Zur Digitalisierung historischer Wörterbücher

Sven Dummer, Frank Michaelis and Michael Schlaefer,
Deutsches Wörterbuch, Akademie der Wissenschaften in Göttingen, Deutschland

Abstract: The textual and structural characteristics of printed historical German dictionaries call for special requirements in converting these works into computer-readable form. The diverse treatment of the articles requires a great deal of follow-up manual work since the often narrative structure of the texts limits automatic processing. The following article describes a series of experiments with Moriz Heyne's "Deutsches Wörterbuch" which were conducted to illustrate the limitations (but also the possibilities) of converting an historical dictionary into electronic media.

Keywords: ELECTRONIC DICTIONARY, HISTORICAL DICTIONARY, ELECTRONIC TEXT ENCODING

Zusammenfassung: Die textuellen und strukturellen Eigenschaften gedruckter historischer deutscher Wörterbücher stellen besondere Bedingungen für die Umsetzung dieser Werke in eine maschinenlesbare Form. Die differenzierte Erfassung der Artikel erfordert einen großen Anteil an manueller Nacharbeit, da die vielfach narrativen Textstrukturen eine automatische Bearbeitung nicht erlauben. Der folgende Beitrag beschreibt eine Reihe von Experimenten mit Moriz Heynes Deutschem Wörterbuch, die mit den Ziel durchgeführt wurden, Grenzen (aber auch Möglichkeiten) einer Umsetzung historischer Wörterbücher ins elektronische Medium zu veranschaulichen.

Stichwörter: ELEKTRONISCHES WÖRTERBUCH, HISTORISCHES WÖRTERBUCH, ELEKTRONISCHE TEXTKODIERUNG

Das Angebot elektronisch nutzbarer Fassungen von Drucktexten hat auf verschiedenen Ebenen in den letzten Jahren eine zunehmende Bedeutung gewonnen. Die z. B. auf CD-Rom verfügbaren Korpora literarischer Texte und Zeitungstexte besitzen inzwischen einen beachtlichen Umfang. Dabei haben die kommerziellen Angebote in ihrem Volumen und vom Niveau ihrer Aufbereitung vielfach die noch vor wenigen Jahren als innovativ geltenden linguistischen Textkorpora wie z. B. die des Instituts für deutsche Sprache in Mannheim überholt. Ein Förderprogramm der Deutschen Forschungsgemeinschaft "Retrospektive Digitalisierung von Bibliotheksbeständen" (1998) läßt für die kommenden Jahre eine wesentliche Verbreiterung der elektronischen Textbasis im wissenschaftlichen Sektor erwarten. Für sprach- und literaturwissenschaft-

liche Forschungen bedeutet eine solche Entwicklung eine qualitative Verbesserung der Forschungsvoraussetzungen insbesondere auf der Ebene der Korpuserstellung und der systematischen Textanalyse.

Die digitale Aufbereitung von Wörterbüchern ist im kommerziellen wie im wissenschaftlichen Bereich in den letzten Jahren ebenfalls verstärkt genutzt worden. Die Frage danach, welcher Wert dieser Art der Wörterbuchaufbereitung zukommt, läßt sich u. a. von den Benutzungsbedingungen elektronischer Wörterbücher ausgehend beantworten. Die praktische Nutzung elektronischer Wörterbücher ist im Unterschied zur üblichen Wörterbuchbenutzung an Arbeitsplätze mit EDV-Ausstattung und eine angemessene technische Verfügbarkeit des Materials gebunden. Die Benutzung elektronischer Wörterbuchfassungen wird unter den z. Z. geltenden technischen Bedingungen allein aus zeitökonomischen Gründen nur dann zu erwarten sein, wenn der Aufwand zur Erreichung des gedruckten Wörterbuchs und zum Nachschlagen von Wörterbuchinformation im Vergleich zur rechnergestützten Nutzung des Wörterbuchs erkennbar höher liegt. Um etwa in einem einbändigen gegenwartssprachlichen Wörterbuch den Artikel *lachen* mit einer punktuellen Frage zur Sprachproduktion nachzuschlagen, erweist sich die Benutzung eines gedruckten Handexemplars am Arbeitsplatz gegenüber dem Starten eines Rechners, dem Einlegen einer CD und der Durchführung einer entsprechenden Suche als der entschieden einfachere und zeitsparendere Weg. Anders dagegen sind die Arbeitsbedingungen zu beurteilen, wenn man im Rahmen einer Arbeitssequenz sehr häufig in einem oder mehreren umfangreichen Wörterbüchern nachzuschlagen hat und die Suchergebnisse arbeitsökonomisch festhalten möchte. Vor allem gilt dies für systematische Wörterbuchbenutzungen, z. B. zur Ermittlung von Wortbildungsreihen, bestimmten Synonymen usw. Bei solchen systematischen Suchen wird man nur mit sehr hohem Leseaufwand im gedruckten Wörterbuch zum Ergebnis kommen. Maschinenlesbare Wörterbücher können demgegenüber für diese Fragestellungen eine effiziente Unterstützung bedeuten.

Unter dem Gesichtspunkt der Literaturversorgung ist die digitale Wörterbuchform vor allem dann von Vorteil, wenn sie sehr umfangreiche oder alte, in den Bibliotheken nur begrenzt vorhandene Werke erschließt. Mit den digital verfügbaren Werken entsteht eine bibliothekarische Situation, die nicht nur den Weg zu verschiedenen Bibliotheken oder gar Fernleihbestellungen erspart, sondern eine generell höhere Arbeitseffizienz vor allem bei intensiver Wörterbuchbenutzung durch verbesserte Literaturversorgung schafft.

Außer in den bislang skizzierten Benutzungssituationen bei der Sprachproduktion oder Sprachbeschreibung kommt digitalisierten Wörterbüchern eine sehr wichtige Rolle in der Lexikographie und in der metalexikographischen Forschung zu. Strukturen von Wörterbüchern und damit letztlich auch deren Aussagewert lassen sich umfassend überhaupt nur mit angemessen aufbereiteten maschinenlesbaren Wörterbuchversionen beurteilen.

Zusammenfassend sind drei Gesichtspunkte zu nennen, unter denen

gegenwärtig digitale Wörterbuchversionen für den Benutzer von Interesse sind: zum einen geht es um die Verbesserung der Literaturversorgung, zum zweiten geht es um die Erschließung effizienter Zugriffe insbesondere bei systematischen Wörterbuchnutzungen, und zum dritten geht es um eine Verbesserung der Grundlagen für Erforschung, Planung und Durchführung von Wörterbüchern.

Die genannten Zielvorstellungen sind von sehr unterschiedlicher Auswirkung auf die Wahl der Digitalisierungsstrategien. Generell lassen sich hier die Möglichkeiten der Image-Erschließung und der sogenannten Volltexterschließung unterscheiden. Die Image-Erschließung beruht auf dem scannergestützten automatischen Erfassen eines authentischen Textbildes. Eine Buchseite z. B. ist dann analog zur xerographischen Kopie als kleinste operationale Einheit verfügbar. Ein inhaltlicher Zugriff ist bei diesem Verfahren nur in dem Umfang möglich, in dem er durch nachträgliche Indizierung der Seitenzahlen, Überschriften oder anderer inhaltlicher Einheiten erschlossen wird. Die Volltexterfassung erschließt demgegenüber jedes Einzelzeichen eines Textes und erlaubt die Suche nach allen im Text vorkommenden Zeichen oder Zeichenkombinationen. Die beiden Verfahren können unter den entsprechenden typographischen Voraussetzungen über das Bindeglied automatischer Texterkennungsprogramme (OCR) kombiniert werden. Komplizierte Druckbilder und viele Frakturschriften sind jedoch gegenwärtig mit Hilfe der automatischen Texterkennung nur sehr unvollkommen zu bearbeiten (Retrospektive Digitalisierung 1998: 46-48). Hier würde die Digitalisierung stets ein Abschreiben und Korrigieren älterer, nur in Buchform vorliegender Werke einschließen und damit eine gegenüber dem Scannen erheblich höhere finanzielle Investition bedeuten.

Da eine image-orientierte Wörterbuchbenutzung nichts wesentlich anderes bietet als die Lesbarkeit des authentischen Textes am Bildschirm, kann sie zwar als Lösung für eine verbesserte Literaturversorgung betrachtet werden. Da sie aber keine strukturierten Suchzugriffe erschließt, muß bei der Digitalisierung von Wörterbüchern zum Zweck systematischer Nutzung oder Analysen die Volltextversion als Standardlösung gelten. Eine Image-Digitalisierung von Wörterbüchern kann nur dann Priorität besitzen, wenn dies bei häufig genutzten Werken durch deren geringe Exemplardistribution oder durch konservatorische Interessen begründet ist. Angesichts der Investitionshöhe für eine Volltexterfassung müssen entsprechende Digitalisierungsvorhaben außer nach ihrem wissenschaftlichen oder praktischen Nutzen spezifisch danach beurteilt werden, mit welchem Aufwand welcher Grad an verbesserten bzw. erweiterten Nutzungsmöglichkeiten zu erreichen ist. Dazu ist kurz auf die bei digitalen Wörterbuchversionen angewandten datentechnischen Aufbereitungsmodi einzugehen.

Die auf dem Markt befindlichen maschinenlesbaren Wörterbücher (Milan 1998) stimmen darin überein, daß sie die Artikeltexte zeichen- und formatgetreu darstellen können. Die Möglichkeiten der rechnergestützten Zugriffe (Textretrival) beschränken sich meist auf die Zeichenebene bzw. zeichen-

abhängige Segmentbildungen. Die Suche kann so z. B. nur eingeschränkt auf die Grobsegmente wie "Stichwort" und "Artikeltext" durchgeführt werden, wenn andere zeichenabhängige Segmente nicht identifizierbar sind. Alle klassischen Suchoperationen werden von dem Programm implementiert: einfache Suche nach Zeichenfolgen, Suche nach Zeichenfolgenmustern (regulären Ausdrücken) und boolschen Operatoren wie UND, ODER, NICHT. Handelt es sich um eine Suche nach einer einfachen Zeichenfolge, so ist vielfach die interaktive Auswahl in einem Wortindex möglich.

Für einen qualifizierten systematischen Wörterbuchzugriff sind darüber hinaus jedoch auch weitere Suchkriterien zu erschließen, und zwar solche, die einen gezielten Zugriff auf Inhaltsstrukturen eines Wörterbuchartikels wie Belege oder Bedeutungsbeschreibungen erlauben.

An den skizzierten Zusammenhängen wird zweierlei deutlich. Zum einen erfordert eine systematische Wörterbuchnutzung die Möglichkeit, gewisse Artikelsegmente (Stichwort, Beleg) anhand bestimmter Kriterien auswählen zu können. Zum zweiten ist es erforderlich, daß diese Segmente in einer vom Artikel losgelösten Form dargestellt werden können. Besonders bei umfangreichen Artikeln, wie sie zum Beispiel das Grimmsche Wörterbuch zu bieten hat, ist diese Reduzierung der als Ergebnis gelieferten Textmenge nicht nur ein wünschenswerter Komfort, sondern eine für die systematische Benutzung notwendige Voraussetzung.

Ein häufiger gewählter Weg, einen Text für differenzierte elektronische Zugriffe vorzubereiten, ist die Textkodierung mittels einer Auszeichnungssprache wie zum Beispiel SGML (Standard Generalised Markup Language). Dabei werden Textsegmente mittels Einklammerung in "Tags" gebildet — was man als "Markup" oder "Text-Auszeichnung" bezeichnet. Die "Tags" entsprechen einem Element des für diesen Text entworfenen Strukturmodells, das in der sogenannten "Document-Type-Definition" (DTD) vereinbart wurde. In der DTD werden die notwendigen und optionalen Elemente festgelegt und ihre Abhängigkeiten zueinander beschrieben. Ferner ist es möglich, für jedes Element Attribute zu vereinbaren. Ein Element "Stichwort" könnte beispielsweise durch ein Attribut "Wortart" näher bestimmt werden.

Die Umsetzung lexikographischer Druckprodukte ins elektronische Medium erfolgt bisher offensichtlich durchweg ohne besondere lexikographische Bearbeitung. Die zugrundeliegenden Strukturmodelle werden auf der Basis dessen entworfen, was technisch mit geringem Aufwand machbar erscheint. Es zeigt sich sehr deutlich, daß zwar durch die Digitalisierung vorhandene lexikalische Datenbestände in ein neues Medium übertragen werden, daß aber die lexikographischen Implikationen der Textversion auch in der elektronischen Version vielfach bestimmend bleiben. Die Digitalisierung bewirkt eine technische Zugriffsverbesserung, nicht die Erstellung neuer Datenbestände und nur sehr begrenzt den Zugriff auf neue lexikographische Organisationsstrukturen. Art und Umfang des verbesserten Zugriffs hängen daher maßgeblich von der gliederungs- und drucktechnischen Aufbereitung des vorhandenen Wörterbuchmaterials ab. Zeigt das Printprodukt konsequente lexikographische Struk-

turen, die sich in der typographischen Textoberfläche angemessen spiegeln, bestehen günstige Voraussetzungen für automatisch erschließbare inhaltliche Zugriffe. Exemplarisch sei hier auf Wörterbücher wie den Robert Électronique (1991) verwiesen. Die Möglichkeit, Artikelstrukturen wie die Gliederungshierarchie auszufiltern oder modulartig isolierbare Paradigmen wie z. B. das aller Stichwörter oder aller Belege zu erstellen, das Vorhandensein entwickelter Variantensuchmöglichkeiten für Wortformen sowie Exportmöglichkeiten für gewünschte Ausschnitte schaffen für die Benutzer auf der datentechnischen Ebene wünschenswert günstige systematische Arbeitsvoraussetzungen, auch wenn man sich vieles, vor allem die Exportmöglichkeiten und die Menüoberflächen, wesentlich komfortabler vorstellen könnte. Andere maschinenlesbare Versionen von Wörterbüchern wie z. B. die des Duden-Universalwörterbuchs (o. J. Version 1.1) bleiben trotz relativ günstig scheinender Strukturbedingungen im Drucktext mit nur sehr beschränkten Such- und Filtermöglichkeiten bei der systematischen Nutzung eher unbefriedigend.

Enthält ein gedrucktes Wörterbuch typographische Polysemien, gering strukturierte, diskursive Artikelbestandteile, implizite bzw. elliptische Darstellungsformen oder metasprachliche Varianten, wird der typographieabhängige maschinelle Zugriff erschwert bzw. durch die Uneindeutigkeit der Informationsklassen so unscharf, daß es nicht mehr sinnvoll ist, eine solche Version ohne Überarbeitung zu benutzen. Offensichtlich aus solchen Gründen ist bei der Erstellung der elektronischen Version des Wörterbuchs von H. Paul in 9. Auflage (1992) auf Strukturierungen weitgehend verzichtet worden. Eine systematische Benutzung der digitalen Version dieses Wörterbuchs ist dadurch nur mit erheblichem Umstand möglich. Der erreichte Standard bleibt gegenüber dem Beispiel des Robert Électronique kaum diskutabel.

Der Zustand der maschinellen Version des Paulschen Wörterbuchs wirft die Frage auf, ob und in welcher Weise die typanalogen wortgeschichtlichen deutschen Wörterbücher überhaupt sinnvoll zu digitalisieren sind. Als Repräsentanten dieses Wörterbuchtyps werden neben dem Paulschen Deutschen Wörterbuch die Werke von J. und W. Grimm (1854-1971) sowie von D. Sanders (1860-1865), F. L. K. Weigand (1909-1910) und M. Heyne (1890-1895) berücksichtigt. Mit Ausnahme des Paulschen Wörterbuchs und den ersten Teilen des Grimmschen Wörterbuchs sind diese Wörterbücher in neuerer Zeit nicht bearbeitet worden. Als materialreiche Hilfsmittel für philologische und sprachwissenschaftliche Arbeit erscheinen sie trotz ihres teilweise nicht unbeträchtlichen Alters und unverkennbarer wissenschaftsgeschichtlicher Bindungen immer noch unverzichtbar. Sie schlagen im Bereich der Verständnissicherung die Brücke von der Gegenwart in ältere Sprachzustände und bieten mit Belegen und Verwendungsbeispielen Anschauung und Materialgrundlage für weitergehende Fragestellungen. Ferner erlauben sie, die einzelnen Wörter und Wortverwendungen u. a. in semasiologischen, etymologischen, kulturgeschichtlichen und morphologischen Zusammenhängen zu betrachten. Die Benutzungsintensität im wissenschaftlichen Bereich ist daher relativ hoch einzuschätzen. Unter je spezifischen Vorstellungen von synchroner oder geschicht-

licher Systematik sind diese Wörterbücher primär für den Zugriff auf Informationen zu einzelnen Wörtern angelegt. Dargestellt wird in der Regel das in der einzelnen Wortgeschichte Spezifische. Das durchaus auch heute noch mit Gewinn nutzbare Inhaltspotential dieser Wörterbücher wird in den Druckversionen nachteilig durch eine im wesentlichen von Vorstellungen des 19. Jahrhunderts geprägte atomistische lexikographische Perspektive bzw. Benutzungserwartung beeinflußt. Die lexikographische Strukturkonsistenz ist ebenso wie die metasprachliche Konsequenz und die Ausführung von Vernetzungen in allen Werken bestenfalls ansatzweise entwickelt. Der Anteil frei umschriebener, elliptischer bzw. impliziter Information erweist sich als hoch. Diskursive Tendenzen überlagern oder durchkreuzen die unterschiedlich entwickelten Gliederungsansätze ebenso, wie die offensichtliche Veralterung vieler beschreibungssprachlicher Formulierungen Barrieren für einen systematischen Zugriff darstellen. Wortbildungsbezogene oder textbezogene Fragestellungen, die vergleichende Suche nach bestimmten Bedeutungsmerkmalen oder die Suche nach Wörtern und Verwendungsweisen u. a. m. sind lektüregestützt nur mit einem ganz erheblichen Such- und Interpretationsaufwand realisierbar. Nicht nur im Fall des Grimmschen Wörterbuchs stößt man bei diesem Verfahren durchaus auch an die Grenze der vernünftigen Relation von Aufwand und Ergebnis. Man befindet sich daher in der unglücklichen Situation, daß zwar ein respektabler Fundus an sprachgeschichtlichen Informationen zur Verfügung stünde, daß sich aber dieser Fundus strukturbedingt einer systematischen Nutzung der Printprodukte gegenüber abweisend verhält.

Die Voraussetzungen für eine Digitalisierung, mit der diese atomistisch-einzelartikelbezogene Benutzungsbarriere überwindbar wäre, erweisen sich angesichts der skizzierten inhaltlichen und formalen Textstrukturen als sehr kompliziert. Mit der bloßen elektronischen Spiegelung von Artikeloberflächen ist eine ernstzunehmende qualitative Verbesserung der Nutzungssituation für historische Wörterbücher nicht zu erwarten. Die Digitalisierung müßte hier kombiniert mit einer Restrukturierung durchgeführt werden. Darunter wird ein Komplex von lexikographischen Eingriffen verstanden, der fehlende oder defekte Segmentbildungen nachträgt bzw. ersetzt und Zugriffsebenen kennzeichnet, ohne die eine elektronische Kodierung der Artikeloberfläche weitgehend ineffizient bleiben muß. Legt man etwa die Standards zugrunde, die aus dem Robert Électronique abzuleiten wären, müßten bis zu 20 Hauptebenen mit zahlreichen Substrukturen segmentiert und klassifiziert werden. Eine solche Bearbeitung ist weder allein auf sprachwissenschaftlicher Grundlage noch allein mit informatischer Kompetenz durchzuführen, sondern erfordert zwingend einen interdisziplinären Ansatz.

Im Rahmen einer Arbeitsgruppe, in der sich Mitarbeiter der Arbeitsstelle Göttingen des Grimmschen Wörterbuchs zusammengefunden haben (H. Albrand, K. Casemir, S. Dummer, U. Härtel, F. Michaelis, M. Schlaefer, M. Schulz), konnten diese interdisziplinären Voraussetzungen geschaffen werden. Die Arbeitsgruppe hat verschiedene Experimente zur Erprobung von Möglichkeiten retrospektiver digitaler Erschließung historischer Wörterbücher durchgeführt.

Dazu wurden Teile des Wörterbuchs von M. Heyne digital erfaßt. Für die Auswahl dieses dreibändigen Wörterbuchs können u. a. seine gegenüber den einbändigen historischen Wörterbüchern höhere Stichwort- und Informationsdichte, die Bearbeitung von einer Hand und die vielfachen Strukturanalogien gegenüber dem Grimmschen Wörterbuch, aber auch gegenüber den anderen genannten historischen Wörterbüchern angeführt werden. Unter dem Blickwinkel der Übertragbarkeit der experimentellen Befunde auf typanaloge Wörterbücher schien das Heynesche Wörterbuch daher am besten geeignet.

Der Text des Heyneschen Wörterbuchs wurde mit einem herkömmlichen Textverarbeitungsprogamm authentisch abgeschrieben. Die benötigten Sonderzeichen waren verfügbar und nur in seltenen Fällen eigens herzustellen. Versuche, den gescannten Text mit Texterkennungsprogrammen (OCR) automatisch umzusetzen, mußten als zu fehleranfällig abgebrochen werden. Der digitalisierte Text wurde in Annäherung an die Originaltypographie formatiert. Angesichts der häufigen typographischen Wechsel ist die Formatierung ebenso wie die Texterfassung zeitaufwendig und fehleranfällig, was besonders sorgfältige Korrekturen erforderte. Das bedeutet etwa eine Verdreifachung der Investitionskosten gegenüber einer in anderen Fällen möglichen automatischen Texterkennung. Trotz der Beobachtungen einer Reihe typographischer Inkonsequenzen wurde zur Simulierung realistischer Arbeitsbedingungen das vorgefundene System beibehalten. Voraussetzungen für eine Konvertierung in ein Nur-Textformat bestehen. Eine exemplarische Gegenüberstellung des Drucktextes und des formatierten maschinenlesbaren Textes zeigen die anschließenden Ausschnitte.

Digitale Fassung

Aal, m. der bekannte Fisch; altes gemeingerm. Wort, ahd. mhd. âl, dunkler Herkunft. Plur. die aale, wenig gebräuchlich die äle: schleimecht fisch und ael Garg. 103; (lasz sie sich wenden wie aele in einer reusze Goethe im Götz, später in aale geändert). Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz. — aal auch von Aufgußtierchen aalähnlicher Form, essigälchen, kleisterälchen. — Zusammensetzungen: **Aalfang,** m. Fang der Aale. — **aalglatt,** ein aalglatter mensch. — **Aalquabbe, Aalraupe,** f. aalähnlicher Fisch. — **Aalreuse,** f. Reuse zum Aalfang. — **Aalstecher,** m. Gabel zum Anspießen der Aale beim Fang. — **Aaltierchen,** n. Aufgußtierchen.

Kopie des Originalartikels

Aal, m. der bekannte Fiſch; altes gemeingerm. Wort, ahd. mhd. âl, dunkler Herkunft. Plur. die aale, wenig gebräuchlich die äle: schleimecht fisch und ael Garg 103; lasz sie sich wenden wie aele in einer reusze Goethe, Götz, ſpäter in aale geändert). Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz. — aal auch von Aufgußtierchen aalähnlicher Form, essigälchen, kleisterälchen. — Zuſammenſetzungen: Aalfang, m. Fang der Aale. — aalglatt, ein aalglatter mensch. — Aalquabbe, Aalraupe, f. aalähnlicher Fiſch. — Aalreuſe, f. Reuſe zum Aalfang. — Aalſtecher, m. Gabel zum Anſpießen der Aale beim Fang. — Aaltierchen, n. Aufgußtierchen.

Die digital verfügbaren Textteile des Wörterbuchs wurden im weiteren analysiert und strukturiert. Resultate dieser Bearbeitung können hier nur exem-

plarisch angeführt werden. Die Beispiele aus dem makro- und mikrostrukturellen Spektrum sollen die Problematik des vorgefundenen Datenbestandes und den zur Kodierung erforderlichen Arbeitsaufwand verdeutlichen.

Die Makrostruktur des Heyneschen Wörterbuchs bietet auf der Stichwortebene neben den abgesetzten Stichwörtern erster Ordnung den Typ unabgesetzter Stichwörter zweiter Ordnung, denen Kompositionsstichwörter zur Einleitung von Nestartikeln, z. T. mit elliptischem Bestimmungswort, gleichgeordnet sind. Ferner wurden Verweisstichwörter von unterschiedlichem Status ermittelt. Unter den Stichwörtern erster und zweiter Ordnung bilden Präfixstichwörter und unmarkierte Homographen jeweils besondere Gruppen. Zu vielen Stichwörtern werden Varianten gebucht. Nicht selten handelt es sich dabei jedoch um eigenständige Wortbildungen. Der Stichwortstruktur sind auch nichtlemmatisierte Weiterbildungen im Artikelfuß zuzuordnen. Die folgende Tabelle listet einige der üblichen Vorkommen im Stichwortbereich auf.

Aal	Einzelstichwort 1. Ordnung
Jahren, jähren	Stichwortvarianten in der Stichwortgruppe
Hohle, f., in älterer Spr. = höhle	Stichwortvariante mit historischer Einordnung im Einleitungsteil
abängstigen (abängsten 17. Jh.)	Stichwortvariante als andere Wortbildungsform mit historischer Einordnung im Einleitungsteil
abfordern (**abfodern,** s. fordern)	Stichwortvariante mit Verweis auf Grundartikel
abkappen, ... abkappen	unmarkierter Homograph
Aalquabbe	Stichwort 2. Ordnung, Kompositionsgruppenwort
=brief	elliptisches Kompositionsgruppenwort
allerwelts=	Präfixstichwort
Angeklagte, m. s. anklagen.	Verweiswort, Verweisstichwort
auch für äsen, s. d.	versteckter Stichwortverweis im Artikeltext
abgelebt, abgelegen, f. ableben, abliegen.	Verweisstichwortgruppe
anderweitige hilfe, thätigkeit, nahrung, vorteile.	nichtlemmatisierte Weiterbildungen im Artikelfuß

Eine Identifikation der Elemente "Stichwort" bzw. "Stichwortverweis" nach typographischen Signalen oder artikelstrukturellen Positionsmerkmalen ist

nach diesem Befund nicht sicher möglich, sondern setzt eine kompetenzgestützte metalexikographische Entscheidung voraus.

Als zweites Beispiel der lexikographischen Strukturierung wird die Erstellung eines mikrostrukturellen Grundmodells der Heyneschen Artikel vorgestellt. Dem Artikelmodell kommt im Restrukturierungsverfahren die Aufgabe zu, Ordnungsrahmen für verschiedene Informationsklassen zu setzen und damit die Möglichkeit zu schaffen, artgleiche Angaben nach ihrem Status innerhalb des Artikels zu gewichten. Solche artgleichen Angaben liegen z. B. mit den Bedeutungsumschreibungen im Artikelkopf und im Artikelkern vor. Mit dem Bezug auf den jeweiligen Artikelteil läßt sich die scheinbare Klassenübereinstimmung jedoch differenzieren. Im Artikelkopf wird im Verständnis des Heyneschen Konzepts eine panchronische Grundbedeutung angegeben, im Artikelkern erscheinen im weiteren Sinn Segmente des polysemen historischen Inhaltsspektrums, die durch Belege oder explizite Angaben anderer Art ihre geschichtliche Konkretion erhalten. Weiterhin trägt die mikrostrukturelle Artikelmodellierung dazu bei, formale Bezugsebenen für die zunächst inhaltlich bestimmten Module zu schaffen und so deren lexikographische Operationalisierbarkeit zu unterstützen.

Das hier entwickelte Modell stützt sich vor allem auf entsprechende Analysen der Grimm-Lexikographie. Es werden autonome Artikel und abhängige Artikel unterschieden. Autonome Artikel enthalten idealtypisch einen Artikelkopf mit Stichwort, Wortartangabe und der Angabe einer generalisierenden Bedeutungsangabe, die im einzelnen einen sehr variablen Status besitzen kann. Dem Artikelkopf schließt sich fakultativ ein Einleitungsteil vorwiegend mit etymologischen, wortgeschichtlichen oder formalen Beschreibungen der Worteinheit an. Der für diesen Artikeltyp obligatorische Artikelkern könnte auch parallel zum Grimmschen Wörterbuch vielfach als "Bedeutungsteil" beschrieben werden, wenn man akzeptiert, daß Bedeutung hier in einem sehr ambivalenten Verständnis verwendet wird und in die Bedeutungsbeschreibung zahlreiche andere Beschreibungsebenen integriert werden. In einem dem Bedeutungsteil folgenden fakultativen Fußteil der Artikel finden sich Aufzählungen, Verweise und Vergleiche.

Aal, m. der bekannte Fisch;	Artikelkopf
altes gemeingerm. Wort, ahd. mhd. âl, dunkler Herkunft. Plur. die aale, wenig gebräuchlich die äle: schleimecht fisch und ael Garg. 103; (lasz sie sich wenden wie aele in einer reusze Goethe im Götz, später in aale geändert).	Einleitungsteil
Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz. — aal auch von Aufgußtierchen aalähnlicher Form,	Artikelkern (Bedeutungsteil)
essigälchen, kleisterälchen. —	Artikelfuß
Zusammensetzungen: **Aalfang,** m. Fang der Aale. —	Wortbildungsgruppe

aalglatt, ein aalglatter mensch. — **Aalquabbe, Aalraupe,** f. aalähnlicher Fisch. — **Aalreuse,** f. Reuse zum Aalfang. — **Aalstecher,** m. Gabel zum Anspießen der Aale beim Fang. — **Aaltierchen,** n. Aufgußtierchen.	

Dem Fußteil ließen sich auch Heynes Kompositionsgruppen zuordnen. Da diese Kompositionsgruppen unbeschadet ihrer Einleitung mit der Überschrift *Zusammensetzungen* jedoch nicht durchgängig Komposita enthalten und nicht alle Artikel auch als abhängige Artikel beschreibbar sind, wird für das im weiteren verwendete Artikelmodell die Wortbildungsgruppe als eigenes Segment behandelt, in dem Artikelmikrostruktur und Makrostruktur der Stichwortreihe verzahnt sind.

Die abhängigen Artikel sind dann deutlich zu erkennen, wenn sie in einer Wortbildungsgruppe angeschlossen werden oder außer dem Stichwort und der Wortartangabe nur rudimentäre Angaben enthalten und nur unter Bezug auf das Stichwort des vorangehenden autonomen Artikels letztlich verständlich sind:

Aaltierchen, n.	Artikelkopf
	Einleitungsteil
Aufgußtierchen.	Artikelkern (Bedeutungsteil)
	Artikelfuß
	Wortbildungsgruppe

In vielen Fällen verschwimmt jedoch diese Grenzziehung. Es erscheint daher am ehesten sinnvoll, alle Artikel, die einer Überschrift *Zusammensetzungen* folgen, den abhängigen Artikeln zuzuordnen. Die Bedeutungsangaben sind nicht sicher in Analogie zu den selbständigen Artikeln zuzuordnen, so daß eine Mehrfachansetzung erwogen werden könnte.

Die mikrostrukturelle Analyse wurde unter dem Blickwinkel der verschiedenen Beschreibungsebenen für Struktur, Status und Verwendung der als Stichwörter angesetzten Sprachzeichen weitergeführt. Eine als Ergebnis dieser Analyse vorgenommene Modellierung des vorgefundenen Informationsprofils kann für das weitere Vorgehen folgende Schichten unterscheiden:

Angaben zu Zeichenkategorien und Zeichenstrukturen:	
Stichwort	Aal
Wortart	m.
Wortbildung	nur noch in Zuss. bergab, hügelab; Zusammensetzungen: Aalfang
Grammatik (Formbildung, Flexion, Syntax)	Plur. in der alten Spr. wie Sing.
Bedeutung	Unsinnlich, zur Bezeichnung eines Schwankens
Betonung	Konfékt

Angaben zum Ursprung, zur Herkunft des Zeichens:	
Etymologie	altes gemeingerm. Wort (...) dunkler Herkunft
Angaben zur Stellung des Zeichens, Zeichengebrauchs im deutschen Diasystem:	
Sprachsoziologie	in beschränktem Gebrauche auch in kaufmänn. Sprache
Stilistik	vielfach gemeines Schimpfwort; In dichterischer Freiheit bei BÜRGER
Sprachgeographie	nach der Zeit nur mundartlich (schwäb., schweiz.) und in Quellen die von der Mundart beeinflußt
Sprachgeschichte	Präp. mit Dativ, = von, jetzt von diesem verdrängt
Angaben zur Zeichenverwendung:	
Phraseologie	Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz.
Frequenz	wenig gebräuchlich die äle
Usualität	das Wort ist im 17. Jh. ungewöhnlich, in der Schriftsprache des 18. Jh. erst allmählich durch Beschäftigung mit dem mhd. wieder aufkommend
Kollokationen	königlicher aar; du junger aar
Dokumentation der Zeichenverwendung:	
Belege	Lessing Nath. 1, 3; alles moralische aas
Verwendungsbeispiele	ab und zu komme ich wohl dahin

Die hier bezeichneten mikrostrukturellen Sachverhalte bilden jeweils Klassen mit z. T. komplexen Subklassifikationsinventaren bzw. stark variablen Bezeichnungen für gleiche Sachverhalte. Bei den Wortarten unterscheidet M. Heyne weitgehend die üblichen morpho-syntaktischen Klassen. Mehrfachklassifikationen bei alternierender Wortart eines Etymons sind aber ebenso üblich wie das Aussparen der Wortartangabe bei Verben, z. T. auch bei Adjektiven. Während im Paradigma der Wortartangaben, abgesehen von den skizzierten Problemen, eine weitgehende Konsequenz in der Terminologie herrscht, werden z. B. stilistische oder diasystematische Sachverhalte vielfach frei diskursiv umschrieben. Der Sachverhalt der sprichwörtlichen Verwendung ist allein im Abschnitt A – Ab mit sechs verschiedenen Bezeichnungen angesprochen worden: *sprichwörtl.; sprichwörtlich; Sprichw.; Rechtssprichwort; im Sprichworte, in Sprichwörtern.*

Eine ähnliche Komplexität und Varianz der Substrukturen zeigt sich bei den Belegen. Idealtypisch bestehen die Belege aus einem objektsprachlichen Zitat, einer Autoren- und/oder Titelnennung und einem Stellenverweis. Der Belegbegriff wird jedoch sehr frei gehandhabt und führt so nicht nur zu einer

Fülle von Typvarianten, sondern auch zur vielfach diskursiv frei eingebetteten Belegform. Zur Veranschaulichung werden einige Beispiele zusammengestellt.

Belegbeispiel	Kommentar	Belegtypbez.
war doch der reiz der groszen arzneiflasche .. bald abgebraucht Immermann Münchh. 4, 121;	vollständiger Beleg mit Zitat, Autorennennung, Werknennung, Band- und Seitenangabe	Standardbeleg
flut abdämmen in einem Bilde Freiligrath 3, 123	Beleg mit Kurzzitat, diskursiver Einflechtung eines Interpretaments, Autorennennung, Band- und Seitenangabe	verkürzter Standardbeleg
als Kunstwort des Festungsbaus 1729 verzeichnet (abdachung, schiefe eines walles Hederich)	Wörterbuchbeleg mit seltener Nennung des Erscheinungsjahres und der sonst üblichen Beschränkung auf die Autorennennung	Wörterbuch-nachweis
die geschlachtete gans, das schwein, ferkel (Seume Spaz. 1, 96)	Verwendungstyp mit Autorennennung, Werknennung und Stellenangabe für einen elliptischen Zitattext	Vorkommens-nachweis nach Verwendungstyp (ohne Zitat)
des lebens mai .. mir hat er abgeblüht Schiller	Zitat mit Autorennennung ohne identifizierende Text- und Stellenangaben	Zitat mit Autorenidentifikation
unterschied zwischen dem letzten thaler, den man borgt, und zwischen dem ersten, den man abbezahlt Goethe Unterh. deutscher Ausgew.	Zitat mit Autorennennung und Textangabe, jedoch ohne identifizierende Stellenangabe	Zitat mit Werkidentifikation
allnächtlich, Adv. alle Nächte vorkommend oder wiederkehrend (Bürger Pfarrers Tocht. v. Taub.).	Autorennennung, Werknennung jedoch ohne Stellenangabe	zitatlose Werkidentifikation

Ein generelles Problem der Restrukturierung ergibt sich für die Segmentbildung überall dort, wo die diskursive Textform in einem syntagmatisch nicht trennbaren Zusammenhang mehrere Informationsklassen verschachtelt. Hier sind zahlreiche Mehrfachklassifikationen nötig, wenn später maschinell Textsegmente gebildet werden sollen, die die nötige Informationsautonomie besitzen. Dazu kommt, daß gerade die diskursive Verknüpfung immer wieder einen

Rückgriff auf den Gesamtartikel nötig macht, wenn die metasprachlichen Textteile narrativ zusammenhängend formuliert wurden und nur in Anmerkungsart durch Belege oder Verwendungsbeispiele unterbrochen sind.

Das komplexe Ergebnis der in Ausschnitten angedeuteten Restrukturierung von Artikeln aus M. Heynes Deutschem Wörterbuch wird im Anschluß in einer TEI-konformen Auszeichnung des Artikels AAL m. demonstriert. Die Richtlinien der Text Encoding Initiative (TEI 1994) bieten auch für die Kodierung von Strukturen im historischen Wörterbuch ein außerordentlich entwickeltes Strukturierungs- und Klassifikationsinventar. Sie erlauben so eine äußerst umfangreiche Anreicherung der Textdaten mit Informationen. In welchem Maße der TEI-Anwender davon Gebrauch macht, bleibt ihm allerdings selbst überlassen. Die Entscheidung über den Umfang der Auszeichnungen hängt in erster Linie davon ab, welche Informationen in die Textdaten eingebracht bzw. welche bereits im Text enthaltenen Informationen explizit gekennzeichnet werden sollen, damit später gezielt auf sie zugegriffen werden kann. Wörterbuchartikel sind hinsichtlich ihrer inhaltsstrukturellen und typographischen Gestaltung ohnehin äußerst komplex; wenn im Falle von hochgradig diskursiven Wörterbüchern wie dem von M. Heyne die Struktur der Artikel zudem keinem festen Schema folgt, so erhöht sich die Schwierigkeit, eine auf alle Artikel anwendbare Auszeichnungsmethode zu finden. Zudem ist eine automatisierte Auszeichnung vieler Informationseinheiten (z. B. etymologischer Erläuterungen) unmöglich. Sind, wie es in diskursiven Texten häufig der Fall ist, verschiedene Informationseinheiten zudem noch ineinander verschachtelt und/oder elliptisch aufeinander bezogen, stellt sich sogar die Frage, inwiefern die Auszeichnung überhaupt sinnvoll ist. Vielfach lassen sich nur Bruchstücke der gesamten Einheit markieren — zumindest, wenn man nicht durch großzügig übergreifende Markierungen hohe Ungenauigkeiten und Redundanzen in Kauf nehmen will und so z. B. als Ergebnis innerhalb einer Markierung "Etymologie" auch noch drei andere, für die Etymologie unerhebliche Informationseinheiten zu finden wären. Ein durch die Markierungen ermöglichter gezielter Zugriff auf solche Fragmente ist oft unbefriedigend, und ihre Extraktion z. B. in einer Suchabfrage "stelle mir alle Etymologie-Angaben zusammen" macht bei einem zu hohen Maß an Fragmentiertheit oder Redundanz überhaupt keinen Sinn.

Enthält der elektronische Text dieselben typographischen Merkmale wie der gedruckte, kann zumindest die Auszeichnung von Artikelanfang und Artikelende sowie einiger typographisch besonders gekennzeichneter artikelinterner Einheiten automatisiert erfolgen. Voraussetzung hierfür ist freilich eine weitgehend genaue und fehlerfreie Erfassung der typographischen Merkmale des gedruckten Textes bei der Digitalisierung; bei Texten, die nicht gescannt werden können und manuell erfaßt werden müssen, bedeutet dies einen erheblichen Mehraufwand. Die Möglichkeiten sind freilich eingeschränkt, da die typographischen Merkmale selten exakt an inhaltliche Merkmale gebunden sind. Zumeist sind zumindest die Stichwörter in einer eigenen Schriftart gedruckt und können damit problemlos automatisch markiert werden, sobald

aber eine Schriftart verschiedene Informationseinheiten auszeichnen kann, stößt die automatisierte Auszeichnung an Grenzen.

Ein großer Teil der wünschenswerten Markierungen könnte also nur manuell erfolgen. Im folgenden sei ein Beispiel für eine sehr elaborierte Auszeichnung gegeben; es sei darauf hingewiesen, daß hier noch nicht einmal alle Möglichkeiten, die die TEI-Guidelines bieten, ausgeschöpft sind. Der eigentliche Text ist zur besseren Übersicht fett gedruckt:

```
<!DOCTYPE TEI.2 system 'tei2.dtd' [
     <!ENTITY % TEI.dictionaries 'INCLUDE' >
     <!-- .... -->
]>

<tei.2>

<!-- ... -->

 <teiHeader>
     <fileDesc>
     <titleStmt>
       <title>Artikel AAL aus dem Heyneschen WB</title>
     </titleStmt>
     <publicationStmt>
       <p>bislang unver&ouml;ffentlicht</p>
     </publicationStmt>
     <sourceDesc>
      <bibl>
         <author>Moriz Heyne<author/>
         <title>Deutsches W&ouml;rterbuch.</title>
         <imprint>
              <pubPlace>Leipzig</pubPlace>
              <publisher>Hirzel Verlag</publisher>
           <biblScope type=volume>Bd. 1.</>
              <biblScope type=edition>2. Aufl.</>
              <date>1905</date>
         </imprint>
       </bibl>
     </sourceDesc>
     </fileDesc>

 </teiHeader>

 <text>
 <body>

 <!-- ... -->

 <entryFree n='1'>

  <form><orth>Aal</orth></form>,
```

```
<gramGrp>
      <gen>m.</gen>
   </gramGrp>

   <sense n='1'>
      <def>der bekannte Fisch</def>;
   </sense>

   <etym>
      altes <lang>gemeingerm.</lang> Wort,
      <lang>ahd.</lang> <lang>mhd,</lang>
      <mentioned>&acirc;1,</mentioned>
      dunkler Herkunft.
   </etym>

   <gramGrp>
      <number>
        Plur. die aale, <usg type=plev>wenig
       gebr&auml;uchlich</usg>
         <eg><cit>
           <q>die &auml;le:
           schleimecht fisch und ael</q>
           <bibl>
               <author>Garg.</author>
                <biblScope type=page>103;</biblScope>
           </bibl>
         </cit></eg>

<eg><cit>
            <q>(lasz sie sich wenden wie aele in einer
             reusze</q>
            <bibl><author>Goethe</author> im
           <title>G&ouml;tz</title></bibl>
         </cit></eg>
             , sp&auml;ter in <q>aale</q> ge&auml;ndert).
      </number>
   </gramGrp>

    <usg type=reg>Redensarten</usg>
       <q>glatt, schl&uuml;pfrig, schleimig wie ein
aal;</q>

    <usg type=reg>Sprichw.</usg>
        <q>wer den aal h&auml;lt bei dem schwanz,
           dem bleibt er weder halb noch ganz.</q>

   &shy;

   <sense n='2'>
```

```
      <form>
        <q>aal</q>
      </form>

        <def>auch von Aufgu&szlig;tierchen
              aal&auml;hnlicher Form,</def>

      <eg>
         <q>essig&auml;lchen, kleister&auml;lchen.</q>
      </eg>

   </sense>

   &shy;

<!-- ... -->

</body>
</text>

</tei.2>
```

Eine solche Anreicherung des Beispieltextes mit Mark-Up ist, wie angedeutet wurde, äußerst zeitaufwendig, und der nötige Aufwand steht ganz sicher in keinem angemessenen Verhältnis zu den dadurch ermöglichten Resultaten. Allein die Festlegung eines für alle Artikel anwendbaren Auszeichnungsschemas würde bei dieser Ausführlichkeit monatelange Planungsarbeiten durch Fachpersonal, das sowohl lexikographisch geschult als auch mit den TEI-Guidelines vertraut ist, erfordern. Es gilt, einen mit vertretbarem Aufwand zu realisierenden Mittelweg zwischen Maximal- und Minimalauszeichnung zu finden und zu evaluieren, ob das, was rechnerunterstützt erreichbar ist, zu brauchbaren Ergebnissen führt.

Es wurde daher in einem zweiten Experimentabschnitt ein Verfahren angewandt, das mit Hilfe einer u.a. auf die typographischen Merkmale gestützten, weitgehend automatischen Auswertung des Wörterbuchtexts Indizes erzeugt, mit denen sich wiederum TEI-konforme Auszeichnungen in den Text einbringen lassen. Abgesehen davon bieten die Indizes ein eigenes Arbeitsinstrument, entscheidend ist aber, daß sie eine TEI-konforme Aufbereitung des Textes ermöglichen, so daß die Vorteile von SGML/TEI zum Tragen kommen — nämlich die Abfassung der Textdaten in einem international standardisierten, plattform- und softwareunabhängigen und alterungsbeständigen Format, aus dem sich mit entsprechenden Hilfsmitteln bei Bedarf andere Formate z. B. für eine Druckvorlage oder die Publikation im WWW erzeugen lassen.

Das modifiziert fortgesetzte Experiment geht weiterhin von einer vorlagengetreuen Abschrift des Heyneschen Wörterbuchs als Grundlage aus. Ziel ist im weiteren eine Strukturierung anhand der typographischen Merkmale.

Dabei kann die Zeichen- von der Formatebene unterschieden werden. Auf der Zeichenebene eignen sich prominente Zeichen, wie zum Beispiel das Leerzeichen oder der Strichpunkt, als Merkmale, auf der Formatebene Absätze, Einrückungen sowie der Wechsel von Schrifttypen.

Anhand des typographischen Merkmals "Absatz" läßt sich der Wörterbuchtext in Segmente zerlegen, die z. T. als Einzelartikel, z. T. aber auch als Reihen von Nestartikeln zu bestimmen sind. Ferner lassen sich bestimmte Zeichen als "Wortende" interpretieren, so daß es möglich ist, den Text automatisch in Wortsegmente zu gliedern. Für jedes Wortsegment gilt nun, daß es in einer und nur in einer Schrifttype dargestellt wird. Deswegen eignet sich das Merkmal "Schrifttype", die Wortsegmente entsprechend der in den Schrifttypen enthaltenen impliziten lexikographischen Informationen zu klassifizieren. Ein Wort ist entweder ein "Stichwort", ein "Verfassername", "Beschreibungssprache" oder "Objektsprache". Es ergibt sich folgende Mengenverteilung der segmentierten Wörter:

Element	gesamt	unterscheidbare Zeichenfolgentypen
Stichwort	8 550	8 490
Beschreibungssprachlich	218 740	23 620
Objektsprachlich	239 500	45 380
Verfassername	18 830	540

Im ausgewerteten Wörterbuchabschnitt A – E können 18 830 Wortsegmente als Verfassernamen identifiziert werden. Davon entfallen z. B. 2 490 auf den Zeichenfolgetyp "Goethe", 4 auf den Zeichenfolgetyp "Goethes". Insgesamt lassen sich 540 verschiedene Zeichenfolgetypen im Bereich der Verfassernamen unterscheiden. Es handelt sich um gerundete Werte, da man je nach den Zeichen, die man als Wortende interpretiert, zu leicht divergierenden Ergebnissen kommt. Außerdem ist zu beachten, daß durch typographische Systemfehler im Drucktext sowie durch Fehlkodierungen bei der Erstellung der digitalen Fassung mit einer Fehlerquote von ca. 15% zu rechnen ist. Das bedeutet bei rein maschineller Ausführung der Auszeichnung des Wörterbuchtextes nach typographischen Merkmalen einen nicht unbeträchtlichen Fehlerfaktor, der hier wie auch bei den im weiteren vorgestellten Verfahren die Grenzen solcher automatisierten Auszeichnungsverfahren zu erkennen erlaubt.

Kombiniert man die bisher angewandten Verfahren, so lassen sich weitere Segmentbildungen erreichen. So ist durch Suchen, die sich auf die Kombination der absatzorientierten und der schriftartbezogenen Auszeichnung beziehen, eine Identifikation der Nestartikel als Einzelartikel möglich. Auch können die Wortartangaben, soweit im Drucktext vorhanden, umgebungs- und formatbezogen ermittelt werden. Die häufigen Lücken in den Wortartangaben zwingen jedoch zum systematischen Prüfdurchgang und zur Ergänzung. Die

aufgrund des skizzierten Verfahrens gewonnenen Artikelsegmente lassen sich etwa in folgender Form darstellen:

Stichwort	Wort-art	Text
A	n.	Ausruf des Ekels (ein ä-geschmack Goethe im Satyros 1); Nachahmung des Kindergeschreis (ders., Künstlers Erdenwallen); des Räusperns, Stockens: viel akzion! viel — ä! ä! — was ich sage! Wieland Abd. 3, 6. vgl. b
Aal	m.	der bekannte Fisch; altes gemeingerm. Wort, ahd. mhd. âl, dunkler Herkunft. Plur. die aale, wenig gebräuchlich die äle: schleimecht fisch und ael Garg. 103; (lasz sie sich wenden wie aele in einer reusze Goethe im Götz, später in aale geändert). Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz. — aal auch von Aufgußtierchen aalähnlicher Form, essigälchen, kleisterälchen.
Aalfang	m.	Fang der Aale.
aalglatt	adj.	ein aalglatter mensch.
Aalquabbe	f.	
Aalraupe	f.	aalähnlicher Fisch.
Aalreuse	f.	Reuse zum Aalfang.
Aalstecher	m.	Gabel zum Anspießen der Aale beim Fang.
Aaltierchen	n.	Aufgußtierchen.

Die automatische Segmentierung des Wörterbuchtextes in Artikel und Wörter und die Auszeichnung der Wörter gemäß ihrer Typologie erlaubt einfache systematische Zugriffe. Es können z. B. alle Artikel bzw. Artikelpositionen aufgesucht werden, die das Wort *Redensart* als beschreibungssprachliches Element enthalten. Diese beschreibungssprachlichen Vorkommen sind von Lemmavorkommen oder objektsprachlichen Vorkommen des Wortes *Redensart* zu unterscheiden. Die bislang weitgehend automatische Segmentierung des Wörterbuchtextes ermöglicht damit eine nicht unerhebliche Vorauswahl. Weitere Experimente der maschinell gestützten Strukturbildung stützen sich auf die Beobachtung systematischer Abfolgen typographischer Elemente.

So erweist es sich als möglich, im Heyneschen Wörterbuchtext das Element "Beleg" teilweise automatisch zu extrahieren, da sich die Standardbelegform durch eine relativ stabile Formatregelung auszeichnet. Dem Zitattext folgt eine Verfassernennung, dieser wiederum eine Stellenangabe jeweils in eigener Typographie. Diese typographische Sequenz ist zwar nicht monosem, aber bei einer automatischen Ausfilterung dieser Formatsequenz überwiegen jedoch die gewünschten Belege. Die manuelle Aussonderung der belegfremden Textfolgen bildet daher kein erhebliches Hindernis hinsichtlich des Arbeits-

aufwandes. Mit diesem maschinell vorbereiteten Segmentierungsgang für die Standardbelege können etwa 40% des Gesamttextes separat erfaßt und inhaltlich als Beleg ausgezeichnet werden. Das Ergebnis des Arbeitsganges ist ein Belegmodul, das sich zudem leicht weiter nach Objektsprache, Verfassern und Werk-/Stellenangabe maschinell gliedern läßt. Die tabellarische Übersicht im Anschluß zeigt das Belegmodul:

Stich-wort	Belegtext	Autor	Werk-angabe
A	ich lêre in daz â bê cê; des enhât er niht mê noch gelernet wan daz â		Pf. Amis 297;*
A	ich bin das a und das o, der anfang und das ende		Offenb. 1, 8;*
A	wolt nit A sagen, auf dasz er nicht müsz B sagen		Garg. 247. 2*
Aal	die äle: schleimecht fisch und ael		Garg. 103*
Aal	lasz sie sich wenden wie aele in einer reusze	Goethe	im Götz*
Aas	wo aber ein asz ist, da samlen sich die adler		Matth. 24, 28*
Aas	wenn fürsten geyer unter äsern sind	Lessing.	Nath. 1, 3*
Aas	alles moralische aas	JPaul	uns. Loge 3, 42
Ab	wenn einmal die wurzel ab sei	JGotthelf*	Schuldenb. 183
Ab	fuhr auf und ab	Freytag	Ahnen 1, 374*
Ab	ging .. im zimmer auf und ab	Freytag	Ahnen 4, 383*
Ab	die auf dem schlosse ab und zu ritten	Eichendorff*	Taugenichts 50
Ab	weiter ab	Lessing	Nath. 1, 4
Ab	kam ab der post ein kistlein	Hebel	2, 101*
Ab	freud und lust an allem ab und an, an und ab dem kleeblatt holder kinder*	Bürger	(60a)
Ä	ein ä-geschmack	Goethe	im Satyros 1
Aas			1. Mos. 15, 11
Ä	viel akzion! viel — ä! ä! — was ich sage!	Wieland	Abd. 3, 6.*

Der nach Ausfilterung der Standardbelege verbleibende Wörterbuchtext stellt als verkürzte Lesefassung ein eigenes Bearbeitungsprodukt dar.

Stichwort	Wort-art	reduzierter Artikeltext
Aal,	m.	der bekannte Fisch; altes gemeingerm. Wort, ahd. mhd. âl, dunkler Herkunft. Plur. die aale, wenig gebräuchlich später in aale geändert). Redensarten glatt, schlüpfrig, schleimig wie ein aal; Sprichw. wer den aal hält bei dem schwanz, dem bleibt er weder halb noch ganz. — aal auch von Aufgußtierchen aalähnlicher Form, essigälchen, kleisterälchen. Zusammensetzungen:
Aalfang,	m.	Fang der Aale.

Aas,	*n.*	totes, faulendes Vieh. Altes westgerm. Wort (ahd. mhd. âs, ags. æs), Ableitung der Wurzel az essen, ursprüngl. als Speise der Raubtiere oder Fütterung für Hunde, Falken gedacht, dann verallgemeinert. Plur. in der alten Spr. wie Sing. später äser (z. B. bei Lessing, Kant), jetzt auch die aase. Obwohl biblisches Wort ist es doch in gewählter und bildlicher Sprache lieber vermieden als gebraucht, da es vielfach gemeines Schimpfwort: du aas Mephist. zur Hexe in Goethes Faust. vgl. raben-, schindaas. Zusammensetzungen

Einer Programm-Oberfläche, die später auf das digitalisierte Wörterbuch aufsetzt, ist es nun möglich, zwei verschiedene Ansichten ein und desselben Artikels zu bieten: einen Kurzartikel, ohne Belege, und einen Vollartikel, mit Belegen. Außerdem kann ein separater Zugriff auf das Belegarchiv ermöglicht werden. Dieses Verfahren, dem Benutzer verschiedene Ansichten eines Artikels anzubieten, sowie der Zugriff auf ein vom Wörtbuch getrenntes Belegmodul haben sich in der Praxis bewährt. Als Beispiel dafür kann wiederum der Robert Électronique gelten.

Das Verfahren der bislang durchgeführten maschinellen Textgliederung führt zu Modulbildungen, die sowohl paradigmainterne Suchen als auch über eine entsprechende Verknüpfung die Rückkopplung auf den Originalartikel gestatten. Das angewandte Verfahren erfordert kontrollierende Nacharbeiten. Diese bleiben aber in einem überschaubaren Umfang und setzen keine umfängliche konzeptionelle oder metalexikographische Kompetenz voraus.

Um einen Überblick über das verwendete Wortmaterial zu erhalten, ist es möglich, die nach ihrem Format ausgezeichneten Wortsegmente je für sich in einem Index zu gruppieren. Um im Hinblick auf die angestrebte Transferierbarkeit des Arbeitsverfahrens einen möglicherweise nur in M. Heynes Wörterbuch vorliegenden Spezialfall der Formatsyntax nicht zum Anlaß unrealistischer Allgemeineinschätzungen werden zu lassen, umfaßt die Indexerstellung in den anschließenden Beispielen den Gesamttext einschließlich der Standardbelege in der Differenzierung nach Objektsprache, Metasprache und Verfassernamen. Alle drei Indizes erschließen den gezielten Weg von den jeweiligen Wortformen zu bestimmten Stichwörtern bzw. zu Artikelpositionen. Als Problem bleiben generell die flektierten Wortformen bzw. historisch variable Wortformen. Ein weiteres generelles Problem liegt in der Belastung der Indizes durch Massenwörter wie Konjunktionen, Präpositionen, Artikel. Im objektsprachlichen Index sind Wörter aus Verwendungsbeispielen nicht von Belegwörtern unterscheidbar. Im metasprachlichen Index sind "echte" Definitionswörter von formulierungstechnisch bedingten Appellativen nicht unterscheidbar. Da die Werkbezeichnungen in den Zitaten typgleich mit der Metasprache sind, ist ein Appellativum *Ahnen* nicht ohne weiteres als Buchtitel zu erkennen. Ein Teil dieser Probleme läßt sich jedoch ausgrenzen, wenn man, wie vorge-

schlagen, die Standardbelege vorab segmentiert oder generell abweichend vom Originaldruck weitere funktionelle Formatunterscheidungen trifft. Eine einfache Möglichkeit zur Straffung der Indizes besteht darin, die Massenwörter durch Stopplisten oder manuelle Nacharbeit zu eliminieren. Zur Veranschaulichung werden jeweils kurze Ausschnitte der genannten Indizes abgebildet.

Objektsprache

Artikel-lemma	Index-wort
<Aal>	aal
<Aal>	aale
<Aal>	aale
<aalglatt>	aalglatter
<Aar>	aar
<Aar>	aar
<Aar>	aar
<Aar>	aar
<Aar>	aare

<Aar>	aaren
<Aar>	aares
<Aar>	aars
<Aas>	aas
<Aas>	aas
<Aas>	aase
<aasig>	aasicht
<aasig>	aasiger
<aasig>	aasiges
<ab>	ab

Metasprache

Artikel-lemma	Index-wort
<Aal>	aalähnlicher
<Aalstecher>	Aale
<Aalfang>	Aale
<Aalreuse>	Aalfang
<Aar>	Aars
<aasig>	Aas
<aashaft>	Aase
<Aasrabe>	Aase
<Aasvogel>	Aase
<aasig>	Aase
<ab>	abgetan
<aasen>	abschaben
<ab>	Ahnen
<aashaft>	ähnlich
<Aar>	allmählich
<A>	Alphabet

<A>	Alphabets
<A>	Alphabets
<Aas>	alten
<ab>	alten
<aasig>	älteren
<Aar>	älterer
<Aal>	altes
<Aas>	Altes
<ab>	Altes
<A>	anders
<A>	Anfang
<A>	anfangen
<Aalstecher>	Anspießen
<A>	aufgekommen
<Aal>	Aufgußtierchen
<Aar>	aufkommend

Autorennamen

Artikel-lemma	Autor
<Aar>	Schlegel
<Aar>	Bürger
<ab>	Bürger
<ab>	Eichendorff
<ab>	Freytag
<Aar>	Gleim
<A>	Goethe
<Aal>	Goethe
<ab>	Goethe
<Aas>	Goethes
<ab>	Hebel
<ab>	Gotthelf
<Aas>	JPaul
<Aas>	Kant
<Aas>	Lessing
<ab>	Lessing
<Aas>	Lessing
<A>	Gerhard
<Aar>	Platen
<ab>	Rückert
<ab>	Scheuchzer
<ab>	Schiller
<ab>	Stieler
<A>	Wieland

Im Rahmen des zweiten Teils des Digitalisierungsexperiments an M. Heynes Deutschem Wörterbuch sind mit den vorgestellten Zugriffen die Möglichkeiten einer automatischen Textauszeichnung weitgehend erschöpft. Das Ergebnis besteht in einer relativ groben inhaltlichen Strukturierung, die eine Reihe von systematischen Zugriffen erleichtert und unterstützt, jedoch vom erreichten Standard nicht an den des Robert Électronique heranreicht. Eine Steigerung des Strukturierungsniveaus ist zwar ausgehend von den typographischen Kennzeichnungen noch mit maschineller Unterstützung möglich, erfordert aber zwangsläufig in wachsendem Maß wieder manuelle Nacharbeiten. Um Möglichkeiten solcher halbautomatischen Ansätze zu demonstrieren, werden in einem dritten Abschnitt des Experiments einige Beispiele vorgestellt.

Der Mangel der Wortformenabhängigkeit im bisherigen Aufbereitungsstatus wirkt sich vor allem im metasprachlichen Index nachteilig aus, da die Mischung der im engeren Sinn definitionssprachlichen gegenüber den im diskursiven Stil formulierungstechnisch geforderten Wörtern, die typographische Gleichheit der historischen und flexivischen Variablen sowie Heynes terminologische Varianten ein gezieltes Arbeiten stören. Um zum Beispiel auf die Phraseologiekennzeichnung *sprichwörtlich* zugreifen zu können, müßte man nach allen variablen Bezeichnungen für diese Kategorie suchen lassen. Wie vorne gezeigt, ergeben sich dabei allein im Abschnitt A – Ab sechs verschiedene Varianten. Das gleiche Problem stellt sich mit den Varianten der flektierten Wörter, insbesondere bei Homographen wie *Macht* f. und *macht* vb. usw. Vor allem im Bereich unregelmäßiger Verben kann die Anzahl der Varianten erhebliche Ausmaße annehmen, so daß sie kaum vom Benutzer kontrolliert werden kann. Diese Sachverhalte lassen sich nur durch manuelle Lemmatisierung und eine weitergehende Klassifikation beheben. Im Rahmen des Experiments wurden dazu einige Weiterbearbeitungen durchgeführt.

Die im engeren Verständnis terminologischen Bezeichnungen für grammatische, semantische oder pragmatische Klassen umfassen ein relativ überschaubares, häufig verwendetes Inventar, das durch die paradigmatische Sortierung im Index überschaubar wird. Varianten lassen sich leicht erkennen und per Lemmatisierung vereinheitlichen. Große Teile der formalen und sprachgeschichtlichen Wortbeschreibung in den Artikeln könnten durch eine Aufbereitung dieser definitionssprachlichen Einheiten gezielt angesteuert werden. Weniger optimistisch ist eine analoge Bearbeitung der bedeutungsbeschreibenden Wörter zu beurteilen. M. Heynes z. T. eigenwilliger, vielfach elliptischer Stil dürfte eine zuverlässige Hyponym/Hyperonym-Suche kaum zulassen, weshalb man sich auf eine bloß kategoriale Kennzeichnung beschränken wird.

Man kann jedoch die Möglichkeiten der Textkodierung zur Erweiterung des Strukturmodells benutzen. Das Element "Beschreibungssprache" wird um die Attribute "Textwortlemma", "Artikellemma" und "Paradigma" erweitert. Das Attribut "Artikellemma" sichert den Rückbezug der beschreibungs- und objektsprachlichen Textwörter zum zugehörigen Artikel. Das Attribut "Textwortlemma" ermöglicht dem Textretrieval eine einheitliche Zugriffsebene. Es muß nicht mehr nach "adv." und "adverb" gesucht werden, sondern nur noch nach dem Lemma, unter dem beide zusammengefaßt sind. Das Attribut "Paradigma" zeigt exemplarisch inhaltliche oder kategoriale Paradigmenzuordnungen, die bei Abfragen als je eigene Inhaltsstrukturen erschlossen werden können. Es ist ausdrücklich darauf hinzuweisen, daß die vorgenommenen Klassifikationen weitestgehend nur manuell und unter Autopsie des Artikeltextes möglich sind. Die anschließende Übersicht zeigt einen entsprechenden Ausschnitt in der alphabetischen Sortierung nach den Einträgen der Spalte "Textwortlemma". (B = Bedeutungsbeschreibung, WB = Wortbildung, WA = Wortart, ST = Sprachstufe, SO = Sprachsoziologie, KA = Kasus, RE = Regionalsprachliche Bindung, SR = bestimmter Sprachraum, NU = Numerus, PH = Phraseologismus, KG = Kompositionsgruppe).

Artikel-lemma	Textwort	Textwortlemma	Paradigma
<Aas>	Ableitung	Ableitung	WB
<ab>	Adv.	Adverb	WA
<ab>	Adverbien	Adverb	WA
<Aal>	ahd.	althochdeutsch	ST
<Aar>	ahd.	althochdeutsch	ST
<Aas>	ahd.	althochdeutsch	ST
<Aas>	ags.	altsächsisch	ST
<ab>	Artikel	Artikel	WA
<Aas>	biblisches	biblisch	SO
<ab>	Dativ	Dativ	KA
<ab>	Dialekten	Dialekt	RE
<ab>	dichterischer	dichterisch	SO

<aasen>	Fischern	Fischer	SO
<Aal>	gemeingerm.	gemeingermanisch	ST
<Aar>	Gen.	Genitiv	KA
<Aar>	Gen.	Genitiv	KA
<Aasseite>	Gerbern	Gerber	SO
<aasen>	Gerberwort	Gerberwort	SO
<Aas>	gewählter	gewählt	SO
<Aar>	goth.	gotisch	ST
<A>	griech.	griechisch	ST
<Aar>	griech.	griechisch	ST
<ab>	griech.	griechisch	ST
<ab>	indogerm.	indogermanisch	ST
<Aasjäger>	Jagenden	Jagender	SO
<aasen>	Jägern	Jäger	SO
<ab>	kaufmänn.	kaufmännisch	SO
<A>	kirchl.	kirchlich	SO
<A>	lat.	lateinisch	ST
<A>	lateinische	lateinisch	ST
<ab>	lat.	lateinisch	ST
<A>	Lernens	Lernen	B
<Aal>	mhd.	mittelhochdeutsch	ST
<Aar>	mhd.	mittelhochdeutsch	ST
<Aas>	mhd.	mittelhochdeutsch	ST
<Aaß>	mhd.	mittelhochdeutsch	ST
<ab>	mhd.	mittelhochdeutsch	ST
<Aaß>	Müller	Müller	SO
<ab>	Mundart	Mundart	RE
<ab>	mundartlich	mundartlich	RE
<ab>	Nomen	Nomen	WA
<ab>	oberd.	oberdeutsch	SR
<ab>	Orts	Ort	B
<Aal>	Plur.	Plural	NU
<Aar>	Plur.	Plural	NU
<Aas>	Plur.	Plural	NU
<ab>	Präp.	Präposition	WA
<Aal>	Redensarten	Redensart	PH
<ab>	sanskr.	Sanskrit	ST
<Aas>	Schimpfwort	Schimpfwort	SO
<Aar>	Schriftsprache	Schriftsprache	SO
<ab>	schwäb.	schwäbisch	SR
<ab>	Schwankens	Schwanken	B
<ab>	schweiz.	schweizerisch	SR
<Aas>	Sing.	Singular	NU
<A>	sprichwörtl.	Sprichwort	PH
<A>	sprichwörtl.	Sprichwort	PH

<Aal>	Sprichw.	Sprichwort	PH
<Ä>	Stockens	Stocken	B
<ab>	Subst.	Substantiv	WA
<ab>	trennbaren	trennbar	WB
<ab>	getrennt	trennen	B
<Aar>	urgerm.	urgermanisch	ST
<Aar>	urverwandt	urverwandt	ST
<ab>	Verb.	Verb	WA
<ab>	Verben	Verb	WA
<ab>	Vorkommens	Vorkommen	B
<Aasjäger>	weidmännische	weidmännisch	SO
<Aas>	westgerm.	westgermanisch	ST
<Aal>	Zusammensetzungen	Zusammensetzung	KG
<Aas>	Zusammensetzungen	Zusammensetzung	KG

Die Textwortlemmatisierung schafft, wie in verschiedenen Fällen erkennbar ist, die Voraussetzung für eine Zusammenfassung terminologischer oder flexivischer Varianten. Es bietet sich an, diese Arbeiten am Index durchzuführen, was eine erhebliche Reduzierung des zu bearbeitenden Wortmaterials mit sich bringt. Die 218 740 beschreibungssprachlichen Elemente reduzieren sich auf 23 620 Elemente im Index, was einer Reduzierung auf rund 10% der ursprünglichen Menge entspricht. Legt man den Wert eines Attributes eines Elementes im Index fest, so ist dies so, als setzte man diesen Wert für alle Elemente, die demselben Zeichenfolgentyp entsprechen. Dieses sehr ökonomische Verfahren ist allerdings nicht unproblematisch. Im Falle von Homographen ist die Arbeit am Index nur bedingt möglich, da nicht wirklich allen Elementen mit dem gleichen Zeichenfolgentyp der gleiche Wert zugeordnet werden darf. Für die 49 beschreibungssprachlichen Elemente *macht* gilt für das Attribut "Textwortlemma", daß ihnen entweder der Wert *machen* vb. oder der Wert *Macht* f. zugeordnet werden muß. Genauso verhält es sich mit Zeichenfolgentypen, die beim Attribut "Paradigma" verschiedene Werte annehmen können. So zeigt sich für das beschreibungssprachliche Element *Müller*, daß diesem entweder der Wert "Definitionswort" oder "sprachsoziologische Markierung" zukommen kann. Für solche Einzelfälle ist es notwendig, jedes Element für sich zu sichten und zu entscheiden.

Deutlich über dem Aufwand für die bisher vorgestellten Kategorialklassifikationen lägen der manuelle Klassifikationsaufwand und die datentechnischen Aufbereitungen für Zugriffe wie den folgenden, in dem zum Phraseologiekennzeichen das objektsprachliche Textelement erfaßt werden soll. Eine automatische Trennung nach Formatabfolgen erweist sich als sehr fehleranfällig.

Artikelstichwort	kategoriale Bezeichnung	objektsprachliches Text
Anfang	sprichwörtlich:	aller anfang ist schwer; guter anfang, halbe arbeit;
Anfangen	sprichwörtlich:	das karnickel hat angefangen,
Anfrage	sprichwörtlich:	eine anfrage ist keine anklage.
Angel	sprichwörtlich und bildlich:	zwischen thür und angel stecken,
Angler	sprichwörtlich:	ein angler musz wissen wann er ziehen soll.
Angreifen	sprichwörtlich:	wer pech angreift, besudelt sich;
Antwort	sprichwörtlich:	keine antwort ist auch eine antwort; es gehört nicht auf alle fragen antwort.
April	sprichwörtlich	april thut was er will;
arg	Rechtssprichwort	kinder folgen der ärgeren hand;
Arm	sprichwörtlich:	grosze herren haben lange arme;
Arm	sprichwörtlich:	er grüszt gern, wo unser herrgott einen arm herausstreckt
armen	im Sprichworte	almosengeben armet nicht;
Armut	in Sprichwörtern:	armut lehrt viel böses; ist ein unwerter gast; der künste mutter; ist keine sünde, schändet nicht

Das Verfahren nähert sich an dieser Stelle freilich wieder den Bedingungen während der Ausgangssituation des Experiments, in der sich der Aufwand für die Restrukturierung einer lexikographischen Überarbeitung nähert.

Noch nicht berücksichtigt wurde bisher die Möglichkeit einer Erschließung der Verbindung von Quellenverzeichnis und Quellenangabe. Diese Verknüpfung erlaubt eine vom Zitiersystem des Wörterbuchs unabhängige, vollständige Zitatnachweisung. Sie erlaubt ferner bei entsprechender Aufbereitung Auskunft über Quellenkorpusstrukturen z. B. im Zusammenhang mit der Einschätzung von Suchzielen. Im speziellen Fall des Heyneschen Wörterbuchs schafft sie auch die Voraussetzung für die Standardisierung der Kurznachweise bei den Zitaten sowie eine Nachdatierung der Belege.

Die vollständige Erfassung des Quellenverzeichnisses zu M. Heynes Wörterbuch erfolgte als eigenes Restrukturierungsmodul. Die Datenstruktur berücksichtigt Vornamen und Familiennamen der Verfasser, Heynes bibliographische Fassung des Publikationstitels und den von ihm angegebenen Kurztitel für die Nachweise bei den Belegen im Wörterbuch. Zusätzlich werden die von Heyne in der Gliederung des Quellenverzeichnisses angelegten Periodenzuordnungen bei jedem Titel notiert. Darüber hinaus enthält der Datenbestand für die Einzeltitel mit Ausnahme der Sammeleditionen entweder ein bibliographisch bzw. aus den vorgefundenen Angaben ermitteltes Entstehungs- oder Ersterscheinungsdatum. Soweit diese Datierungen nicht aus M. Heynes eigenen Angaben stammen, sind sie anhand bibliographischer Hilfsmittel

rekonstruiert worden. Die rekonstruierten Datierungen erscheinen in Klammern. In der Rubrik "Texttyp" wird zwischen Editionen (Ed.), Originalausgaben (Oa.) und Wörterbücher (Wb.) unterschieden. Da M. Heyne in den Quellenangaben für die Belege bei literarischen Werkausgaben oft die Einzeltitel angibt, müssen diese in einer Substruktur separat erfaßt und datiert werden. Ein nicht unbeträchtliches Problem für den Arbeitsaufwand ergibt sich auch aus der z. T. wenig konsequenten Form der angegebenen Kurztitel.

	Vorname	Name	Titelansatz Heyne	Kurztitel Heyne	Texttyp	Datierung	Periodenzuordnung
		O.Vf.	Beowulf, herausgegeben von M. Heyne. 7. Auflage, besorgt von Adolf Socin, Paderborn und Münster 1903.	Beowulf	<Ed>	[8./9. Jh.]	ahd., as., ae.
		O.Vf.	Ezzos Leich s. Müllenhoff-Scherer	Ezzo	<Ed>	[11.Jh.]	ahd., as., ae.
		O.Vf.	Heliand, herausgegeben von Moritz Heyne. 4. Auflage. Paderborn 1905.	Heliand	<Ed>	[<822/40>]	ahd., as., ae.
4	E.	Wildenbruch, von *	Wildenbruch, E. v.*, Die Quitzows, Schauspiel in 4 Acten, 1888.	Wildenbruch	<Oa.>	1888	nhd. heutige Zeit
	...						
5	E.	Wildenbruch, von *	Wildenbruch, E. v.*, Der Generaloberst, Trauerspiel im deutschen Vers, 3. Aufl., 1890.	Wildenbruch	<Oa.>	1890	nhd. heutige Zeit

Eine Redatierung der Belege im Artikelkontext könnte innerhalb des markierten Materials durch Zeichenfolgenaustausch vorgenommen werden.

Die Digitalisierungsversuche mit M. Heynes Deutschem Wörterbuch zeigen, daß die digitale Erfassung vorliegender historischer Wörterbücher mit einer Reihe von spezifischen wörterbuchtypischen Gegebenheiten zu rechnen hat. Bereits eine erste Stufe der Digitalisierung mit Datenerfassung, Textformatierung und erforderlichen Korrekturen erfordert aufgrund fehlender Automatisierbarkeit eine beträchtliche Investition. Ein solches Produkt wäre nur unter dem Blickwinkel der Literaturversorgung nach dem für seine Herstellung erforderlichen Aufwand zu teuer. Für systematische Nutzungen ist es in dieser Aufbereitungsstufe aufgrund der Unstrukturiertheit weitgehend ungeeignet. Um Zugriffe zu ermöglichen, die mit elaborierten digitalen Wörterbuchversionen möglich sind, müßten erhebliche metalexikographische Strukturierungsarbeiten geleistet werden. Aufgrund der diskursiven Textstruktur und verschiedener Stilmerkmale stößt eine solche Strukturierung des Heyneschen Wörterbuchs auch sachlich an Grenzen. Vor allem ist eine konsequente Unter-

scheidung aller Inhaltsebenen kaum möglich, ohne verändernd in den bestehenden Text einzugreifen. Aber auch die sachlich vertretbaren Strukturierungen erfordern einen Arbeitsaufwand, der sich dem für eine inhaltliche Überarbeitung nähert. Die rechnergestützt möglichen Strukturierungen des Wörterbuchtextes erlauben eine Unterscheidung derjenigen Wörterbuchbestandteile, die in der Makrostruktur bzw. in der Mikrostruktur durch typographische Merkmale bestimmt sind. Außer den Artikelgrenzen sind dies nach der Originaltypographie des bearbeiteten Werks vor allem die Ebenen der Stichwörter, der Verfassernamen in Belegen sowie die Ebenen der Objekt- und Metasprache. Die Standardbelege sind zudem aufgrund einer bestimmten Formatsyntax identifizierbar. Eine automatische Kennzeichnung dieser Schichten ist möglich, verlangt aufgrund der verschiedenen Fehlerquellen jedoch nachträgliche manuelle Prüfgänge. Dieses Ergebnis automatischer Textauszeichnung bleibt deutlich hinter den beobachteten elaborierten Standards zurück. Sie kann jedoch schichtspezifische Suchen wirksam unterstützen und auf diese Weise einen Einstieg in systematische Wörterbuchbenutzungen eröffnen. Gleichzeitig bietet sie den Vorteil einer alterungsbeständigen, transferierbaren und plattformneutralen Datenkodierung im TEI-Format. Weitergehende Formen der Restrukturierung und maschinenlesbaren Auszeichnung des Wörterbuchtextes sind nur noch in sehr beschränktem Maß automatisierbar. Hier wäre als eine Möglichkeit der Rechnerunterstützung die temporäre Indexbildung zur weiteren Subklassifikation zu nennen. Generell ist bei solchen weiterführenden Arbeiten mit erheblichem Arbeitsaufwand zu rechnen, der vor allem im Hinblick auf die wissenschaftsgeschichtlichen Implikationen der vorliegenden historischen Wörterbücher eher im Rahmen lexikographischer Überarbeitungen zu diskutieren wäre. Insgesamt zeigen die Experimente am Heyneschen Wörterbuch, daß angesichts der "Goldrauschstimmung", die gegenwärtig Teile der Diskussion um die Erstellung retrospektiv digitalisierter Wörterbücher bestimmt, vor einer Überschätzung der Möglichkeiten gewarnt werden muß. Es war deutlich zu zeigen, daß eine retrospektive Digitalisierung historischer Wörterbücher kaum ohne sprachwissenschaftliche und lexikographische Bearbeitung und in ausschließlicher Stützung auf informatische Kompetenz zu einem befriedigenden Ergebnis führen kann. Es wäre bedauerlich, wenn die sicher zukunftsträchtigen Perspektiven digitaler Wörterbuchbearbeitungen und Wörterbuchnutzung durch negative Erfahrungen mit quick-and-dirty-Produkten dauerhaft beeinträchtigt würden.

Literaturhinweise

Deutsches Wörterbuch. 1854-1971. Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, I-XVI. Leipzig: S. Hirzel.

Deutsches Universalwörterbuch. o. J. Deutsches Universalwörterbuch A-Z. 3. Aufl. Version 1.1, PC-Bibliothek. Mannheim/Wien/Zürich: Duden.

Guidelines. 1994. Guidelines for Electronic Text Encoding and Interchange. Hg. v. C. M. Sperberg-McQueen und L. Burnard. Chicago/Oxford, 1994. Vgl. bes. Kap. 12: Print Dictionaries.

Heyne, M. 1890-1895. *Deutsches Wörterbuch*, I-III. Leipzig: S. Hirzel.

Milan, C. 1998. *Elektronische Lexikographie am Beispiel der Wörterbücher romanischer Sprachen*. Erscheint im Internet unter der Adresse http://www.uni-bamberg de.ba4hi99/milan.htm

Paul, H. 1992. *Deutsches Wörterbuch*. 9. neubearbeitete Aufl. v. H. Henne und G. Objartel unter Mitarbeit v. H. Kämper-Jensen. Tübingen: Niemeyer [CD-Version].

Retrospektive Digitalisierung. 1998. Retrospektive Digitalisierung von Bibliotheksbeständen. Berichte der von der Deutschen Forschungsgemeinschaft einberufenen Facharbeitsgruppen "Inhalt" und "Technik", dbi-materialien 166. Berlin: dbi.

Le Robert Électronique. 1994. Le Robert Électronique Dos-Macintosh-Windows (CD-ROM). Paris: Robert.

Sanders, D. 1860-1865. *Wörterbuch der deutschen Sprache*, I-II. 2. Aufl. Leipzig: O. Wigand.

Weigand, F.L.K. 1909-1910. *Deutsches Wörterbuch*, I-II. 5. Aufl. Gießen: A. Töpelmann.

Nachbemerkung

K. Casemir und M. Schulz danken wir für kritische Durchsicht und Diskussion des Manuskripts, F. M. Wohlers und R. Bohne für die Korrektur und Übersetzung.