Gregor Meder

Zur maschinellen Unterstützung lexikographischer Arbeiten[1]

1. Lexikographische Problemstellungen in der LDV

Zu den ersten Arbeiten der Linguistischen Datenverarbeitung (LDV) gehören auch Versuche im lexikographischen Bereich. Ergebnisse dieser Arbeiten sind (auch heute noch genutzte) maschinenlesbare Textkorpora sowie Indizes und Konkordanzen zu diesen Korpora.

Nach diesen Arbeiten spielt die Lexikographie innerhalb der LDV als eigenständiges Problem nur noch eine untergeordnete Rolle. Im Forschungsüberblick zur LDV von Lenders (1980) kommen lexikographische Arbeiten nur als sog. "Wortkorpora" (auf Datenträger verfügbare Wörterbücher) oder als Dokumentation des in den "Sprach- oder Autorenkorpora" vorkommenden Belegmaterials vor. (vgl. Lenders 1980, 231-234). Auch im HSK-Band "Computerlinguistik" (Bátori/Lenders/Putschke 1989) wird der Bereich "Computer-Aided Lexicography" nur unter den Aspekten "Konkordanzen und Indizes" (Jones/Sondrup 1989) und "Wort-Datenbanken und maschinelle Wörterbücher" (Calzolari 1989) abgehandelt, wobei sich Calzolari (1989, 518f) am Ende ihrer Überlegungen Gedanken macht über den Nutzen lexikalischer Datenbanken für die Lexikographie.

Eine Durchsicht der einschlägigen Literatur zeigt, daß eine Diskussion im Problemfeld "Computerlexikographie" überwiegend in folgenden Bereichen stattfindet:

(1) Theoretische Fundierung des Computereinsatzes in der Lexikographie.
Hier werden insbesondere Fragen der Begründung und Praktikabilität von Datenbanken im lexikographischen Arbeitsprozeß diskutiert (vgl. Wiegand 1986). Nur Schaeder (1986) untersucht Fragen des Einsatzes von Computern in der klassischen Lexikographie.

---

[1] Bei dem vorliegenden Beitrag handelt es sich um eine erweiterte Fassung eines Vortrags im "Lingvistisk Kollokvium" der Handelhojskolen i Århus im Jahre 1988. Erste Ideen habe ich bereits 1984 für einen DFG-Antrag (vgl. Mugdan 1985) entwickelt und auf der Euralex Jahrestagung in Zürich 1986 vorgetragen.

(2) Lexikographische Datenbanken
Diskutiert werden hier Fragen der Anlage von lexikographischen Datenbanken. Immer wieder wird untersucht, wie Daten über Sprache in Datenbanken abgelegt und wieder aufgefunden werden können (vgl. z.B. Domenig 1987). Ausgangspunkt sind oft die Daten, die durch die Auswertung von maschinenlesbaren Textkorpora gewonnen werden (vgl. Brückner 1982).

(3) Lexikographische Komponenten in sprachverarbeitenden Systemen.
Sprachanalyse- und -syntheseprogramme sowie Programmsysteme zur maschinellen Übersetzung verfügen über lexikographische Komponenten. Fragen der Organisation dieser Komponenten ist derzeit Hauptgegenstand der Diskussion innerhalb der Computerlexikographie. Methoden der Produktion von Wörterbüchern sind hier Ausgangspunkt nicht jedoch Ziel der Überlegungen (vgl. Heid 1988).

(4) Konkordanzen und Indizes
Damals wie heute ist der Computer interessant für die Herstellung und Verwaltung von Konkordanzen und Indizes größerer Textkorpora. Im Lichte verfeinerter Programmiertechniken, immer leistungsfähigerer Computer und größerer Speicherkapazitäten werden stets neue Korpora und dazugehörige Konkordanzen und Indizes vorgestellt.

(5) Berichte aus Wörterbuchprojekten
Die Wörterbuchmacher unserer Tage erkennen die Zeichen der Zeit und setzen bei der Herstellung ihrer Wörterbücher Computer ein. Von der (meta-) lexikographischen Diskussion, die in dieser Hinsicht keine praktischen Ergebnisse gezeitigt hat, und der LDV allein gelassen, entwickeln sie für ihre Wörterbuchprojekte spezifische Arbeitsumgebungen auf ihren Computern, die auf andere Projekte meist nicht übertragbar sind.

Diese kleine Übersicht zeigt, daß systematische Überlegungen zum Computereinsatz in der Lexikographie sowohl innerhalb der (meta-) lexikographischen Diskussion als auch innerhalb der LDV derzeit fehlen. Allein Calzolari (1989) gibt einen kleinen Ausblick darauf, wo der Einsatz von Computern den Lexikographen entlasten könnte. Gerade die Arbeiten der letzten Gruppe zeigen jedoch, wie sehr systematische Überlegungen zu diesem Thema ein Desiderat darstellen.

2. Computer in der Lexikographie

Die Berichte aus den Wörterbuchprojekten und die internationale Übersicht von Keitz (1982) zeigen, daß der Computer zu unterschiedlichen Zwecken in Wörterbuchprojekten eingesetzt wird. Neben der Herstellung von Konkordanzen und Indizes wird im einfachsten Fall der Computer wie eine Schreibmaschine zum Schreiben der Wörterbucheinträge genutzt. Oft werden die Wörterbucheinträge in einem Datenbanksystem gespeichert. Im Vordergrund steht dabei immer die Möglichkeit, mithilfe des Computers einzelne Teile des Wörterbuchs (selbst kurz vor dem Druck noch) ohne großen Aufwand überarbeiten zu können. Eingesetzt wird dabei i.d.R. sog. Standardsoftware, die, wenn überhaupt, nur mühsam an die Erfordernisse des lexikographischen Arbeitsprozesses angepaßt werden kann.

Im Folgenden sollen Ideen für ein Programmsystems entwickelt werden, das über den bisherigen Einsatz von Computern als reines Speichermedium für das fertige Manuskript (Textverarbeitung) oder als Datenbank für die Einträge hinaus den Computer zum *interaktiven Hilfsmittel* schon bei der *Erstellung* des Wörterbuchs macht.

Dabei geht es, das soll an dieser Stelle nochmals betont werden, um die maschinelle Unterstützung bei der Erstellung konventioneller Wörterbücher (aus Papier und in Form von Büchern). Diskutiert werden soll demnach ein Programmsystem innerhalb einer *Lexikographie mit dem Computer* im Unterschied zu einer *Lexikographie für den Computer* wie sie z.B. im Rahmen der Bereiche (2) und (3) der Übersicht betrieben wird (zu dieser Terminologie vgl. Schaeder 1986, 253)[2].

2.1 Lexikographische Arbeitsschritte

Die Arbeit an einem Wörterbuch verläuft in verschiedenen Phasen. Ein Programmsystem, das den Lexikographen bei seiner Arbeit unterstützen soll, ohne daß dieser sich dabei von seiner bisherigen Praxis weit entfernt, muß diese Arbeitsschritte abbilden. Die Arbeit an einem Wör-

[2] Dies soll nicht bedeuten, daß die Daten, wenn sie einmal maschinenlesbar vorliegen nicht auch in solchen Systemen benutzt werden können. Der umgekehrte Weg scheint allerdings schwerlich möglich. — Im Unterschied zu Schaeder (1986) wird hier die Meinung vertreten, daß Lexikographie mit dem Computer immer auf Realisationen von Wörterbüchern außerhalb von Rechnersystemen zielt. Wörterbücher als Datenbanken sind (derzeit noch) für menschliche Benutzer uneffektiv. Insofern sollte die Aufmerksamkeit vielmehr auf die maschinelle Unterstützung bei der Produktion herkömmlicher Wörterbücher gerichtet werden.

terbuch läßt sich in folgende Arbeitsschritte gliedern:

(a) Datenerhebung
Gleichgültig vor welchem theoretischen Hintergrund ein Wörterbuch entsteht: Am Anfang der lexikographischen Arbeit steht das Sammeln der Daten. Im einfachsten Falle wird man sich auf die Auswertung vorhandener Wörterbücher stützen. Selbsterhobenes Material ist meist eine Sprachkartei, in der unter dem entsprechenden Lemma Belege verzettelt sind, die ein bestimmtes sprachliches Phänomen illustrieren sollen. Eine weitere Form der Datenerhebung für lexikographische Zwecke ist die Erstellung und Auswertung von Textkorpora[3].

(b) Daten vorbereiten
Die Daten werden je nach Art der Erhebung gesichtet und ggf. gezählt; Frequenzlisten sind für Wörterbücher auf kontrollierter empirischer Basis unerläßlich. Ggf. werden Vollformenlisten aus Textkorpora lemmatisiert.

(c) Daten auswählen
In diesem Arbeitsschritt werden die Stichwörter des geplanten Wörterbuchs ausgewählt. Das wird im Allgemeinen auf der Grundlage der wie auch immer erhobenen und aufbereiteten Daten erfolgen.

(d) Daten beschreiben
Erst in diesem Arbeitsschritt beginnt die lexikographische Tätigkeit im engeren Sinne. Jetzt werden die Einträge geschrieben und auf der gewählten Datengrundlage die grammatischen Angaben verfaßt, die Bedeutungen beschrieben und Belege bzw. Beispiele ausgewählt. Dies alles geschieht auf der Grundlage der vorher erstellten Beschreibungskonventionen.

(e) Datenpräsentation
Schließlich wird das Wörterbuch, meist außerhalb der Kontrolle des Lexikographen, gedruckt.

Diese Arbeitsschritte lassen auf ganz unterschiedliche Weise den Einsatz von Computern zu. Vor der Entwicklung eines lexikographischen

---

[3] Eine Beurteilung der einzelnen Formen der lexikographischen Datenerhebung soll an dieser Stelle nicht geleistet werden. Hier sei nur der Hinweis erlaubt, daß dem Mangel an empirischer Fundierung mancher Wörterbücher durch den Einsatz von Computern in der Datenerhebung begegnet werden kann. Zur grundsätzlichen Notwendigkeit einer korpusorientierten Lexikographie vgl. z.B. Mugdan (1985, 196ff).

Programmsystems steht deshalb die gründliche Analyse dieser Arbeitsschritte im Hinblick auf die Möglichkeit ihrer maschinellen Unterstützung.

2.2 Automatisierung und maschinelle Unterstützung lexikographischer Arbeiten

Unterzieht man die unter 2.1 genannten Arbeitsschritte einer näheren Betrachtung, so bietet sich der Einsatz eines Computers an verschiedenen Stellen in unterschiedlicher Intensität an.

Die Erstellung von Textkorpora (Arbeitsschritt (a), Datensammlung) kann heute (fast) vollautomatisch geschehen. Moderne OCR-Geräte und Texterkennungssoftware bringen jede Art von gedruckten Texten in maschinenlesbare Form. Daneben liegen viele Texte schon vor ihrer Publikation in maschinenlesbarer Form vor. Für Wörterbuchprojekte kann die Datenbasis in Form von Textkorpora leicht erstellt werden. Auch vorhandene Belegsammlungen sind auf diese Weise leicht in maschinenlesbare Form zu bringen[4].

Auch die Vorbereitung der Daten für lexikographische Zwecke kann bis zu einem gewissen Grade vollautomatisch geleistet werden. Dies gilt vor allem für die Erstellung von Frequenzlisten und Satzkonkordanzen. Dafür steht heute eine ganze Reihe von Software zur Verfügung[5]. Bei Erstellung dieser Frequenzlisten und Satzkonkordanzen ist darauf zu achten, daß die auf diese Weise aufbereiteten Daten ohne Umkodierungsaufwand in die nächsten Arbeitsschritte übernommen werden können.

Zur *Datenaufbereitung* gehört in gewisser Weise auch die Lemmatisierung der Vollformen. Diese Arbeit kann nicht vollautomatisch erfolgen (vgl. Willée 1979). Selbst mit großem programmtechnischem Aufwand und unter Berücksichtigung des syntaktischen Kontextes bleibt immer ein Rest von Wortformen, die sich nicht eindeutig einer Grundform zuordnen lassen.

---

[4] Mit dem derzeitigen Stand der Technik ist dies nur mit maschinengeschriebenen Karteikarten effektiv möglich. — Ob die Übertragung von Belegsammlungen in Datenbanken sinnvoll ist, wenn mit geringerem Aufwand große Textkorpora erstellt werden können, ist fraglich.

[5] Eine Zusammenstellung solcher Software z.T. für Großrechner enthält Brustkern/Lenders/Willée (1981). Eine Darstellung des Textretrievalprogramms “WordCruncher” gibt Kammer (1989). Weitere Konkordanz- und Indexsoftware stellen Jones/Sondrup (1989) vor.

Da es bei diesem *lexikographischen* Arbeitsschritt nicht darum geht, die Wörter eines Textkorpus restfrei ihrem Lexem zuzuordnen, sondern darum, die Datenbasis für das Wörterbuch zur Verfügung zu stellen, scheint es sinnvoller, die Wortform-Lexem-Zuordnung der Datenbasis erst nach der Datenauswahl vorzunehmen[6].

Die *Datenauswahl* ist in viel geringerem Maße automatisierbar als die Erstellung von Textkorpora und Konkordanzen. Neben der Auswahl der Informationen, die in den Einträgen stehen sollen, ist die Frage nach den Lemmata im Wörterbuch der wichtigste Teil der Datenauswahl. Die Frage, wie dies geschieht und welches die Gründe für oder gegen die Aufnahme bestimmter Lemmata in ein Wörterbuch sind, gehört zu den großen Geheimnissen in der Lexikographie[7].

Der Einsatz eines Computers erlaubt zumindest für das Wörterbuch relevante Wortlisten (die zuvor ggf. mithilfe von OCR-Geräten in maschinenlesbare Form gebracht werden) zu vergleichen und das Ergebnis dieses Vergleiches der Entscheidung bei der Lemmaauswahl zugrunde zu legen[8]. Durch geschickte Markierung einzelner Lemmata kann sichergestellt werden, daß bestimmte Wortgruppen, wie Wochentage, Monatsnamen etc. vollständig im Wörterbuch verzeichnet sind.

Steht die Lemmaliste für das Wörterbuch fest, so kann auch das Problem der Wortform-Lexem-Zuordnung auf ökonomische Weise gelöst werden: Damit für die spätere eigentliche lexikographische Arbeit für jedes Lemma alle entsprechenden Informationen aus dem Korpus zur Verfügung stehen, müssen zu jedem für das Wörterbuch vorgesehenen Lemma die Belegstellen und die Frequenzen für die Wortformen in einer Datenbank zusammengefaßt werden. Diese Daten müssen für jedes Lemma einzeln abrufbar sein.

---

[6] Effektiver wäre es natürlich, wenn das Datenmaterial eines Textkorpus unabhängig von einem lexikographischen Projekt lemmatisiert zur Verfügung stehen würde. Bei der Datenaufbereitung für ein einzelnes Wörterbuchprojekt ist es jedoch wenig effektiv, das ganze Korpus zu lemmatisieren.

[7] Eine exemplarische Kritik zu diesem Thema ist nachzulesen in Bergenholtz/Mugdan (1986), grundsätzlich wird dieses Thema behandelt in Bergenholtz (1989) und lobenswerterweise offengelegt werden die Überlegungen zur Lemmaselektion für ein konkretes Wörterbuchprojekt in Bergenholtz (1990).

[8] Manuell ist es sehr mühsam, lange Wortlisten miteinander zu vergleichen, wie dem Werkstattbericht zur Lemmaauswahl für das Deutsch-Madagassische Wörterbuch (Stegemann 1988) zu entnehmen ist.

Bei der lexikographischen Arbeit i.e.S., der *Datenbeschreibung*, kann der Computer als strukturierendes Hilfsmittel dienen. Das Programm kann die vereinbarte Mikrostruktur des Wörterbucheintrags vorgeben und für die Einhaltung der Kodierungsvereinbarungen sorgen. Das Programm kann dem Lexikographen beispielsweise vorgeben, welche Wortarten im gerade bearbeiteten Wörterbuch zugelassen sind und welche Kodierungen dafür vorgesehen sind.

Neben diesen eher formalen Hilfen scheint es mir wichtiger, daß bei diesem Arbeitsschritt dem Lexikographen ohne große Mühe der Zugriff auf die Datenbasis ermöglicht werden muß: Auf Tastendruck müssen Frequenzlisten, Belege und Korpusauszüge zur Verfügung stehen. Der Lexikograph muß die Arbeit am Eintrag unterbrechen können und Belege markieren, klassifizieren und zählen (lassen) können. Zu diesem Zweck müssen hier eine Reihe von Datenbank- und Statistikfunktionen zur Verfügung stehen. Die Ergebnisse dieser Analyse und bestimmte Belege müssen ohne größeren Aufwand, d.h. ohne sie "eintippen" zu müssen, in den Wörterbucheintrag übernommen werden können.

Für die lexikographische Arbeit mit dem Computer ist also ein elektronischer Arbeitsplatz ("workbench") zu entwickeln, der es einerseits erlaubt, den Wörterbuchartikel zu schreiben und andererseits die Arbeit am sprachlichen Material vereinfacht.

Die Vorbereitung des Wörterbuchs zum Druck wird schon seit längerem mit Computerunterstützung vorgenommen. Hier werden die vom Lexikographen zugeordneten Informationsklassen in bestimmte Textauszeichnungsmerkmale (Schrifttyp, -schnitt, -größe etc.) übersetzt. Dadurch, daß der Lexikograph die Textattribute nicht selbst zuordnet, sondern sie in einer Art (kontextfreien) Grammatik den einzelnen Informationsklassen zugeordnet werden, läßt sich in der Darstellungsweise eine vollständige Konsistenz erreichen.

3. Ein modulares Modellsystem zur Unterstützung lexikographischer Arbeiten

Nach den theoretischen Überlegungen stellt sich natürlich die Frage, wie ein solches lexikographisches Programmsystem konkret aussehen kann. Diese Frage kann hier nicht endgültig beantwortet werden. Es soll vielmehr ein Modellsystem vorgestellt werden, das bislang vollständig noch nicht existiert und folglich auch noch nicht in einem konkreten lexikographischen Projekt eingesetzt werden konnte. Es stellt vielmehr eine Konkretisierung der theoretischen Überlegungen dar, die in der Ausein-

andersetzung mit der lexikographischen Praxis entstanden ist. Teile des Programmsystems wurden allerdings programmiert und leisten gute Dienste im Rahmen von Wortschatzuntersuchungen.

Für die Überlegungen ist es jedoch nicht wichtig, ob das Modellsystem in der Weise schon realisiert wurde oder ob es sinnvoll ist, die Module in genau dieser Aufteilung zu realisieren. Im Folgenden sollen die Probleme eines lexikographischen Programmsystems vielmehr exemplarisch diskutiert werden.

Das Modellsystem besteht aus fünf Moduln, die sich den einzelnen lexikographischen Arbeitsschritten zuordnen lassen.

Datenaufbereitung	MakeFRQ
	MakeKWIC
Datenauswahl	MakeLEM
Datenbeschreibung	MakeLEX
Datenpräsentation	MakeWB

Das Zusammenspiel der Komponenten des System wird in Abb. 1 dargestellt.

Fig. 1
(45% formindskelse)

Abb. 1 Zusammenwirken der Systemkomponenten

3.1. Komponenten des Systems und ihre Zusammenarbeit

Die Moduln **MakeFRQ** und **MakeKWIC** gewinnen aus einem Textkorpus bekannter Struktur die Vollformenfrequenzlisten und die Satzkonkordanzen. Bis für das Programmsystem eigene Programme zur Herstellung von Frequenz- und Beleglisten zur Verfügung stehen, können vorhandene Programme, deren Ausgabeformat bekannt ist, genutzt werden. Wichtig ist nur, daß die Ausgabe der Programme im nächsten Arbeitsschritt von MakeLEM verwertet werden können. Die beiden Komponenten sind abhängig vom Format des verwendeten Textkorpus, so daß es günstiger sein kann, die ggf. mit dem Korpus gelieferte Auswertungssoftware zu verwenden. Wird ein eigenes Korpus aufgebaut, so ist darauf zu achten, daß diese beiden Systemkomponenten das Korpus möglichst effektiv auswerten und für die Eingabe in MakeLEM vorbereiten können.

Fig. 2
(45% formindskelse)

Abb. 2 Struktur des Moduls MakeLEM

Das Modul **MakeLEM** (vgl. Abb. 2) unterstützt den Lexikographen bei zwei wichtigen Arbeiten. Zum einen wird durch Rückgriff auf verschiedene (maschinenlesbare) Wortlisten und die Vollformenfrequenzliste die Lemmaliste für das Wörterbuch festgelegt. Zum anderen werden hier die Satzkonkordanzen und die Frequenzlisten lemmatisiert.

Die Selektion eines Lemmas für das Wörterbuch kann dann abhängig gemacht werden vom Vorkommen des Lemmas in bestimmten Wortlisten oder in einer bestimmten Anzahl von Wortlisten und einer bestimmten Häufigkeit im Textkorpus.

Letzteres zeigt das Dilemma dieses Arbeitsschrittes: Die Häufigkeit bestimmter Lexeme im Textkorpus kann erst dann angegeben werden, wenn die Wörter des Textkorpus ihrem Lexem zugeordnet sind. Andererseits sollte aus ökonomischen Gründen erst dann lemmatisiert werden, wenn festgelegt ist, welche Lemmata ins Wörterbuch aufgenommen werden. Diese Entscheidung sollte jedoch mithilfe der Korpusdaten erleichtert werden. Hier hilft für den Anfang (bis genügend lemmatisierte Textkorpora zur Verfügung stehen[9]) nur, daß zunächst aufgrund anderer Kriterien eine vorläufige Lemmaliste erstellt wird. Diese Lemmata werden den Wortformen aus dem Korpus zugeordnet, um dann ggf. nach Häufigkeitskriterien zu entscheiden, welches Lemma ins Wörterbuch kommt.

Das Modul arbeitet in zwei Schritten: Zunächst werden die Vollformenfrequenzliste und die Satzkonkordanz interaktiv lemmatisiert: Der Benutzer erhält vom System Lemmatisierungsvorschläge, die akzeptiert oder verändert werden können. Es wird dann eine Datenbank erzeugt, in der über das Lemma auf die Belege zu den Wortformen, auf deren Frequenzen und auf die Belegstelle im Korpus zugegriffen werden kann.

Im zweiten Schritt wird eine Liste aller Lemmata in den zur Verfügung stehenden Wortlisten und der lemmatisierten Frequenzliste verzeichnet sind, erzeugt. Der Benutzer kann dann interaktiv eine Lemmaliste nach bestimmten Kriterien zusammenstellen lassen. Es wird dann eine Lemmaliste zur weiteren Bearbeitung mit MakeLEX und bei Bedarf eine Statistik der Lemmaliste (Anzahl der Lemmata, Häufigkeit der Lemmata im Textkorpus, Verteilung der Lemmata auf die Quelllsiten etc.) erstellt.

Die erzeugte Datenbank aus Beleg- und Frequenzliste wird in ein Format überführt, das es erlaubt, die Belege für ein Lemma nach verschiedenen Kriterien zu makieren, zu sortieren und zu zählen.

---

[9] In diesem Zusammenhang scheint es mir sinnvoll über Sprachdatenbanken (lexikalische oder lexikographische Datenbanken) nachzudenken: Sprachliche Daten (Frequenzlisten und Satzkonkordanzen) in einer Datenbank stehen unabhängig von bestimmten Wörterbuchprojekten für die lexikographische Auswertung zur Verfügung.

Abb. 3 Struktur des Moduls MakeLEX

Mit der Komponente **MakeLEX** (vgl. Abb. 3) wird die eigentliche lexikographische Arbeit unterstützt. Im Wechsel zwischen zwei Arbeitsumgebungen wird der Wörterbuchartikel erstellt. Der Lexikograph arbeitet sich durch die Lemmaliste und kann in der Arbeitsumgebung COMPILER in der Datenbank die Belege und Frequenzlisten bearbeiten: Abhängig vom bearbeiteten Wörterbuch wird das sprachliche Material durchgesehen, geordnet, gezählt und nach semantischen Gesichtspunkten (für die Bedeutungserklärungen und die semantische Gliederung der Wörterbuchartitels) und nach sytaktischen oder morphologischen Gesichtspunkten (für die grammatischen Angaben) klassifiziert. Kollokationen können aufgefunden und zugeordnet werden. Diese Komponente entlastet den Lexikographen vom vielfachen Aufsuchen und Umsortieren seines Materials. Zu jeder Zeit stehen dem Lexikographen alle Informationen zur Verfügung. Anstrebenswert ist, daß auch Material, das für andere Wörterbücher erstellt wurde, hier zur Verfügung steht.

Die Arbeitsumgebung EDITOR gibt die Mikrostruktur des Wörterbuchartikels vor. Dem Lexikographen werden für die einzelnen Lemmatypen Eingabemasken vorgegeben und Informationen abhängig von den definierten Strukturpositionen abverlangt. Es wird dafür gesorgt, daß die Wörterbuchartikel einheitlich aufgebaut sind.

Die für den Wörterbucheintrag ausgewählten Belege werden direkt aus der Datenbank übernommen, ohne daß sie erneut geschrieben werden

müssen. Auch Häufigkeitsangaben werden direkt aus dem Datenmaterial übernommen[10].

Selbstverständlich besteht für den Lexikographen hier die Möglichkeit, freien Text in den Wörterbucheintrag einzufügen, um z.B. Belege zu adaptieren. Er kann seinen Eintrag für spätere Zwecke kommentieren und ihn auszudrucken.

Der Lexikograph braucht sich zu diesem Zeitpunkt allerdings keine Gedanken zu machen über die Kodierung der im Wörterbucheintrag enthaltenen Informationen.

MakeLEX erzeugt schließlich das Wörterbuch aus drei Teilen, die von MakeWB als Wörterbuch ausgedruckt werden können: .IND enthält den Zugriffsindex für das Wörterbuch, .BBS enthält die Belege und Beispiele mit Verweis auf den zugehörigen WB-Artikel und .LEX enthält die einzelnen Wörterbucheinträge (ohne die Belege und Beispiele).

Aus den mithilfe des Moduls MakeLEX erzeugten Dateien mit den Wörterbucheinträgen und den zugeordneten Belegen und Beispielen sowie den Kodierungsvereinbarungen erstellt **MakeWB** das Wörterbuch. MakeWB ist ein Modul, das die Ausgabe für bestimmte Drucker steuert und den Informationsklassen im Wörterbucheintrag bestimmten Textauszeichnungsmerkmalen zuordnet. Diese Komponente wurde ausgelagert, damit die Druckausgabe sowohl unabhängig vom verwendetet Ausgabemedium als auch unabhängig von der eigentlichen lexikographischen Arbeit mit MakeLEX erfolgen kann.

4. Probleme und Ausblick

Soweit die Vorstellung des Modellsystems. Es sollten hier nicht alle Aspekte und Probleme eines Programmsystems zur Unterstützung lexikographischer Arbeiten dargestellt werden. Insbesondere kommt er hier nicht darauf an, die Probleme der Programmierung eines solchen Systems in einer bestimmten Programmiersprache zu diskutieren. Trotzdem sollen hier einige grundsätzlichen Probleme bei der Realisierung genannt werden.

Ein Problem besteht in der enorm hohen Speicherkapazität, die die Verfügbarkeit der sprachlichen Daten (Korpus, Frequenz- und Beleglisten) erforderlich macht. Auch die Speicherung der Wörterbuchartikel

---

[10] Zu den Einzelheiten des Zusammenspiels von EDITOR und COMPILER vgl. auch Meder (1990).

kann leicht zu Kapazitätsproblemen führen. Es ist davon auszugehen, daß für die Realisierung des Systems einige hundert Megabyte externer Speicherkapazität zur Verfügung stehen muß.

Ein weiteres Problem stellt der nicht vorhersagbare Umfang eines Wörterbucheintrags dar. Neben dem Problem der variablen Datensatzlänge (beliebig groß, jedoch nur soviel gespeichert, wie benötigt wird) ist das Problem der rekursiven Eintragstruktur zu lösen. Es zeigt sich nämlich, daß Wörterbucheinträge für jede Semem- oder Strukturposition dieselben Informationsklassen enthalten (ein Wörterbucheintrag im Wörterbucheintrag). Dies ist für den Aufbau der Datenstruktur ein Problem, weil nicht vorhersagbar ist, wie tief sich die Strukturen verschachteln.

Bei großen Wörterbüchern tritt schließlich das Problem der Zugriffsgeschwindigkeit auf. Hier bietet die Informatik jedoch geeignete Such- und Sortieralgorithmen an, die ihre Arbeit in akzeptabler Zeit verrichten.

Zum Schluß sei noch das Problem der Benutzeroberfläche genannt. Es versteht sich im Zeitalter der graphischen Benutzeroberflächen für Computerprogramme fast von selbst, daß ein lexikographisches Programmsystem einfach zu bedienen sein muß und vor allem durch seinen Einsatz nicht von der eigentlichen lexikographischen Arbeit ablenkt. Besonders im Modul MakeLEX muß der Übergang zwischen den Programmteilen problemlos zu bewältigen sein.

Falls mehrere Lexikographen an einem Wörterbuch arbeiten sollen wäre es sinnvoll, mit einem netzwerkfähigen System zu arbeiten, damit alle Bearbeiter gleichzeitig Zugriff auf alle Daten haben können, ohne daß die Daten mehrfach abgespeichert werden müssen.

Es zeigt sich also, daß der Entwurf und die Realisierung eines Programmsystems zur maschinellen Unterstützung lexikographischer Arbeiten ein so komplexes Unternehmen ist, daß nur Programmierer, Informatiker und Lexikographen gemeinsam daran arbeiten können. Versuche von Einzelnen haben in der Vergangenheit zwar brauchbare Teillösungen hervorgebracht, eine gute elektronische Arbeitsumgebung kann jedoch nur von kompetenten Fachleuten gemeinsam entwickelt werden.

5. Literatur

Bátori, István S./Lenders, Winfried/Putschke, Wolfgang (1989): *Computational Linguistics. Computerlinguistik. An International Handbook on Computer Oriented Language Research and Applications*. Berlin/New York: de Gruyter.

Bergenholtz, Henning (1989): "Probleme der Selektion im allgemeinen einsprachigen Wörterbuch." In: Hausmann, F. J. / Reichmann, O. / Wiegand H. E. (Hrsg.): *Wörterbücher. Dictionaries. Dictionaires. Ein internationales Handbuch zur Lexikographie*. Berlin/New York: de Gruyter 1989, 772-779.

Bergenholtz, Henning (1990): "Lemmaselektion in zweisprachigen Wörterbüchern." In: Meder, G./Dörner, A. (Hrsg.): Worte, *Wörter Wörterbücher. Lexikographische Beiträge zum Essener Linguistischen Kolloquium*. Erscheint in *Lexicographica Series Maior*. Tübingen: Niemeyer 1990. Im Druck.

Bergenholtz, Henning/Mugdan, Joachim (1986): "Der neue Super-Duden. Die authentische Darstellung des deutschen Wortschatzes?" In: Wiegand, H.E. (Hrsg.): *Studien zur neuhochdeutschen Lexikographie VI.1*. Hildesheim u.a.: Olms. 1986 (= Germanistische Linguistik 64-86, 1986), 1-149.

Brückner, Tobias (1982): "Der interaktive Zugriff auf die Textdatei der Lexikographischen Datenbank (LEDA)." In: *Sprache und Datenverarbeitung*, Heft 6 (1982), 16-36.

Brustkern, Jan/Lenders, Winfried/Willée, Gerd (1981): *Handbuch der Programmbibliothek zur linguistischen und philologischen Textverarbeitung*. Opladen: Westdeutscher Verlag. (= Forschungsberichte des Landes NRW, Nr. 2939)

Calzolari, Nicoletta (1989): "Computer-Aided Lexicography: Dictionaries, and Word Data Bases." In: Bátori, István S./Lenders, Winfried/Putschke, Wolfgang (1989), 510-519.

Domenig, Marc (1987): *Entwurf eines dedizierten Datenbanksystems für Lexika*. Tübingen: Niemeyer.

Heid, Ulrich (1988): "Wörterbücher in sprachverarbeitenden Systemen, speziell maschineller Übersetzung." In: LDV-Forum 5 (1988), 64-84.

Jones, Randall L./Sondrup, Steven P. (1989): "Computer-Aided Lexicography: Indexes and Concordances." In: Bátori, István S./Lenders, Winfried/Putschke, Wolfgang (1989), 490-509.

Kammer, Manfred (1989): "WordCruncher: Problems of Multilingual Usage." In: *Literary and Linguistic Computing* 4 (1989), 135-140.

Keitz, Wolfgang von (1982): "Projekte zur maschinellen Lexikographie." In: *Sprache und Datenverarbeitung* Heft 1/2 (1982), 11-27.

Lenders, Winfried (1980): "Linguistische Datenverarbeitung — Stand der Forschung." In: *Deutsche Sprache* 8 (1980), 213-264.

Meder, Gregor (1990): "Flexionsangaben in Wörterbüchern. Zur rechnergestützten Erstellung von Flexionsangaben in Wörterbüchern." Figge, U.L. (Hrsg.): *Akten der Internationalen Fachkonferenz "Zweisprachige Lexikographie: Deutsch Portugiesisch/Portugiesisch-Deutsch."* Erscheint in *Lexicographica Series Maior*. Tübingen: Niemeyer 1990. Im Druck.

Mugdan, Joachim (1985): “Pläne für ein grammatisches Wörterbuch. Ein Werkstattbericht.” In: Bergenholtz, H./Mugdan, J. (Hrsg.): *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch*. Tübingen: Niemeyer 1985, 187-224.

Schaeder, Burkhard (1986): “Die Rolle des Rechners in der Lexikographie.” In: Wiegand, H.E. (Hrsg.): *Studien zur neuhochdeutschen Lexikographie VI.1*. Hildesheim u.a.: Olms. 1986 (= Germanistische Linguistik 64-86, 1986), 243-277.

Stegemann, Sabine (1988): *Werkstattbericht: Lemmaselektion für das DtMaWb. DFG-Forschungsbericht*. Ms. Essen.

Wiegand, Herbert Ernst (1986): “Metalexicography. A Data Bank for Contemporary German.” In: *Interdisciplinary Science Reviews* 11 (1986), 122-131.

Willée, Gerd (1979): “LEMMA — Ein Programmsystem zur automatischen Lemmatisierung deutscher Wortformen.” In: *Sprache und Datenverarbeitung* Heft 1/2 (1979), 45-60.