Niederdeutsches Wort

BEITRÄGE ZUR NIEDERDEUTSCHEN PHILOLOGIE

begründet von

WILLIAM FOERSTE †

herausgegeben von

JAN GOOSSENS

Band 18
1978

ASCHENDORFF · MÜNSTER

Das NIEDERDEUTSCHE WORT wird veröffentlicht von der Kommission für Mundart- und Namenforschung des Landschaftsverbandes Westfalen-Lippe unter Mitarbeit der Niederdeutschen Abteilung des Germanistischen Instituts der Universität Münster.

Die Zeitschrift erscheint jährlich in einem Band.

Herausgeber: Prof. Dr. JAN GOOSSENS
Redaktionelle Arbeiten: Dr. GUNTER MÜLLER

Magdalenenstr. 5, 4400 Münster

Copyright © 1979 by Kommission für Mundart- und Namenforschung Westfalens, Magdalenenstraße 5, 4400 Münster

Alle Rechte vorbehalten, insbesondere die des Nachdrucks, der fotomechanischen oder fontechnischen Wiedergabe und der Übersetzung. Ohne schriftliche Zustimmung des Verlages ist es auch nicht gestattet, aus diesem urheberrechtlich geschützten Werk einzelne Textabschnitte, Zeichnungen oder Bilder mittels aller Verfahren wie Speicherung und Übertragung auf Papier, Transparente, Filme, Bänder, Platten und andere Medien zu verbreiten und zu vervielfältigen. Ausgenommen sind die in den §§ 53 und 54 URG genannten Sonderfälle.

Printed in Germany

Aschendorffsche Buchdruckerei, Münster Westfalen, 1979

ISSN 0078-0545

Inhalt des 18. Bandes (1978)

Gunter Müller, Münster

BERICHT ÜBER DIE RECHNERUNTERSTÜTZTE BEARBEITUNG DER WESTFÄLISCHEN TOPONYMIE IN MÜNSTER : DIE FLURNAMEN (I)

0. Anfänge des Projekts

Die umfänglichen Datenmengen, die bei der Aufarbeitung großlandschaftlicher Mikrotoponymien anfallen und der hohe Anteil mechanisierbarer Abläufe innerhalb des Bearbeitungsprozesses haben seit 1969 zu Überlegungen geführt, wie die Möglichkeiten der Linguistischen Datenverarbeitung für onomastische Zwecke genutzt werden könnten. Im Namen des Vereins für Niederdeutsche Sprachforschung hatte deshalb J. Hartig 197o zu einem vorbereitenden Kolloquium unter dem Rahmenthema "Prüfung der Einsatzmöglichkeiten datenverarbeitender Maschinen im Bereich der Flurnamenforschung" eingeladen. Das Kolloquium, an dem Namenforscher sowie an der Datenverarbeitung interessierte Sprachwissenschaftler aus Amsterdam, Groningen, Göttingen, Hamburg, Kiel, Münster und Schleswig teilnahmen, fand dann am 22.5. desselben Jahres im Anschluß an die Pfingsttagung des Vereins in Aurich statt. Die Leitvorstellungen, die in der Auricher Diskussion hervortraten, orientierten sich an den damals schon weiter fortgeschrittenen Thesen zur teilautomatischen Bearbeitung von Dialektwörterbüchern und an dem daraus entwickelten Projekt "Zentrale Speicherung nichthochsprachlichen Materials des Deutschen", das die Einbringung des gesamten bisher in verschiedenen deutschen Wörterbucharchiven lagernden mundartlichen, appellativen Wortmaterials in einen gemeinsamen, maschinell gespeicherten und abrufbaren Datenpool vorsah[1]. Entsprechend

---

1 H. KAMP, *Methoden zur Herstellung und Auswertung von Dialekt-Wörterbüchern mit Hilfe der elektronischen Datenverarbeitung*, NdW 9 (1969) 73-96; G. KESELING, *Mundartwörterbücher und Datenverarbeitung*, ZDL 36 (1969) 31o-326; G. KESELING - B.U. KETTNER - W. KRAMER - W. PUTSCHKE - M. RÖSSING-HAGER - U. SCHEUERMANN, *Richtlinien zur Ablochung*

bemühte man sich in Aurich und am 3.7.197o in Münster in einem erweiterten Kreis um die Festlegung einheitlicher Konventionen für die Erfassung von Flurnamendaten auf maschinenlesbare Datenträger, um nach der Überführung der alten Zettelarchive in elektronisch gespeicherte Dateien eine Vergleichbarkeit der Flurnamensammlungen aus dem niederdeutschen und niederländischen Sprachraum und die Möglichkeiten ihrer zentralen Speicherung und Auswertung gewährleisten zu können. Die vereinbarten Konventionen, die auf einer noch im Oktober 197o in Göttingen durchgeführten Arbeitstagung weiter präzisiert und nach einer längerfristigen Erprobung auf der Anlage des Rechenzentrums der Universität Münster im Oktober 1971 endgültig festgelegt wurden, haben sich während der gesamten, Anfang 1978 abgeschlossenen Überführung des älteren Zettelmaterials des Westfälischen Flurnamenarchivs[2] auf maschinenlesbare Datenträger (d.h. in Münster Datenerfassung auf Lochkarten) als praktikabel erwiesen. Änderungen an den Ablochkonventionen, die in Münster aufgrund später gewonnener Erfahrungen durchgeführt wurden, sind durchweg peripherer Art und heben die Kompatibilität mit Daten, die nach den ursprünglichen Richtlinien angeordnet sind, nicht auf.

Unterschiede in der personellen und technischen Ausstattung der verschiedenen deutschen, niederländischen und belgischen Arbeitsstellen, auch unterschiedliche wissenschaftliche Zielvorstellungen und damit verbunden unterschiedliche Vorstellungen über den Stellenwert der Linguistischen Datenverarbeitung innerhalb des Bearbeitungsablaufs[3] haben

---

und zentralen Speicherung mundartlichen Wortmaterials des Deutschen, GL 2 (197o) 179-242; vgl. auch U. SCHEUERMANN, *Linguistische Datenverarbeitung und Dialektwörterbuch* (Beihefte der ZDL, NF 11), Wiesbaden 1974.

2 Das Westfälische Flurnamenarchiv ist eine Einrichtung der Kommission für Mundart- und Namenforschung Westfalens und befindet sich in den Arbeitsräumen der Kommission, Magdalenenstraße 5, 44oo Münster.

3 Über die ganz anders ausgerichtete EDV-Unterstützung bei der Bearbeitung der südniederländischen Toponyme in Belgien am Instituut voor Naamkunde in Leuven vgl. H. DRAYE, *Naamkundig Repertorium. Een pro-*

verhindert, das entworfene Konzept zu realisieren . Es ist somit dem Vorbildprojekt "Zentrale Speicherung des nichthochsprachlichen Materials des Deutschen" vergleichbar, das ebenfalls an den technologischen und konzeptionellen Divergenzen verschiedener Wörterbuchstellen gescheitert ist.

1. Die Ablochkonventionen

Das ältere Zettelarchiv war so organisiert, daß das gesamte Belegmaterial in drei Serien vorlag. Eine davon ordnete die toponymischen Daten nach der Herkunft, d.h., soweit das möglich war, nach den Gemeinden, in denen sich die mit den jeweiligen Namen belegten Flurorte befanden. Die zweite und dritte Serie, als Kopien der ersten angefertigt, waren nach den 'Bestimmungswörtern' und den 'Grundwörtern' der Namen sortiert. Es ist einsichtig, daß mit diesem Sortiersystem wichtige Informationskategorien, die üblicherweise mit toponymischen Daten verbunden sind, nicht unmittelbar zugänglich waren. Die - unter bestimmten Fragestellungen sehr erwünschte - Möglichkeit, Daten eines bestimmten Überlieferungszeitraums, einer bestimmten Überlieferungsform (z.B. nur Mundartaufzeichnungen), einer bestimmten Namenträgerkategorie (Waldnamen, Wegenamen, Gewässernamen usw.) aus dem Gesamtmaterial aussondern zu können, war somit nicht gegeben, obwohl Datierung, Quellenart, Beschaffenheit des Flurortes usw. nahezu regelmäßige Begleitinformationen der Namenbelege sind. Auch der Zugriff zum Namen selbst war mit der zweifachen Aufschlüsselung nach 'Bestimmungs-' und 'Grundwort' unzureichend, weil damit Fragen nach der syntaktischen Strukturierung[4], nach Präpositionen und Ortsadverbien oder nach den Mittelsegmenten in drei- und mehrgliedrigen Toponymen nur schwer beantwortet werden konn-

---

ject voor de machinale bewerkung van de onuitgegeven toponymische documentatie uit Nederlandstalig België, Naamkunde 2 (197o) 231-235; J. MOLEMANS, *De machinale bewerkung van het onuitgegeven Zuidnederlands toponymisch materiaal*, Naamkunde 4 (1972) 26o-277.

4 Vgl. etwa K.-F. HILLESHEIM - W. HÜLS - G.MÜLLER - H. TAUBKEN, *Zur Struktur westfälischer Flurnamen*, NdW 13 (1973) 88-99.

ten. Die neue Datenorganisation mußte daher, sollte sie einen Fortschritt bringen, gegenüber der bisherigen dahingehend verändert werden, daß ein insgesamt verbesserter Informationszugriff erreicht wurde. Dies erforderte eine Klassifizierung der mit einem Flurnamen üblicherweise verbundenen Informationen zu verschiedenen Kategorien und eine konsequente Verschlüsselung der Informationen. Die ausführlichen Kodierungsvereinbarungen brauchen hier nicht explizit vorgeführt zu werden, Hinweise sollen genügen.

1.1. *Die Informationskategorien*

Kat.1: Adresse 1. Es ist eine Nummer, die in Verbindung mit Kategorie 2 für jeden Beleg des Archivs eine eigene, unverwechselbare Adresse ergibt, mit der dieser Beleg identifiziert werden kann.

Kat.2: Lokalisierung des Flurnamens durch Angabe der Gemeinde nach dem Stand der Kommunalgliederung vom 1.6.1961[5]. Die Gemeinde wird kodiert mittels einer dreistelligen Kreis- und zweistelligen Ortssigle nach dem Abkürzungsprinzip des Westfälischen Wörterbuches[6].

Kat.3: Adresse 2. Sämtliche Belege e i n e s Namens erhalten eine gemeinsame, aus einer Zahl oder einer Ziffern-Buchstabenkombination bestehende Adresse.

Kat.4: Sammlersigle. Sind die Daten nicht direkt durch das Archiv erhoben, sondern von einem freiwilligen Mitarbeiter dem Archiv zugesandt worden, so wird eine den Sammler identifizierende Sigle in die Belegeinheit aufgenommen. Sie verhindert, daß das von auswärts dem Archiv zur Verfügung gestellte Material 'namenlos' im Gesamtmaterial aufgeht, und erlaubt auch eine gewisse Einschätzung der Belegqualität, da die Genauigkeit einzelner Sammler bei der Wiedergabe archivalischer Belege sowie ihre Verschriftungspraxis bei mundartlichen Namenformen sehr unterschiedlich ist.

Kat.5: Datierung. Neben die Datierung des Namenbeleges können hier auch - was vor allem bei mittelalterlicher Überlieferung wichtig ist - Angaben über originale oder kopiale Überlieferung, die Art der kopialen Tradition, über urkundlich bezeugte oder erschlossene Datierung u.ä. eingetragen werden.

Kat.6: Angaben zum Bewuchs, zur Nutzung bzw. Bebauung des benannten Geländes (z.B. A = Acker; B = Haus, Hof; G = Grasland; F = Wald; S = Weg, Straße; T = Garten).

---

5 *Amtliches Verzeichnis der Gemeinden und Wohnplätze (Ortschaften) in Nordrhein-Westfalen*, Düsseldorf 1962.

6 *Westfälisches Wörterbuch. Beiband*, bearb. v. F. WORTMANN, Münster 1969, S.21-46.

Kat.7: Angaben zum Relief und zur sonstigen Beschaffenheit des benannten Geländes (z.B. B = Bruch, Marsch; E = ebenes Gelände; H = Höhenlage, Berg; R = fließendes Gewässer).

Kat.8: Überlieferungstyp (z.B. M = von einem Gewährsmann oder Explorator aufgezeichnete, mündlich gebrauchte Namenform; S = schriftliche Überlieferung in Akten, Katasterbüchern u.ä.).

Kat.9: Quellenangaben. Bei bereits veröffentlichten Namen werden hier Angaben über den Druckort gemacht. Bei nicht publizierten Materialien können hier, soweit bekannt, Angaben über den archivalischen Fundort (Archivsiglen) eingetragen werden.

Kat.1o: Belegteil. Der Namenbeleg wird in der Form eingetragen, in der er in der Quelle vorliegt bzw. von Sammlern aufgezeichnet worden ist. Ein vereinbartes System diakritischer Zeichen stellt sicher, daß auch phonetische Notationen der Vorlage - Angaben zu Quantität/Qualität von Vokalen, Stimmhaftigkeit/Stimmlosigkeit, Betonung usw. - und Besonderheiten älterer Schriftsysteme mit einem Lochkartenlocher kodiert und von den üblicherweise zur Verfügung stehenden 6o-Zeichen-Druckerketten der Schnelldrucker dargestellt werden können[7].

Kat.11: Der Lemmateil. Zu ihm → 3.5. und 3.6.

Die Informationen der Kategorien 2 - 1o, die maschinell nicht generiert werden können, wurden so formalisiert und angeordnet, daß sie auf der 8o-spaltigen Normallochkarte Platz finden. Die Adresse 1 (Kat.1) wird bei Übertragung der Lochkartendaten auf Magnetband automatisch festgelegt, die Lemmatisierung (Kat.11) so weit wie möglich maschinell erstellt (→ 3.5.). In keinem Fall werden jedoch Informationen der Kategorie 11 bereits bei der Datenersterfassung auf der Lochkarte mitgeteilt.

1.2. *Die Belegsegmentierung*

Obwohl die Durchführung automatischer Lemmatisierung (AL)[8]

---

7 Vgl. entsprechende Kodierungsvereinbarungen bei KESELING - KETTNER (wie Anm.1) S.18of.

8 Die Bemühungen, zu automatisierten Lemmatisierungsverfahren zu gelangen, standen zunächst im unmittelbaren Zusammenhang mit der Entwicklung automatischer Syntaxanalyse und Sprachübersetzung. Sie erhielten innerhalb der deutschen Linguistik einen wesentlichen Motivationsschub durch das Ungenügen an reinen Wortformenindices, die die Frühphase der Linguistischen Datenverarbeitung beherrschten, vgl. SCHEUERMANN (wie Anm.1) S.7f.; W. LENDERS, *Lexikographische Arbeiten zu Texten der älteren deutschen Literatur mit Hilfe von Datenverarbeitungsanlagen*, ZdPh 9o (1971) 321-336, hier S.324ff. Grundlegend

von orthographisch nicht normiertem Wortmaterial zum Teil auf erhebliche Schwierigkeiten stößt und bei bestimmten Problemlagen wohl auch nicht sinnvoll ist[9], war man sich bei Festlegung der Ablochkonventionen darüber einig, für die weitere Bearbeitung der mikrotoponymischen Daten zumindest ein teilautomatisiertes Lemmatisierungsverfahren anzustreben. Dabei war auch klar, daß ein maschinelles Parsing[10] bzw. eine Morphemanalyse[11], wie sie für die AL orthographisch standardisierter Wortmaterialien möglich ist, für mund-

---

zur automatischen Lemmatisierung ist: R. DIETRICH, *Automatische Textwörterbücher. Studien zur maschinellen Lemmatisierung verbaler Wortformen des Deutschen*, Tübingen 1973; vgl. weiter H. EGGERS u.a., *Elektronische Syntaxanalyse der deutschen Gegenwartssprache*, Tübingen 1969, S.64ff.; H.D. MAAS, *Homographie und maschinelle Sprachübersetzung* (Linguistische Arbeiten des Germanistischen Instituts und des Instituts für Angewandte Mathematik der Universität des Saarlandes), Nr.8, Saarbrücken 1969; W. KLEIN - R. RATH, *Automatische Lemmatisierung* (Linguistische Arbeiten des Germanistischen Instituts und des Instituts für Angewandte Mathematik der Universität des Saarlandes, Nr.1o), Saarbrücken 1971; R. RATH, *Vorschläge zur Automatischen Lemmatisierung (AL) deutscher Adjektive*, Linguistische Berichte 12 (1971) 53-59; H.D. MAAS, *Homographie und Maschinelle Sprachübersetzung*, Beiträge zur Linguistik und Informationsverarbeitung 21 (1971) 7-35; H.H. ZIMMERMANN, *Zur Auflösung von Mehrdeutigkeiten bei einer maschinellen Analyse des Deutschen*, ebd. S.36-49; W. v. HAHN - W. HOEPPNER, *HAM2 - Ein Algorithmus zur Lemmatisierung deutscher Verben*, in: *Neue Forschungen in Linguistik und Philologie. Aus dem Kreise seiner Schüler L.E. Schmitt zum 65. Geburtstag gewidmet* (ZDL, Beiheft 13), Wiesbaden 1975, S.151-171; W. v. HAHN - H. FISCHER, *Über die Leistung von Morphologisierungsalgorithmen bei Substantiven*, ebd., S.13o-15o.

9 So ging auch das Projekt "Zentrale Speicherung nichthochsprachlichen Materials" von der Voraussetzung aus, daß die normierten Wortansätze von den Bearbeitern nach bestimmten Richtlinien für jedes Belegwort einzeln festgelegt und nicht über einen Lemmatisierungsalgorithmus automatisch generiert werden sollten, KESELING - KETTNER u.a. (wie Anm.1)S.197f.;vgl. SCHEUERMANN (wie Anm.1) S.8; vgl. auch J. SPLETT, *Verfahrensweisen zur grammatikalischen Auswertung althochdeutscher Glossen mit Hilfe elektronischer Datenverarbeitung*, in: W. LENDERS - H. MOSER (Hrg.), *Maschinelle Verarbeitung altdeutscher Texte II: Beiträge zum Symposium Mannheim 15./16.Juni 1973*, Berlin 1978, S.147-181, hier S.15off.

1o W. KLEIN, *Parsing. Studien zur maschinellen Satzanalyse mit Abhängigkeitsgrammatiken und Transformationsgrammatiken*, Frankfurt 1971.

11 Vgl. v. HAHN - FISCHER (wie Anm.8).

artliches, historisches oder sonst unnormiertes Sprachgut nicht in Frage kam[11a]. Andrerseits führt eine kontextfreie, auf syntaktische Merkmale - wozu hier auch Wortbildungsmerkmale gerechnet werden sollen - nicht rekurrierende AL zumindest bei toponymischem Wortmaterial mit Sicherheit zu unbefriedigenden Ergebnissen. Deshalb mußten mit den Daten zugleich auch Hilfen zur Erkennung syntaktischer Strukturen eingegeben werden. Solche Hilfen wurden für das Ablochen der Flurnamenbelege in Form von Steuersignalen vereinbart, die die Stellung eines "Namenwortes" (→ 2., S.145) innerhalb der proprialen Nominal- oder Präpositionalphrasen bzw. innerhalb proprialer Komposita markieren sollten:

Beleg	Darstellung im Belegteil (Kat.1o)
(1) *die rode bieke*	DIE <RODE >BIEKE
(2) *op der lannwer*	¬OP DER >LANN/WER
(3) *bierbaumsgorn*	>BIER/BAUMS/GORN
(4) *boven der kalten hovestat*	¬BOVEN DER <KALTEN >HOVE/STAT
(5) *Koerts höffken*	<KOERTS >HO"FFKEN[12]
(6) *Koldehoffs graute wolflage*	<KOLDE/HOFFS <GRAUTE >WOLF/LAGE

Dabei wurde das Steuerzeichen '>' zur Markierung des Namenkerns, '<' für Attribute (sowohl Adjektiv- wie Substantivattribute, vgl. (1),(4),(5) und (6)), '¬' für Präpositionen und Orts-/Richtungsadverbien (vgl. (2) und (4)) und '/' zur Markierung von Fugen bei Zusammensetzungen vereinbart. Eine Segmentierung von Wortbildungssuffixen wie in >BUSCHEI , >HO"FFKEN , >HO"RSTING oder >DO"RNTE wurde nicht verein-

---

11a Zwar scheinen auch hier Fortschritte möglich, wie ein von A. LOHR entwickeltes Programm zur automatischen Lemmatisierung (Veröffentlichung Onoma 1979, im Druck) altdeutscher Personennamen zeigt. Das Programm analysiert ohne Hilfen bei der Dateneingabe (Hilfssegmentierung, s. im folgenden den Haupttext) die Syntax von Personennamen (etwa Namenwort + Fugenzeichen + Namenwort + Flexiv; Namenwort + Suffix(e) + Flexiv usw.). Ob es sich auch für die variablere und daher weniger leicht formalisierbare Syntax der Flurnamen mit ihrem der Appellativlexik gegenüber offeneren und vieldeutigen Wortschatz adaptieren ließe, muß abgewartet werden.

12 Umlaute werden nach den geltenden Ablochkonventionen durch nachgesetzte doppelte Hochkommata (A", O", U") umschrieben.

bart; dasselbe gilt für Präfixableitungen und Partikelkomposita (Präpositional- und Adverbialkomposita) wie >GEHEGEDE, >ANWENDE , >ANEWEIDE , >AFGUNST , >UNLAND , >UMMEGANG usw., sofern sie nicht deutlich erst sekundär aus einer Zusammenrückung entstanden sind: >ACHTER/WEIDE ← *achter der Weide*, >TOM/BRINK ← *to dem Brink*. Zur Begründung → 2., S.148. Artikel (vgl. (1), (2) und (4)) wurden nicht markiert.

2. Voraussetzungen proprialer Segmentierung und Lemmatisierung

Der Terminus Lemma wird nicht einheitlich verwendet. Einerseits gilt er synonym für 'Stichwort', also für ein Label, mit dem eine Anzahl zusammengehöriger Wortformen angesprochen werden kann[13], andrerseits wird Lemma als eine Menge von Wortformen mit übereinstimmenden semantischen und paradigmatischen Merkmalen definiert[14]. Für das Label, das diese Menge benennt, ist die Bezeichnung Lemmaname üblich geworden. Diese terminologische Regelung wird im folgenden übernommen, wobei für einen onomastischen Lemmabegriff das Kriterium der Klassifizierung allerdings nicht in der Identität semantischer Merkmale liegen kann. Eine Lemmadefinition, die für die Elemente eines Lemmas die semantische Gleichheit voraussetzt, ist ohnehin nur für streng synchronische und auf ein einziges sprachliches System bezogene Wörterbücher/Wortschatzbeschreibungen brauchbar, während historische und mundartliche Wörterbücher Morpheme und Morphemkombinationen zu einem Lemma zusammenfassen müssen, bei denen zumindest die semantischen Merkmale fast immer voneinander differieren. Ihnen liegt ein diasystematischer Lemmabegriff zugrunde, der alle die Wortformen zu einem Lemma zusammenzufassen erlaubt, die sich aufgrund phonologischer, morphologischer und seman-

13 Vgl. G. WAHRIG, *Anleitung zur grammatisch-semantischen Beschreibung lexikalischer Einheiten*, Tübingen 1973, S.41f.; DIETRICH (wie Anm.8) S.1f.

14 RATH (wie Anm.8) S.54f.; MAAS 1971 (wie Anm.8) S.9ff.; DIETRICH (wie Anm.8) S.1f.

tischer Regeln auf eine gemeinsame - üblicherweise historische - Bezugsgröße projizieren lassen. Ein propriales Lemma könnte man dann definieren als eine Menge proprialer Einheiten, deren gemeinsames Merkmal darin besteht, daß sie alle von Elementen desselben appellativen Lemmas[15] abgeleitet sind. Diese proprialen lexikalischen Einheiten sind dabei nicht als Morpheme oder Morphemkombinationen mißzuverstehen, da der Morphembegriff auf Propria schwer anwendbar ist. Das Fehlen einer lexikalischen Eigenbedeutung proprialer Zeichen ist der Grund dafür, warum sich ein propriales Lemma nicht autonom, sondern nur im Rekurs auf ein appellatives Lemma definieren läßt.

Die areale und historische Staffelung des toponymischen Wortschatzes setzt dabei voraus, daß bei seiner Lemmatisierung die zugrunde gelegten appellativen Lemmata diasystematisch gefaßt werden, denn die Elemente eines proprialen Lemmas werden in der Regel in verschiedenen Zeiten/Räumen abgeleitet sein. Onomastisches Lemmatisieren ist somit ein Verfahren, daß dem Etymologisieren sehr nahe kommt, mit diesem aber nicht verwechselt werden darf[16].

Nomina propria gelten als bedeutungslos. Aber auch dort, wo ihnen eine wie immer definierte semantische Valenz zugeordnet wird[17], geht man von einer nicht zerlegbaren Be-

15 "Appellativ" hier umfassend im Sinn von "nicht-proprial" gebraucht.

16 Vgl.T. WITKOWSKI, *Zu einigen Problemen der Bedeutungserschließung bei Namen*, Onoma 18 (1974) 319-336.

17 Zur Diskussion um die "Bedeutung" von Eigennamen: P. v. POLENZ, *Name und Wort. Bemerkungen zur Methodik der Namendeutung*, Mitteilungen für Namenkunde, Heft 8 (1960/61) 9; H. VATER, *Eigennamen und Gattungsbezeichnungen*, Muttersprache 75 (1965) 208; F. DEBUS, *Aspekte zum Verhältnis Name - Wort*, Groningen 1966, S.14; B. SCIARONE, *Proper Names and Meaning*, Studia Linguistica 21 (1967) 73-86; W. FLEISCHER, *Die deutschen Personennamen. Geschichte, Bildung und Bedeutung*, Berlin ^{2}1968, S.7; F. ZABEEH, *What is a name ?* The Hague 1968, S.1of.; W.P. SCHMID, *Skizze einer allgemeinen Theorie der Wortarten* (Abhandlungen der Akademie der Wissenschaften und der Literatur Mainz, geistes- u. sozialwiss. Kl., Jg.197o, Nr.5), Wiesbaden 197o, S.15f.; R. WIMMER, *Der Eigenname im Deutschen. Ein Beitrag zu seiner linguistischen Beschreibung* (Linguistische Arbeiten, 11), Tübingen 1973, S.1off.; dazu W. VAN LANGENDONCK, *Über das Wesen des Eigennamens*, Onoma 18 (1974) 337-361. Das Problem ist zuletzt zusammenfassend diskutiert (mit Angabe weiterer Literatur) bei H. KALVERKÄMPER, *Textlinguistik der Eigennamen*, Stuttgart 1978, S.58ff., bes.62ff.

deutung aus, d.h., Morphemstatus kann höchstens dem Eigennamen selbst[18], nicht aber Teilen von ihm zugebilligt werden: Propria wie *Von der Mühl* oder *Weißenburg* gelten als morphologisch nicht segmentierbare Simplicia, Amalgame vorpropriale Morphemketten[19].

Andrerseits läßt sich nicht leugnen, daß solche "Amalgame" wie *Von der Mühl* und *Weißenburg* sich in kleinere Einheiten zerlegen lassen und dies nicht nur mit sprachhistorisch-etymologischen Analyseverfahren, sondern mit der Kompetenz jedes deutschen Sprechers. Für diese Einheiten hat sich bis jetzt im terminologischen Inventar der strukturellen Sprachbeschreibung keine verbindliche Benennung durchgesetzt. Ihre Bezeichnung schwankt zwischen Namenstamm[20], Namenwort[21], Namenelement[22], Namenbestandteil[23], Eigennamen-Teil[24], Namenglied[25] und Namensegment[26]. Hier wird im folgenden S e g m e n t

18 Vgl. WIMMER (wie Anm.17) S.19ff.

19 WIMMER (wie Anm.17) S.19ff., 47ff.; O. LEYS, *Der Eigenname in seinem formalen Verhältnis zum Appellativ*, BNF NF 1 (1966) 118; die Verwendung des Terminus M o r p h e m für Eigennamenteile bei E. EICHLER, *Zur morphematischen Struktur der Substratonomastik*, in: *Probleme der strukturellen Grammatik und Semantik*, Leipzig 1968, S.243-252, ist daher trotz aller vorgetragenen Einschränkungen unglücklich, vgl. die Kritik bei WIMMER, S.48ff.

2o Vgl. etwa E. FÖRSTEMANN, *Altdeutsches Namenbuch I: Personennamen*, Bonn ^{2}19o1, Vorwort S.V; H. KAUFMANN, *Ergänzungsband zu E. Förstemann, Altdeutsche Personennamen*, München Hildesheim 1968, passim; J.M. PIEL - D. KREMER, *Hispano-gotisches Namenbuch*, Heidelberg 1976, passim.

21 Vgl. etwa H. KAMP, *Ein Algorithmus zur automatischen Lemmatisierung von Personennamen*, in: *Die Klostergemeinschaft im früheren Mittelalter*, hrg. v. K. SCHMID, Bd.1, München 1978, S.9off.; D. GEUENICH, *Die Personennamen der Klostergemeinschaft von Fulda im früheren Mittelalter*, München 1976, S.24 u.ö.; D. GEUENICH, *Die Lemmatisierung und philologische Bearbeitung des Personennamenmaterials*, in: *Die Klostergemeinschaft von Fulda im früheren Mittelalter* (wie oben), S.38ff.

22 Vgl. etwa G. MÜLLER, *Studien zu den theriophoren Personennamen der Germanen* (Niederdeutsche Studien, 17), Köln 197o, passim; GEUENICH, *Personennamen* (wie Anm.21) S.3o.

23 Vgl. etwa EICHLER (wie Anm.19) S.244.

24 WIMMER (wie Anm.19) S.47.

25 Vgl. etwa KAMP (wie Anm.21) S.88ff.

26 Vgl. etwa EICHLER (wie Anm.19) S.245, 249f.; KAMP (wie Anm.38) S.3.

als neutralste und mit Konnotationen am wenigsten belastete Bezeichnung eingeführt.

Sieht man von morphonologischen Regularitäten ab, die die Grenze zwischen zwei Morphemen auf der Ausdrucksseite markieren können und deren Kenntnis auch auf die Segmentierung proprialer Phonemsequenzen angewandt werden kann[27], so beruht die Segmentierbarkeit komplexer Propria auf einem Phänomen, das als semantische Motivation[28], Transparenz[29] oder Durchsichtigkeit[30] beschrieben worden ist. Einem Proprium können Appellative im Bewußtsein der Sprachteilnehmer kopräsent sein, wobei diese Kopräsenz vom Sprachbenutzer als derivationelle Abhängigkeit empfunden wird[31]. Die Wörter *zwölf*, *Scheffel*, *Saat* sind dem Eigennamen *Zwölfscheffelsaat* kopräsent und dieser wird auch "von diesen Wörtern herkommend" aufgefaßt. Die Kopräsenz kann so stark sein, daß sich propriale Segmente grammatisch völlig konform zu den Appellativen verhalten, auf die sie zurückweisen. Die in den Mikrotoponymien häufigen, aber auch sonst geläufigen "Mehrwortnamen" flektieren innerhalb des Namens regelrecht (*Rotes Meer*, zum *Roten Meer*, das *Rote Meer*) und stellen somit sowohl für die Auffassung vom Proprium als e i n e m Morphem wie auch für alle Versuche, zu einer akzeptablen Wortdefinition zu gelangen, eine Crux dar[32]. Jedenfalls verhalten sich innerpropriale Flexive funktional wie Flexions-

---

27 Ein solches intuitives Anwenden von Regeln der Lautkombinatorik lag etwa vor, wenn Hilfskräfte, die das Ablochen von Flurnamendaten durchführten, Belege wie *Doigtrüen*, *Hanckfoete* und *Wuirlauke* durchaus richtig als DOIG/TRU"EN, HANCK/FOETE und WUIR/LAUKE segmentierten, obwohl die Belege keineswegs zu "durchschauen" waren.

28 Für Eigennamen z.B. EICHLER (wie Anm.19) S.248; HILLESHEIM u.a. (wie Anm.4).

29 Für Eigennamen z.B. KALVERKÄMPER (wie Anm.17) S.11of.

30 H.-M. GAUGER, *Wort und Sprache*, Tübingen 197o, S.113ff.; H.-M. GAUGER, *Durchsichtige Wörter. Zur Theorie der Wortbildung*, Heidelberg 1971.

31 GAUGER, *Wort und Sprache* (wie Anm.3o) S.115; GAUGER, *Durchsichtige Wörter* (wie Anm.3o) S.8ff.

32 O. REICHMANN, *Deutsche Wortforschung*, Stuttgart 1969, S.3; O. REICHMANN, *Germanistische Lexikologie*, Stuttgart 1976, S.6.

morpheme[33]. Die grammatische Affinität der "Mehrwortnamen" zu normalen nominalen Syntagmen ist oft so groß, daß man die Segmentketten vereinfachend mit demselben Begriffsapparat (Nominalphrasen, Präpositionalphrasen usw., → 1.2.) beschreiben kann. Daß ein durchsichtiges Verhältnis Proprium ← Appellativum ein Faktor bei der Strukturierung des Lexikons ist, zeigen die vielen Fälle, in denen nach Verlust der Homophonie Proprium - vorpropriales Appellativum oder bei Entlehnung fremdsprachlicher Eigennamen durch Umformung eine neue "Durchsichtigkeit" erreicht wurde, ein Vorgang, der teils als "Volksetymologie"[34], teils als "Resemantisierung" beschrieben wurde. Was entsteht, ist die Neuerstellung einer derivationellen Abhängigkeit, nicht die einer "Bedeutung" des Propriums. Umformungen wie *Aubruch* → *Ehebruch*, *Vatrosiby* → *Wassersuppe*, *Clebeloc* → *Knoblauch*[35] lassen erkennen, daß die Eigenschaften des Objekts, auf das sich die Namen referentiell bezogen, für die Umformung nur beschränkt eine Rolle spielten.

Die als derivationelle Abhängigkeit verstandene Durchsichtigkeit kann unterschiedlich ausgeprägt sein. Sie reicht von vollständiger über mehr oder weniger gestörte Homophonie (solange die Störung die Durchsichtigkeit noch nicht aufhebt)[36] bis hin zu unterschiedlichen Wortbildungen (z.B. Prop. *Buschei* ← App. *Busch*), die man etwa in der Gaugerschen Terminologie als ausgreifende oder variierende Bil-

---

33 Das gilt auch für andere Repräsentationen grammatischer Morpheme in proprialen Segmentketten (Artikel, Fugenzeichen, Steigerungsmorpheme usw,). Es gibt innerhalb von Segmentketten auch Partikel, die zumindest in der selben Stellung appellativ nicht vorkommen und dennoch voll die Funktion grammatischer Morpheme ausfüllen, vgl. EICHLER (wie Anm.19) S.248.

34 Vgl. W. SANDERS, *Zur deutschen Volksetymologie 3: Volksetymologie und Namenforschung*, NdW 15 (1975) 1-5.

35 Vermittelt über mnd. *kloflōk* 'Knoblauch'. Die Beispiele nach LEYS (wie Anm.19) S.115; EICHLER (wie Anm.19) S.247; R. FISCHER, *Brandenburgisches Namenbuch 4: Die Ortsnamen des Havellandes*, Weimar 1976, S.144f., 227.

36 Zu den Ursachen gestörter Homophonie vgl. etwa LEYS (wie Anm.19) S.118.

dungen beschreiben könnte[37].

Wird onomastisches Lemmatisieren - in Präzisierung der obigen Definition (S.144) - als Klassifizieren proprialer Einheiten mit identischer derivationeller Abhängigkeit verstanden (was impliziert, daß die erwähnten "Resemantisierungen" eine Überführung des Propriums in einen anderen Lemmaverband bewirken), dann dürfen die als Hilfen für eine automatische Lemmatisierung eingeführten Segmentmarkierungen die Propria nach Möglichkeit nicht in quasi-morphemische Einheiten untergliedern, sondern müssen Ableitungseinheiten absondern. Daraus folgt, daß Präfixableitungen und Partikelkomposita (*Anwende, Gehegede, Afgunst*) nicht morphologisch (AN/WENDE , GE/HEGEDE , AF/GUNST) zu segmentieren waren, da sie offensichtlich auf appellative Ableitungen (*ānewand, gehi^{e}gede, afgunst*) zurückweisen (→ 1.2.). Etwas anders verhält es sich bei den Wortbildungssuffixen. Hier gibt es echte propriale Suffixe, die Ableitungen ergeben können, welche niemals eine unmittelbare appellative Grundlage hatten, so z.B. bei den Hofnamen des Typus *Büsching*. Sie sind aber von solchen, die direkt aus einer appellativen Suffixableitung übernommen wurden, vielfach nicht auszusondern, so daß es sinnvoller ist, auch auf eine Suffixmarkierung zu verzichten (→ 3.4.). Das Problem stellt sich ebenfalls bei den Zusammensetzungen. Auch hier gibt es toponymische Verbindungen ohne idiomatisierte Appellativvorbilder (*Heister/weg, Ssiegen/hiege, Laisch/bieke*) und Komposita, die direkt aus Appellativzusammensetzungen entwickelt sind (*Bierbaum, Helweg, Hopfenblome* (zu *hoppenblō^{1}me* 'Zaunwinde')) und bei denen man erwägen könnte, nicht zu segmentieren (und sie damit als Elemente e i n e s Lemmas zu interpretieren). Auch hier scheitert ein differenziertes Verfahren an der Unmöglichkeit, die beiden Gruppen ausreichend voneinander trennen zu können. Nimmt man den Aufwand hinzu, den ein solches differenziertes Behandeln der Kompositionsfugen durch die vielen, beim Belegsegmen-

37 GAUGER (wie Anm.3o) S.7o.

tieren anfallenden philologischen Entscheidungen zur Folge hätte, dann bleibt nur übrig, sämtliche Kompositionsfugen zu markieren, was überdies noch Vorteile hat (dazu → 5.1.; vgl. 3.2.5.).

Bisher war nur von der Ableitung Appellativ → Proprium die Rede. Hinzu tritt die in Mikrotoponymien verbreitete Ableitung Proprium → Proprium. Gemeint sind Flurnamen, von denen ein Teil aus einem vorgegebenen Anthroponym (*Koerts höffken*), einem Hydronym (*Geithebruch*, zum Flußnamen *Geithe*) oder einem Siedlungsnamen (*Soestweg*) abgeleitet ist. Von diesen Anthropo-, Hydro- und Toponymen sind die Flurnamen derivationell direkt abhängig und von ihnen her sind deshalb auch die Lemmata zu definieren: KONRAD={*Conrad, Koert, Kordt, Käort* ...}, GEITHE={*Geite, Gaite, Geete* ...}, SOEST={*Soest, Soost, Saust* ...}. Das muß dann nicht nur für die Fälle gelten, in denen die propriale Lemmabasis etymologisch verdunkelt erscheint, sondern auch dann, wenn sie durchsichtig ist. Ein Flurname *Holthuser Feltkamp* (in der Nähe von Holthausen, Kr. Steinfurt) ist demnach nicht zu HOLZ, HAUS, FELD, KAMP, sondern zu HOLTHAUSEN, FELD, KAMP zu lemmatisieren. Daraus folgt weiter (zumindest für Siedlungs- und Gewässernamen), daß die Derivate verschiedener, aber sprachlich identischer Propria (etwa Holthausen, Kr. Steinfurt, und Holthausen, Ennepe-Ruhr-Kreis) nicht einem gemeinsamen, sondern zwei verschiedenen Lemmata (HOLTHAUSEN1, HOLTHAUSEN2) zuzuordnen sind. Dazu weiter → 3.2.1. und 3.2.5.1.

3. Die automatische Lemmatisierung (AL)[38]

3.1. *Voraussetzungen der AL*

Der Entscheidung, eine AL von Flurnamen durchzuführen, haben zwei Vorfragen vorauszugehen:

---

38 Die folgenden Darlegungen zur AL verdanken außerordentlich viel dem am Sonderforschungsbereich 7 der Universität Münster durchgeführten Projekt 'Personen und Gemeinschaften', in dem unter Zusammenarbeit von Philologen, Historikern und Mitarbeitern des Rechenzentrums der

1. Ist ein solches Verfahren gegenüber der traditionellen "manuellen" Lemmatisierung arbeitsökonomisch vertretbar ?
2. Lassen sich die Kriterien, nach denen ein geschulter Bearbeiter Lemmazuweisungen für Flurnamen(segmente) vornimmt, soweit formalisieren, daß die Ergebnisse der AL sich qualitativ mit der manuellen Lemmatisierung vergleichen lassen ?

Ad 1. Die Motivation dafür, die Lemmatisierung von einem Rechner maschinell durchführen zu lassen, geht vor allem von der Voraussetzung aus, daß die verwendeten toponymischen Segmente in der Regel in mehr als nur einem Flurnamen vorkommen, d.h., daß die Entscheidung, ein bestimmtes Segment S einem Lemma L zuzuweisen, normalerweise mehrmals gefällt werden muß. Ein Verfahren, das es erlaubt, eine einmal getroffene linguistische Entscheidung - hier eine Lemmazuweisung - maschinell beliebig oft zu wiederholen, wird gegenüber der manuellen Reproduktion umso rationeller sein, je größer die Zahl der Wiederholungen ist. Zwar handelt es sich bei der mikrotoponymischen Lexik - verglichen mit anderen proprialen Subsystemen - um einen der Appellativlexik gegenüber jederzeit recht offenen Wortschatz[39], doch ist nicht zu verkennen, daß der mikrotoponymische "Kernwortschatz" relativ klein ist und einer hohen Gebrauchsfrequenz unterliegt. *Kamp*, *feld*, *heide* usw. kommen in einer begrenzten Menge von Schreibungen/Lautungen fast unzählbar häufig in der westfälischen Toponymie vor. Die dominierende Stellung des begrenzten Kernwortschatzes läßt sich an den folgenden Daten ungefähr

---

Universität Münster die linguistischen Grundlagen und programmtechnischen Voraussetzungen für die AL von Namen (frühmittelalterlichen Personennamen) geschaffen worden sind. Besonders danken möchte ich dabei Hermann Kamp vom Rechenzentrum der Universität Münster, der das Programm zur Personennamenlemmatisierung entwickelt hat und der den Aufbau des Programms zur Flurnamenlemmatisierung, das infolge der unterschiedlichen Problemlage anders konzipiert werden mußte, immer mit Rat und Geduld unterstützt hat. An Literatur zum münsterschen Projekt sind zu nennen: H. KAMP, *Die automatische Lemmatisierung frühmittelalterlicher Personennamen*, Phil. Diss. Münster 1976; KAMP (wie Anm.21); GEUENICH, *Lemmatisierung* (wie Anm.21); GEUENICH, *Personennamen* (wie Anm.21).

39 Ein recht geschlossenes propriales System, in das kaum Neuerungen aus dem Appellativwortschatz mehr eindringen konnten, bildeten z.B. die altdeutschen Personennamen.

ablesen.

Das der Sammlung von Schoppmann[4o] entnommene Material des Flurnamenarchives für den Kreis Soest umfaßt 15.24o Einzelbelege mit 28.o25 Segmenten (ohne Präpositionen und Adverbien, → 3.5.1.), unter denen sich 6.365 unterschiedlich geschriebene Segmente (= V a r i a n t e n)[41] befinden. Jede Variante wiederholt sich in der angegebenen Datei also durchschnittlich 4,4mal. Die folgende Liste und die Kurven S.152 zeigen allerdings, daß die tatsächliche Häufigkeit einzelner Varianten und ihr prozentualer Anteil an der Gesamtzahl aller Segmente von diesem Mittelwert sehr stark differieren können.

lfd. Nr.	Häufigkeit	Variante
1	1o6o	KAMP
2	45o	WEGE
3	34o	WEG
4	337	KAMPE
5	298	WIA"GE
⋮		⋮
1o	15o	LANDE
⋮		⋮
2oo	2o	BRU"CKE

Die Liste gibt in Ausschnitten den nach Häufigkeit ihres Vorkommens geordneten Bestand der 2oo im Soester Material geläufigsten Varianten wieder[42], die durchgezogene Kurve S.152 stellt den prozentualen Anteil dieser 2oo Varianten am Gesamtbestand der Segmente dar[43]. Aus ihr ergibt sich, daß

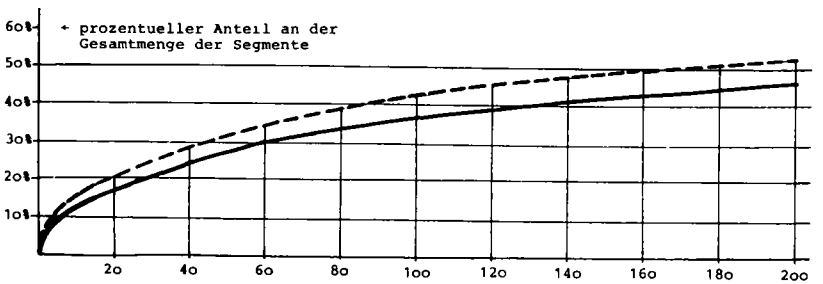
4o H. SCHOPPMANN, *Die Flurnamen des Kreises Soest*, 2 Bde, Soest 1936-194o.

41 Im folgenden sollen alle gleich geschriebenen Elemente (gleich geschrieben nach Abzug der diakritischen Zeichen, → 3.2.3.) als Repräsentanten einer Variante gelten. Nach dieser terminologischen Regelung enthalten die Belege >BOHNEN/FELD und >FELD/KAMP vier Segmente und drei Varianten. Zu einer präziseren Definition von Variante → 3.2.4.

42 Eine vergleichbare Anordnung von Varianten nach der Häufigkeit ihres Auftretens findet sich bei KAMP (wie Anm.38) S.1o7-111, vgl. S.73ff.

43 Auf der x-Achse sind die nach der Häufigkeit ihres Auftretens geordneten Varianten durchgezählt, auf der y-Achse ist ihre prozentuale Summe am Gesamtbestand der Segmente eingetragen. Eine vergleichbare Summenkurve für altdeutsche Personennamen findet sich bei KAMP (wie Anm.38) S.113.

- vereinfachend formuliert - ein Wörterbuch, das 2oo Varianten mit Angabe des jeweils dazugehörigen Lemmas enthält, in der Lage sein könnte, fast 46 % des gesamtem Segmentbestandes zu lemmatisieren, und daß bereits ein Wörterbuch mit 1oo Varianten einen Lemmatisierungserfolg von über 36 % hätte. Bedenkt man, daß diese Varianten auch in Flurnamen anderer Gebiete auftreten, ein Wörterbuch dieser geringen Größenordnung also schon in der Lage wäre, eine Vielzahl von Belegen zu interpretieren, dann läßt sich daraus folgern, daß zumindest ein teilautomatisiertes Lemmatisierungsverfahren angebracht ist.



Die 2oo häufigsten Varianten, nach ihrer Häufigkeit absteigend angeordnet

Ad 2. Eine AL wäre einfach durchzuführen, wenn eine eindeutige Beziehung zwischen Variante und Lemma vorhanden wäre, mit anderen Worten, wenn es nicht das Problem der Homographen gäbe[44]. Ein kleines, aus dem Archiv ausgewähltes Belegkorpus, in dem Namen mit den Lemmata BODEN (zu *bo^{a}dem* 'Boden, Grund, Ackergrund'), BAUM (zu *bō^{2}m*) und BOHNE (zu *bō^{2}ne*) zusammengestellt sind, verdeutlicht das Ausmaß von Homographien in der schriftlichen Flurnamenüberlieferung:

1. ¬IM >BODEN; 2. ¬IM >BUOM; 3. >BOHNEN/KAMP - >BA"ONEN/KAMP[44a]; 4.>KIRSCH-

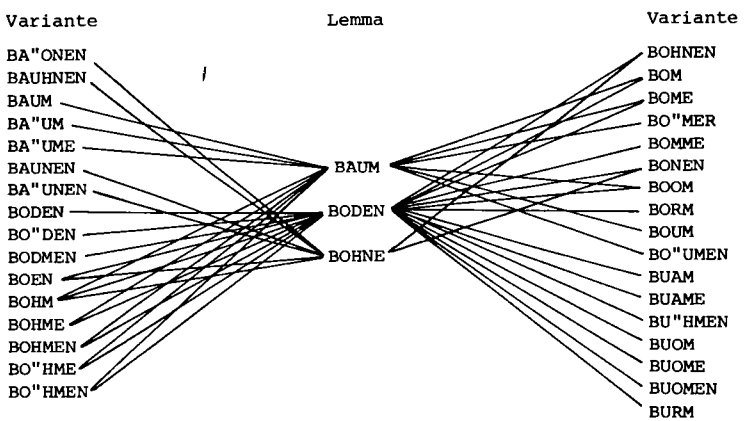
44 Hier gelten im folgenden zwei zeichengleiche Varianten, die verschiedenen Lemmata zugewiesen werden können, als homograph, z.B. OLE (entweder zu *ǫld* 'alt' oder *ō^{2}l* 'Flußwiese'): DE <OLE >KAMP, -BIM >WERD/OLE.

44a Verschiedene Belege desselben Namens sind unter einer Nummer zusammengefaßt.

/BA"UME; 5. >BOHNEN/GRUND - >BAUNEN/GRUND; 6. >KIRCH/BUOM; 7. ¬AM >BIRN/BAUM - ¬UNTER DEM >BEHR/BOHME; 8. ¬OPM >BIA"RN/BOUM; 9. >NOTTE/BOHM; 1o. >BONEN/WINKEL - >BA"ONEN/WINKEL; 11. ¬IM >BOME - ¬IM >BUOME; 12. ¬IM >BOHMEN - ¬IM >BOHME - ¬IM >BUOME; 13. ¬AN DIA"N >RIEGN/BO"UMEN - >REYEN/BO"HMEN; 14. >KIEPEN/BO"HMEN - >KIEPEN/BU"HMEN; 15. ¬IM >BODEN - ¬IME >BOMME; 16. >RENNE/BAUM - >RENNE/BOHM; 17. >SURKE/BOM; 18.>BA"UNEN/KAMP; 19. >BURM - >BO"HME; 2o. ¬IM >BODMEN; 21. ¬AM <BLINDEN >BAUM - <BLINNEN >BO"HME - ¬AM >BLINDEN/BO"MER/WEGE; 22. >BAUM/HOF; 23. ¬IM >BUAM; 24. >BAUHNEN/KAMP; 25. ¬AUF DEM >BOHM - ¬AUF DEM >BO"HME - ¬OPM >BUAM; 26. >BOME/DAHL - >BAUM/TAL; 27. ¬IM >BOM - ¬IM >BODEN; 28. ¬AM >BAUM/WEGE - ¬AM >BA"OM/WIA"GE; 29.>BOHNEN/KAMP - >BOEN/KAMP - >BAUNEN/KAMP; 3o. ¬OPM >BOMME - ¬AUF DEM >BOME; 31. >BOM/ACKER - >BORM/ACKER; 32. >BA"UM/BIEKE - >BOM/BACH; 33. >BOOM/GORDEN; 34. >BOHM/BRAOKE ->BOHNEN/BRACHE; 35. >BOM/GORDEN; 36. >BOOM/KAMP - >BUOMEN/KAMP; 37. >BODEN/WIESE; 38. >BOHM/BERG - >BAUM/BERG; 39. >BOHNEN/WIESE - >BOHM/WIESE; 4o. >BOUM/HUAF; 41. >BODEN/BRUCH - >BOEN/BRAUK; 42. ¬OBERN >BO"HMEN - ¬IN DEN>BO"DEN ; 43. >BOHM - >BONEN .

Obwohl das Wörterbuch, das sich aus den Varianten des obigen Korpus bilden läßt, zahlreiche Homographen aufweist, kann die Zuweisung zu einem der drei Lemmata dennoch in allen 43 Fällen eindeutig erfolgen (BODEN: 1f., 6, 11f., 14f., 19f., 23, 25, 27, 3of., 36f., 39, 41-43; BAUM: 4, 7-9, 13, 16f., 21f., 26, 28, 32f., 35, 38, 4o; BOHNE: 3, 5, 1o, 18, 24, 29, 34):

(SWB1)



Für die eindeutige Lemmatisierung des obigen Korpus ist neben der Eindeutigkeit einiger Varianten (BODMEN, BUOM, BO"UMEN) auch konstitutiv die Berücksichtigung

1. der semantischen Durchsichtigkeit der im Kompositum auftretenden Kookkurrenten (NOTTE/BOHM 'Nußbaum', BIRN/BAUM, SURKE/BOM 'Holzapfelbaum'),
2. sich wechselseitig erhellender Schreibungen (BOM - BORM, BO"HMEN - BU"HMEN, BOHNEN - BOEN),
3. der Wortstellungen (¬IM >BONEN käme z.B. als Variante für das Lemma BOHNE nicht in Frage, im Gegensatz etwa zu >BONEN/WINKEL),
4. der Lautungen/Graphien kookkurrierender Segmente (BOHM ist als Variante für das Lemma BOHNE nur unmittelbar vor Labialen wahrscheinlich zu machen, vgl. >BOHM/BRAOKE, nicht dagegen in Fällen wie >BOHM/ACKER).

Daraus ergibt sich, daß ein Wörterbuch, das keine anderen Informationen enthält als eine Liste von Varianten und eine auf diese bezogene Liste von Lemmanamen, keine ausreichende Grundlage für ein Lemmatisierungsverfahren bildet. Ein AL-Verfahren wird nur erfolgreich sein, wenn es wenigstens zum Teil gelingt, die in den Punkten 1-4 aufgeführten Entscheidungskriterien so zu formalisieren, daß sie für einen Lemmatisierungsalgorithmus nutzbar gemacht werden können. Welche Möglichkeiten sich dafür anbieten, soll u.a. in den Abschnitten 3.2. bis 3.5. und 3.7. geklärt werden.

3.2. *Das Subwörterbuch (SWB)*

Voraussetzung für die AL ist ein dem Rechner während des Lemmatisierungsvorganges zur Verfügung stehendes Wörterbuch, in dem Varianten ihrem Lemma verbunden sind. Für dieses alfabetisch geordnete Varianten-Lemma-Inventar, dessen einfachste Form als SWB1 auf S.153 dargestellt ist, ist die Bezeichnung S u b w ö r t e r b u c h geprägt worden. Der Terminus, der auch hier im folgenden Anwendung findet, gründet auf der Vorstellung, daß die Menge aller im Archiv zur Verfügung stehenden Namenbelege das eigentliche "Wörterbuch" bilde, während das SWB ein Teilwörterbuch darstelle, das in der Lage sein soll, die Sprachdaten des Gesamtwörterbuches zu interpre-

tieren[45].

3.2.1. Die Festlegung der Lemmanamen

Um im SWB und in Kategorie 11 (Lemmateil, → 3.6.) Varianten und Lemmanamen unterscheiden zu können, sind einige Vereinbarungen zu deren Form nötig. Diese ist prinzipiell arbiträr, doch haben motivierte Lemmanamen den Vorzug, Hinweise auf das Lemma selbst geben zu können. Am informativsten wären somit Lemmanamen, die den Ansätzen des Westfälischen Wörterbuches entsprächen, allerdings werden für diese diakritische Zusatzzeichen (*bō²ne*, *brāke*, *lē¹f* usw.) verwendet[46], die als Zeichenfolgen (z.B. BO+2NE , BRA@+KE , LE+1F) zu umschreiben wären. Da diakritische Sonderzeichen(folgen) das erforderliche alfabetische Sortieren der Lemmanamen zwar nicht verhindern, aber umständlicher machen, ist es ratsam, auf Sonderzeichen möglichst zu verzichten. So bot sich alternativ der Ansatz hochdeutscher Wortformen an, also BOHNE statt BO+2NE , BRACHE statt BRA@+KE , LIEB statt LE+1F usw. Diese Entscheidung entbehrt nicht innerer Berechtigung, da die schriftliche Überlieferung, vor allem das katastrale Quellengut, die niederdeutschen Flurnamen überwiegend in hochdeutscher Umformung oder gar Übersetzung fixiert und man deshalb durchaus von einer zweisprachigen Tradierung der Namen sprechen kann (→ 3.3.; 3.4.). Allerdings sollten konstruierte Ansätze nach Möglichkeit vermieden werden[47]. Der Lemmaname wurde nur dann nach den Prinzipien hochsprachlicher Orthographie/Lautung festgelegt, wenn eine solche, soweit überblickbar, auch tatsächlich in der schriftlichen Überlieferung Verwendung fand, sonst wurde eine andere schriftliche Tradierungsform gewählt, etwa LIEKEDE für *līkede* 'Ebene'. Eine Lemmanamenkonkordanz zwischen den An-

---

45 Zum Terminus Subwörterbuch s. KAMP (wie Anm.21) S.93; KAMP (wie Anm.38) S.46.

46 Zu den Prinzipien des Ansatzes vgl. *Beiband* (wie Anm.6) S.62f.

47 Vgl. die Überlegungen zu einem hochdeutschen Lemmaansatz bei KESELING - KETTNER u.a. (wie Anm.1) S.197; vgl. auch LENDERS (wie Anm.8) S.329.

sätzen des Flurnamenarchivs und denen des Westfälischen Wörterbuches muß die nötige Kompatibilität sichern (→ 4.8.). Jeder Lemmaname wird mit einer Ziffer eingeleitet und einem Punkt abgeschlossen (1LIEKEDE. , 1BOHNE. , 1BRACHE. usw.). Die einleitende Ziffer, die ebenso wie der abschließende Punkt zur formalen Lemmanamenbegrenzung und -identifizierung dient (→ 3.2.5.; 3.6.), hat darüber hinaus auch die Funktion, das Lemma einer bestimmten Lemmakategorie zuzuweisen (→ 2, S.149): Lemmabasis ist bei 1 ein Appellativ (z.B. 1GARTEN.), 2 ein Siedlungsname (z.B. 2SOEST.), 3 ein Ländername (z.B. 3PREUSZEN.), 4 ein Anthroponym (z.B. 4MARIA.) und 5 ein Hydronym (z.B. 5LIPPE.).

3.2.2. Die Notation der Flexive im SWB

Der Umfang des SWB sollte - nicht zuletzt zur Reduktion von Speicherplatz und Rechenzeit - so klein wie möglich gehalten werden. Ein SWB in der Form SWB1 (S.153) ist deshalb unökonomisch, weil in ihm Flexive und Fugenzeichen den Varianten zugeordnet sind, für die Belege

(7) <SCHWARTZE >BRACHT
(8) <SCHWARTZER >BOM
(9) <SCHWARTZES >HOLZ
(1o) >SCHWARTZ/BIEKE
(11) ¬AM <SCHWARTZEN >SIEPEN

also allein sechs Varianten des Lemmas 1SCHWARZ. zur erfolgreichen Lemmatisierung benötigt werden. In Verbindung mit einer dem eigentlichen AL-Programm vorgeschalteten Prozedur zur Flexivanalyse läßt sich die Variantenzahl im SWB allerdings drastisch drücken. Eine der Lemmatisierung vorausgehende Flexivanalyse empfiehlt sich auch deshalb, weil die Zahl der Flexive und ihrer Schreibungen in der Mikrotoponymie außerordentlich gering ist. In Frage kommen als Flexivendungen und Fugenzeichen eigentlich nur E, EN, N, ER, R, ES und S. Als Fugentrenner kommen dieselben Zeichen vor (>WOL/BECKER/FELD , >DORNS/HEIDE , >BOHN/KAMP neben >BOHNE/KAMP und >BOHNEN/KAMP), so daß die Notwendigkeit einer gesonderten Analyse von Fugen und Endungen entfällt. Der Wechsel zwischen EN/N, ER/R, ES/S (DEHLEN - DEHLN, LO"HER - LO"HR, HOLTES -

HOLTS) setzt dabei voraus, die *e*-haltigen Formen als Varianten der *e*-losen Flexive N, R und S aufzufassen. EM/M ist als Flexiv so selten, daß es wie einige andere spärlich vertretene Endungstypen vernachlässigt und dem jeweiligen Wortstamm zugeordnet bleiben kann (z.B. DAIPM 'tiefen' statt DAIP-M).

Die Flexivanalyse interpretiert Segmente, die auf E, (E)N, (E)R oder (E)S enden, als mit einem Flexiv/Fugenzeichen versehen auf und trennt dieses ab. Damit wird z.B. erreicht, daß die Belege (7) bis (11) mit einer einzigen Variante (SCHWARTZ) dem Lemma (1SCHWARZ.) zugeordnet werden können.

Probleme bei der Flexiv-/Fugenanalyse ergeben sich dadurch, daß zwischen Endungen und zum Wortstamm gehörigem -E, -(E)N, -(E)R, -(E)S nicht unterschieden werden kann. Eine formale Differenzierung etwa von

(12) >ALDEN/HEESTER/FELD (*hē²ster* 'junge Eiche oder Buche'),
(13) >HU"LS/WENDE (*hŭls(e)* 'ilex'),
(14) >KLOIN/KAMP (*kleine* 'klein') und

(15) >ALDEN/BECKER/FELD (*bieke* 'Bach'),
(16) >HOLTS/WENDE (*holt* 'Holz'),
(17) >BREIN/KAMP (*brē²d* 'breit')

ist nicht ohne weiteres möglich. Das Problem wird noch kompliziert durch Mehrdeutigkeiten wie

(18) >HAMMER/WEG (*hāmer* 'Hammer' oder *Hamm* 'Stadt in Westf.'),
(19) >FLACHS/RAUTE (*flas* 'Flachs') und
(2o) <FLACHES >MEER (*flak* 'flach').

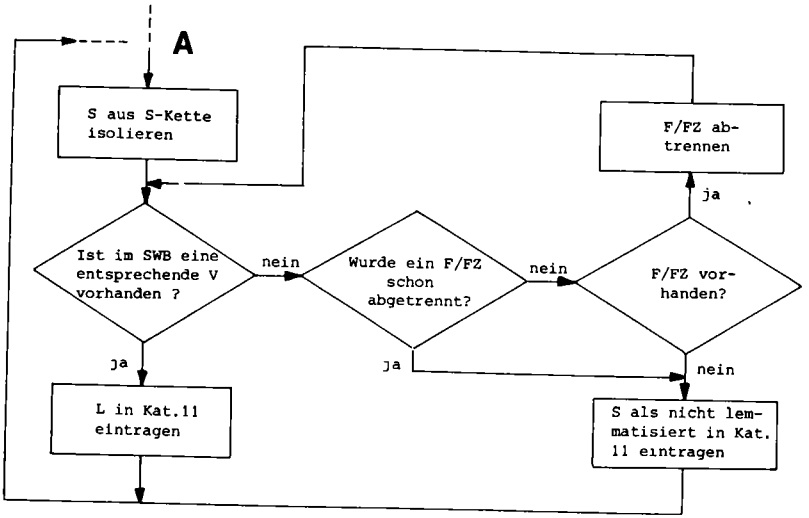
Es gibt zwei Lösungswege, die durch die beiden Flußdiagramme A und B veranschaulicht werden können[48](S.158):

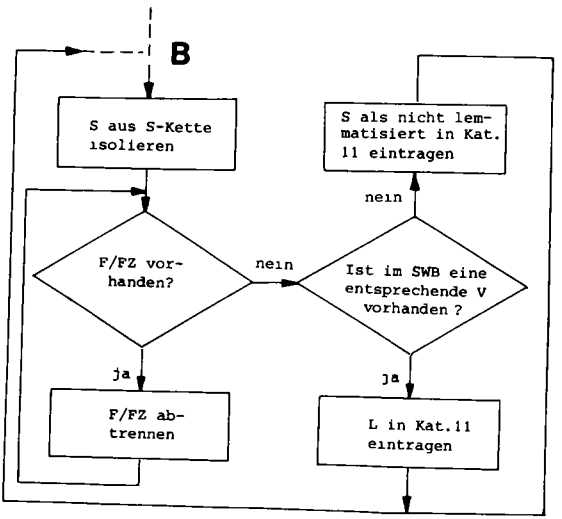
Beim Verfahren A wird die Flexivanalyse an einem Segment erst durchgeführt, wenn die Suche nach einer passenden Variante im SWB erfolglos geblieben ist. Bei einem gegebenen SWB

(SWB2)

ALD ——— 1ALT.
BECKE ——— 1BACH.
BREI ——— 1BREIT.
FELD ——— 1FELD.
HEESTER ——— 1HEISTER.

48 In den beiden Diagrammen A und B sind folgende Abkürzungen verwendet: F=Flexiv, FZ=Fugenzeichen, L=Lemmaname, S=Segment, V=Variante.


A
S aus S-Kette isolieren
Ist im SWB eine entsprechende V vorhanden ?
nein
Wurde ein F/FZ schon abgetrennt?
nein
F/FZ vorhanden?
ja
F/FZ abtrennen
ja
L in Kat.11 eintragen
ja
nein
S als nicht lemmatisiert in Kat. 11 eintragen


B
S aus S-Kette isolieren
F/FZ vorhanden?
nein
Ist im SWB eine entsprechende V vorhanden ?
nein
S als nicht lemmatisiert in Kat. 11 eintragen
ja
F/FZ abtrennen
ja
L in Kat.11 eintragen

HOLT ———— 1HOLZ.
HU"LS ———— 1HU"LS.
KAMP ———— 1KAMP.
KLOIN ———— 1KLEIN.

würde die SWB-Suche nach den für die Segmente HEESTER , KLOIN und HU"LS (vgl. (12) bis (14)) zutreffenden Lemmata sofort erfolgreich sein und eine Flexivanalyse könnte entfallen. Bei BREIN, BECKER, ALDEN und HOLTS ((12), (15) bis (17)) wäre der erste Versuch, die Varianten im SWB aufzufinden, erfolglos, erst nach Durchführung einer Endungsabtrennung gelänge über BREI, BECKE, ALD und HOLT eine Lemmatisierung.

Ein solches Verfahren ist linguistisch einwandfrei[49], aber nicht unbedingt rationell bei einem Material, in dem flexivhaltige Segmente überwiegen. Bei Flurnamen ist dies der Fall. Fast alle adjektivischen Segmente sind mit einem Flexiv verbunden - ¬IM <SCHWATTEN >RUOTT, DE <SCHWATTE >WIA"G, >SCHWARTEN/FELD, <SCHWARZES >FELD usw. ist wesentlich häufiger als z.B.>SCHWATT/MECKE oder >SCHWARZ/BACH -, d.h., eine zunächst erfolglose SWB-Suche wäre die Regel. Ein ähnliches Problem ist bei den vielen femininen, auf *-e* auslautenden Substantiven gegeben. Da *bieke*, *brāke* oder *heide* in "Grundwortstellung" das *-e* vorwiegend bewahren (>BREN/BIEKE,>TWERS/BRAOKE, >HOLT/HEIDE), in der Fuge dagegen öfter aufgeben (>BIEK/FELD, >HEID/KAMP, >BRAOK/WIESE), müßte entweder das SWB vergrößert - 1BACH=:{BIEK, BIEKE}, 1BRACHE=:{BRAOK, BRAOKE} usw. - oder (bei ausschließlichem Ansatz der kürzeren Varianten BIEK, BRAOK, HEID) die Zahl der zunächst erfolglosen SWB-Suchgänge drastisch erhöht werden.

Es erscheint deshalb zweckmäßiger, Lösung B anzustreben und die Flexivanalyse vor die SWB-Suche zu legen. Jedes auslautende E, (E)N, (E)R und (E)S wird dann als Endung gewertet, gleichgültig, ob es sich innerhalb des deutschen morphologischen Systems um ein Flexions-/Fugenzeichen handelt oder nicht.

---

49 Dieses Verfahren beschrieben bei KAMP (wie Anm.38) S.53ff.; KAMP (wie Anm.21) S.94ff.

Ein solches Verfahren impliziert allerdings die Aufnahme von Angaben über erforderliche bzw. zulässige Endungen in das SWB. Sie können etwa wie folgt formuliert werden:

(SWB3)
AFTEK-ØEN ——— 1APOTHEKE.
AFTEK-R ——— 1APOTHEKER.
FLACH-S ——— 1FLACHS.
FLACH ——— 1FLACH.

AFTEK-ØEN beschränkt die möglichen Segmentausformungen für 1APOTHEKE. auf AFTEK (Ø als Kennzeichnung fehlender Endung)[50], AFTEKE und AFTEK(E)N, während AFTEK-R für 1APOTHEKER. zwingend AFTEK(E)R und FLACH-S für 1FLACHS. die Segmentform FLACH(E)S voraussetzt. Ist keine restriktive Angabe vorhanden, gelten alle Endungen als zugelassen: Die Variante FLACH deckt somit die Segmentformen FLACH, FLACHE, FLACHER, FLACHES, FLACHEN, FLACHR, FLACHN, FLACHS[51] ab.

Version A geht von einer einmaligen Anwendung der Regeln zur Endungsabtrennung aus. Fälle wie <AFTEKERS >HOLT/WISCH oder ¬IM >FLACHSE zwingen jedoch, im Falle B die Regeln zur Endungsabtrennung rekursiv zu formulieren. Im übrigen wird durch das Vorhandensein echter Flexivkombinationen (Plural- + Kasusflexiv wie in ¬OP DEN >HU"SERN) sowie durch die vor allem in älterer schriftlicher Überlieferung geläufige Fugenzeichenkumulation (>RODENE/FELD,>BONENN/KAMP, >HAHNENS/KAMP) auch im Falle A eine - wenn auch durch Bedingungen einzuschränkende - Wiederholbarkeit von Abtrennungsregeln von Vorteil sein.

Die der rekursiven Anwendung der Abtrennungsregeln zugrunde liegende Annahme von Endungsketten hat zur Konsequenz, daß auch im SWB Angaben über notwendige oder zulässige Endungsfolgen gemacht werden können. In einer verbesserten Version kann daher das SWB umformuliert werden zu

---

50 Die Notation -ØEN ist aufzufassen als "Ø oder E oder N oder EN folgt".

51 FLACHR, FLACHN sind aufgrund der oben formulierten Konvention (R= R oder ER, N=N oder EN) möglich, obwohl sie de facto kaum vorkommen werden.

(SWB4)

```
AFTEK-ØEN-Ø ———1APOTHEKE.
AFTEK-R-ØSN-Ø —1APOTHEKER.
FLACH-S-ØES-Ø —1FLACHS.
FLACH-ØESN-Ø ——1FLACH.
      \R-ØERSN-Ø
```

Die Praxis zeigt, daß die in SWB4 vorgeführte explizite Formulierung der Endungsfolgen nur in wenigen Fällen zur ausreichenden Trennung ähnlicher Lemmata notwendig ist. Fehllemmatisierungen aufgrund unzureichender Endungsspezifizierung im SWB halten sich in Grenzen (→ 3.2.6.2. und 3.5.3.).

Eine uneingeschränkte Rekursivität der Abtrennungsregeln, wie in B vorgesehen, hätte zur Folge, daß z.B. das erste Segment in >RENNE/WEG als reine Endungskette (Ø-R-N-N-E) aufgefaßt würde. Deshalb wurde für endungshaltige Segmente eine Mindestzeichenlänge vereinbart, wobei sich die Annahme einer Segmentlänge von mindestens vier Zeichen als praktikabel erwies. Dies führt zum Abbruch der Endungsabtrennung nach dem fünften Zeichen der zu analysierenden Kette und damit beim obigen Beispiel zu einer Abtrennung RENN-E. Die Vereinbarung einer Mindestzeichenlänge hat zwar den Nachteil, daß gelegentlich echte Flexive in die Varianten aufgenommen werden müssen - etwa 1WEG.= {WEG, WEGE ...}, 1HAUS.= {SEN, SER ...}[52] - doch hält er sich bei der geringen Zahl solcher Kurzvarianten in Grenzen.

Der Vorteil einer mit der SWB-Suche verbundenen Endungsabtrennung (nach dem Verfahren B und mit der eben beschriebenen Zusatzbedingung) läßt sich wieder an dem schon 3.1. herangezogenen Soester Flurnamenmaterial demonstrieren. Betrug für die 28.o25 Segmente die Zahl der für ein vollständiges SWB benötigten Varianten ohne Endungsabtrennung 6.365, so sinkt dieser Wert mit Endungsabtrennung auf 4.948. Die Lemmatisierungsleistung eines Kern-SWB mit den 2oo häufigsten Varianten steigt gleichzeitig von 46 % auf fast 53 % an. Die genauen Werte lassen sich an der gestrichelten Kurve auf S.152 ablesen.

---

52 Vgl. >HOLT/SEN, >HOLT/SER/BRUCH (←*Holthusen, Holthuser Bruch*).

3.2.3. Diakritische Zeichen im SWB

Der Belegteil enthält diakritische Zeichen, die z.T. direkt aus der Vorlage übernommen sind (>SAUST-/WIEG für *Saust-wieg*), teils Besonderheiten der Vorlage umschreiben: Kürzungsauflösungen (<GR(OSZE) >WIESE für *gr. Wiese*), Hinweise auf nicht sicher lesbare Zeichen (SIEG?EN), bestimmte Schreibweisen (>WU*ORT für *wůrt*) usw. Mit Ausnahme der Diakritika für Umlaut (") und Vokallänge (+), die für die Lemmazuweisung von großer Bedeutung sein können, werden sämtliche Sonderzeichen vor der SWB-Suche in den Segmenten getilgt. Dies bedeutet, daß im SWB an Sonderzeichen nur '"' und '+' bei den Varianten vorzukommen brauchen, nicht aber z.B. andere phonetische Diakritika (für Vokalöffnung, Stimmhaftigkeit, Betonung usw.), die sich nicht als lemmatisierungsentscheidend erwiesen haben.

Segmente	reduzierte Segmente	Varianten im SWB
STRO"+,TKEN/:K(AM)PE	STRO"+TKEN/KAMPE	KAMP STRO"+TK

3.2.4. Die Notation paradigmatischer Merkmale im SWB

Unter paradigmatischen Merkmalen sind hier sowohl Eigenschaften von Varianten bezüglich ihrer Positionierung innerhalb proprialer Syntagmen (als Präposition, Attribut usw.) als auch Wortbildungseigenschaften von Varianten (Fähigkeit zur Bildung eines Grundwortes, ausschließliches Vorkommen als Erst- oder Mittelglied von Zusammensetzungen usw.) verstanden.

Das bereits Anm.44 gegebene Beispiel (21) DE <OLE >KAMP (1ALT.) - (23) ¬BIM >WERD/OLE (1OHL. 'Flußwiese') zeigt, daß Wortstellungs- und Wortbildungsangaben im SWB zur Homographendifferenzierung beitragen können. Am wichtigsten erscheint dabei eine Unterscheidung von attributiver Verwendung einer Variante und der Verwendung als Basis von Namenkernen. Das Merkmal "attributiv" (vereinbarte Kodierung im SWB '-') beschränkt dabei eine Variante in ihrem Vorkommen auf Attribute und Vorderglieder von Namenkernen[53], das Merk-

53 Diese beiden Eigenschaften sind in der Regel gleichzeitig gegeben,

mal "Namenkernbasis" (vereinbarte Kodierung '+') beschränkt auf Grundwortstellung und auf Simplicia in Namenkernen[54]. Ist keine Verwendungseinschränkung intendiert, entfällt eine Merkmalsangabe. Ein gegebenes SWB

(SWB5)
KAMP-ØES ——————— 1KAMP.
OLE - ——————— 1ALT.
OLE + ——————— 1OHL.
WERD-ØES ——————— 1WERT. (*wērd* 'Flußufer')

erlaubt so eine eindeutige Lemmatisierung der Belege (21), (22) >OLE/KAMP, (23), (24) >KAMP/WERDE und (25) ¬IM >OLE zu (21') und (22') 1ALT. und 1KAMP. , (23') 1WERT. und 1OHL. , (24') 1KAMP. und 1WERT. sowie (25') 1OHL.[55]

Die Belege (26) >MERGE/LOH
(27) >BOHM/MERGE
(28) >HAM/BIEKE
(28a) >HAGEN/BIEKE
(29) >BIEK/HAM

verdeutlichen die Notwendigkeit weiterer Differenzierung. MERG-ØE ist als Variante von 1BERG. geläufig, tritt aber nur als Folgeglied in Zusammensetzungen auf, da der Wechsel M < B einen vorausliegenden assimilierenden Faktor (hier MM < MB) erfordert. Umgekehrt setzt HAM (in (28)) als Variante von 1HAGEN. unbedingt Folgelaute in einem Kompositum voraus, die den Wechsel N > M verursacht haben. Eine entsprechende Merkmalskodierung ('1'=Variante kommt nur als Vorderglied vor; '2'=Variante kommt nur als Hinterglied vor; '3'=Variante kommt nur als Simplex vor) erlaubt eine weitere Homographendifferenzierung. Ein gegebenes SWB

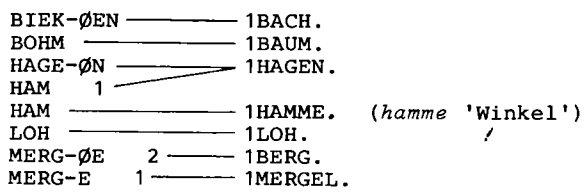
---

da jedes Attribut durch eine Zusammenrückung ein vorderes Glied im Kompositum eines Namenkernes werden kann, vgl. DE <OLE >KAMP und >OLE/KAMP, ¬BIM <LANGEN >OLE und >LANGEN/OLE.

54 Auch diese Eigenschaften sind in der Regel gekoppelt, da Simplicia durch Anrücken von Attributen Teil eines Kompositums werden können, vgl. die Beispiele Anm.53.

55 Eine Differenzierung von OLE - → 1ALT. und OLE + → 1OHL. ist deshalb möglich, weil 1ALT. nur attributiv vorkommt, 1OHL. zwar generell nicht eingeschränkt erscheint, wohl aber seine Variante OLE auf Grund- und Simplexstellung. Bei attributivem Gebrauch fehlt das E bzw. tritt S ein: OL(S)/KAMP.

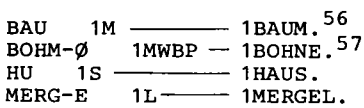
(SWB6)



lemmatisiert (26) bis (29) zu (26') 1MERGEL. und 1LOH., (27') 1BAUM. und 1BERG. , (28') 1HAGEN. oder 1HAMME. (zur Mehrfachlemmatisierung → 3.3.) und 1BACH. , (28a') 1HAGEN. und 1BACH. sowie (29') 1BACH. und 1HAMME.

Schreibungen wie (28) >MERGE/LOH (← *mergel-loh*) beruhen auf Vereinfachung von Doppelkonsonanz in der Kompositionsfuge. Ähnliche Fälle sind (3o) >HU/STEDE (← *hus-stede*) oder (31) >BAU/MERG (← *baum-merg*). Da in den Ablochkonventionen vereinbart war, Zeichen an der Fugengrenze, an denen beide der komponierten Segmente Anteil haben, jeweils dem zweiten Segment zuzuordnen, treten solche defekte Varianten jeweils in Vordergliedposition auf. Um zu verhindern, daß Schreibungen wie (32) >HU/LAND oder (33) >BAU/SCHULTE/WEG fälschlich zu 1HAUS. oder 1BAUM. lemmatisiert werden, können Bedingungen über die nach der Kompositionsfuge folgenden Zeichen in das SWB aufgenommen werden. Die Möglichkeit, Angaben über die "Hintergliedanfangszeichen" in das SWB zu schreiben, kann auch dazu benützt werden, um für Vorderglieder, die durch die Lautumgebung verursachte Abweichungen aufweisen, einschränkende Vorkommensbedingungen zu formulieren:

(SWB7)



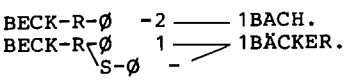
Die vorgeführten Merkmale können, soweit sie sich nicht widersprechen ('+' und '1'; '+' und Angaben zu Hinterglied-

56 Die Notation ist zu lesen: BAU ist eine Variante von 1BAUM., wenn ein mit M anlautendes Segment im Kompositum folgt.

57 Die Notation ist zu lesen: BOHM-Ø ist eine Variante von 1BOHNE., wenn ein mit M, W, B oder P anlautendes Segment im Kompositum folgt.

anfangszeichen) oder redundant sind ('-' und '1'), kombiniert werden:

(SWB8)



lemmatisiert (15) >ALDEN/BECKER/FELD , (15a) <ALDEN/BECKER >FELD , (34) <BECKERS >HOLT/WISCH , (35) >BECKER/STRAOTE zu (15') 1BACH. oder 1BÄCKER. (zur Mehrfachlemmatisierung → 3.3), (15a') 1BACH., (34') bis (35') 1BÄCKER.

In Präzisierung der in Anmerkung 41 gegebenen Definition soll V a r i a n t e im folgenden verstanden werden als eine einem Lemma zugeordnete Zeichenkette im SWB einschließlich der mit dieser Zeichenkette verbundenen Angaben über notwendige oder zulässige Endungen, über paradigmatische Eigenschaften und über die erforderliche "lautliche" Umgebung.

3.2.5. Komposita im SWB

3.2.5.1. Im Belegteil segmentierte Komposita

Das hier vorgestellte AL-Verfahren geht davon aus, daß Komposita möglichst nach ihren Einzelsegmenten lemmatisiert werden. Da einzelne Segmente häufig, was ihre Lemmazugehörigkeit betrifft, mehrdeutig sind, ist dabei die Möglichkeit der Mehrfachlemmatisierung zugelassen. Sie ist z.B. bei (36) >BEHR/KAMP , das ein SWB etwa zu (36') 1BA"R1. (*bāre* 'Bär'; möglich unter Annahme einer hochdeutschen Umsetzung) oder 1BA"R2. (*bār* 'Eber') oder 1BEERE. (*bi^{e}re* 'Beere') oder 1BIRNE. (*bi^{e}re* 'Birne') und 1KAMP. lemmatisieren könnte, durchaus erwünscht, da auch ein Bearbeiter bei einem isolierten Beleg dieser Art (und ohne weitergehende Sachinformation) kaum eine exaktere Entscheidung treffen könnte. Sie ist aber störend, wo Kookkurrenten eine eindeutige Bestimmung zulassen, so bei (37) >BEHR/BOHM , das deutlich *bi^{e}re-bō^{2}m* 'Birnbaum' entspricht.

Da Segmente hinsichtlich ihrer Lemmazugehörigkeit durch ein im Kompositum kookkurrierendes Segment näher bestimmt werden können, mußte die Möglichkeit vorgesehen werden,

nicht nur einzelne Varianten, sondern auch Variantenkomposita in das SWB aufzunehmen:

(SWB9)

```
BEHR-ØEN ———— 1BA"R1.
BEHR-ØEN ———— 1BA"R2.
BEHR-ØEN ———— 1BEERE.
BEHR-ØEN ———— 1BIRNE.
BEHR/BOHM ———— 1BIRNE.    &1BAUM.
KAMP ———————— 1KAMP.
KERS-ØENS ———— 1KIRSCHE.
KERS-ØENS ———— 1KRESSE.
KERS/BOM ———— 1KIRSCHE.   &1BAUM.
```

[58]

Unter der Voraussetzung, daß ein im Belegteil auftretendes Kompositum zunächst im SWB aufgesucht wird und erst nach erfolgloser SWB-Suche die Zerlegung in die Einzelsegmente stattfindet, kann SWB9 die Belege (36), (37), (38) >KERS/BOM und (39) >KERS/KAMP zu (36') 1BA"R1. oder 1BA"R2. oder 1BEERE. oder 1BIRNE. und 1KAMP. , (37') 1BIRNE. und 1BAUM. , (38') 1KIRSCHE. und 1BAUM. sowie (39') zu 1KIRSCHE. oder 1KRESSE. und 1KAMP. lemmatisieren.

Es gibt in der Toponymie immer wieder auftretende Komposita (die in der Regel aus Appellativkomposita abgeleitet sind), deren Einzelsegmente für sich genommen unklar oder mehrdeutig wären:>HEL/WEG , >HIEL/WECH , >WOL/MEINE , >WOL/MAI, >LANN/WER usw. Es ist sinnvoll, sie als Komposita in das SWB aufzunehmen, während bei den in ihren Einzelsegmenten eindeutigen Formen derselben Kompositionstypen (>WALD/MEINDE , >LANT/WEHR) dies überflüssig ist.

3.2.5.2. Im Belegteil nicht segmentierte Komposita

Es gibt Fälle, bei denen die Segmentierung auf der Belegebene Schwierigkeiten bereitet. Während ein Beleg *Viehwech* leicht segmentiert - >VIEH/WECH - und über seine Einzelbestandteile zu 1VIEH. und 1WEG. lemmatisiert werden kann, wird mundartliches *Faiwe*, sofern es isoliert steht, bei der Materialerfassung kaum als Kompositum erkannt und behandelt werden (>FAI/WE). Sollten umständliche Nachkorrekturen im Belegteil vermieden werden, so mußte die Möglichkeit vorge-

58 Zur Darstellung der Kompositionsfuge auf der Lemmaebene ('&') → 3.6.

sehen werden, einfache Segmente der Belegebene als Komposita auf der Lemmaebene darzustellen. Denn nur bei einer Darstellung von >FAIWE als Zusammensetzung von 1VIEH. und 1WEG. ist die nötige Kompatibilität mit >VIEH/WECH herzustellen. Hinzu kommt, daß ungleiche Segmentierungsergebnisse nicht nur durch mangelnde Übersicht bei der Datenerfassung verursacht sind. Dort, wo die Segmente eines Kompositums durch Assimilation und Kontraktion eng verschmolzen sind, empfiehlt sich von einer Segmentierung abzusehen. *Bremke* 'Breitenbach' ist im Gegensatz etwa zu *Bredenbeke* weder als >BRE/MKE noch als >BREM/KE befriedigend segmentierbar. Und eine Segmentierung >LAF/FER ('Landwehr') ist - im Gegensatz zu >LANT/WEHR , >LAND/WIER oder >LANN/WER - noch problematischer. Um (4o) >LAFFER und (41) >LANT/WEER auf der Lemmaebene kompatibel zu machen, kann im SWB einer Variante ein Lemmakompositum verbunden werden:

(SWB1o) LAFF-R-Ø ——— 1LAND. :1WEHR.[59]
LANT-ØS-Ø ——— 1LAND.
WEER-ØE ——— 1WEHR.

lemmatisiert (4o), (41) zu (4o'), (41') 1LAND. und 1WEHR.

3.2.6. Zur Optimierung des SWB

3.2.6.1. Die Entstehung der ersten Fassung des SWB

Ein Teil des bisherigen Zettelarchives war noch von J. Hartig "handlemmatisiert" und nach den Stichwortansätzen des Westfälischen Wörterbuchs sortiert worden. Da in dieses lemmatisierte Korpus Namendateien aus verschiedenen westfälischen Gebieten eingegangen sind, ergab seine Durchsicht einen guten Überblick über die häufigeren Lemmata und die ihnen zugeordneten Varianten. Ihre Zusammenstellung lieferte die Basis für eine erste Version des SWB nach den 3.2.1. bis 3.2.5. dargelegten Prinzipien. Von ihm durfte man annehmen, daß es nicht nur eine maschinelle Wiederholung der bereits von Hand geleisteten Lemmatisierungsarbeit, sondern auch die Inter-

59 Zur Darstellung der Kompositionsfuge auf der Lemmaebene (':') → 3.6.

pretation neuen, noch nicht lemmatisierten Namenmaterials bis zu einem gewissen Umfang leisten konnte. Zusätzlich wurden von verschiedenen bereits abgelochten, aber noch nicht lemmatisierten Dateien alfabetisch geordnete Segmentindices hergestellt, die einen erweiterten Überblick über vorkommende Varianten und Lemmata, über Flexivgebrauch, Fugenmarkierungen und Wortstellungsregularitäten ermöglichten. Die dabei gewonnenen Erkenntnisse sind mit in die erste Fassung des SWB eingegangen.

3.2.6.2. SWB-Korrektur und SWB-Ergänzung

Die weitere stufenweise Optimierung des SWB ging von den Daten aus, die bei der AL neuer Namenmaterialien anfielen. Zunächst ließen mißglückte Lemmatisierungen Fehler im SWB erkennen. Diese Fehler sind im wesentlichen zwei Gruppen zuzuordnen:

1. Unzureichende oder zu weit gefaßte Angaben über zulässige Endungen und zulässige Wortstellung,
2. zu restriktiv formulierte Angaben über Endungen und Wortstellung.

So beruhen etwa die Fehllemmatisierungen <VATTERS >LANDE → 1FASZ. und ¬AM <DICKEN >HUCHT → 1TEICH. auf den zu weit gefaßten SWB-Einträgen

```
VATT ———————— 1FASZ.
DICK ———————— 1TEICH.
```

Eine präzisere Notation

```
VATT-ØES-Ø —— 1FASZ.
DICK-ØE ————— 1TEICH.
     N-Ø +
     S-Ø -
```

hätte diese und ähnliche Fehlsteuerungen vermeiden können. Umgekehrt war der SWB-Eintrag

```
GRO"-N - ——— 1GRU"N.
```

von der Auffassung ausgegangen, daß das Farbadjektiv nur attributiv oder als vorderes Segment in Namenkernen auftrete, was, wie eine zunächst unlemmatisiert gebliebene Substantivierung ¬IM >GRO"NEN erwies, unrichtig ist. Eine Umformulierung des SWB-Eintrages in

GRO"N - N-ØEN + ⟶ 1GRU"N.

erweiterte den Anwendungsbereich der Variante.

Mängel dieser Art müssen bei Herstellung einer verbesserten SWB-Version durch Ergänzung bzw. Korrektur beseitigt werden.

Auch wenn man annehmen darf, daß sich die Zahl solcher Fehlerquellen im SWB mit der Zeit verringern wird, werden sich die Mängel doch nie vollständig beseitigen lassen. Die Notwendigkeit einer nicht-automatischen Lemmatisierungskorrektur (→ 3.5.3.) wird daher immer bestehen bleiben.

Unvollständig wird das SWB auch immer hinsichtlich seiner lexikalischen Einträge bleiben. Jeder AL-Lauf erbringt eine mehr oder weniger große Zahl von lemmatisierbaren Varianten und von Lemmata, die im SWB noch nicht vorhanden waren. Ihre kontinuierliche Hinzufügung wird zwar das SWB in seinem Bestand immer mehr dem tatsächlich vorhandenen toponymischen Gesamtwortschatz annähern, aber es wird diesen dadurch nie erreichen können.

3.2.6.3. Temporäre SWB

Im übrigen ist es auch nicht erstrebenswert, jede bei einem AL-Lauf auftretende Variante, die einem Lemma zugeordnet werden kann (→ 3.5.1.) und die noch nicht im SWB ist, sofort in dieses aufzunehmen. Hält man sich vor Augen, daß 5o % (2449) aller im Flurnamenmaterial des Kreises Soest auftretenden Varianten (4948) dort nur je einmal verwendet sind (→ 3.2.2.), dann kann man abschätzen, daß auch innerhalb des toponymischen Gesamtwortschatzes der Prozentsatz der bloß ein- oder zweimal vertretenen Varianten/Lemmata nicht gering sein wird. Der Versuch, möglichst alle lemmatisierbaren Varianten in das SWB aufzunehmen, würde dieses unerhört anschwellen lassen und damit zu Verarbeitungsproblemen führen.

Es ist daher sinnvoll, neben einem dauernd für die AL zur Verfügung stehenden Kern-SWB, das durch Fehlerkorrektur und Varianten-/Lemma-Ergänzung laufend optimiert wird, auch tem-

poräre Subwörterbücher zu bilden, die nur zur AL bestimmter Dateien und damit nur zur zeitweisen Ergänzung des Kern-SWB herangezogen werden sollen. So können die beim ersten Lemmatisierungslauf (→ 3.5.1.) einer Datei X neu anfallenden Varianten/Lemmata, von denen angenommen werden darf, daß sie sehr selten oder gar singulär sind, in einem temporären SWB zusammengestellt werden, das nur der Ergänzungslemmatisierung (→ 3.5.2.) von X dient. Gewässer- und Siedlungsnamenlemmata (→ 2., S.149; 3.2.1.) werden grundsätzlich in temporäre SWB aufgenommen. Denn es wäre wenig effektiv, etwa die Varianten des Flußnamenlemmas 5WESER. in einem SWB permanent bereitzustellen, daß über längere Zeit nur dazu verwendet wird, Dateien aus dem westlichen Westfalen zu lemmatisieren.

(Fortsetzung folgt in Band 19)