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.

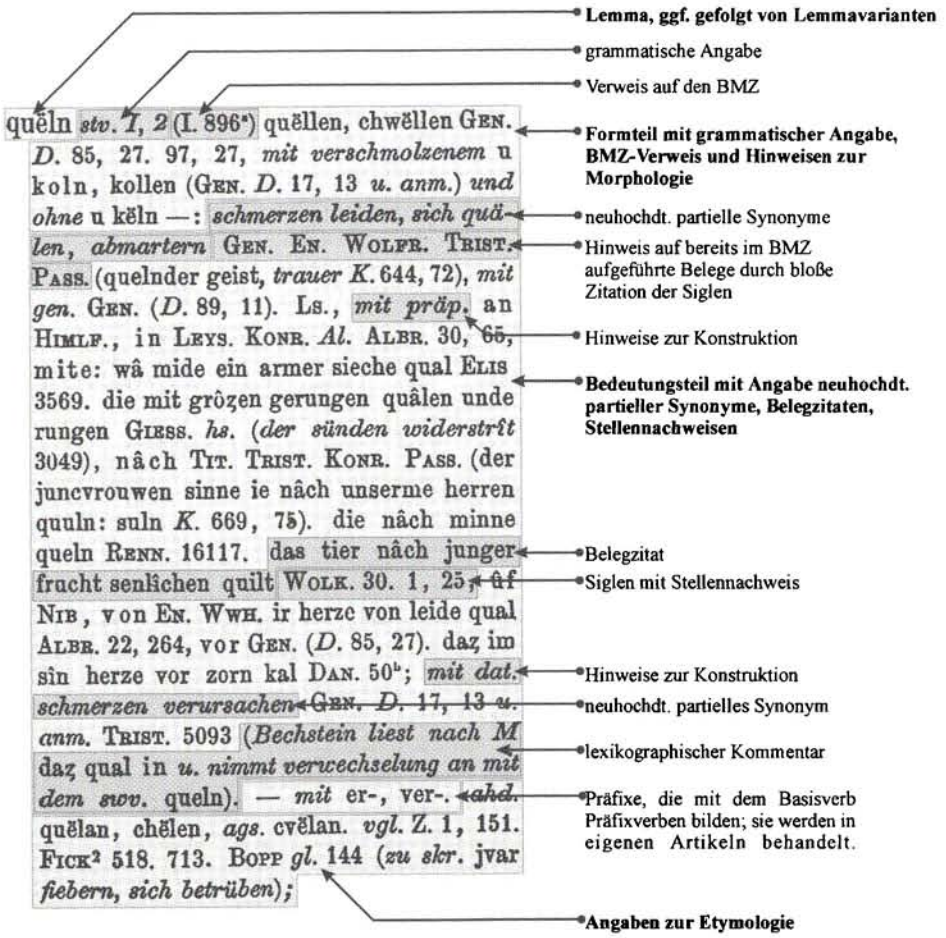

Abb. 1: Artikel *quëln* (Lexer II, 321)

---

[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.

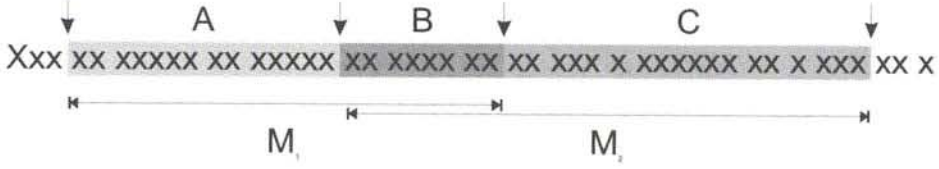

Abb 3: Kodierung überlappender Strukturen M_1 und M_2

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.

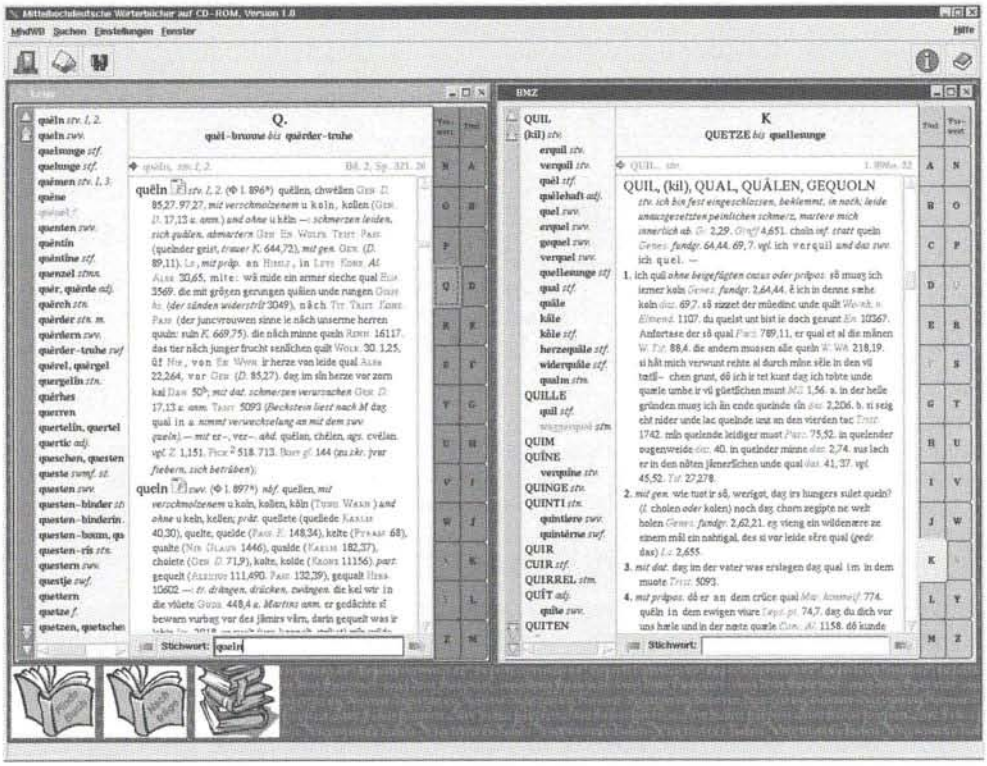

Abb. 4: Artikel *quëln* in der Darstellung der CD-ROM

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