahren Selektionspräferenzen für das Deutsche ermittelt werden. Für lexikographische Zwecke ist es hierbei wichtig, daß sich die ermittelten Konzepte auf einer angemessenen Generalisierungsebene befinden. Angenommen, wir stoßen u.a. auf folgende Korpusbelege für das Prädikat *essen*:

(4a) Meine Tochter ißt gern Käsekuchen,
(4b) Max hat schon drei Äpfel gegessen.
(4c) Muslime essen kein Schweinefleisch.

Die Komplemente sollen einerseits möglichst kompakt, andererseits empirisch adäquat repräsentiert werden. Das Konzept, das die Objekte in den genannten Beispielen angemessen zusammenfaßt, ist *Nahrungsmittel*. *Gegenstand* wäre zu generell; Konzepte wie *Backwaren*, *Obst*, *Fleisch*, etc. würden dem Kompaktheitsdesiderat zuwiderlaufen.

Das Problem der angemessenen Generalisierung wird von den oben genannten Verfahren nicht befriedigend gelöst. Abe/Li (1996) nehmen zwar für sich in Anspruch, einen informations-theoretisch motivierten Ansatz zu implementieren, der die angemessene Generalisierungsebene liefert. Jedoch haben eigene Experimente gezeigt, daß der ermittelte Generalisierungsgrad von der Größe des untersuchten Korpus sowie der Häufigkeit des untersuchten Prädikats abhängt: Bei häufigen Verben wird tendentiell untergeneralisiert, bei seltenen Verben tendentiell übergeneralisiert. Dieses Verhalten ist zumindest für lexikographische Zwecke nicht akzeptabel. Hier sind also Modifikationen der Verfahren notwendig.

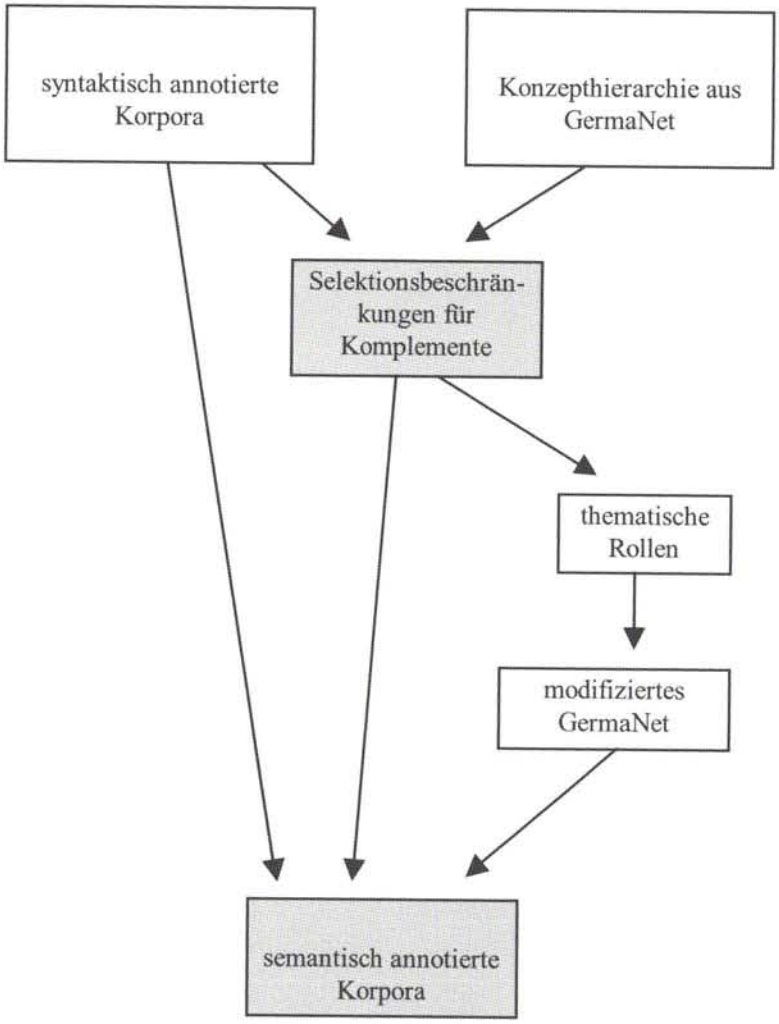


Abb. 7: Selektionspräferenzen und semantische Annotation

4.4.3 Kodierung von Selektionspräferenzen in GermaNet

Die oben genannten Verfahren ermitteln selektionale Präferenzen für syntaktische Verbkomplemente. Jedoch sind Selektionspräferenzen primär mit thematischen Rollen wie Agens, Patiens oder Instrument assoziiert, die auf unterschiedliche Weise syntaktisch realisiert werden können. So kann der Patiens von *kochen* sowohl als Nominativ- als auch als Akkusativergänzung realisiert werden:

(5a) Der Küchenchef kocht die Suppe.
(5b) Die Suppe kocht.

Solche Korpusbelege können dazu führen, daß *Nahrungsmittel* sowohl für das Subjekt als auch das direkte Objekt von *kochen* als präferiertes Konzept ermittelt wird. Beiden Sätzen liegt jedoch zugrunde, daß *kochen Nahrungsmittel* für seine Patiens-Rolle präferiert.

Um das Selektionsverhalten eines Verbs adäquat zu modellieren, müssen also die für seine syntaktischen Komplemente ermittelten Präferenzen auf die zu Grunde liegenden thematischen Rollen abgebildet werden. Um diesbezüglich geeignete Verfahren zu entwickeln, bietet sich z.B. der Ansatz von McCarthy/Korhonen 1998 an.

Die gewonnenen Präferenzen für thematische Rollen sollen in GermaNet kodiert werden. Hierbei soll auf das Inventar entsprechender Relationen zurückgegriffen werden, das im Rahmen der EuroWordNet-Spezifikation festgelegt wurde. Beispielsweise soll die Relation *kochen* **involved_patient** *Nahrungsmittel* in GermaNet aufgenommen werden.

4.4.4 Semantische Annotierung von Korpora

Das um thematische Relationen erweiterte GermaNet soll u.a. für die Erstellung semantisch annotierter Korpora (genauer: Korpora, deren Wörter semantisch disambiguiert sind) genutzt werden, die wiederum für bisher genannte Anwendungen von großem Interesse sind. Nicht nur die Hyponymiebeziehungen sind hierbei nützlich (vgl. 4.1), sondern auch die thematischen Rollenrelationen. So kann Ball in Beispiel (2) durch eine **involved_patient** Relation disambiguiert werden, die zwischen *treten* und der physikalischen Lesart von *Ball* besteht.

Aber auch die statistisch ermittelten Selektionspräferenzen selbst (s. Abschnitt 4.4.2) bilden eine wichtige Informationsquelle für die semantische Disambiguierung, da z.B. die Präferenzstärke, die nicht in GermaNet kodiert werden soll, essentiell sein kann (vgl. Beispiel (3)).

Die von uns vorgesehene Anwendung, das GermaNet zusammen mit syntaktisch annotierten Korpora für die Ermittlung von Selektionspräferenzen und die semantische Korpusannotierung einzusetzen, ist in Abbildung 7 schematisch dargestellt:

5 Schlußwort

In diesem Aufsatz haben wir den Aufbau und die grundlegenden Eigenschaften des GermaNet, eines lexikalisch-semantischen Wortnetzes für das Deutsche, beschrieben und seine Anwendungsperspektiven für die Computerlinguistik dargelegt. Wir haben gezeigt, wie wir von der Kooperation im Rahmen eines multilingualen Projektes in bezug auf die qualitative als auch quantitative Abdeckung profitieren konnten. Die mittlere Größe des Netzes und die Qualität der Daten bieten eine empirische Basis sowohl für theoriebezogene Fragestellungen als auch für praktische Anwendungen.

Der Ausbau unserer Ressource umfaßt nicht nur die korpusbasierte Erweiterung des repräsentierten Wortschatzes, sondern auch die Adaption neuer innersprachlicher Relationstypen, wie sie durch thematische Rollen gegeben sind. So kann GermaNet noch effizienter die Ermittlung von Selektionspräferenzen sowie die semantische Annotierung von Korpora unterstützen.

6 Literatur

Abe, Naoki und Li, Hang (1996): Learning Word Association Norms Using Tree Cut Pair Model. In: Proc. of 13th Int. Conf. on Machine Learning.

Buenaga Rodríguez, Manuel de und Gómez-Hidalgo, José-María und Díaz-Agudo, Belén (1997): Using WordNet to Complement Training Information in Text Categorization. In: Proc. of 2nd Int. Conf. on Recent Advances in NLP.

Buitelaar, Paul (1998): CORELEX: Systematic Polysemy and Underspecification. PhD thesis. Brandeis University.

COLLINS German-English, English-German Dictionary. Hg. Peter Terrell, Veronika Schnorr, Wendy V. A. Morris, Roland Breitsprecher. Glasgow: Harper-Collins ^{2}1991.

Cruse, Alan (1986): Lexical Semantics. Cambridge: Cambridge University Press.

DEUTSCHER WORTSCHATZ. EIN WEGWEISER ZUM TREFFENDEN AUSDRUCK. Hgg. H. Wehrle und H. Eggers. Stuttgart: Ernst-Klett-Verlag 1961.

Dowty, Donald (1988): On the Semantic Content of the Notion Thematic Role. In: G. Chierchia, B. Partee und R. Turner (eds.): Property Theory, Type Theory and Natural Language Semantics. Dordrecht: Kluwer.

DUDEN 8: SINN- UND SACHVERWANDTE WÖRTER. Hgg. Günther Drosdowski, Wolfgang Müller, Werner Scholze-Stubenrecht, Matthias Wermke. Mannheim: Dudenverlag ^{2}1986.

EuroWordNet: Building a multilingual database with wordnets for several European languages. University of Amsterdam. 31. Mai 1999, http://www.let.uva.nl/ewn/.

Fellbaum, Christiane (1998): WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press.

GermaNet. Universität Tübingen. 31. Mai 1999, http://www.sfs.nphil.uni-tuebingen.de/lsd/.

Gonzalo, Julio und Verdejo, Felisa und Chugur, Irina und Cigarrán, Juan (1998): Indexing with WordNet synsets can improve text retrieval.

Hamp, Birgit (1997): German Particle Verbs in GermaNet. Unveröffentlichtes Arbeitspapier.

– und Feldweg, Helmut (1997): GermaNet – A lexical-semantic net for German. In: Proc. of ACL/EACL-97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid.

Harley A. (1994): Cambridge Language Survey: semantic tagger. Technical report. Acquilex II Working Paper 39. Cambridge: Cambridge University Press.

McCarthy, Diana und Korhonen, Anna (1998): Detecting Verbal Participation in Diathesis Alternations. In: Proc. of 36th Annual Meeting of the ACL, Montreal, Vol. 2, 1493–1495.

McCawley, James D. (1968): The role of semantics in grammar. – In: E. Bach, R. Harms (eds.): Universals in Linguistic Theory. 125–169. New York: Holt, Rinehart & Winston.

Miller, George und Beckwith, Richard und Fellbaum, Christiane und Gross, Derek und Miller, Katherine (1990): Five Papers on WordNet. CSL Report, Vol. 43. Cognitive Science Laboratory, Princeton University.

– und Leacock, Claudia und Tengi, Randee (1993): A semantic concordance. In: Proc. of ARPA Human Language Technology Workshop 303–308.

Peters, Wim und Peters, Ivonne und Vossen, Piek (1998): The Reduction of Semantic Ambiguity in Linguistic Resources. In: Proc. of 1st Int. Conf. on Language Resources and Evaluation, Granada.

Resnik, Philip Stuart (1993): Selection and Information: A Class-Based Approach to Lexical Relationships. PhD thesis. University of Pennsylvania.

Ribas, Francesc (1994): An experiment on learning appropriate selectional restrictions from a parsed corpus. In: Proc. of COLING, Kyoto.

Rosch, Eleanor (1978): Principles of Categorization. – In: E. Rosch, B. Lloyd (eds.): Cognition and Categorization. Hillsdale: Lawrence Erlbaum Associates. 27–48.

Stetina, Jiri und Kurohashi, Sadao und Nagao, Makoto (1998): General Word Sense Disambiguation Method Based on Full Sentential Context. In: Proc. of COLING/ACL'98 Workshop on Usage of WordNet for NLP, Montreal.

Vossen, Piek (1997): EuroWordNet-2. Extending EuroWordNet with other languages. Annex I to Telematics Application Programme. LE 4–8328.

VERBEN IN FELDERN. VALENZWÖRTERBUCH ZUR SYNTAX UND SEMANTIK DEUTSCHER VERBEN. Hg. Helmut Schumacher. Berlin: Walter de Gruyter 1986.

WÖRTERBUCH ZUR VALENZ UND DISTRIBUTION DEUTSCHER VERBEN. Hgg. Gerhard Helbig, Wolfgang Schenkel. Leipzig: VEB Bibliographisches Institut 1969.

Yarowsky, D. (1992): Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In: Proc. of 15th Int. Conf. on Computational Linguistics. Vol II, 454–460.

Claudia Kunze, Tübingen
Andreas Wagner, Tübingen

Lothar Lemnitzer

Das Internet als Medium für die Wörterbuchbenutzungsforschung

1 Wozu Benutzerforschung?

Wann und wie Wörterbücher tatsächlich benutzt werden, dies ist eine der am schwierigsten zu beantwortenden Fragen in der Wörterbuchforschung. Die experimentelle Benutzerforschung erfordert einen hohen Aufwand, der die Datenerhebung bisher nur im Rahmen akademischer Forschung vertretbar erscheinen läßt.[1]

Bei der Herstellung von Wörterbüchern muß von einem Idealtypen von Benutzer ausgegangen werden. Entsprechend vage sind die Angaben zu den Zielgruppen in vielen kommerziellen Printwörterbüchern.[2]

Dabei hängt der Erfolg eines Wörterbuches in einer echten Konkurrenzsituation – mehrere, im Preis vergleichbare und auf die gleiche Zielgruppe ausgerichtete Wörterbücher sind verfügbar – von der Zufriedenheit der Benutzer ab. Zufriedenheit stellt sich dann ein, wenn möglichst viele Benutzungshandlungen zum gewünschten Ziel führen, oder – wenn nicht – die Gründe des Mißerfolgs erkennbar sind. Für die Produzenten von Wörterbüchern wäre es besonders wichtig zu erfahren, welche Benutzungshandlungen erfolglos waren. Solche Daten können in den Arbeitsprozeß, der zu einer Neuauflage eines Wörterbuches führt, einfließen.[3]

Das World-Wide Web – die Hypertext-Oberfläche für den Zugriff auf das Internet – bietet nun neue und effektive Möglichkeiten der Benutzerforschung. Die Möglichkeiten, an eine genauere Beschreibung des Kollektivs der Benutzer eines bestimmten Wörterbuchs zu gelangen, sind auch für die Vertreiber kommerzieller Wörterbücher interessant. Eine unabdingbare Voraussetzung dafür ist natürlich, daß die Wörterbuchdaten mediengerecht aufbereitet, also zumindest digitalisiert sind. Des weiteren muß das vertragliche Verhältnis zwischen Wörterbuchproduzent und Wörterbuchbenutzer auf eine neue Ebene gestellt werden: Der Wörterbuchbenutzer erhält zunächst eine Leistung umsonst und liefert im Gegenzug dem Produzenten Daten, die zu einer genaueren Bestimmung des Kollektivs der

---

[1] Vgl. Hartmann 1987, Hartmann 1989, Wiegand 1987, Wiegand 1998.

[2] Vgl. Püschel, 1989, S. 128.

[3] „Der einfache Grundgedanke aller Wörterbuchbenutzungsforschung läßt sich so wiedergeben: Wenn man Kenntnisse, insbesondere empirische, über den Wörterbuchbenutzer und die Wörterbuchbenutzung hat, kann man den Nutzungswert künftiger Wörterbücher erhöhen“ (Wiegand 1987, S. 179; ähnlich Wiegand 1998, S. 259).

Benutzer und zu einer Kategorisierung erfolgreicher und erfolgloser Benutzungshandlungen führen.

Die ZERES GmbH, mein – mittlerweile: bisheriger – Arbeitgeber, vertreibt mehrere zweisprachige Wörterbücher – jeweils eines für jede Richtung der Sprachenpaare deutsch-englisch, deutsch-französisch, deutsch-italienisch und deutsch-spanisch. Die Lemmaauswahl erfolgte für jedes Wörterbuch unabhängig und im Wesentlichen auf die Auswertung großer Textkorpora gestützt. Die jeweiligen Teams von Lexikographinnen hatten Lemmalücken, die sich aus Abdeckungslücken bei den jeweiligen Korpora ergaben, zu schließen. Es war uns wichtig, von den Benutzern unserer Wörterbücher zu erfahren, welche Suchen aufgrund fehlender Einträge in den Wörterbüchern erfolglos verliefen. Die Wörter wiederum, bei denen die Benutzer aber zu Recht davon ausgehen konnten, daß diese in einem Wörterbuch des angegebenen Umfangs vorhanden sind, wurden fortlaufend in die bestehenden Daten eingearbeitet.

Des weiteren waren wir an einer Verbesserung der Benutzeroberfläche unserer Produkte interessiert. Die Wörterbücher werden zwar auf CD-ROM unter einer Oberfläche vertrieben, die mit der Oberfläche von Web-Browsern nur entfernt etwas zu tun hat. Dennoch erhofften wir uns besonders von den erfolglosen Benutzungshandlungen, die auf „Fehler" in der Bedienung zurückzuführen waren, Aufschlüsse. Wir verfuhren dabei nach der Devise, daß ein „Fehler" in der Bedienung nicht ein Fehler des Benutzers ist, sondern eine Unzulänglichkeit der Benutzeroberfläche.[4]

Dafür schlossen wir implizit den folgenden Vertrag mit den potentiellen Benutzern. Der Zugriff auf die Daten der Wörterbücher deutsch-englisch (beide Richtungen) und deutsch-französisch (beide Richtungen) erfolgt kostenlos und ohne vorherige Anmeldung. Dafür erstellten wir ein anonymisiertes Benutzungsprotokoll und boten die Möglichkeit eines Feedback per E-Mail.[5]

Im folgenden werde ich darstellen, welche interessanten Ergebnisse die Auswertung der Benutzungsprotokolle ergab und welche Folgerungen wir daraus für die Gestaltung unserer Wörterbücher und Benutzeroberflächen gezogen haben.

2 Der Versuchsaufbau

Im März 1996 stellten wir drei zweisprachige Wörterbücher – deutsch-englisch , deutsch-französisch und französisch-deutsch – auf unserem WWW-Server (http://www.zeres.de/dict) zum „Nachschlagen" zur Verfügung.[6]

Die Lemmalisten umfaßten damals 50 000 Stichwörter (deutsch-englisch) bzw. 25 000 Stichwörter (deutsch-französisch, französisch-deutsch). Zu jedem Stichwort werden minimale grammatische Angaben gemacht. Gibt es mehrere Übersetzungsäquivalente zu einem Stichwort, die verschiedene Bedeutungen oder Gebrauchsweisen des durch das Stichwort

---

[4] Der Aufwand hat sich letztlich gelohnt. Wir haben für ein Produkt, das eine Reihe einsprachiger niederländischer Wörterbücher eines niederländischen Verlages mit unserer Benutzeroberfläche verbindet, den niederländischen Software Preis für die beste Sprachsoftware 1998 gewonnen.

[5] Diese Möglichkeit wurde freilich kaum genutzt. Die etwa 15 Kommentare, die in zwei Jahren eingingen, waren zu jeweils einem Drittel konstruktive Kritik, Fragen danach, ob und wie die Wörterbücher zu erwerben seien, und Schmähbriefe.

[6] Dort stehen diese Wörterbücher auch jetzt noch zur Verfügung.

repräsentierten sprachlichen Zeichens wiedergeben, dann werden diese Bedeutungen/ Gebrauchsweisen durch Glossen illustriert. Ein Eintragsbeispiel finden Sie in Abbildung 1.

abandon (m) # (combat) Aufgabe (f);
(biens) Preisgabe (f);
(de soi-même) Aufopferung (f);
(d'un enfant) Aussetzung (f);
(renoncement) Verzicht (m);

Abb. 1: Beispieleintrag französisch-deutsch

Das „Nachschlagen" selber besteht aus der Eingabe eines Suchwortes und der Auswahl des Wörterbuches, in dem gesucht werden soll. Dies geschieht über ein Eingabeformular.

Jeder Suchvorgang wurde von unserem Server protokolliert. Dabei wurden der Server, von dem die Anfrage kam, das Wörterbuch, in dem gesucht wurde und das eingegebene Stichwort in ein „Logbuch" eingetragen. Es wurde ebenfalls vermerkt, wann der Anfragende eine Fehlermeldung erhielt. Diese Fehlermeldungen hatten verschiedene Ursachen, auf die unten noch ausführlich eingegangen wird. Abbildung 2 zeigt Teile eines Logbucheintrags.

```
Remote_Addr: 134.106.80.3
.....
TIME: 31. JAN 1998 um 16:10:22
Request: Eng-dts.idx diabetes
RESULT:diabetes#n#null#Zuckerkrankheit#f#null
.....
```

Abb. 2: Logbucheintrag einer erfolgreichen Suche im englisch-deutschen Wörterbuch

Die Benutzer hatten außerdem die Möglichkeit, eine „Benutzeranleitung" aufzurufen und der Wörterbuchredaktion einen elektronischen Brief zu schicken.

Die Untersuchung wurde in zwei Phasen aufgeteilt: Die erste Phase endete im Juli 1997, umfaßte also 15 Monate, in denen der Server bis auf wenige Tage Ausfallzeit permanent zur Verfügung stand. Die zweite Phase begann im Januar 1998 und umfaßte 7,5 Monate.[7]

Nach der ersten Phase wurden die Benutzeroberfläche und das darunterliegende Suchprogramm modifiziert. Die Änderungen an der Benutzerschnittstelle werden weiter unten ausführlicher dargestellt. Außerdem wurden aktualisierte Versionen der oben genannten drei Wörterbücher eingespielt (das deutsch-englische Wörterbuch umfaßt nun 60 000 Einträge, die beiden deutsch-französischen jeweils 50 000 Einträge) sowie ein 50 000 Einträge umfassendes englisch-deutsches Wörterbuch erstmals zur Verfügung gestellt.

[7] Es war zunächst geplant, die Ergebnisse im März 1998 vorzutragen. Dieser Vortrag kam nicht zustande. Die Aufteilung der Untersuchungsphasen wurde dennoch so belassen, da der Vortrag, auf dem dieser Aufsatz fußt, im September 1998 gehalten wurde und deshalb eine zweite 15monatige Untersuchungsphase nicht möglich war.

3 Auswertung

3.1 Quantitative Auswertung

Die quantitative Auswertung für beide Phasen der Untersuchung ist in Tabelle 1 dargestellt. Diese Tabelle läßt auf den ersten Blick die folgenden bemerkenswerten Tendenzen erkennen.

	Phase 1 (Apr. 96 – Juli 97)	**Phase 2 (Jan. 98 – Aug. 98)**
Zugriffe gesamt	23 175	126 655
dts-eng	5 782	8 376
eng-dts	11 646	14 289
dts-frn	-	5 839
frn-dts	5 470	97 799
erfolgreiche Zugriffe	8 730	57 715
erfolglose Zugriffe	14 445	68 940

Tabelle 1: quantitative Auswertung der beiden Untersuchungsphasen

Wir beobachteten zwei interessante Trends: 1. Die Zahl der fehlgeschlagenen Suchen ist erstaunlich hoch, vor allem in der ersten Phase der Untersuchung.[8] 2. Die Suchanfragen richten sich in der zweiten Phase verstärkt an die deutsch-französischen Wörterbücher.

Der zweite Trend dürfte recht einfach zu erklären sein. Es steht meines Wissens bis heute kein weiteres deutsch-französisches oder französisch-deutsches Wörterbuch zur Verfügung. Die Zunahme der Suchanfrage ist also sicherlich auf die Zunahme des Bekanntheitsgrades dieser – immerhin kostenlosen – Informationsquelle zurückzuführen. Außerdem sind anhand des Profils der zugreifenden Server „Stammkunden" auszumachen. Im Bereich deutsch-englischer Wörterbücher hingegen gibt es ausreichend gute und ebenfalls kostenlose Informationsangebote.

Der erste Trend – die enorme Zahl an erfolglosen Suchen – hatte eine genauere Analyse verdient. Es wurde ein Sample von 500 erfolglosen Suchen ausgewertet. Die Ergebnisse nach Kategorien:

- Das Suchwort wurde auf irgendeine Weise falsch geschrieben: 233 Fälle
- Das Suchwort fehlte im entsprechenden Wörterbuch : 164 Fälle
- Es gab Probleme mit der Auswahl der Grundform / des Lemmas: 54 Fälle
- Es wurde das falsche Wörterbuch ausgewählt: 34 Fälle

Die restlichen erfolglosen Suchen waren keiner dieser Kategorien zuzuordnen.

Diese Ergebnisse der ersten Analyse gingen in die Neugestaltung der Oberfläche wie folgt ein.

---

[8] Daß die Wörterbücher trotz der hohen Zahl erfolgloser Suchen kontinuierlich weiter benutzt wurde, stärkt die Vermutung von Püschel, daß sich Wörterbuchbenutzer mit dem Gegebenen arrangieren, vgl. Püschel 1989, S. 130

- Es wurde klarer hervorgehoben, daß vor Eingabe des Suchbegriffs zunächst das richtige Wörterbuch auszuwählen ist.
- Die Suchmaschine wurde fehlertoleranter gestaltet. Groß- und Kleinbuchstaben wurden auf ein gemeinsames Symbol abgebildet.
- Es wurden ausführlichere Benutzungshinweise verfaßt, und zwar in deutsch, englisch und französisch. In diesen Benutzerhinweisen wird auf die Verwendung von HTML-Entities für Umlaute und diakritische Zeichen, auf die Möglichkeit der Suche mit Wildcards und auf einige spezielle Formen der Lemmatisierung hingewiesen.
- In den Benutzungshinweisen wird auf einen Terminologieserver hingewiesen. Dieser kann in den Fällen angesteuert werden, wo das Suchwort zu speziell für ein Wörterbuch des gegebenen Umfangs ist.

Die Änderungen an der Suchmaschine mußten gering ausfallen. Parallel zu der auch zur Zeit noch verfügbaren Suchmaschine wurde eine radikal neue, datenbankbasierte Suche entwickelt. Diese Oberfläche löst einige der zeichenbasierten Probleme, war aber zur Zeit der Abfassung dieses Artikels noch in Arbeit.

Dennoch haben sich diese Änderungen gelohnt. Die Fehleingaben machen in der zweiten Untersuchungsphase nur noch etwas weniger als 1 Prozent der erfolglosen Suchen aus.

3.2 Qualitative Auswertung

Für die qualitative Auswertung beschränke ich mich auf die deutschen Suchanfragen an das deutsch-englische und das deutsch-französische Wörterbuch. Es werden hier alle Suchanfragen einbezogen, die sinnvolle und orthographisch korrekte Suchwörter bilden, unabhängig davon, ob zu diesen Suchausdrücken im ausgewählten Wörterbuch ein Eintrag vorhanden war.

Wir haben zunächst die Liste der am häufigsten angefragten Einträge in Sach- oder Diskursbereiche eingeteilt.

Mit jedem Diskursbereich wird eine Handlungshypothese verbunden, die in Klammern hinter dem Diskursbereich angegeben wird.

3.2.1 Alltagswelt (Handlungen: Kommunikation in der anderen Sprache; Schließen von Lemmalücken)

Buch (84), *Liebe* (75), *Auto* (54), *Hund* (43), *Freund* (28), *Tag* (31), *Maus* (25), *Frau* (23), *Regen* (22), *Blume* (20), *Katze* (17), *Mann* (16), *Stuhl* (12), *Herr* (9); *gehen* (108), *essen* (96), *suchen* (40), *wollen* (33), *leben* (26), *schlafen* (17), *lieben* (16)

In Klammern die Anzahl der Suchen nach dem jeweiligen Wort. Diese Wörter sind problemlos und alle in den Wörterbüchern aufzufinden.

3.2.2 Sexual- und Fäkalwortschatz (Handlung: Wörterbuch testen)

Scheiße (40), *ficken* (35), *Schwanz* (23), *Arsch* (23), *Arschloch* (15), *Sex* (15), *vögeln, bumsen* (6)

Die Suche nach diesen Wörtern gehört sicher zu einer typischen Testsituation. Sie sollten in den Wörterbüchern vorhanden sein und dort mit der gebotenen Sensibilität beschrieben werden.

3.2.3 Arbeitswelt (Handlung: geschäftlich/beruflich korrespondieren)

Buch (84), *Dienst* (38), *Zeit* (33), *Arbeit* (23), *eignen* (22), *senden, arbeiten* (beide 20), *erstellen* (17), *schreiben, Ort, Name* (alle 16), *Adresse* (14), *Schule, Praktikum, Auftrag* (alle 13), *bewerben* (10)

Es kann davon ausgegangen werden, daß die meisten tatsächlichen Benutzer dieses Wörterbuchs dieses an ihrem Studien- oder Arbeitsplatz tun. Entsprechend häufig gefragt und wichtig sind die Termini der Arbeitswelt, insbesondere das Vokabular für Bewerbungen. Es ist unbedingt ratsam, auch schwierige Terme ohne echtes Äquivalent, wie *Vordiplom* oder *Abitur*, aufzunehmen und für diese Terme Gebrauchsäquivalente anzugeben, also entweder eine ungefähre Entsprechung des Sachverhaltes im Land der Zielsprache oder eine Umschreibung.

3.2.4 Benutzeranleitung (Handlung: die Benutzeranleitung verstehen)

Haus (232), *Buch* (84), *suchen* (40), *Wörterbuch* (15), *Suchausdruck* (6)

Diese Wörter fanden sich in den minimalen, ausschließlich deutschen, Benutzungshinweisen. *Haus* wurde dort als Beispiel angegeben und stand somit unmittelbar als Testwort zur Verfügung. Die anderen Suchwörter zeigten ein Bedürfnis nach Benutzungsanleitungen in den Zielsprachen an. Diesem Bedürfnis sind wir vor der zweiten Phase entgegengekommen.

Eine weitere interessante Sicht auf die Daten ergab sich durch die Einteilung aller sinnvollen (und im Wörterbuch belegten) Suchausdrücke nach ihren Wortarten.

Substantive sind im Verhältnis zu ihrem Anteil an den Wörterbuchlemmata (ca. 75%) bei den Suchanfragen unterrepräsentiert (ca. 55%). Hingegen sind Adverbien, und hier vor allem die Zeitadverbien, und Gesprächspartikel / Interjektionen überrepräsentiert. Die Vermutung liegt nahe, daß es bei den Zeitadverbien um deren korrekte Verwendung in Bewerbungsschreiben und ähnlichen Texten handelt. Bei den Gesprächspartikeln sind die Einwortfloskeln (*hallo, bitte, danke*) „gefragt".

3.3 Lexikalische Lücken

Eines der wichtigen praktischen Ziele der hier dargestellten Untersuchung war es, lexikalische Lücken in den angebotenen Wörterbüchern aufzuspüren. Lexikalische Lücken taten, sich, wie erwartet, dort auf, wo das von Benutzern Gewünschte vom Wortschatz unserer Datenbasis abweicht: Genannt seien hier beispielhaft der Sexualwortschatz, Floskeln und andere feste Redeformeln (Grüße, Glückwünsche) und einige Bereiche der Studien- und Arbeitswelt.

Dennoch wurde in den Benutzerhinweisen auch auf einen anderen mehrsprachigen Terminologieserver als alternative Informationsquelle hingewiesen.

4 Zusammenfassung und Ausblick

Es handelt sich bei diesen Beitrag um eine explorative Studie über die Möglichkeiten, mit Hilfe eines neuen Mediums neue Einsichten und Erkenntnisse zu einer schwierigen, aber wichtigen Frage der Wörterbuchforschung zu liefern. Die präsentierten Ergebnisse sind beschränkt, was zum einen an der Art des Vertrages mit den Benutzern, zum zweiten an der praktischen Ausrichtung der Fragestellung und zum dritten an der flachen Struktur der Wörterbuchdaten liegt.

Der Ansatz scheint mir jedoch gerade für kommerzielle Wörterbuchverlage, wenn Sie denn Interesse an der Verbesserung ihrer Produkte haben oder durch eine Konkurrenzsituation dazu gezwungen sind, vielversprechend zu sein, wenn die folgenden Rahmenbedingungen verbessert werden:

- Vertrag mit dem Benutzer: der Anbieter der Daten kann die Anmeldung des Benutzers verlangen oder dem Benutzer „Cookies" senden. Die zweite Methode halte ich für moralisch etwas fragwürdig. Wenn die Daten zumindest teilweise re-individualisiert werden können, dann können auch genauere Aussagen über komplexere Handlungsmuster als Folge erfolgloser Suche genauer bestimmt werden (aufgeben, Suchwort ändern, andere Suchstrategie wählen etc.)
- Struktur der Wörterbuchdaten: wenn Wörterbuchartikel mit einem komplexeren Informationsprogramm angeboten werden, dann läßt sich die Anwendung so einrichten, daß bestimmte, zusammenhängende Informationsteile (Formkommentar, grammatische Angaben, semantischer Kommentar, Angaben zur Etymologie) nur nach einer weiteren Aktion des Benutzers präsentiert werden. Auf diese Weise läßt sich auch der Bedarf nach bestimmten Informationstypen quantifizieren.

Eng damit verbunden ist der Aspekt der Bereitstellung und Präsentation der Daten im World-Wide Web. Der momentan verwendete Publikationsstandard HTML fördert die Tendenz, relativ große Informationseinheiten vorzuhalten und auf Abruf zu präsentieren (meistens ganze Wörterbuchartikel). Der kommende Publikationsstandard XML, der die Möglichkeit vorsieht, kleinere Informationseinheiten zu modellieren, diese zu verknüpfen und erst auf ausdrückliche Anforderung der Benutzer anzuzeigen, bietet auch für die Benutzungsforschung wesentlich bessere Möglichkeiten. Ein Wörterbuchartikel kann in funktionale Textsegmente zerlegt und diese können miteinander verknüpft werden. Der Benutzer erhält bei einer Abfrage zunächst ein Minimum an Informationen und kann weitergehende Informationen anfordern. Es läßt sich somit bis auf die Ebene einzelner Informationseinheiten nachvollziehen, welcher Anteil an Benutzern diese Information tatsächlich „nachfragt". (Zu XML vergleiche St. Laurent 1998, zur Informationsmodellierung Lobin 2000).

5 Literatur

Hartmann, R. R. K. (1987): Four Perspectives on Dictionary Use: A Crititical Review of Research Methods. In: A. P. Cowie (Ed.): The Dictionary and the Language Learner. Papers from the EURALEX Seminar 1985. Tübingen 1987, S. 11–28.

– (1989): Sociology of the Dictionary User: Hypotheses and empirical studies. In: Wörterbücher – Dictionaries – Dictionnaires. Ein internationales Handbuch zur Lexikographie. Berlin / New York 1989 (= HSK Band 5.1), S. 102–111.

Kühn, Peter (1989): Typologie der Wörterbücher nach Benutzungsmöglichkeiten. In: Wörterbücher – Dictionaries – Dictionnaires. Ein internationales Handbuch zur Lexikographie. Berlin/New York 1989 (= HSK Band 5.1), S. 111–127.

Lobin, Henning (2000): Informationsmodellierung in XML und SGML. Berlin et al. 2000.

Püschel, Ulrich (1989): Wörterbücher und Laienbenutzung. In: Wörterbücher – Dictionaries – Dictionnaires. Ein internationales Handbuch zur Lexikographie. Berlin/New York 1989 (= HSK Band 5.1), S. 128–133.

St. Laurent, Simon (1998): XML. A Primer. Foster City 1998.

Wiegand, Herbert Ernst (1987): Zur handlungstheoretischen Grundlegung der Wörterbuchbenutzungsforschung. In: Lexicographica 3. Tübingen 1987, S. 178–227.

– (1998): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. Berlin / New York 1998.

Lothar Lemnitzer, Tübingen

Abstracts

Gregor Büchel, Bernhard Schröder

The choice of coding systems for complex text structures depends on numerous relevant criteria. In most cases the choice of an SGML or XML based coding system will meet these criteria. The data can be managed with the help of database management systems. For this purpose the text objects and relations have to be mapped onto the expressive means of the database type chosen. It depends strongly on the way the documents are structured and on the way the data will be used, which kind of database modelling is most adequate for retrieval and maintenance purposes.

Ingrid Schmidt, Carolin Müller

Dictionaries that are currently available in electronic form only exploit the potential offered by the medium in a very restricted way. It is clear the high quality of dictionaries in printed form is often lost in the electronic version. Since the requirements of the media landscape are constantly shifting, it is necessary to develop a new starting point for dealing with lexicographic content. To this end, we first explore various aspects of SGML, multiple media publishing, TEI, and meta lexicography. We then use this as a basis for discussing a fresh attempt to develop a durable lexicographic model that will outlive rapidly outdated media and with which we are thus in a position to react flexibly to market requirements.

Angelika Storrer

The paper examines the new prospects that hypertext technology opens up for creating innovative electronic dictionaries. It first discusses the ideas which are central to the concept of hypertext: The non-sequential storage and presentation of the data, the integration of text, sound, images and video, and the tools for searching and filtering lexical information that are provided by hypertext management systems. It will be argued that most existing electronic dictionaries are more or less digital copies of paper dictionaries which only utilize a small portion of the technical possibilities available.

For this reason, the main part of the paper will establish seven theses and discuss ways of using the value-adding features of hypertext technology in future electronic dictionaries more efficiently.

Ingrid Lemberg

The Internet delivers a multifunctional medium for publishing, production and communication. The essential components of Internet-dictionaries are their global and around-the-clock availability (Chapter 4.1), doing away with a limited amount of printed pages, the possibility for continual corrections and updates of dictionary articles that have already been published (Chapter 2.2), hypertextualization and the use of multimedia integration.

Within this new medium, dictionaries can be adapted and expanded using other forms of presentation, for example by creating direct links between the dictionary and the text-corpus that it is based upon or between other informational systems (Chapter 2.1.1). Using multimedia integration is considered to be particularly suitable for the presentation of lexical or lexicographic knowledge with different types of expressive mediums thus contributing to an efficient conveyance of knowledge (Chapter 2.1.2).

Internet availability also means that lexicography has a totally new production technology at its disposal: making lexicographic databases available on the Internet makes it possible to create homogenous and completely networked dictionaries from different locations spread out all over the world (Chapter 3). Being a medium for communication, the Internet makes it possible to establish direct contact between the user and the lexicographer: users of an Internet-dictionary can request comprehension support via e-mail, depending on the type of situation that users find themselves in. In addition, users can also make valuable and developmental contributions to the informational status of the dictionary (Chapter 4.2).

Annette Klosa

This article starts from a fundamental definition of "quality" and discusses which quality characteristics for CD-ROM dictionaries are mentioned in reviews. On this basis a list of quality criteria for electronic dictionaries is developed, which, finally, is compared with current practice in publishing houses.

Ulrike Haß-Zumkehr

This article contains reflections on the design of an abstract microstructure for a hypertext information system for the German language, which results from a lexicographical project at the Institut für Deutsche Sprache, Mannheim. The new medium confronts lexicographers with different conditions and consequences. Among these are the separation of data structure and data presentation, the creation of several ways of accessing the entry via various dimensions of information (semantics, grammar etc.), the context-free formulation of information elements, the removal of most of the text compression, the explicit naming of information elements to help the user select what he wants to know, the creation of a typology of links and the decision between a structure pivoting on the word form or on the single word meaning.

Thomas Gloning, Rüdiger Welter

Documenting a vocabulary as an electronic data base offers some considerable benefits as compared with the one-dimensionality of alphabetically arranged works of reference: it is a more adequate representation of the complex interior structure of a vocabulary concerning questions of onomasiology, text types, style(s), the history of ideas and concepts etc.; it provides ways of data retrieval that are differentiated and flexible in selection and combination; it is continuously open to amendments and revisions.

The article starts by explaining how the complex structure of a vocabulary (parts 1 and 2), in particular the interior structure of Goethe's vocabulary (part 3), may be characterized. Part 4 discusses some of the new possibilities the Goethe Dictionary as a SGML-coded data base offers its users, e.g. multiple use of the data, improved methods of analysis, and links to other related documents. It is argued that on the basis of these data the Goethe-Wörterbuch should be made available free of charge via the internet.

Thomas Burch, Johannes Fournier

The most important dictionaries of Middle High German presently are the ‚Mittelhochdeutsches Wörterbuch' by Benecke/Müller/Zarncke, the ‚Mittelhochdeutsches Handwörterbuch' by Lexer, its ‚Nachträge' and the ‚Findebuch zum Mittelhochdeutschen Wortschatz'. These dictionaries are closely interconnected and have to be used simultaneously. They therefore are ideal candidates for the composition of an electronically interlinked dictionary. Additionally the lexicographical information within the entire contents of all four works of reference will be accessible via a database supplied with complex retrieval options. The application of SGML, defined as ISO standard, provides the ground for an encoding which does not depend on specific hardware or software, neither for the CD-ROM nor for the Internet. Moreover this markup will guarantee the longevity of the data concerned. The application of TEI Guidelines which have already been successfully employed in a variety of projects was essential to the steady progression of the retrospective digitisation.

The essay discusses problems that occurred when applying TEI Guidelines to the electronic dictionaries. It is however apparent that problems did not stem from the application of TEI Guidelines as such, but instead were primarily due to the fact that the articles of BMZ and Lexer's dictionary do not always follow a well defined overall structure, which has made automatic SGML encoding an often difficult task. In many cases only manual markup led to TEI compliant documents. Nevertheless the results achieved so far fully justify the decision in favour of TEI Guidelines.

Ralf Plate, Ute Recker

This essay discusses current possibilities and future prospects of the computer-aided composition of a historical citation dictionary (*Belegwörterbuch*), as they present themselves to the authors after five years of preliminary work towards a new Middle High German dictionary. The phase of elaboration of the dictionary, that is to say, the writing of entries, will in all probability begin in 2001. A period of twenty years will see the completion of the dictionary, the printed version consisting of four volumes, containing approximately 1000 pages each. The new dictionary is meant to assume the role hitherto fulfiled by its predecessors dating from the 19th Century, for the language of the literary sources from the period 1050 to 1350. Being the most recent of current projects concerning historical citation lexicography, it was possible to base all of the work on the new Middle High German dictionary entirely on electronic data processing from the very beginning. The compilation of data for the new dictionary is based on an comprehensive electronic archive containing all texts of the basic corpus in full, as well as a list of potential headwords (*Lemmakandidatenliste*) consisting of app.ly 80,000 lemmata, obtained from the

preceding dictionaries. After preparing the digitized text sources with specific markings, the lemmatized archive of citational evidence is then produced by employing a system of programmes devised for semi-automated lemmatization, drawing informational data from the texts. Additionally, the programming also inputs to the archive all word forms as derived independently from the texts by way of an especially devised lemmatization component. Thus the electronic archive of citational evidence will serve as the basis for the future dictionary, complete with an additional system of programmes which will support the composition of entries, from the coordination and presentation of citational data through to the stage of typesetting the finished entries.

By entirely basing the new dictionary on data processing, it will be possible to publish an electronic version alongside the printed edition of the dictionary, which will appear periodically in separate issues/installments. With little extra preparation the electronic publication will furthermore provide access to the digitized text material itself as well as offer a wide range of further data retrieval opportunities.

The future prospects of computer-aided lexicography as discussed in the contribution with regard to the new dictionary of Middle High German refer to the chance of improving traditional lexicographical research strategies by employing computers for the tasks of reorganizing and economizing the process of information retrieval. The greater part of former lexicographical endeavor used to consist of the copying, editing, and lemmatizing of individual pieces of citational evidence. When employing the aid of computers and semi-automated lemmatization software, the resources previously invested in this time-consuming process of compiling data can be used for the production of once-and-for-all carefully edited digitized texts which will be electronically processed for all further applications. Proceeding in this way means that the author of dictionary entries him/herself is retains full control over the choice of citational evidence, rather than being limited to the result of an earlier excerpt, and can thus make his/her selection on lexicographical grounds.

The composition of the entries themselves is supported by a system of programmes designed to perform most of the mechanical tasks such as the sorting, writing and controlling of the data concerned, thereby enabling the lexicographer to focus attention entirely on the analysis of the citational evidence and the formulation of lexicographical information.

Additionally, due to the obligatory formatting routines devised in Trier, the entry files produced will be supplied with a certain (minimal) amount of encoding as it relates to the overall structure of the electronic dictionary, thus allowing the later realization of sophisticated lexicographic information retrieval strategies.

Other prospects for computer-based lexicography can be discerned in the fact that much of the data which is compiled in order to produce the dictionary will not only be employed and made accessible in combination with the published work but can also be made accessible for independent use in support of other linguistic and philological research interests. In the case of the new Middle High German dictionary, additional advantages take the form of an electronic textarchive, the electronic archive of citational evidence, and, finally, an efficient instrument in the semi-automated production of lemmatized indices and concordances of any given Middle High German text.

The prospects discussed in the contribution refer to the particular advantages of publishing electronic dictionaries as well as to questions pertaining to lexicographical concepts. These questions are basically concerned with the nature of the relationship between the electronic publication of the dictionary and the data material it is based on (particularly the text- and citation archive), should they be published in combination and thus present the opportunity of proceeding directly from the article into the archive. We have also addressed

the question of whether and to what extent the dictionary should – in addition to its essential functions – be designed to meet the demands of new classes of users for whose purposes various kinds of structured listings or registers might be required.

Gerd Richter[1]

Electronic forms of presentation such as hypertext provide new opportunities for the design of toponymic repertories in the fields of presentation, retrieval and analysis. Based on the printed version of the *Südhessisches Flurnamenbuch* (SHFLNB), this article shows the advantages brought about by the transformation of the traditional book version into a hypertext. A model analysis of the SHFLNB which includes an analytical description of its medial structure will show the hypertextual potential of a toponymic repertory. An example realisation as a hypertext will then serve as a basis for a thorough description of the potential new modes of presentation for toponymic repertories, especially in comparison with the traditional print version.

Krzysztof Petelenz

The availability of reference works on CD-ROM has led to the practical and commercial significance of hypertext for lexicography. The development of the WWW has lead to improved qualities for linked information; open systems, topical contents, easier communication between the authors and their audience and, consequently, an improved consideration of user needs.

It therefore appears appropriate to consider ways of improving on-line dictionaries. This paper attempts to lay the foundations for the presentation of a bilingual dictionary in high-quality hypertext. The structuring and storage strategies for dictionary data will be dealt with in particular and described using the example of a section of a Polish-German dictionary. First, however, we will discuss the position of bilingual computer-dictionaries within hypermedial systems and special features of bilingual lexicography which are significant for the structuring of bilingual on-line dictionaries. Particular attention will be paid to use and categorisation of pictures which can finally be given adequate space in this new medium. In conclusion, some "added value" features of an on-line dictionary will be discussed, such as the networking with other web-projects and the continuous expansion of the macro- and microstructure.

Claudia Kunze, Andreas Wagner

This paper presents GermaNet, a lexical-semantic network and on-line thesaurus which covers the German basic vocabulary. GermaNet encodes the basic semantic relations like synonymy, hyponomy, antonymy, and the causal relation that hold among the lexical items being implemented. This semantic network constitutes an important resource for word

[1] Für die Übersetzungen danke ich Frau Sabine Prechter. – G. R.

sense disambiguation which is a prerequisite for various applications within natural language processing like information retrieval and the semantic annotation of corpora.

Lothar Lemnitzer

Sound knowledge of user needs is crucial for the planning and realization of new and revised dictionaries. However, such data are costly to acquire. Nowadays the World Wide Web enables a new kind of contract with users. They are allowed to use dictionary data free of charge but in turn agree that their requests are logged and analysed. The present study presents such an approach with a set of bilingual dictionaries which are available on the net. User requests have been logged for more than a year. A quantitative and qualitative analysis of these data is presented.

Résumés

Gregor Büchel, Bernhard Schröder

Le choix d'un système de codage pour les structures complexes de texte dépend de nombreux critères. Dans la plupart des cas le choix d'un système de codage basé sur SGML or XML répond à ces critères. Les données peuvent être contrôlées avec l'aide des systèmes de gestion de base de données. A cette fin les objectes de texte et leurs relations doivent être représentées par les moyens expressifs du type de base de données choisi. Il dépend de la manière dans laquelle les documents sont structurés et qu'on utilisera les données quel model de base de données est le plus adéquat pour la recherche et la gestion.

Ingrid Schmidt, Carolin Müller

Les dictionnaires électroniques actuellement disponibles sur le marché n'utilisent qu'une partie très limitée de leur potentiel. Il est notamment frappant de constater que les versions électroniques de bons dictionnaires imprimés laissent souvent à désirer sur le plan de la qualité. Si l'on observe les exigences sans cesse changeantes du paysage médiatique, on remarque la nécessité de nouvelles bases dans les rapports avec les contenus lexicographiques. C'est avec cet objectif en tête que nous nous penchons d'abord sur certains aspects des thèmes suivants: SGML, Multiple-Media-Publishing, TEI, métalexicographie. Nous débattons ensuite sur cette base de l'élaboration d'un modèle lexicographique opposant la longévité à la rapidité des médias et avec lequel il doit être justement possible de réagir avec flexibilité aux exigences du marché.

Angelika Storrer

L'article rend compte des perspectives que la technologie hypertexte ouvre à l'élaboration de dictionnaires électroniques innovateurs.

Il explique d'abord les caractéristiques essentielles de l'idée de hypertexte: la liaison des unités d'information par des liens hypertextes, l'integration de texte, images, son et vidéo dans le même logiciel, l'accès flexible aux données lexicaux. Il est démontré que la plupart des dictionnaires électroniques en vente aujourd'hui n'utilisent pas encore le véritable potentiel de la technologie hypertexte.

Pour cette raison, la partie principale de l'article présente sept thèses et discute les voies d'une meilleure utilisation de la plus-value de la technologie hypertexte dans les dictionnaires électroniques à l'avenir.

Ingrid Lemberg

Internet nous offre un moyen multifonctionnel de publication, de production et de communication. Les dictionnaires sur Internet possèdent les avantages suivants: liberté d'utilisation à partir de n'importe quel lieu et pendant une durée illimitée (chap. 4.1), possibilité de

correction et de mise à jour continuelles d'éléments du lexique déjà publiés (chap. 2.2), hypertextualisation et utilisation multimedia. Il est possible, grâce à ce nouveau moyen, de présenter les dictionnaires sous d'autres formes et de les élargir en créant par exemple des liens entre le dictionnaire et les textes de référence ou d'autres systèmes d'information (chap. 2.1.1). L'utilisation multimedia est considérée comme particulièrement appropriée pour présenter des connaissances lexicales ou lexicographiques avec divers moyens d'expression et contribuer par conséquent à une communication efficace du savoir (chap. 2.1.2). Internet fournit également à la lexicographie une toute nouvelle technique de production: l'accès à des banques de données lexicographiques sur Internet permet la création de dictionnaires en réseau, complets et homogènes, à partir de sites répartis dans le monde entier (chap. 3). Internet est enfin un moyen de communication qui permet le contact entre l'utilisateur et le lexicographe : les utilisateurs d'un dictionnaire sur Internet peuvent demander par E-Mail des compléments d'explication tout comme ils peuvent eux-mêmes apporter des renseignements supplémentaires et précieux au dictionnaire (chap. 4.2).

Annette Klosa

L'auteur de l'article entend d'abord donner une définition fondamentale de ce qu'est „la qualité" d'un dictionnaire sur CD-ROM pour discuter ensuite les critères qualitatives énoncés par les critiques de dictionnaires électroniques. Cette comparaison permet ensuite à l'auteur de développer ses propres critères de qualité pour la production d'un dictionnaire électronique avant de les confronter à la réalité des conditions de production éditoriale.

Ulrike Haß-Zumkehr

L'article contient des réflexions sur l'esquisse d'un programme concernant les microstructures d'un système hypertextuel d'informations concernant le vocabulaire allemand contemporain. Ces réflexions sont ancrées sur un projet concret qui est en train d'être réalisé à L'Institut für Deutsche Sprache à Mannheim. Ce nouvel outil de travail met les lexicographes dans une situation et devant des conséquences nouvelles, comme par exemple devant la nécessité de séparer structuration et présentation des données, de créer plusieurs accès à l'article de dictionnaire à partir d'informations différentes mais de valeur égale (concernant la grammaire, la sémantique etc.), de formuler les informations hors contexte, de changer la structure souvent trop dense des textes explicatifs, de nommer explicitement les sortes d'informations afin de faciliter leur accès pour l'utilisateur. Ce nouvel outil nécessite également l'élaboration d'une typologie des renvois et des chaines d'information ainsi qu'un choix entre une structure d'article basée sur la forme du mot ou basée sur ses emplois.

Thomas Gloning, Rüdiger Welter

La documentation d'un lexique sous forme de base de données offre plusieurs avantages par rapport à l'uni-dimensionalité des dictionnaires structurés de manière alphabétique: elle permet de saisir le vocabulaire de divers points de vue structurels et analytiques (p. ex.

réseaux onomasiologiques, différentes matières, types de textes, diachronie et évolution biographique, rendement fonctionnel, histoire des idées, etc.) tout en ouvrant à l'utilisateur des possibilités subtiles d'accès et de consultation suivant des critères combinés. En outre, les données électroniques peuvent, à tout moment, être corrigées et mises à jour.

Notre contribution explique d'abord les principes de l'architecture complexe et certains aspects fondamentaux de la structure lexicale (§§ 1 et 2) avant de fournir des précisions sur la structure interne et les particularités du vocabulaire de Goethe (§ 3). La partie principale (§ 4) traite des possibilités d'accès étendues qu'offre le dictionnaire de Goethe en tant que base de données SGML traitée à l'aide de marquages lexicologiques, p. ex.: usages variés des données de base, multiples possibilités d'analyse et de recherche en faveur des utilisateurs, mise en réseau avec d'autres documents. Importante du point de vue de la politique culturelle sera la publication gratuite des données sur la toile mondiale.

Thomas Burch, Johannes Fournier

Les dictionnaires du moyen haut allemand les plus significatifs sont actuellement le «Mittelhochdeutsches Wörterbuch» de Benecke/Müller/Zarncke, le «Mittelhochdeutsches Handwörterbuch» de Lexer, ses «Nachträge» et enfin le «Findebuch zum Mittelhochdeutschen Wortschatz». Ces dictionnaires sont étroitement reliés entre eux et doivent pour cela être utilisés simultanément. C'est pourquoi ils sont des candidats idéaux pour la composition d'un dictionnaire relié électroniquement. De plus, les informations lexicographiques contenues dans l'ensemble des quatre ouvrages de référence seront accessibles dans une banque de données pourvue d'options de recherche complexes. L'application du SGML, défini comme un standard ISO, fournit la base pour un encodage qui ne dépend pas d'un logiciel ou d'un matériel spécifique, que ce soit pour un CD-ROM ou pour Internet. En outre, cette codification garantira la longévité des données concernées. L'application des lignes directrices TEI qui ont déjà été utilisées avec succès dans divers projets s'est avérée essentielle pour la progression constante de la numérisation rétrospective.

La dissertation traite des problèmes qui surviennent lors de l'application des lignes directrices TEI des dictionnaires électroniques. Toutefois, il est évident que les problèmes ne proviennent pas de l'application des lignes directrices TEI en tant que telle, ils résultent plutôt principalement du fait que les articles du BMZ et du dictionnaire de Lexer ne suivent pas toujours une structure générale bien définie, ce qui fait de l'encodage automatique du SGML une tâche souvent difficile. Dans des nombreux cas, seulement l'encodage manuel permet d'obtenir des documents TEI maniables. Néanmoins, les résultats obtenus jusqu'à présent justifient pleinement la décision en faveur des lignes directrices TEI.

Ralf Plate, Ute Recker

L'article discute les chances et les perspectives de la rédaction d'un dictionnaire de référence (Belegwörterbuch) historique soutenue par l'ordinateur ainsi qu'elles se représentent pour le nouveau Dictionnaire du Moyen Haut Allemand au bout de cinq ans de travail préparatoire. L'élaboration de ce dictionnaire dont on a prévu une dimension de quatre volumes (chacun à 1000 pages environ) commencera en l'année 2001 et devrait être terminée dans l'espace de vingt années. Le nouveau dictionnaire remplacera ses prédecesseurs du XIX^{e} siècle qui se refèrent aux sources de la période entre 1050 et 1350. Il sera basé dès

le début sur le traitement des données électroniques ce qui lui procure un caractère innovatif parmi les grands projets de la lexicographie historique en Allemagne. Le recueil des matériaux se fonde d'un part sur des archives électroniques qui contiennent les textes de toutes les sources du corpus de base et d'autre part d'un index d'environ 80.000 mots compilés à part des enregistrements dans les dictionnaires précédents. Sont produits ensuite des archives de référence lemmatisés à partir de deux composantes: d'un côté des textes préparés d'abord à l'aide d'un système informatique pour la lemmatisation semi-automatique, d'autre côté d'une liste qui comprend toutes les occurrences des textes déjà enregistrés. Le dictionnaire sera donc élaboré sur la base de ces deux archives et à l'aide d'autres systèmes informatiques qui soutiennent le travail à partir du premier arrangement de la référence jusqu'à la composition de l'article. Le dictionnaire ainsi achevé servira à une double fonction: À cause de son construction aux moyens électroniques, il sera utilisable non seulement dans sa forme traditionnelle comme livre imprimé mais encore comme instrument électronique qui offre des possibilités de recherche plus étendues.

Les chances de la lexicographie soutenue par l'ordinateur consistent en la possibilité – ici démontrée à l'exemple du nouveau Dictionnaire du Moyen Haut Allemand – de réorganiser et de rationnaliser avec l'assistance des moyens électroniques les opérations lexicographiques traditionnellement utilisées. Il est ainsi possible d'économiser une grande partie du travail jadis nécessaire pour copier, corriger et lemmatiser chaque référence particulière. Une grande avantage est représentée par le fait que la choix des références anciennement entreprise au cours de l'excerption rentre maintenant dans le domaine de l'auteur d'un article qui est instruit en lexicographie. Il peut se pencher sur l'interprétation du matériau tandis que le système informatique achève les principaux travaux nécessaires pour ranger, écrire et contrôler les articles. En plus, on trouve des informations sur la structure du dictionnaire directement dans l'arrangement du fichier de chaque article ce qui rend possible des recherches exactes dans la version électronique du dictionnaire. Au-delà de la préparation du dictionnaire même, il y existe une autre chance présentée par les résultats faits quasiment en passant qui pourront avoir du valeur pour d'autres projets linguistiques ou philologiques. Dans le cas du dictionnaire décrit ici, ce sont p.ex. les archives électroniques contenant les textes et les références ou le système électronique pour l'établissement des index et des concordances qui peut être utilisé pour n'importe quel texte en moyen haut allemand.

Les perspectives discutées dans l'article se refèrent surtout aux avantages de la publication électronique du dictionnaire et aux questions de la conception d'un tel dictionnaire. D'un part, ces questions concernent le rapport entre le dictionnaire et les matériaux de base au moment où ils sont publiés ensemble ce qui offre la possibilité d'un accès direct de l'article aux archives. D'autre part, la question se pose dans quelle mesure le dictionnaire devrait, apart de sa fonction usuelle, être accessible pour des prétentions ultérieures comme p.ex. la production des registres.

Gerd Richter[1]

Les formes électroniques de présentation, comme celle du hypertexte, offrent de nouvelles possibilités dans la présentation aussi bien que l'utilisation des répertoires toponymiques.

[1] Für die Übersetzungen danke ich Frau Sabine Prechter. – G.R.

Basé sur la version traditionnelle – imprimée – du *Südhessisches Flurnamenbuch* (SHFLNB), cet article va présenter les avantages et les potentiels supplémentaires de la version hypertexte face aux restreintes de la version traditionnelle. Une analyse modèle du SHFLNB qui inclura aussi une description détaillée de sa structure médiale servira comme point de départ pour une présentation du potentiel hypertextuel du genre. Finalement, une réalisation exemplaire du SHFLNB en tant que hypertexte montrera les possibles nouvelles formes de présentation offertes aux auteurs de répertoires toponymiques, surtout en comparaison avec les formes traditionnelles en tant que livre imprimé.

Krzysztof Petelenz

L'importance pratique et commerciale de la technologie hypertextuelle dans le domaine de la lexicographie est une conséquence de la vulgarisation d'ouvrages de référence sur CD-ROM. Le développement du World Wide Web, quant à lui, a contribué à améliorer la qualité des informations reliées par des liens ainsi que de promouvoir les systèmes ouverts, les contenus thématiques, une communication plus aisée entre les auteurs et leur public et, par conséquent, une meilleure prise en compte des besoins de l'utilisateur. Ainsi, il est utile de réfléchir sur les améliorations possibles dans les dictionnaires en ligne. Le présent article essaie de créer les bases pour la présentation d'un dictionnaire bilingue reposant sur un hypertexte de haute qualité. Les stratégies de la structuration et de l'encodage des données lexicographiques seront discutées en détail et illustrées à partir d'un échantillon d'un dictionnaire polonais-allemand. En préliminaire, nous situerons les dictionnaires électroniques bilingues dans le contexte des systèmes hypermédia et nous évoquerons les particularités de la lexicographie bilingue qui déterminent la conception de dictionnaires bilingues en ligne.

L'intégration d'images, auxquelles ce nouveau média permet enfin d'accorder l'espace qui leur est dû, ainsi que leur catégorisation feront l'objet d'une étude approfondie. Finalement, nous traiterons quelques caractéristiques du dictionnaire en ligne qui apportent une «survaleur», comme par exemple l'interconnectabilité avec d'autres projets Web et l'expansion suivie et de la macrostructure et de la microstructure.

Claudia Kunze, Andreas Wagner

Nous présentons dans cet article GermaNet, réseau lexico-semantique et thésaurus «en ligne», couvrant le vocabulaire allemand de base. GermaNet intègre les relations sémantiques élémentaires comme la synonymie, l'hyponymie, l'antonymie, ainsi que les rélations casuelles des éléments lexicaux traités. Ce réseau sémantique offre une ressource importante pour le traitment de la désambiguïsation des mots, pré-requis pour de nombreuses applications en traitment automatique des langues, comme la recherche d'information ou l'annotation sémantique de corpus.

Lothar Lemnitzer

Les connaissances des besoins de l'utilisateur sont vitales pour la réalisation de nouveaux dictionnaires et leurs mises à jour. Cependant, ces données sont très difficiles à acquérir.

Heureusement, le Oueb nous permet maintenant de mettre en place un nouveau partenariat avec l'utilisateur. Celui-ci peut utiliser des dictionnaires sans frais, la contre-partie étant un enregistrement et une analyse de sa recherche. Cet article présente une telle utilisation avec un ensemble de dictionnaires bilingues consultables sur Internet. Les résultats présentés illustrent les avantages d'une telle recherche.

Register