Niederdeutsches Wort

KLEINE BEITRÄGE ZUR NIEDERDEUTSCHEN MUNDART- UND NAMENKUNDE

begründet von
WILLIAM FOERSTE †

herausgegeben von
DIETRICH HOFMANN

Band 9 · Heft 1/2
1969

VERLAG ASCHENDORFF · MÜNSTER

Das NIEDERDEUTSCHE WORT erscheint als Organ der Volkskundlichen Kommission, Abt. Mundart- und Namenforschung (Westfälisches Wörterbuch, Westfälisches Flurnamenarchiv), in Münster/Westfalen mit Unterstützung der Niederdeutschen Abteilung des Germanistischen Instituts der Universität Münster. Die Zeitschrift wird jährlich in einem Band oder zwei Heften von insgesamt 120-130 Seiten herausgegeben.

Herausgeber: Prof. Dr. DIETRICH HOFMANN
Redaktionelle Arbeiten: Dr. IRMGARD SIMON

44 Münster, Domplatz 20

© Aschendorff, Münster Westfalen, 1970 • Printed in Germany
Alle Rechte, auch die des auszugsweisen Nachdrucks, der fotomechanischen und tontechnischen Wiedergabe und die der Übersetzung, vorbehalten.
Aschendorffsche Buchdruckerei, Münster Westfalen, 1970

Inhalt des 9. Bandes (1969)

AUFSÄTZE

MISZELLEN

Hermann Kamp, Münster

Methoden zur Herstellung und Auswertung von Dialekt-Wörterbüchern mit Hilfe der elektronischen Datenverarbeitung

Vorbemerkungen

Der Ausgangspunkt der Untersuchung ist nicht die Frage, ob durch den Einsatz eines Computers in der Lexikographie Zeit und Arbeitskraft eingespart werden können; beides sind Tatsachen, die sich heute kaum noch bestreiten lassen. Als vielleicht deutlichstes Beispiel dafür wäre die Sortierung von großen Materialmengen zu nennen, wie sie ja gerade bei Wörterbüchern oft anfällt. Es ist schwierig, generelle Angaben über die maschinelle Sortiergeschwindigkeit zu machen, da sie sehr von der Länge der Belege wie auch von der Größe und Organisation der verwendeten Maschine abhängt; zum Vergleich sei jedoch angegeben, daß eine Anzahl von 20000 Belegen durchaus innerhalb von 8–12 Minuten sortiert werden kann. Bereits zu diesem Zweck dürfte das Ablochen von noch nicht sortiertem Material lohnend sein. Viel wichtiger erscheint mir jedoch, daß die von den Wörterbüchern gebotene Information erheblich verbessert werden könnte. Bei der bisherigen lexikographischen Praxis verschwindet ein Belegzettel zunächst unter seinem Lemma im Archiv, ungeachtet seiner sonstigen Aussagen in lautlicher, grammatischer oder syntaktischer Hinsicht. Auch bei einer späteren Publizierung bleibt eine Auswertung nach diesen Gesichtspunkten eine Zufälligkeit, da der Benutzer wohl kaum ein vielbändiges Lexikon nach bestimmten lautlichen oder grammatischen Kriterien durchforschen kann, was jedoch maschinell ohne Schwierigkeiten möglich ist. Obwohl auch die Ersparnis von Zeit und Aufwand ein wichtiges Argument für die Datenverarbeitung sein mag, liegt meines Erachtens ihr eigentlicher Vorteil erst darin, daß ein einmal abgelochtes Wortmaterial maschinell nach den verschiedensten sprachwissenschaftlichen Fragestellungen ausgewertet werden kann. Einige Beispiele dazu sollen im späteren Teil dieses Aufsatzes erläutert werden.

Zum Austesten der Programme habe ich Belege aus dem Archiv des Westfälischen Wörterbuches, Münster, verwendet; abgelocht

wurden jeweils ein niederdeutsches Wort, dessen hochdeutsche Übersetzung, ein Grammatiksigle sowie ein Belegsigle. Weitere Informationen, wie etwa die Bezeichnung einer Sammelstelle oder Angaben zur Datierung eines Belegs, können selbstverständlich miteingespeichert werden.

Bei den niederdeutschen Wörtern wurde nicht nur die ursprünglich belegte mundartliche Form, sondern auch das vom Bearbeiter angesetzte Lemma abgelocht. Galt für Wörter unterschiedlicher Herkunft das gleiche Lemma, diente ein entsprechender Zusatz als Unterscheidungsmerkmal für die spätere Sortierung. Lautschriftliche Belege wurden in normale Typen umgesetzt und zu Anfang des Belegsigles besonders gekennzeichnet.

Sofern ein Belegwort mehrere hochdeutsche Übersetzungen hatte, wurde jede Bedeutung einschließlich ihres Kontextes auf eine eigene Lochkarte übertragen. Die Bedeutungsangabe erfolgte stets in Normalform, d. h. für Substantive im Nominativ Singular, für Verben im Infinitiv, für die übrigen Wortarten in der unflektierten Form. Das angesetzte Grammatiksigle bezog sich dagegen nicht auf das Lemma, sondern auf die in der Quelle stehende Form des betreffenden Wortes. Das Belegsigle gab an, aus welchem Kreis und aus welchem Ort der Archivbeleg gemeldet worden war[1].

Die Programme zur Verarbeitung dieses Wortmaterials wurden in der Sprache PL/1 verfaßt; bei einigen Schritten fanden jedoch in anderen Sprachen geschriebene Standardprogramme der IBM Verwendung, die mir eine wesentliche Verkürzung der Rechenzeit ermöglichten. Bei der Programmierung wurde davon ausgegangen, daß das Datenmaterial von Lochkarten eingelesen und auf Magnetbändern gespeichert wird, die dann für die späteren Sortier- oder Untersuchungsprogramme benutzt werden. Bei einzelnen Testläufen wurde jedoch nur eine geringe Datenmenge verwendet, wodurch der Gebrauch externer Speicher, also etwa eines Magnetbandes oder einer Platte, nicht erforderlich war.

Die Programme wurden auf einem Computer des Typs IBM 360/50 im Rechenzentrum der Universität Münster ausgetestet.

[1] Die Anordnung des Datenmaterials entspricht weitgehend den *Richtlinien zur Ablochung und zur zentralen Speicherung mundartlichen Wortmaterials des Deutschen*, als Manuskript hrsg. von der vorläufigen Kommission für ein Arbeitsprogramm zur zentralen Speicherung mundartlicher Wortsammlungen des Deutschen, Göttingen-Marburg 1968.

Über technische Einzelheiten der Maschine soll hier nicht näher referiert werden. Eine Beschreibung der Grundeinheiten eines Computers sowie deren Arbeitsweise geben K. GANZHORN und W. WALTER in ihrem Aufsatz *Technik der Datenverarbeitung*[2]; auch die Abhandlung von S. M. LAMB, *The Digital Computer as an Aid in Linguistics*[3], bietet technische Informationen, die heute jedoch z. T. überholt sind.

Über die Geschwindigkeit der benutzten Ein- und Ausgabeeinheiten sei angegeben, daß der Kartenleser pro Minute 1000 Lochkarten lesen kann und der Schnelldrucker im gleichen Zeitraum 1100 Zeilen voll ausdruckt.

Beschaffung von Belegmaterial

Als maschinelle Hilfen zu einer Aufbereitung von Texten für lexikographische Zwecke lassen sich vor allem die Wortindex- und Konkordanzprogramme nennen; beide können – je nach der beabsichtigten Verwendung des Materials – in sehr verschiedenen Formen organisiert sein. Unter einem Wortindex wird hier eine alphabetisch geordnete Liste von allen in einem Text vorkommenden Wörtern verstanden. Diese Grundform des Wortindex kann durch eine Hinzunahme anderer Kriterien erweitert werden, so etwa durch die Angaben, wo in der Quelle das Belegwort zu finden ist, wie häufig es im Text vorkommt und aus wieviel Buchstaben es besteht. Für das Ablochen ergibt sich dabei keine nennenswerte Mehrarbeit, da nur der Beginn einer neuen Seite oder Zeile durch ein Sonderzeichen zu markieren ist. Auch die Programmierung der genannten Zusätze bietet keinerlei Schwierigkeiten. Für die Angabe der Belegstelle wird zu jedem Wort der von der Maschine errechnete Seiten- und Zeilenzähler miteingespeichert, während ein zweiter Zähler die Häufigkeit des Belegs vermerkt. Zur Feststellung der Buchstabenzahl genügt ein Aufruf der built-in function LENGTH; steht z. B. im Speicher Z das Wort *brennen*, so wird durch das statement: Y=LENGTH(Z); für Y der Wert 7 zurückgegeben.

Gemessen an der geringen zusätzlichen Arbeit bietet der erweiterte Wortindex eine erheblich größere Information, da er nicht

[2] Studium Generale 21 (1968) 828–858.
[3] Language 37 (1961) 382–412.

nur nach der alphabetischen Ordnung, sondern auch nach der zu- oder abnehmenden Wortlänge bzw. Worthäufigkeit ausgedruckt werden kann[4].

Für Aufgaben der Stilanalyse oder der Sprachstatistik dürfte der Wortindex sicherlich ein wichtiges Hilfsmittel sein; für lexikographische Zwecke ist er dagegen nur wenig verwendbar, da er keinen Kontext der Belegwörter enthält. DE TOLLENAERE urteilt darüber:

„Dem Lexikographen liefert der Wortindex Belegstellen, weiter aber nichts. Die Stellen ersparen ihm allerdings eine recht umständliche, zeitraubende und unvollständige Exzerpierarbeit. Auf Grund der Belegstellen lassen sich Belegzettel herstellen, in denen das Wort in seinem Kontext vorkommt"[5].

Damit fiele dem lexikographisch genutzten Wortindex vor allem die Aufgabe zu, ein gezieltes Exzerpieren aus einem größeren Text zu ermöglichen. Eine Schwierigkeit ergibt sich jedoch dadurch, daß die ausgedruckten Wortlisten nicht lemmatisiert sind, also auch keine Scheidung der Homographen stattfindet. Bei den Versuchen einer mechanischen Übersetzung entscheidet das Programm diese Fragen durch eine Prüfung des syntaktischen Zusammenhangs, wozu jedoch nicht nur ein entsprechend organisiertes Wörterbuch, sondern auch eine programmierte Grammatik erforderlich sind. Für historisch oder mundartlich orientierte Wörterbücher sind diese Voraussetzungen nur schwer zu erfüllen, da Syntax und Grammatik innerhalb des Belegzeitraumes oder des Beleggebietes oft sehr uneinheitlich sind. So entscheidet sich auch DE TOLLENAERE dafür, die Lemmatisierung zumindest vorläufig noch manuell durchführen zu lassen, sieht aber einen gewissen Grad der Automation als erreichbar an[6].

Erwägt man den Nutzen des Wortindex für lexikographische Zwecke, dürfte auch der Aspekt der Wirtschaftlichkeit nicht unberücksichtigt bleiben; das Ablochen der Texte erfordert einen hohen Zeitaufwand, der nur lohnen wird, wenn das gleiche

[4] Hinweise über die maschinelle Herstellung eines Wortindex gibt z. B. A. J. T. COLIN, *The Automatic Construction of a Glossary*, Information And Control 3 (1960) 211ff.

[5] F. DE TOLLENAERE, *Lexikographie mit Hilfe des elektronischen Informationswandlers*, Zeitschrift für Deutsche Sprache 21 (1965) 6.

[6] Ebd. S. 7ff.

Material auch zu weiteren maschinellen Untersuchungen verwendet werden soll[7]. Für die Phase der Kompilation eines Wörterbuchs bietet sich hier die Herstellung von Konkordanzen an, die auf den Ergebnissen des vorausgegangenen Wortindex fußen können.

Im Gegensatz zum Wortindex wird bei der Konkordanz jeder Beleg in seinem Kontext wiedergegeben. Der Umfang des Kontextes kann z. B. so gewählt werden, daß immer eine feste Anzahl von Wörtern vor und nach dem betreffenden Beleg in der später alphabetisch ausgedruckten Liste verzeichnet ist. Will man jedoch das Satzgefüge des Textes berücksichtigen, so erscheint es günstiger, nur eine Höchst- und Mindestzahl der in beiden Richtungen aufzunehmenden Wörter festzusetzen und die in diesem Bereich vorkommenden Satzzeichen als Begrenzung zu benutzen. Wie groß die Mindestmenge des Kontextes für lexikographische Belange sein muß, läßt sich nicht pauschal entscheiden; für zweisprachige Wörterbücher dürfte im allgemeinen jedoch eine Gesamtzahl von ca. 40–50 Wörtern ausreichen, um eine eindeutige Übersetzung des Belegs zu ermöglichen. Die Aufnahme einer wesentlich höheren Kontextmenge wird in den meisten Fällen nicht lohnen, da der relativ geringen zusätzlichen Information ein erheblicher Mehrbedarf an Speicherplatz und Rechenzeit gegenübersteht.

Würden alle im Text vorkommenden Wörter bei der Herstellung der Konkordanz berücksichtigt, so dürften gerade die lexikalisch unbedeutenden Bindewörter einen großen Raum darin einnehmen; K. Baldinger weist darauf hin, daß bei der Bibelkonkordanz des Paters Ellison 127 „mots outils" ausgeschieden wurden, die aber 60% des insgesamt 800000 Wörter umfassenden Textes ausmachten[8]. Entsprechend schlägt D. Hays vor, häufig vorkommende unwichtige Wörter vorher einzuspeichern und ihre Konkordanzbelege zu unterdrücken; falls andererseits nur eine geringe Wortmenge von Interesse ist, könnte auch diese vorher eingegeben und die Konkordanz ausschließlich auf sie beschränkt werden[9].

[7] In begrenztem Umfang können bereits Klarschriftleser eingesetzt werden, wodurch ein Ablochen der Texte entfällt.

[8] *Automation und Lexikologie*, Zeitschrift für romanische Philologie 75 (1959) 543ff.

[9] *Introduction to Computational Linguistics*, New York 1967, S. 171f.

Vom lexikographischen Standpunkt betrachtet, bietet die Konkordanz eine erheblich größere Information als der Wortindex, da durch den mitangegebenen Kontext ein semantischer Vergleich der Belegstellen möglich ist. Sofern im Einzelfall die Kontextmenge dazu nicht ausreicht, kann der Beleg aufgrund der ausgedruckten Seiten- und Zeilenangabe leicht in der Quelle überprüft werden. Zusätzliche Angaben über Wortlänge oder Worthäufigkeit entsprechen dem Wortindexprogramm[10].

Bei der für lexikographische Zwecke benutzten Konkordanz mag es als unbequem empfunden werden, daß die Belege auf Endlospapier ausgedruckt sind und sich daher nicht in das sonst übliche Zettelschema einfügen. Zum Teil dürfte dies zwar eine Sache der Gewöhnung sein, aber andererseits wäre es vorteilhaft, wenn maschinell hergestellte Konkordanzbelege auch in bereits bestehende Zettelarchive übernommen werden könnten. Über einen entsprechenden Versuch des Goethe-Wörterbuches berichten G. STICKEL und M. GRÄFE in einem Aufsatz[11], der hier kurz referiert werden soll. Als Aufgabenstellung geben die Autoren an:

> Für einen fortlaufenden Text ist zu jedem vorkommenden Wort, welches nicht als insignifikant definiert ist, ein DIN A6-Zettel anzulegen, welcher das Lemma, seine genaue Stellenangabe (mit Seite und Zeile), eine Kennzeichnung des betreffenden Textes und den Wortlaut der Stelle, das heißt eine gewisse Menge Kontext, enthält. Die Zettel sollen diese Angaben in einer Anordnung zeigen, die der lexikalischen Konvention möglichst weitgehend entspricht (das heißt Lemma oben links ausgeworfen usw.), und sie sollen in alphabetischer Ordnung stehen[12].

Im ersten Teil des Programms wird eine Konkordanz hergestellt, die die Belege in nicht lemmatisierter Form enthält. Anhand dieser ausgedruckten Listen erfolgt dann eine manuelle Normalisierung, die abgelocht und in den ursprünglichen Datensatz übertragen wird. Das alphabetisch sortierte Material wird zu gleichen Teilen auf vier Magnetbänder geschrieben, wobei auf dem ersten Band

[10] Einzelheiten über technische Fragen finden sich in dem Aufsatz von F. BERNHARD/H. REUL/F. SCHULTE-TIGGES/H. SUNKEL, *Erstellung von Konkordanzen zu Sanskrit-Texten durch elektronische Rechenanlagen*, Linguistics 22 (1966) 5–23.

[11] *Automatische Textzerlegung und Herstellung von Zettelregistern für das Goethe-Wörterbuch*, Sprache im technischen Zeitalter (1966) 247–257.

[12] Ebd. S. 248.

das erste Viertel der Daten steht, auf dem zweiten Band das zweite Viertel usw. Auf eine Druckseite ist dann fortlaufend je eine Eintragung von allen vier Bändern im gewünschten Format auszudrucken. Das Endlospapier wird anschließend maschinell in einzelne Bogen aufgetrennt und dabei gleichzeitig richtig nacheinander gelegt. Zerschneidet man die Druckerausgabe in vier Blöcke, so ergeben sie hintereinandergelegt ein alphabetisch sortiertes Zettelregister.

Für Mundart-Wörterbücher könnte das Verfahren nicht nur bei einer Übernahme gedruckter Quellen verwendet werden, sondern auch zur maschinellen Exzerpierung laienschriftlicher Aufzeichnungen dienen. Besonders geeignet dafür wären zur Wortschatzerforschung verschickte Fragebogen, bei denen kürzere Sätze in die mundartliche Entsprechung zu übertragen sind. Während beim konventionellen Verzetteln der gleiche Satz mehrfach geschrieben werden muß, würde hier ein einmaliges Ablochen ausreichen. Auch die technische Regelung ist einfach, da die Bedeutung durch die mitabgelochte Fragenummer gekennzeichnet ist und als Kontext jeweils der gesamte Satz ausgedruckt werden kann. Das Ansetzen des Lemmas bliebe jedoch auch hier Aufgabe des Bearbeiters, da selbst bei einem bereits eingespeicherten Wörterbuch wegen der vor allem bei Diphthongen stark differierenden Laienschreibung eine maschinelle Lemmatisierung nicht durchführbar wäre.

Ablochkonventionen

Bei der Schreibung mundartlichen Wortmaterials sind häufig Länge, Kürze oder Öffnung eines Vokals anzugeben. Die dazu benutzten lautschriftlichen Zeichen sind jedoch auf dem Locher nicht vorhanden, so daß eine Umsetzung in andere Typen erforderlich ist. Das gleiche gilt für die Umlaute *ä, ö, ü,* die nicht als AE, OE, UE abgelocht werden können, da sie sonst nicht von den entsprechenden Diphthongen zu unterscheiden wären. Alle Umschreibungen sollten jedoch so gewählt werden, daß auch der mit der Ablochkonvention nicht vertraute Benutzer die ausgedruckten Belege ohne Schwierigkeiten lesen kann.

Bei den zu Testzwecken abgelochten Daten wurde die Länge eines Vokals durch ein nachgestelltes Pluszeichen angegeben, z. B.:

STE+N	STEIN
O+GE	AUGE
DRU+ST	STRAUCH

Die Hervorhebung besonderer Kürze erfolgte durch ein Minuszeichen:

TA—KKE	ZWEIG
KO—P	KOPF

Die Öffnung eines Vokals wurde durch einen Stern wiedergegeben:

HO*F	HOF
DRE*GGEN	DREHEN

Zur Kennzeichnung des Umlauts dienten nachgesetzte Anführungsstriche:

WU"RKEN	WEBEN
MAUE	A"RMEL

Traten zu einem Vokal mehrere Zeichen, so galt die Reihenfolge Umlaut vor Öffnung vor Quantität. Als zusätzliche grammatische Bestimmung der Belege stand bei praefigierten Wörtern am Ende des Praefixes eine Abschlußklammer, während die einzelnen Glieder eines Kompositums nach dem Fugenelement durch einen Schrägstrich voneinander getrennt wurden, z. B.:

AF)HE+LEN	AB)HEILEN
STU+F/WI+DE	KOPF/WEIDE
BO"KEN/HA+GEN	HAIN/BUCHEN/HECKE

Sortierverfahren

Das Sortieren von Datenmaterial kann nach sehr unterschiedlichen Methoden erfolgen; die meisten von ihnen sind jedoch hinsichtlich der benötigten Rechenzeit derart aufwendig, daß sie für eine Verarbeitung großer Datenmengen nur noch von theoretischem Interesse sind. Im folgenden sollen daher nur einige ausgewählte Verfahren näher erläutert und kritisiert werden. Zuvor ist jedoch noch auf eine Schwierigkeit hinzuweisen, die bei allen Sortierprogrammen durch die bei den Daten mitabgelochten Sonderzeichen entsteht. Diese werden von der Maschine als Buchstaben

aufgefaßt, deren alphabetischer Wert kleiner A ist[13]; so wird z. B. das Wort *Ärmel* vor *Aal* eingeordnet. Um dieses zu vermeiden, habe ich zu dem jeweiligen Wort einen Hilfsspeicher eingerichtet, worin das betreffende Wort ohne die darin enthaltenen Sonderzeichen aufgenommen wird. Für die spätere Sortierung dient dann dieser Vergleichsspeicher, wodurch z. B. *Ärmel* als *Armel* sortiert wird, was auch der lexikalischen Konvention entspricht. Die in dem Hilfsspeicher stehende Form des Wortes braucht selbstverständlich nicht abgelocht zu werden, sondern wird von der Maschine dort eingetragen. Beim späteren Ausdrucken kann dieser Hilfsspeicher ignoriert werden; er dient also nur dem rein internen Rechenvorgang, um eine gewünschte alphabetische Folge der Wörter zu erreichen.

a) Die Intervallschachtelung

Vergegenwärtigt man sich zunächst, wie das manuelle Nachschlagen eines Wortes im Lexikon geschieht, so könnte man den Vorgang auf folgende Weise beschreiben: aufgrund der Kenntnis des Alphabets wird abgeschätzt, wo das gesuchte Wort in etwa stehen könnte; das Lexikon wird entsprechend aufgeschlagen und die Differenz zwischen der tatsächlichen Aufschlagung und dem gesuchten Wort durch ein erneutes Aufschlagen zu korrigieren versucht. Von der dann gefundenen Stelle nähert man sich durch weitere Intervallschachtelungen dem gesuchten Wort.

Ein mechanisches „Nachschlagen" kann auf sehr ähnliche Weise erfolgen, wobei von der Maschine jedoch nicht geschätzt, sondern gerechnet wird. Aus der Gesamtzahl der Lexikoneintragungen wird das in der Mitte stehende Wort mit dem nachzuschlagenden Wort verglichen; ist dieses kleiner, braucht nur noch in der oberen Hälfte weitergesucht zu werden. Hiervon wird wieder die Mitte gebildet und das an dieser Stelle stehende Wort mit dem Suchwort verglichen; ist dieses größer, kann es nur in der unteren Hälfte liegen, wovon wieder die Mitte gebildet wird usw. Bei jedem Schritt vermindert sich die Zahl der noch zu untersuchenden Wörter um die Hälfte, sinkt also jeweils um eine Zweierpotenz. Anders ausgedrückt, ist damit die Anzahl der notwendigen Intervallschritte

[13] Bei anderen Maschinen sind die Sonderzeichen oft als größer Z eingeordnet; das Problem bleibt jedoch dasselbe.

gleich dem Logarithmus zur Basis 2 von der Gesamtzahl der im Lexikon stehenden Eintragungen. Bei 1000 Wörtern würden also 10 Schritte, bei 8000 Wörtern 13 und bei 30000 Wörtern 15 Schritte erforderlich sein.

Das gleiche Verfahren kann auch zur Einsortierung neuer Belege verwendet werden. Zunächst wird die Stelle ermittelt, an der das Wort gemäß der alphabetischen Reihenfolge stehen müßte; von hier sind dann alle im Speicher stehenden Eintragungen um einen Platz nach hinten zu rücken, so daß das neue Wort an der freigewordenen Stelle eingefügt werden kann[14].

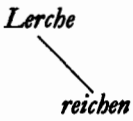
Als Vorteil dieser Methode könnte man vielleicht nennen, daß jedes Wort schon beim Einspeichern alphabetisch sortiert wird, wodurch die eingetragenen Belege bereits während des Sortiervorgangs für Zwischenuntersuchungen zu verwenden sind. Nachteilig ist jedoch, daß das ständige Verschieben der Daten eine sehr hohe Rechenzeit beansprucht. Will man das Verfahren daher überhaupt anwenden, so dürfte es nur bei speziellen Problemen oder relativ geringen Datenmengen lohnend sein. Als obere Grenze kämen vielleicht ein- bis zweitausend Wörter in Frage, da spätestens dann die etwas höhere Systemzeit anderer Programme durch ihre erheblich größere Sortiergeschwindigkeit ausgeglichen wird.

b) Der Sorting Tree

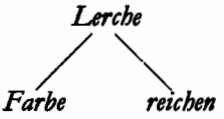
Das Verfahren des sorting tree[15] soll zunächst graphisch an einem Beispiel dargestellt werden, wobei folgende Wörter zu sortieren sind: *Lerche*, *reichen*, *Farbe*, *Pilz*, *helfen*, *Erde*, *Brunnen*, *Spule*, *Arm*, *Zapfen*, *Saum*. Dazu wird das erste Wort auf die Mitte der Seite gestellt und mit dem folgenden Wort verglichen; da *reichen* größer *Lerche* ist, wird es mit einem nach rechts weisenden Pfeil darunter gestellt:

[14] Die unter a und b geschilderten Verfahren wurden in dem Seminar „Organisationsprinzipien automatischer Wörterbücher" behandelt, das von Herrn Dr. K. Brockhaus, Seminar für vergleichende Sprachwissenschaften der Universität Münster, im SS 1968 gehalten wurde.

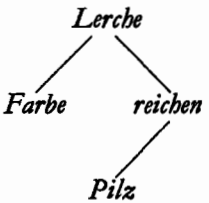
[15] A. D. Booth und A. J. T. Colin behandeln den sorting tree in ihrem Aufsatz *On the Efficiency of a New Method of Dictionary Construction*, Information and Control 3 (1960) 327–334. Ein weiteres Beispiel findet sich bei M. Levison, *The Computer in Literary Studies*, in: Machine Translation, hrsg. v. A. D. Booth, Amsterdam 1967, S. 177ff.



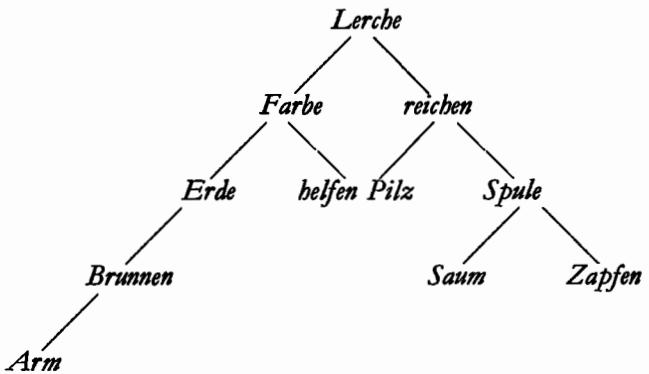
Anschließend wird das dritte Wort mit dem ersten verglichen; da es kleiner ist, erhält es einen nach links weisenden Pfeil:



Das vierte Wort, *Pilz*, kommt alphabetisch nach *Lerche*; da der nach rechts weisende Pfeil jedoch schon besetzt ist, wird es mit dem Wort am Ende dieses Pfeiles verglichen. *Pilz* ist kleiner *reichen*, wird also nach links angeschlossen:



Das fünfte Wort, *helfen*, kommt vor *Lerche* und nach *Farbe*, ist also von dort nach rechts zu setzen. Werden die folgenden Wörter in der gleichen Weise angeordnet, so ergibt sich:



Der dargestellte sorting tree läßt sich in einzelne sub-trees zerlegen, die jeweils aus einem Zentralwort bestehen müssen und zusätzlich ein links oder rechts darunter stehendes Nebenwort haben können. Für die Sortierung ist bei dem am weitesten links stehenden

sub-tree anzufangen und zuerst sein linkes Nebenwort, dann das Zentralwort und schließlich das rechte Nebenwort auszudrucken, sofern es kein linkes Nebenwort hat und dadurch selbst Zentralwort eines eigenen sub-tree wird. Bei dem gegebenen Beispiel hat der unterste sub-tree das Zentralwort *Brunnen*; zuerst ist das linke Nebenwort *Arm* auszudrucken, dann das Zentralwort. Da ein rechtes Nebenwort fehlt, wird der nächste sub-tree genommen und in der Reihenfolge *Erde – Farbe – helfen* ausgedruckt usw.

Im Programm werden die für „größer" oder „kleiner" verwendeten Pfeile durch Zahlenspeicher ersetzt, worin nach entsprechendem Vergleich die Adresse jedes neu zu sortierenden Wortes einzusetzen ist. Zusätzlich wird ein Rückverweis aufgenommen, der angibt, mit welcher Eintragung das neu eingespeicherte Wort als letztes verglichen wurde. Als Beispiel seien wieder die obigen Wörter verwendet:

Das erste Wort wird in den Speicher eingelesen (s. Beispiel 1).

Das zweite Wort wird eingelesen (Adresse Nr. 2) und mit dem ersten verglichen; da es größer ist, wird seine Adresse im Speicher „größer" des ersten Wortes eingetragen (s. Beispiel 2).

Das dritte Wort ist kleiner als das erste, seine Adresse (Nr. 3) wird also im Speicher „kleiner" dieses Wortes eingetragen (s. Beispiel 3).

Das vierte Wort ist größer als das erste; der betreffende Speicher ist jedoch schon besetzt, so daß mit dem Wort verglichen werden muß, dessen Adresse (Nr. 2) hier angegeben wird. Da *Pilz* kleiner *reichen* ist, wird seine Adresse im Speicher „kleiner" von *reichen* eingetragen (s. Beispiel 4).

Das fünfte Wort ist kleiner als das erste; da im Speicher „kleiner" jedoch schon die Zahl 3 steht, ist der Vergleich bei diesem Wort fortzusetzen. Das Wort *helfen* ist größer *Farbe*, seine Adresse wird also dort im Speicher „größer" eingetragen (s. Beispiel 5).

Die folgenden Wörter werden auf die gleiche Weise eingeordnet, so daß sich am Ende die in Beispiel 6 dargestellte Tabelle ergibt.

Bei dem hier beschriebenen Verfahren wird ebenfalls bereits während des Einspeicherns die Sortierfolge der Belege festgestellt; dabei entfällt jedoch die beim Intervallschachtelungs-Verfahren notwendige Verschiebung der Daten, so daß die zur Sortierung benötigte Zeit hier kürzer sein dürfte. Bei größeren Datenmengen

ist jedoch ein relativ häufiges Vergleichen der einzelnen Wörter notwendig, da bei jedem neu einzusortierenden Wort auch die bereits besetzten Speicherplätze für den Vergleich zu berücksichtigen sind. Spätestens bei Benutzung eines Magnetbandes wird das Verfahren unwirtschaftlich, da das ständige Vor- und Zurückspulen des Bandes eine unvertretbar hohe Rechenzeit beanspruchen würde.

	Adresse	Wort	kleiner	größer	Rückverweis
Beispiel 1	1	LERCHE			0
Beispiel 2	1	LERCHE		2	0
	2	REICHEN			1
Beispiel 3	1	LERCHE	3	2	0
	2	REICHEN			1
	3	FARBE			1
Beispiel 4	1	LERCHE	3	2	0
	2	REICHEN	4		1
	3	FARBE			1
	4	PILZ			2
Beispiel 5	1	LERCHE	3	2	0
	2	REICHEN	4		1
	3	FARBE		5	1
	4	PILZ			2
	5	HELFEN			3
Beispiel 6	1	LERCHE	3	2	0
	2	REICHEN	4	8	1
	3	FARBE	6	5	1
	4	PILZ			2
	5	HELFEN			3
	6	ERDE	7		3
	7	BRUNNEN	9		6
	8	SPULE	11	10	2
	9	ARM			7
	10	ZAPFEN			8
	11	SAUM			8

c) Der Sort/Merge

Unter dem Begriff Sort/Merge versteht man ein Programm, das vom jeweiligen Maschinenhersteller geschrieben ist, um ein optimales Sortieren von Daten auf der betreffenden Maschine zu erreichen. Während bei anderen Sortierprogrammen die benötigte

Zeit bei großen Datenmengen sehr stark zunimmt, steigt der beim Sort/Merge nötige Aufwand mit etwa nlog n, was einem fast linearen Verlauf entspricht. Über die Menge der in einem Sort-Lauf zu bearbeitenden Wörter läßt sich kaum eine Zahlenangabe machen, da sie sehr von der Länge der Belege wie auch von der Größe und Organisation der verwendeten Maschine abhängig ist; als Vergleichszahl könnte man etwa angeben, daß auf einer größeren Anlage durchaus 100000 Belege zu je 60 Buchstaben in einem Arbeitsgang sortiert werden können. Sind wegen größerer Datenmengen mehrere Einzelsortierungen notwendig, so erfolgt die endgültige alphabetische Ordnung durch ein nachfolgendes Merge-Programm.

Bei den von mir durchgeführten Testläufen wurde zur Eingabe der Daten ein Magnetband benutzt, worauf der zu jedem Wort gehörende Vergleichsspeicher bereits beim Einlesen verzeichnet worden war. Die Anzahl der auf einem Band zu speichernden Belege betrug etwa 400000, was jedoch durch eine andere als die gewählte Datenorganisation[16] verändert werden kann.

Änderung des Maschinen-Codes

Die in der Maschine bestehende Konvention, daß das Alphabet nur aus den 26 normalen, nicht erweiterten Buchstaben besteht, kann durch ein Programm geändert werden, so daß nicht nur den Buchstaben und Sonderzeichen, sondern auch den Kombinationen aus beiden ein beliebiger alphabetischer Wert zugeordnet werden kann. Dadurch ist es möglich, die Menge der zur Sortierung benutzten Zeichen erheblich zu vergrößern und eine eigene alphabetische Folge zu bestimmen. Darin könnte z. B. festgesetzt werden, daß der Vokal *å*, abgelocht als A*, unmittelbar nach kurzem *a* (A) kommt und dann von langem *a* (A+), umgelautetem *a* (A") und schließlich von *b* gefolgt wird. Entsprechend können auch die lautlichen Varianten der anderen Vokale als selbständige Einzelbuchstaben definiert werden, deren alphabetischer Wert durch die zur Qualitäts- oder Quantitätskennzeichnung benutzten Sonderzeichen eindeutig bestimmt ist. Je nach der bearbeiteten Sprache muß ein solches System individuell verschieden sein und kann des-

[16] Beleglänge = 104 Buchstaben, Blockung zu 3120 Bytes.

halb auch nicht von vornherein in der Maschine einprogrammiert sein; der dort bestehende Code ist daher durch einen eigenen, neu zu schreibenden Code zu ersetzen. Für Versuchszwecke habe ich ein Alphabet von insgesamt 54 Buchstaben festgesetzt und entsprechend codiert, wofür im folgenden ein kurzes Beispiel gegeben werden soll. Links stehen die in den Lemmata gebräuchlichen Typen, daneben die beim Testmaterial getroffene Ablochkonvention und rechts der zugewiesene Code:

Typus	abgelocht als	Code
a	A	1011 0000
ā	A+	1011 0001
ä	A"	1011 0010
ǟ	A"+	1011 0011
å	A*	1011 0100
å̄	A*+	1011 0101

Das geschilderte Beispiel mag vielleicht den Eindruck erwecken, daß bei einem derartig erweiterten Alphabet sehr viele zusätzliche Abfragen nötig sind, wodurch die benötigte Rechenzeit stark ansteigen könnte. Daß diese Abfragen mit großer Geschwindigkeit durchgeführt werden, wäre noch kein Argument, da sie sich bei großen Datenmengen trotzdem summieren würden; entscheidend ist jedoch, daß beim Einlesen eines Buchstabens die Zentraleinheit der Maschine weitgehend unbeschäftigt wartet und während dieser Zeit durchaus die erforderlichen Abfragen durchführen kann. Die Rechenzeit dürfte daher in den meisten Fällen nicht erhöht, sondern nur besser ausgenutzt werden[17].

Da in dem beschriebenen Code die einzelnen Vokalvarianten als unterschiedlich groß definiert sind, werden bei der Sortierung zuerst alle mit kurzem *a* beginnenden Wörter ausgedruckt, also etwa von *ab* bis *awwer;* danach kommen die mit langem *a* anfangenden Wörter, etwa von *Ābel* bis *āwig*, anschließend die Wörter mit kurzem *ä*, etwa von *ächter* bis *Ätte* usw. Im Wortinnern gilt das gleiche Prinzip, so daß *Bast* vor *ba"bbelen* steht. Für semantische Zwecke ist diese Methode sicherlich nachteilig, da zusammengehörende Wörter wie *Angst* und *ängstlich* in ihrer lexikalischen

[17] Herrn H. Hölsken vom Rechenzentrum der Universität Münster danke ich für seine Beratung bei dem Umcodierungs-Programm.

Folge auseinandergerissen werden. Andererseits bieten sich günstige Möglichkeiten für einen lautlichen Vergleich, da z. B. alle mit einem bestimmten Vokaltypus anfangenden Wörter alphabetisch geordnet nacheinander stehen. Sucht man nach einem Benennungskriterium für diese Sortierung, so könnte man sie vielleicht als lautorientiert bezeichnen; für den Aufbau eines Mundart-Wörterbuchs dürfte sie weniger geeignet sein, aber für Zwecke einer strukturellen Untersuchung der betreffenden Sprache einen wesentlichen Vorteil bieten. Da die gleiche Ablochkonvention wie bei der normalen Sortierung eingehalten wird, wäre es technisch einfach, in einem zweiten Vergleichsspeicher die umcodierte Form eines jeden Wortes unterzubringen, so daß sie bei Bedarf für die betreffenden Untersuchungen zur Verfügung steht.

Aufbau des Wörterbuchs

Die alphabetisch sortierten Belege entsprechen noch nicht der bei Mundart-Wörterbüchern üblichen Folge, da hier gewöhnlich alle mit Praefixen gebildeten Wörter unter ihrem Grundwort zusammengefaßt werden; es erscheinen also die Verben *abbrennen, ausbrennen, entbrennen, verbrennen* alle unter dem Lemma *brennen*. Die gleiche Regelung gilt für die anderen Wortarten, so daß etwa das Substantiv *Behausung* unter den Ableitungen des Wortes *Haus* zu suchen ist. Für den Benutzer hat diese Methode den Vorteil, daß er alle zu einer bestimmten Wortgruppe gehörenden Belege zusammen findet und ihm ein zeitraubendes Blättern in anderen Bänden des Lexikons erspart wird.

Mit Hilfe der als Praefixkennzeichen abgelochten Klammer läßt sich die gewünschte Sortierfolge auch maschinell herstellen[18]. Beim Einlesen der Wörter wird zunächst untersucht, ob eine Abschlußklammer darin vorkommt; falls ja, wird das davorstehende Praefix in dem zur Sortierung benutzten Vergleichsspeicher an das Ende des betreffenden Wortes gesetzt, so daß z. B. *ab)brennen* als *brennenab* sortiert würde. In einigen Fällen könnten dabei jedoch andere

[18] Sofern bereits ein Wörterbuch eingespeichert ist, das zum Vergleich benutzt werden kann, ist das Ablochen der Praefixklammer nicht mehr erforderlich. Stattdessen kann der Maschine eine Liste aller bestehenden Praefixe eingespeichert werden, anhand derer sie feststellt, zu welchem Grundwort das betreffende praefigierte Wort zu stellen ist.

Wörter zwischen das Stammverb und seine praefigierten Formen treten; so würde etwa der Ortsname *Laufen/burg* zwischen den Verben *laufen* und *zu)laufen* stehen. Um eine solche alphabetische Folge zu vermeiden, wird vor dem nachgestellten Praefix ein Sonderzeichen eingespeichert; da dieses bei der Sortierung als kleiner A gewertet wird, sind auch die praefigierten Formen eines Grundwortes kleiner als seine nachfolgenden Belege. Als Beispiel für den Unterschied zwischen der ausgedruckten Form und ihrer im Vergleichsspeicher stehenden Entsprechung seien noch einmal die zuletzt genannten Wörter aufgeführt:

ausgedruckter Beleg	interner Sortierbegriff
LAUFEN	LAUFEN
ZU)LAUFEN	LAUFEN)ZU
LAUFEN/BURG	LAUFENBURG

Ein internes Nachstellen des Praefixes hat den weiteren Vorteil, daß innerhalb der zu einem Stammverb gehörenden praefigierten Formen automatisch die richtige alphabetische Reihenfolge hergestellt wird, da *brennen)ab* kleiner *brennen)ent* kleiner *brennen)ver* ist usw. Die genannten Formen stehen jedoch nur in dem zur Sortierung benutzten Vergleichsspeicher, während im normalen Speicher, der später ausgedruckt wird, das Praefix selbstverständlich am Wortanfang erscheint.

Dennoch kann auch die jetzt hergestellte Sortierung noch nicht endgültig sein, da in den meisten Wörterbüchern die zu einem bestimmten Wort gehörenden Übersetzungen nicht nur alphabetisch geordnet sind, sondern auch in bedeutungsmäßig zusammengehörende Gruppen gefaßt werden. Vor allem bei sehr umfangreichen Artikeln in einem Wörterbuch, die sich oft über mehrere Seiten erstrecken, ist eine solche Anordnung für den Benutzer eine wesentliche Hilfe. Zumindest zum gegenwärtigen Zeitpunkt jedoch ist eine vollmechanische Einteilung in Bedeutungsgruppen durch die Maschine noch nicht möglich, und ich bezweifle auch, daß sie jemals perfekt möglich sein wird. Als Ausweg käme die Lösung in Frage, daß der Bearbeiter anhand der alphabetisch ausgedruckten Listen eine manuelle Einteilung vornimmt und daß diese durch ein Umsortierungsprogramm in die ursprüngliche Belegfolge über-

tragen wird[19]. Dies könnte so geschehen, daß die Bedeutungsgruppe, die später als erste aufgeführt werden soll, ein A als Kennzeichen erhält, die zweite Gruppe ein B, die dritte ein C usw. Diese Kennbuchstaben sind dann zu dem jeweiligen Wort, etwa in den Vergleichsspeicher, einzutragen und bei der neuen Sortierung als zweites Sortierkriterium anzugeben, wodurch die als zusammengehörig gekennzeichneten Belege auch zusammen ausgedruckt werden. Als drittes Sortierkriterium dienen die hochdeutschen Übersetzungen, so daß die Belege auch innerhalb der einzelnen semantischen Gruppen in alphabetischer Reihenfolge stehen.

An dieser Stelle muß deutlich darauf hingewiesen werden, daß das vom Computer ausgedruckte Material keineswegs eine vorläufige Wortliste sein darf, die nur den Rohbau des späteren Artikels angäbe und worin z. B. die in der Quelle stehenden Satzbeispiele oder Erläuterungen fehlen könnten. In diesem Fall wäre beim Verfassen des Artikels doch wiederum das konventionelle Zettelarchiv zu berücksichtigen, was natürlich bedeutet, daß dessen Belege auch weiterhin konventionell sortiert werden müßten. Es ist deshalb zu fordern, daß das ausgedruckte Material bereits die endgültige Fassung des späteren Wörterbuchartikels haben muß, worin sowohl die in der Quelle stehenden Erläuterungen als auch die vom Bearbeiter gegebenen Kommentare oder Literaturhinweise enthalten sind.

Es ist meines Erachtens jedoch günstig, die oft sehr langen Anmerkungen nicht in das Wörterbuchmaterial selbst zu übernehmen, sondern sie auf einem eigenen Kommentarband zu speichern. Nach der Umsortierung steht die endgültige alphabetische Folge und damit auch die Adressennummer der Belege fest; unter Angabe der Nummer des Belegs, auf den sie sich beziehen, werden die Kommentare abgelocht und auf Band gespeichert. Vor dem Ausdrucken wird jeweils abgefragt, ob zu dem betreffenden Wort ein Kommentar vorhanden ist; falls ja, ist zuerst das Datenband und dann das Kommentarband auszudrucken, falls nein, ist nur das Datenband zu drucken und der Vergleich beim nächsten Beleg fortzusetzen. Durch die Wahl eines verschiedenen Ausgabeformats (Leerzeile,

[19] Den Hinweis auf die Umsortierung verdanke ich Herrn Prof. Dr. KESELING, Marburg.

Einrücken usw.) ist gewährleistet, daß Wörter und Kommentare auch optisch sofort voneinander zu unterscheiden sind.

Auswertung des eingespeicherten Materials

Eine einfache Form der Auswertung eines zweisprachigen Wörterbuchs dürfte die automatische Herstellung eines Registers sein. Vor allem für den Benutzer eines mundartlichen Wörterbuchs ist ein beigegebenes Register eine wesentliche Hilfe, da er, von einem beliebigen hochdeutschen Wort ausgehend, alle in der betreffenden Mundart dafür bekannten Ausdrücke findet. Zugleich ergibt sich für den Bearbeiter der Vorteil, daß er zu jedem Lemma eine vollständige Synonymaliste erhält und diese bereits beim Aufbau der Artikel berücksichtigen kann. Da für Registerzwecke weder die einzelnen mundartlichen Formen noch das miteingespeicherte Grammatik- oder Belegsigle von Bedeutung sind, kann sich die Bearbeitung auf die niederdeutschen Stichwörter und ihre verschiedenen hochdeutschen Übersetzungen beschränken. Das Material wird zunächst nach den hochdeutschen Wörtern alphabetisch sortiert; sobald dasselbe Wort zum zweiten Mal auftritt und seine niederdeutsche Übersetzung gleich der vorhergehenden ist, wird der Beleg nicht in die auszudruckende Liste übernommen. Sofern es erwünscht ist, kann stattdessen ein Häufigkeitszähler eingeführt werden, der die zu jedem Stichwort vorhandene Anzahl der Belege angibt. Zusätzliche Ablocharbeiten sind für die Herstellung des Registers nicht erforderlich.

Die eigentliche Auswertung des eingespeicherten Materials bezieht sich jedoch auf eine Untersuchung der lautlichen und grammatischen Phänomene der betreffenden Sprache, wofür im folgenden einige Beispiele gegeben und erläutert werden sollen. Eine angenehme Hilfe bei der Programmierung boten die verwendeten built-in functions, deren Arbeitsweise kurz beschrieben wird. Eine Kenntnis der Programmiersprache PL/1 wird dabei jedoch nicht vorausgesetzt.

Sollen z. B. aus dem eingespeicherten Material alle niederdeutschen Wörter aufgesucht werden, in deren Lemma der Vokal $\bar{e}^2$, abgelocht als E+2, vorkommt, so wäre der Programmablauf:

Von einem Magnetband werden die Daten blockweise in den Puffer der Maschine geholt und von dort nacheinander zur Aus-

wertung in den Kernspeicher eingelesen. Nimmt man an, daß im Speicher S1 das jeweilige niederdeutsche Lemma steht, so wird die Untersuchung ausgeführt:

I=INDEX (S1, 'E+2');

Mit dem Begriff INDEX wird eine built-in function aufgerufen; sie prüft, ob das zweite angegebene Argument in dem ersten enthalten ist, d. h. in diesem Fall, ob der Vokal E+2 in dem im Speicher S1 stehenden niederdeutschen Lemma erscheint; falls ja, wird für I die Stelle zurückgegeben, an der E+2 zum ersten Mal vorkommt; falls nein, wird der Wert Null zurückgegeben. Ist also der Wert für I von Null verschieden, soll das Lemma mit den dazu gehörenden Angaben ausgedruckt werden. Die Untersuchung aller folgenden Belege geschieht nach der gleichen Methode. Zweckmäßigerweise wird das Verfahren auf ein bereits alphabetisch sortiertes Material angewendet, so daß auch die ausgedruckten Wörter automatisch in alphabetischer Reihenfolge stehen. Da zu jedem niederdeutschen Lemma auch seine mundartliche Form angegeben wird, ergibt sich am Ende eine genaue Beschreibung, wie der Vokal $\bar{e}^2$ innerhalb des Beleggebietes behandelt wird.

Das gleiche Schema kann auch für eine Untersuchung nach beliebigen anderen Lauten verwendet werden, so daß sich dadurch das exakte Material für die Herstellung einer Lautgrammatik des Beleggebietes gewinnen läßt. Um für diesen Zweck die praktische Auswertung zu erleichtern, können die Belege durch einen neuen Sort/Merge-Lauf nach ihrer geographischen Verteilung umsortiert werden. Ebenfalls ist es möglich, die Untersuchung nur auf bestimmte Mundartgebiete oder grammatische Kategorien zu beschränken.

Eine derartig umfassende Auswertung dürfte jedoch nur für Spezialuntersuchungen erforderlich sein, da die vom Material gegebene Information von einer gewissen Belegdichte an nur noch unwesentlich zunimmt. Es würde daher im Normalfall ausreichend sein, von jedem Ort etwa 1–3 Belege zu berücksichtigen.

Das im vorigen beschriebene Verfahren untersucht, ob das gegebene Kriterium an einer beliebigen Stelle im Wort enthalten ist. Eine etwas andere Problemstellung wäre daher die Frage, ob das Untersuchungsmerkmal an einer bestimmten Wortstelle, etwa am

Ende vorkommt. Als Beispiel sei gegeben, daß aus dem eingespeicherten Material alle hochdeutschen Substantiva zu finden sind, die auf *-e* enden.

Der erste Teil des Programmablaufs entspricht dem vorigen, d. h. die Belege sind vom Band nacheinander in den Kernspeicher einzulesen. Dann ist jedoch anhand des Grammatiksigles zuerst zu prüfen, ob es sich bei dem jeweiligen hochdeutschen Wort um ein Substantiv handelt; falls nein, kann sofort der nächste Beleg eingelesen werden; falls ja, ist die Wortlänge festzustellen und zu prüfen, ob der Buchstabe an der letzten Stelle des Wortes gleich dem gesuchten Buchstaben E ist.

Die zuletzt beschriebene Untersuchung kann durch die built-in function SUBSTRING erfolgen; sie hat drei Argumente, von denen das erste angibt, mit welchem Wort die folgende Operation auszuführen ist, das zweite, an welcher Stelle sie beginnen und das dritte, über wieviele Buchstaben des Wortes sie sich erstrecken soll. So ist z. B. der SUBSTRING von ('HAUS', 1,2) = HA, der SUBSTRING von ('HAUS', 2,2) = AU usw.

Nimmt man an, daß im Speicher S2 jeweils das hochdeutsche Wort steht und die vorher errechnete Stelle, nämlich der letzte Buchstabe des Wortes, der Maschine unter dem Wert Y bekannt ist, lautet die zur Untersuchung nötige Abfrage:

```
IF SUBSTRING(S2, Y,1) = 'E'  THEN DO;
```

Ist die gestellte Bedingung erfüllt, soll der Beleg ausgedruckt werden.

Nach der gleichen Methode können Untersuchungen über die Endungen anderer Wortklassen, etwa der Adjektiva oder Pronomina, durchgeführt werden.

Als dritte Variante der verschiedenen Auswertungsmöglichkeiten sei noch genannt, daß das Untersuchungskriterium an einer bestimmten Stelle, vom Wortanfang gerechnet, stehen soll; in diesem Fall kann die zuletzt angegebene SUBSTRING-Funktion verwendet werden, wobei jedoch statt der Variablen Y eine Konstante eingesetzt wird, die angibt, an welcher Stelle des jeweiligen Wortes die Untersuchung beginnen soll.

Für eine Auswertung des Belegmaterials nach grammatischen Kriterien lassen sich im wesentlichen die gleichen Untersuchungs-

programme verwenden, wie sie im vorigen Abschnitt beschrieben sind. In Frage käme z. B. eine Untersuchung, welche Praefixe in einer bestimmten Mundart noch lebendig oder wortbildungsmäßig aktiv sind; ebenfalls könnten die eingespeicherten Komposita nach der gleichen Fragestellung ausgewertet werden. Für lexikographische Zwecke ergibt sich dabei der Vorteil, daß zu jedem Wort festgestellt werden kann, wo und wie oft es als zweites Kompositionsglied auftritt; für das Verfassen des Artikels *Hucht* z. B. würde der Bearbeiter die Information erhalten, daß das betreffende Lemma auch noch in den Komposita *Geilhucht*, *Kophucht*, *Stüfhucht* usw. belegt ist. Dadurch wäre es möglich, die Komposita nicht entweder nur nach ihrem Grundwort oder nur nach ihrem Bestimmungswort einzuordnen, sondern beide Kriterien zu berücksichtigen.

Sofern bei dem eingespeicherten Grammatiksigle auch Kasusangaben gemacht sind, können mit Hilfe der SUBSTRING-Funktion alle Wörter aufgesucht werden, die in einer bestimmten flektierten Form belegt sind, wodurch sich eine monographische Beschreibung des Kasussystems der betreffenden Sprache ermöglichen ließe.

Die Datenverarbeitung ermöglicht einen neuen Typ von Wörterbüchern

Es mag sein, daß die im vorigen Kapitel beschriebene Auswertung von einem Teil der Wörterbuch-Bearbeiter zurückgewiesen wird mit dem Argument, daß derartige Untersuchungen über den Rahmen der Lexikographie hinausgehen würden. In der Tat sind Auswertungen dieser Art bisher kaum erfolgt, was jedoch in dem früher dafür benötigten Zeit- und Arbeitsaufwand begründet ist. Betrachtet man einzelne Punkte der genannten Untersuchungsmöglichkeiten, so zeigt sich, daß z. B. die Registerherstellung oder die Aufnahme der zweiten Kompositionsglieder einen lang gehegten Wunsch der Lexikographie erfüllen könnten. Schwieriger ist dagegen die Frage, ob auch eine Auswertung nach einzelnen sprachlichen Phänomenen erfolgen sollte. Ich glaube, daß diese Frage bejaht werden kann, da hierdurch ohne nennenswerte Mehrarbeit eine lautliche oder grammatische Beschreibung des gespeicherten Materials ermöglicht wird. Die Ergebnisse würden den verschie-

densten Bereichen der Sprachwissenschaft zugute kommen, da sich für viele oft noch ungeklärte Fragen ein umfassendes Belegmaterial für die spätere Interpretation gewinnen ließe.

Zugleich erhält der Benutzer des Wörterbuchs die Möglichkeit, anhand der nach bestimmten Kriterien geordneten Archivbelege zu einer eigenen Deutung sprachlicher Probleme zu gelangen. Welche Untersuchungen durchzuführen und innerhalb des Wörterbuches zu veröffentlichen sind, kann nur im Einzelfall entschieden werden; für den Bereich der Dialektologie etwa würde mir eine Auswertung des Laut -und Flexionssystems lohnend erscheinen, die in einem zweiten, problemorientierten Teil des Wörterbuchs veröffentlicht werden könnte.

Hinweise auf Setzmöglichkeiten

Ein Wörterbuch der beschriebenen Art würde eine erheblich größere Information bieten als die konventionell hergestellten Lexika, zugleich aber auch einen beträchtlich größeren Umfang haben. Aus diesem Grunde dürfte für den Druck zumindest der manuelle Bleisatz als zu langwierig und zu teuer ausscheiden. Es bieten sich nun verschiedene Verfahren an, ein durch die Datenverarbeitung aufbereitetes Material mechanisch setzen zu lassen, die hier kurz erwähnt werden sollen. Als erste Möglichkeit sei genannt, daß die vom Drucker der Maschine ausgegebenen Seiten abphotographiert und dann im Offsetdruck vervielfältigt werden. Sofern einige technische Voraussetzungen, wie z. B. eine gute Justierung des Druckers, erfüllt sind, ist das Druckbild sauber und gut lesbar, wie die von MARVIN SPEVACK herausgegebene Shakespeare-Konkordanz[20] beweisen mag. Bei diesem Verfahren ist allerdings der Zeichenvorrat begrenzt; außerdem sind Fett-, Halbfett- oder Kursivdruck nicht möglich, was jedoch – z. B. für die Auswertungsprogramme – auch nicht erforderlich sein würde.

Zweitens wäre es möglich, daß die Ausgabe der Maschine nicht auf Papier, sondern auf Lochstreifen erfolgt, der dann als Steuerung für eine konventionelle Setzanlage benutzt wird. Da ich praktische Versuche dieser Art nicht durchführen konnte, muß ich für eine

[20] *A Complete and Systematic Concordance to the Works of Shakespeare*, Bd. 1ff., Hildesheim 1968ff.

nähere Information an die jeweiligen Rechenzentren bzw. Verleger verweisen.

Als dritte und wohl eleganteste Möglichkeit käme das Verfahren des Lichtsatzes in Frage. Die folgende Beschreibung bezieht sich auf die Information eines Herstellers (Fa. Dr.-Ing. HELL, Kiel)[21]. Man kann hierbei jede in der Praxis vorkommende Schrift, also auch jedes beliebige Sonderzeichen, programmieren und speichern. Die Schriftzeichen sind in einem besonderen Magazin, das herkömmlichen Setzmaschinen vergleichbar wäre, materielos, d. h. nur magnetisch abgespeichert. Die Eingabe des zu setzenden Textes erfolgt über Lochstreifen oder Magnetband, kann aber auch in direkter Verbindung mit dem Computer geschehen. Wird ein Buchstabe eingelesen, so projiziert ein Kathodenstrahl die im Schriftspeicher enthaltene Form dieses Buchstabens auf eine Fernsehröhre, von der sie über eine Optik auf lichtempfindliches Papier oder Film übertragen wird. Der in einer Dunkelkammer stehende Film-Entwicklungsautomat liefert Filme von exakter Schärfe, die sehr schnell weiterverarbeitet werden können. Die theoretische Höchstgeschwindigkeit für den Satz beträgt 3000 Zeichen pro Sekunde; in der Praxis verlangsamt sich die Geschwindigkeit jedoch durch unterschiedliche Schriftzeichen, Filmtransport usw. Nachträgliche Korrekturen des gesetzten Materials sind selbstverständlich möglich. Es sei abschließend darauf hingewiesen, daß die mechanischen Setzverfahren vor allem für den problemorientierten Teil von Wörterbüchern einen großen Vorteil bieten, da in diesem Fall ein Korrekturlesen der einzelnen Auswertungen nicht mehr erforderlich ist.

[21] Form 50 T 1 – 4 – 6903 (618).