Erschienen in: Lemberg, Ingrid/Schröder, Bernhard/Storrer, Angelika (Hrsg.): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. - Tübingen: Niemeyer, 2001. S. 29-52. (Lexicographica Series maior 107)

Ingrid Schmidt, Carolin Müller

Entwicklung eines lexikographischen Modells: Ein neuer Ansatz[1]

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

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

1 Die Grenzen meiner Vorstellungen bedeuten die Grenzen meiner Möglichkeiten

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

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

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

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

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

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

---

[2] Feldweg (1997), 110.

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

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

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

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

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

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

2 Auf dem Weg zu einem neuen Ansatz

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

---

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

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

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

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

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

2.1.1 Ein neues Publikationsmodell

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

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

---

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

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

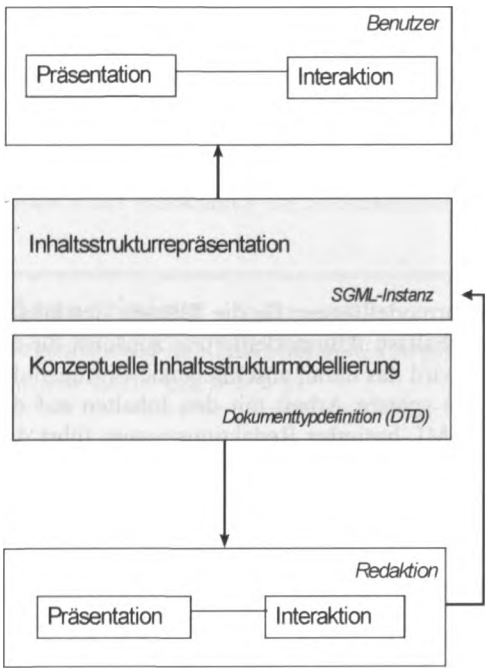

Abb. 1: Publikationsmodell (Schema)

2.1.1.1 Ebene der Inhaltsstrukturmodellierung

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

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

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

2.1.1.2 Redaktionsebene

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

2.1.1.3 Die Ebene der Inhaltsstrukturrepräsentation

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

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

2.1.1.4 Benutzerebene

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

---

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

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

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

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

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

Fläche (Weltrang: 8.): 2 780 400 km²

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

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

Amtssprache: Spanisch

Bruttosozialprodukt 1996 je Einw.: 8380 $

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

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

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

Abb. 2: Argentinien (Ausschnitt)

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

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

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

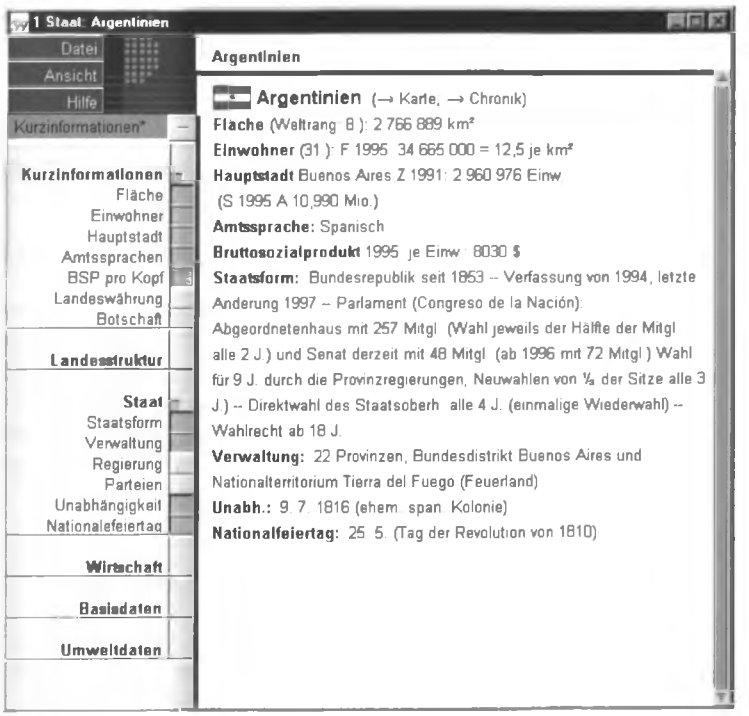

Abb. 3: Benutzerdefinierte Sicht auf Argentinien

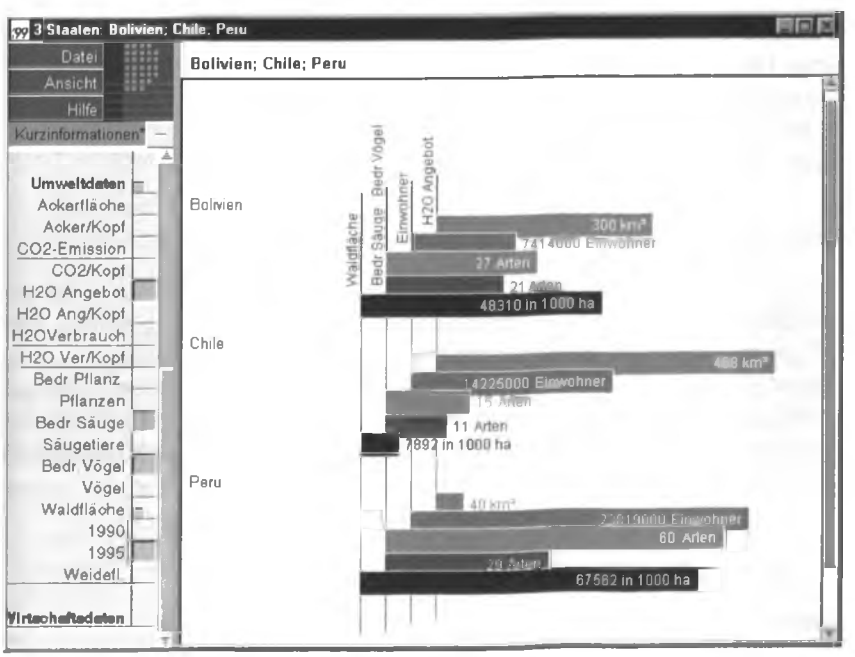

Abb. 4: Grafische Sicht nach Staaten geordnet

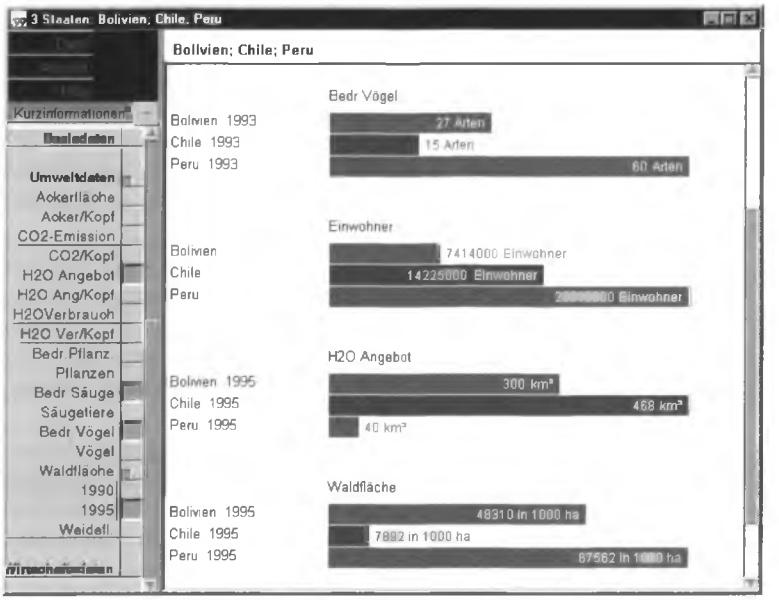

Abb. 5: Grafische Sicht nach Kategorien geordnet

2.1.2 Die zentrale Rolle der Inhaltsstrukturmodellierung

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

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

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

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

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

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

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

2.2.1 Flexibilität der hierarchischen Struktur

Die TEI-Wörterbuchstruktur unterscheidet drei Artikeltypen:

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

Daneben gibt es eine Sonderform des Artikels:

`<re>`	der verwandte Artikel

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

---

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

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

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

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

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

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

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

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

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

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

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

---

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

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

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

2.2.2 Flexibilität der inhaltlichen Struktur

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

Stark unterstrukturiert sind:

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

Weniger stark unterstrukturiert sind:

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

Nicht weiter unterstrukturiert sind:

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

2.2.2.1 Zum Beispiel: Informationen zur Wortform und Grammatik

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2.2.2 Zum Beispiel: Verdichtung

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

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

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

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

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

2.2.3 Verschiedene Sichten auf Wörterbücher

Die TEI unterscheidet grundsätzlich folgende verschiedene Sichten auf Wörterbücher:

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

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

Redaktionelle Sicht mit allen Zeichen für die typographische Darstellung:

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

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

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

Redaktionelle Sicht ohne die Zeichen für die typographische Darstellung:

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

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

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

Lexikographische Sicht:

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

Redaktionelle Sicht (Hauptsicht) mit lexikographischer Sicht:

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

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

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

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

2.3 Mikrostrukturen nach H. E. Wiegand

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

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

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

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

---

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

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

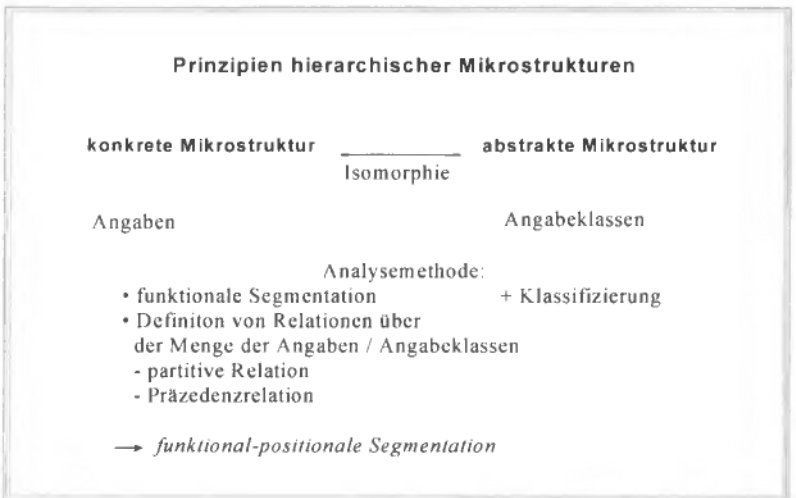

Abb. 6: Prinzipien hierarchischer Mikrostrukturen nach H. E. Wiegand

3 Ein neuer Ansatz

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

3.1 Der Standardisierungs- und Multiple-Media-Aspekt

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

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

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

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

3.2 Der Modularitäts- und Flexibilitätsaspekt

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

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

---

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

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

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

3.3 Der Aspekt der Inhaltsstrukturanalyse

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

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

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

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

---

[24] Vgl. Fußnote 9.

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

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

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

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

4 Vom Ansatz zum Modell

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

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

5 Literatur

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Reichmann, Oskar; Wiegand, Herbert Ernst; Zgusta, Ladislav. Berlin, New York (Handbücher zur Sprach- und Kommunikationswissenschaft 5.1), 462–501.

- (1996): Über die Mediostrukturen bei gedruckten Wörterbüchern. In: Symposium on Lexicography VII. Proceedings of the Seventh Symposium on Lexicography May 5–6, 1994 at the University of Copenhagen. Ed. by Arne Zettersten, Viggo Hjørnager. Tübingen 1996 (Lexicographica, Series Maior), 11–43.

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

Ingrid Schmidt, Heidelberg
Carolin Müller, Heidelberg