Linguistische Arbeiten 2

Herausgegeben von Herbert E. Brekle, Hans Jürgen Heringer, Christian Rohrer, Heinz Vater und Otmar Werner

Rainer Dietrich

Automatische Textwörterbücher

Studien zur maschinellen Lemmatisierung verbaler Wortformen des Deutschen

Max Niemeyer Verlag
Tübingen 1973

ISBN 3-484-10167-9

INHALTSVERZEICHNIS

Wenige Monate, nachdem die vorliegende Untersuchung, eine vorwiegend bibliographisch erweiterte Fassung meiner Dissertation begonnen wurde, beauftragte die Deutsche Forschungsgemeinschaft die Arbeitsgruppe 'Elektronische Sprachforschung' unter Leitung von Hans Eggers, ein Verfahren zur maschinellen Lemmatisierung von Texten der deutschen Gegenwartssprache zu entwickeln. Die Aufgabenstellung sowie arbeitstechnische Erwägungen legten es nahe, arbeitsteilig vorzugehen. Die Untersuchungsergebnisse der Teilgruppen wurden in regelmäßigen Sitzungen koordiniert.

Die zahlreichen Erörterungen, die die im folgenden behandelten Probleme oft unmittelbar betrafen, trugen in vielen Punkten zur Bereicherung meiner Überlegungen bei. Ich danke daher allen Mitarbeitern und Gästen der Arbeitsgruppe für die Anregungen, die sich für mich aus vielen Gesprächen ergaben. Insbesondere gilt mein Dank Hans Eggers, der die Arbeit betreut hat, sowie Arnim von Stechow für neunzehn kritische Anmerkungen und dem Verleger Robert Harsch-Niemeyer für seine Großzügigkeit und Geduld.

Helena Peltonen hat das druckfertige Manuskript und Manfred Thiel die Zeichnungen dazu angefertigt; ohne ihre umsichtige und selbständige Arbeit...

R. D.

A GESCHICHTE, SYSTEMATISCHER ORT UND ZIELSETZUNG DER MASCHINELLEN LEMMATISIERUNG

1 Notwendigkeit einer maschinellen Lemmatisierung

1.1 Der Terminus 'Lemma'

'Lemma' ist ein Terminus, der in der Linguistischen Datenverarbeitung (LDV) selten und uneinheitlich verwendet wird. Stickel und Gräfe, Mitarbeiter am Goethe-Wörterbuch, bezeichnen damit signifikante Wort- und Grundformen in Text und Wörterbuch[1], während die Herausgeber der 'Indices zur deutschen Literatur' die im Lexikon verzeichnete Grundform darunter verstehen[2]. Ähnlich definiert Busa Lemma als für das Paradigma repräsentatives Stichwort des Lexikons: "cequi, dans les lexiques, réprésente toutes les formes réunies dans un même paradigme"[3]. Dies ist für Verben zum Beispiel die erste Person, Singular, Präsens, Indikativ, Aktiv. Nach Maas schließlich ist Lemma bestimmt als "ein Paar (M_x, I_x), wobei M_x eine Menge von Wortformen und I_x die ihnen gemeinsame grammatische Information ist. Die Elemente von M_x sind nicht zufällig ausgewählt, sondern es sind die verschiedenen Formen eines Wortes (des Lemmas x), die dieses je nach seiner Funktion, die es im Satz erfüllt, annehmen muß"[4]. Fast synonym zu dem unten S. 22 in (8) entwickelten Lemmabegriff definieren Klein und Rath in 'Automatische Lemmatisierung', S. 2 - 3. Dort umfaßt Lemma alle, hier nur

---

1 Stickel, G. und Gräfe, M., 'Automatische Textzerlegung...'
Die genauen bibliographischen Angaben zu der zitierten Literatur finden sich in der Bibliographie, S.226

2 Schwerte, H. und Schanze, H., 'Indices...',Bd.1, S. VII

3 Busa, R. und Zampolli, A., 'Centre pour l'automation de l'analyse linguistique'

4 Maas, H.D., 'Homographie und maschinelle Sprachübersetzung', S. 3

alle im Text belegten Formen mit den definierten Merkmalen.[1]

Diese terminologische Vielfalt, von der hier nur ein Ausschnitt wiedergegeben wurde, ist jedoch älter als die linguistische Datenverarbeitung. Als der Name Lemma - in Anlehnung an den lexikologischen Gebrauch[2] - in die Sprachwissenschaft Eingang fand, schuf einerseits die Assoziation mit dem philologischen Begriff[3] Verwirrung, andererseits wurde es notwendig, ihm in der Diskussion um die Wortdefinition einen begrifflichen Ort zuzuweisen, was bislang noch zu keiner allgemein akzeptierten Lösung geführt hat. Die Orientierung an der lexikologischen Bedeutung verhilft nur zu der Klarheit, daß Lemma ein Stichwort einer lexikalischen Einheit, eines Wörterbuchartikels, bezeichnet. Sowohl die Stichwörter selbst als auch die Artikel unterscheiden sich erheblich, je nach Intention und Gattung der einzelnen Wörterbücher. Sowohl finite als auch infinite Verbformen können als Stichwort gewählt werden[4], Vollformen ebenso wie 'Stämme'. Die unter einem Stichwort zusammengetragenen Informationen können phonetische Transskriptionen (Aussprachewörterbuch) oder entwicklungsgeschichtliche Erklärungen sein (Etymologisches Wörterbuch), Belegstellenhinweise (Indizes) oder bedeutungsgleiche Wörter (Synonymwörterbuch), zur gleichen Sachgruppe gehörende Begriffe ebenso wie Adressen oder Telefonnummern. Wie jede Gattung spezifischen Ansprüchen genügen muß, so bestimmt sich auch die Struktur von Maschinenwörterbüchern aus den Zielsetzungen und den Gegebenheiten der linguistischen Datenverarbeitung.

---

1 Teildarstellungen enthalten: Rath, R.: 'Vorschläge zur Automatischen Lemmatisierung (AL) deutscher Adjektive' und Rath, R.: 'Probleme der automatischen Lemmatisierung'

2 Daß Lemma in der Mathematik und der Philosophie ebenfalls ein - allerdings genau definierter - Begriff ist, sei nur nebenbei erwähnt.

3 Texteinheit als Stichwort in Verbindung mit kritischen Anmerkungen des Herausgebers.

4 Vgl. z.B. das 'Mittelhochdeutsche Wörterbuch' von Benecke, Müller und Zarncke gegenüber dem 'Mittelhochdeutschen Handwörterbuch' von Lexer. Gegenüber diesen beiden wiederum unterscheidet sich Pokornys 'Indogermanisches etymologisches Wörterbuch', indem es als Stichwörter vorwiegend Wurzeln enthält.

In der Geschichte der maschinellen Untersuchung von Texten führten zwei im Ansatz völlig verschiedene Richtungen zur Entwicklung von Lemmatisierungsverfahren: die Erstellung von Indizes zu literarischen Texten und die maschinelle Sprachübersetzung.

1.2 Lemmatisierte Textwörterbücher

In immer größer werdender Anzahl werden zu Teiltexten, zu geschlossenen Werken einzelner Schriftsteller oder zur Literatur von Zeiträumen sortierte Wörterverzeichnisse maschinell erstellt. Sie unterscheiden sich zwar voneinander in der Anlage sowie im Anteil von Maschinenarbeit verglichen mit dem der menschlichen Bearbeiter[1], doch verbindet sie erstens der gemeinsame Ausgang von literarischen Texten, zweitens das Bearbeitungsprinzip, dessen Hauptteile Segmentieren, Zählen und Ordnen[2] der einzelnen Graphemfolgen zu Listen bilden und schließlich das Ziel, Texteinheiten zu Sprachinventaruntersuchungen aufzubereiten. Segmentierungs-, Zähl- und Ordnungsanweisungen stellen das Regelsystem, den Algorithmus, der maschinellen Bearbeitung dar; alle zusätzlichen Ergebnisse, wie Angaben über syntaktische oder semantische Merkmale (Homographie, Homonymie usw.) müssen in einer Vor- oder Nachbereitung hinzugefügt werden - für jeden Text. Dies ist neben den zahlreichen Vorteilen ein Nachteil[3]. Ihm kann man teilweise begegnen, indem man zu jedem Token eine bestimmbar große Kontextmenge[4] mit in die Liste übernimmt, gewissermaßen als mittelbare

---

1 Dieser Anteil ist beispielsweise beim Goethe-Wörterbuch wesentlich größer als bei den Aachener 'Indices'.

2 Als Grenzsymbole der Segmente gelten Leer- und Satzzeichen. Der Bezugsrahmen der Segmentzählung ist unterschiedlich, bei Prosatexten meist Seite und Zeile, bei Gedichtsammlungen Seitenzahl, Gedichtnummer und Zeile, bei Dramen Seite, Akt, Szene und Vers oder Zeile. Als Ordnungsprinzip gilt das Alphabet.

3 Von vielen, die dieser Unzulänglichkeit Ausdruck verleihen, sei nur W. Müller erwähnt: 'Gedanken zur automatischen Analyse von Normen...', S. 312 - 314.

4 In Wisbeys 'Vollständiger Konkordanz zur "Wiener Genesis"..' wird beispielsweise zu jedem Textwort der jeweilige Vers ausgedruckt.

Informationen[1].

Was allerdings die Termini 'Index' und 'Konkordanz' betrifft, so ist hier der Sprachgebrauch nicht weniger schwankend als beim Lemma, wie etwa ein Vergleich zwischen den 'Indices zur deutschen Literatur' und Spevacks 'Complete Concordance to the Works of W. Shakespeare'[2] zeigt. In der weiteren Darstellung wird für Verzeichnisse ohne Kontextausdruck der Name Index[3], für die andere Klasse der Name Konkordanz[4] benutzt.

Werke beider Gattungen sind nützliche Hilfsmittel für Stilanalysen und -vergleiche einerseits und die interpretatorische Erforschung von Einzelproblemen andererseits. Für beide Verwendungszwecke aber stellen sie noch bei weitem nicht die Ideallösung dar, wie manche Kritik und Selbstkritik erkennen läßt: im Vorwort zu Band I bedauern die Herausgeber der 'Indices...', "vorerst auf die Vollständigkeit der linguistischen Analyse zu verzichten"[5]. In einer Rezension der Racine-Konkordanz von Freeman und Batson[6] beklagt Brody die fehlende Lemmatisierung, die Quemada[7] in seiner Ausgabe zum gleichen Text manuell durchgeführt hatte. Nahezu beschwörend klingt eine Formulierung aus Lloyds Besprechung von Wisbeys[8] Wiener-Genesis-Konkordanz: "... one can still look forward to the millennium, when parsed concordances will be less forbidding undertakings."[9]

---

1 Daß Konkordanzen darüberhinaus noch Aufschlüsse eigener Art ermöglichen und beabsichtigen, steht außer Frage, ist hier aber nicht von Bedeutung.

2 Dies sind nur Beispiele für eine ständig wachsende Zahl ähnlicher Projekte.

3 Mit Index ist also nicht das Ergebnis eines Indexing-Programms verstanden, dessen verschiedene Variationen H. Borko in seiner Arbeit 'Indexing and Classification' aufführt; S. 99 - 125

4 Über die einzelnen Phasen eines Konkordanzprojektes informiert die interessante Arbeit von Hines, Harris und Lewy, 'An Experimental Concordance Program'.

5 Schwerte, H. und Schanze, H., 'Indices...',Bd.1, S. VII

6 Freeman, B.C. und Batson, A., 'Concordance du théatre...'

7 'J.Racine, Phèdre, Concordances, Index et Relevés statistiques'

8 siehe Anm. 4, S. 3

9 in: Computers and the Humanisties, Vol.3, 1969, S. 182

Konkordanzen und Indizes weisen - wie zu erkennen ist - drei grundsätzliche Mängel auf:

- die Trennung von inhaltlich zusammengehörigen Texteinheiten
- die Zusammenfassung von inhaltlich ganz Verschiedenem
- die Gleichbehandlung von primär funktionsorientierten (grammatischen) und primär semantisch relevanten Texteinheiten.

Wo immer von Datenverarbeitungsanlagen gelieferte Ergebnisse unbefriedigend bleiben, können die Fehler - abgesehen von technischen Defekten der Maschine - in zwei Bereichen liegen: den eingegebenen Daten oder der Bearbeitungsvorschrift. Da aber einem Text nicht vorgeworfen werden kann, daß er so ist, wie er ist, müssen die Grundsätze der Index- und Konkordanzprogramme geprüft werden.

Wie schon gesagt, sind ihre grundlegenden Teilprozesse Segmentieren, Zählen und Ordnen. Indizes und Konkordanzen lassen sich mithin nur weiterentwickeln, wenn

- die Segmentierung nicht ausschließlich auf der strengen und oberflächlichen Definition von Wort als Graphemfolge zwischen zwei Leerzeichen basiert,
- somit die im Grunde genommen sprachinadäquate Ordnung nach dem Alphabet nicht als oberstes oder gar einziges Kriterium für die Reihenfolge bei der Auflistung von Wortformen gilt.

Diese noch sehr allgemeinen Einsichten sind es, die im Bereich der Literaturwissenschaft die automatische Lemmatisierung notwendig machen. Indizes und Konkordanzen liefern die fehlenden Informationen mittelbar über einen Belegstellenhinweis, der es ermöglicht, in einer Nachbereitung, die allerdings vom Benutzer geleistet werden muß, die Lemmatisierung der gesuchten Wörter durchzuführen[1]. Er ist dazu imstande aufgrund seiner Kenntnis des Sprachsystems. Aus diesem Zusammenhang heraus erhellt der interessante systematische Ort der dritten Teiloperation, des Zählens und Errechnens der Belegstellen. Eine automatische Lemmatisierung wird

---

1 Die Nachbereitung kann ein äußerst zeitraubendes Unterfangen werden, wenn das gesuchte Lemma beispielsweise einen trennbaren Verbzusatz aufweist und das entsprechende Simplex häufig belegt ist.

jeweils in dem Maße gelingen, wie es möglich ist, die dazu notwendigen Kenntnisse des Sprachsystems zu mechanisieren.

1.3 Lemmatisierung als Teilprozess der maschinellen Sprachübersetzung

Betrachtet man im Bewußtsein der Ergebnisse von Kap. 1.2 die Entwicklung der scheinbar völlig anders gelagerten maschinellen Sprachübersetzung, zeigen sich aufschlußreiche Parallelen. Es sei darauf jedoch nur soweit eingegangen, wie es für die Lösung der in dieser Arbeit thematisierten Problematik nützlich ist.

Das Ziel der MT (machine translation) ist klar: Kommunikationsinhalte innerhalb von mindestens zwei sprachlichen Systemen, L_1 und L_2, zu übertragen:

(1) entweder: $L_1 \longrightarrow L_2$
oder : $L_1 \rightleftharpoons L_2$

Die Hypothesen, daß der formalen Einheit Satz (Graphemfolgen zwischen zwei Begrenzungszeichen[1]) in Bezug auf die zu übermittelnden Inhalte eine gewisse Geschlossenheit und Selbständigkeit zukomme und daß die ebenso formal definierten Wortformen und ihre Reihenfolge die konstituierenden Elemente von Sätzen seien, führten zu den bekannten Wort-für-Wort-Übersetzungen[2] und zu den ebenso bekannten Kritiken an der Möglichkeit maschineller Übersetzung überhaupt. Wie die Erfolgsaussichten der MT-Projekte zu beurteilen sind, kann hier nicht diskutiert werden. Wir können nur feststellen, daß sie eng verbunden sind mit der adäquaten Analyse der Ausgangssprache. Dazu müssen in einem ersten Schritt die Wortformen der Sätze der Ausgangssprache mit grammatischen Informationen aus einem geeigneten Verzeichnis ausgestattet werden. Geschieht diese 'Textvorbereitung', die selbst schon Teil der automatischen Bearbeitung ist, mit Hilfe eines Stammwörterbuchs, so umfaßt sie im wesentlichen die gleichen Operationen, wie sie für eine Lemmatisierung entwickelt werden müssen. Aus dieser Sicht bezeichnet

---

1 Punkt, Ausrufezeichen und Fragezeichen z.B.

2 siehe dazu die ausführlichen Kapitel 1.1 und 2 der Dissertation von W. Klein, 'Parsing...'

zum Beispiel Busa die Lemmatisierung als Prozess der Informationszuordnung zu Textsegmenten[1].

2 Allgemeine Zielsetzung

Nun ist aber unsere Aufgabe nicht, eine vollständige Analysegrammatik zu erstellen und in ein Programm umzuarbeiten, sondern, wie schon gesagt, eine mehr lexikologische, die aber nicht ohne gewisse Elemente einer Satzanalyse zu bewerkstelligen ist.

Die Ausführungen in den Kapiteln 1.2 und 1.3 ermöglichen es, die Anforderungen, die an das Verfahren zu stellen sind, zu explizieren:

(1)[2] Das Ergebnis eines Lemmatisierungsprogramms zu einem gegebenen Text soll sein: eine geordnete Liste von Lexikoneinheiten (Lemmata), die die zur Einheit gehörenden und im Text belegten Elemente (Wortformen) enthalten und durch einen Namen (Lemmaname) identifiziert werden können. Als Namen wählen wir die unflektierten Grundformen des jeweiligen Paradigmas (Adjektive und Substantive) und bei den Verben jeweils den Infinitiv. Zu den Lemmata sollen die Wortklasse[3] und die Gesamthäufigkeit angegeben sein und zu den Elementen des Lemmas weitere Informationen wie grammatische, Belegstellenhinweise, Häufigkeit und andere.

Um als Output ein solches Verzeichnis zu erhalten, müssen folgende Schritte der Textbearbeitung möglich sein:

(2) 1) Eine Folge von Graphemen zwischen Blanks muß als Element oder Elementteil identifiziert werden.[4]

2) Elementteile müssen zu einem Element zusammengeführt werden.

---

1 vgl. Anm. 3, S. 1

2 Sätze, Definitionen und Beispiele werden durch in () gesetzte Ziffern gekennzeichnet und kapitelweise durchnumeriert.

3 Bei idiomatischen Wendungen kann eine entsprechende Information als Wortklassenangabe gelten.

4 Das Element *wird geliebt haben* besteht aus drei Elementteilen; der Lemmaname ist *lieben*.

3) Elemente müssen zu einem Lemma zusammengeführt werden.

4) Elemente müssen durch a) grammatische Informationen
b) Belegstellenhinweise
charakterisiert werden.

Wir betrachten nach (1) Lemma als eine Menge von Graphemfolgen des Alphabets A mit bestimmten gemeinsamen Eigenschaften, E_1, E_2, ... E_n, formal dargestellt:

(3) $$y = \{x | E_{n_A}(x)\}$$

Zum Lemma y gehört jede Graphemfolge x, welche die über der Grundmenge A definierten Eigenschaften E_n aufweist. E_1, E_2, ... E_n müssen exakt definiert sein. Sie könnten beispielsweise lauten:

(4) E_1: x ist eine sprachrichtige Wortform.
E_2: das dritte Graphem von x ist ein B. Eigenschaften dieser Art sind nicht so sinnlos, wie sie auf Anhieb scheinen[1]. Für eine Lemmatisierung könnte E_2 modifiziert werden in
E_3: die letzten Grapheme von x sind *keit*.
E_4: x ist im Text T belegt.

Offenbar unterscheidet sich E_4 von den übrigen in (4) genannten Eigenschaften. Sie beinhaltet eine primäre und quantitative Aussage. Ihr Bezugsrahmen sind die vom jeweiligen Autor gesetzten Textgrenzen. E_4 ist neben der automatischen Zähloperation die einzige Grundlage eines in keiner Weise vor- oder nachbereiteten Werkindex'. E_1 bis E_3 richten sich nach sprachlichen Kriterien. Sie gewinnen für die Lemmatisierung erst an Bedeutung, wenn E_4 nachgewiesen ist. Die weitere Aufgabe besteht nun aus zwei Teilen: erstens, einen Katalog von Eigenschaften der Art E_1 bis E_3 aufzustellen, der den für jedes LDV-Projekt geltenden Postulaten nach linguistischer Adäquatheit und Mechanisierbarkeit genügen muß, zweitens, einen Algorithmus zu entwickeln, mit dessen Hilfe für jede Graphemfolge entschieden werden kann, welche Eigenschaften

---

1 vgl. dazu den Trakl-Index von Klein und Zimmermann, dem nicht ein Graphem- sondern ein Phonemkode zugrunde liegt.

sie aufweist.

Der erste Teil der Aufgabenstellung muß signifikante Eigenschaften aller möglicherweise in Texten repräsentierten Types[1] berücksichtigen. Er wird im folgenden Spezifikationsteil (ST) genannt. Da dem Lemmatisierungs- beziehungsweise Analysealgorithmus die Aufgabe der Identifizierung der E für die jeweilige Graphemfolge zufällt, soll er Identifikationsteil (IT) heißen[2].

3 Mögliche Verfahrensweisen

Jedes Lemmatisierungsprojekt muß drei Faktoren berücksichtigen und sinnvoll aufeinander abstimmen: Textvorbereitung (TV) und die Analyseteile ST und IT. Unter ihnen besteht folgende Relation:

(1) Die Strukturen von ST und IT charakterisieren den automatisch ablaufenden Teil, das Programm.

(2) Das Verhältnis TV zu (ST, IT) charakterisiert das gesamte Lösungsverfahren.

Jenachdem, ob in (2) das Schwergewicht auf TV oder (ST,IT) liegt, heißt ein Verfahren inputintensiv oder programmintensiv. Jenachdem, ob in programmintensiven Verfahren die Eigenschaften einer Graphemfolge vorwiegend durch direkte Einzelinformationen erkannt werden oder durch einen Analysalgorithmus auf der Basis linguistischer Regeln, heißt das Programm informations- oder regelintensiv. Die Vor- und Nachteile der drei Lösungswege sollen im folgenden kurz dargestellt und diskutiert werden.

3.1 Inputintensive Verfahren

Inputintensive Verfahren sind, wie wir gesehen haben, so angelegt, daß im Zuge der manuellen Textvorbereitung Arbeitsgänge mitvollzogen werden, die Teile der sprachlichen Problematik der Lemmatisierung lösen oder deren Lösung durch das Programm bis zu einem gewissen Grad unterstützen.

---

1 Unter Type wird zunächst ganz allgemein die jeweilige Klasse nicht durch Leerzeichen unterbrochener Graphemfolgen verstanden.

2 Die Termini finden sich in einem analog auf Satzanalyse übertragenen Sinne bei W. Klein,'Eine Analysegrammatik',S. 15-16.

Der meines Wissens noch nicht in der Praxis durchgeführte Extremfall wäre, zu jeder Graphemfolge des Textes einen Verweis einzugeben, wie sie zu lemmatisieren ist. Das kann durch direkte Angabe des Lemmanamens oder über eine kodierte Information als unmittelbare Bearbeitungsrichtlinie geschehen. Der Beispielsatz *Verse entstehen, die mehr tönen als sagen wollen*[1] hätte dann bei der Ablochung die Form (1) oder (2).

(1) *Verse* nom Pl (*Vers*) SUB[2]
entstehen 3.Pl PRÄS IND AKT (*entstehen*) VRB; FOLGT KOMMA
die REL nom Pl MASK
mehr ADV
tönen INF (*tönen*) VRB
usw.

(2) *Verse* nom Pl SUB / LEMMANAME: ELIMINIERE GRAPHEM 5
entstehen 3.Pl PRÄS IND AKT VRB / LEMMANAME: GLEICH TOKEN
usw.

Die in (1) und (2) angedeuteten Vorbereitungen nehmen die gesamte Lemmatisierung voraus, wie man sieht. Input (1) oder (2) kann mit einem entsprechend erweiterten Indexprogramm[3] bearbeitet werden. Die Vorteile eines solchen Verfahrens liegen in der verhältnismäßig einfachen Programmierung und der äußerst niedrigen Fehlerquote, die theoretisch gleich null sein kann, sofern in der menschlichen Bearbeitung kein Fehler unterläuft. Die eingegebenen Informationen können beliebig erweitert werden, zum Beispiel auf dem semantischen Bereich, sodaß auch polyseme Formen wie *Schloss*, *Druck* und *Fuchs* getrennt lemmatisiert werden.

---

1 Viele Beispiele sind - wie dieses - aus den Korpora gewählt, die dem unter Leitung von H. Eggers entwickelten Saarbrükker Analyseprogramm als Material dienten. Eine ausführliche Beschreibung findet sich in Eggers, H., 'Zur Syntax der deutschen Sprache der Gegenwart'.

2 Alle Zeichen und Abkürzungen sind im Abkürzungsverzeichnis, S. 232 - 233 erklärt.

3 Die Berechnung der relativen Häufigkeit der Belege sowie die Summierung der Gesamthäufigkeit sind auch - allerdings von der linguistischen Fragestellung unabhängige - Erweiterungen von Indexprogrammen. Man vergleiche die Konkordanz von Spevack.

Jedoch sprechen zwei schwerwiegende Nachteile dagegen, diese Methode weiter in Betracht zu ziehen. Der Aufwand an intellektueller Arbeit, der für jeden Text gleich bleibt, ist unökonomisch groß. Zweitens stellt die Lösung für die Erforschung maschineller Sprachbearbeitung keinen Fortschritt dar.

Solange allerdings keine automatischen Lösungsverfahren entwickelt sind, der Bedarf an lemmatisierten Textwörterbüchern aber immer nachdrücklicher laut wird, bleibt der Weg über die Textvor- bzw. Nachbereitung durch menschliche Bearbeiter der einzige überhaupt, wenn man nicht ganz auf die Unterstützung durch Maschinen verzichten will[1].

So kann man die in ihrer 'Generation' nicht mehr zu verbessernden Indizes als Ausgangsbasis für Lemmatisierungsprojekte ansehen, indem der noch sehr hohe Anteil manueller Arbeit Zug um Zug abgebaut wird. Einen Schritt in dieser Richtung gingen Klein und Zimmermann, die den schon erwähnten Trakl-Index[2] durch ein rationelles 'man-machine-interaction'-System lemmatisierten[3]. In einer zwischengeschobenen Phase werden zum Beispiel homographe Wortformen von Hand bearbeitet.

3.2 Programmintensive Verfahren

Programmintensive Verfahren erfordern ein Minimum an Textvorbereitung. Als fast utopisch anmutendes Fernziel kann man sich ein Programm vorstellen, das über Klarschriftleser den Text einliest und ein lemmatisiertes Wörterbuch ausgibt. Weder die technischen noch die linguistischen Voraussetzungen dazu sind zur Zeit gegeben. Die Entwicklung entsprechender technischer Geräte hat mit den Normschriftlesern allerdings schon einen Stand erreicht, der die Vollendung deutlich ahnen läßt.

---

1 Ein Exempel für in der Nachbereitung lemmatisierte Textwörterbücher ist der schon erwähnte Racine-Index Quemadas (vgl. Anmerkung 7, S. 4).

2 vgl. Anm. 1, S. 8.

3 Klein, W. und Zimmermann, H., 'Index zu Georg Trakl. Dichtungen'
ähnlich übrigens: Wetzel, H., 'Konkordanz zu den Dichtungen Georg Trakls . Salzburg 1971.

Zur Lösung der sprachlichen Probleme werden im folgenden zwei mögliche Wege diskutiert, die allerdings wohl nur in Verbindung miteinander befriedigende Ergebnisse liefern können. Sie werden getrennt nacheinander behandelt, damit der systematische Unterschied klarer hervortritt.

3.2.1 Informationsintensive Verfahren

Der oben (S. 10) beschriebene, vorbereitete Text kommt zustande, indem jeder Wortform vom Bearbeiter aufgrund seiner Sprachkenntnis alle nötigen Informationen zugeordnet werden. Der Versuch, diesen Prozess zu simulieren, indem die Angaben nicht von einem menschlichen Bearbeiter, sondern von einem umfangreichen Wörterbuch bezogen werden, führt zu einem informationsintensiven Lemmatisierungsverfahren. Das Wörterbuch gehört zum Spezifikationsteil innerhalb des Analysesystems und ist nicht mit dem Output, dem Lemmawörterbuch (LWB) zu verwechseln. Es wird im weiteren ST-Wörterbuch (ST-WOBU) genannt. Jede seiner Einheiten umfaßt drei Komponenten, eine Graphemfolge (g), den Namen des Lemmas, zudem g gehört (L) und grammatische Informationen (h). Das ST-Wörterbuch hat also die allgemeine Form:

(1) (g_1, L_1, h_1)
(g_2, L_2, h_2)
... (g_n, L_m, h_o)

Das Prinzip läßt die Voraussetzungen, allerdings auch die Schwierigkeiten erkennen. In der Spalte g $(g_1, g_2, \ldots, g_n)$ müssen möglichst alle Wortformen der zu behandelnden Texte aufgenommen sein, was trotz des beachtlichen Umfangs theoretisch möglich wäre, da der Wortformenbestand einer Sprache Grenzen hat. Komplikationen treten erst durch die Tatsache auf, daß zwischen g, L und h nicht das Verhältnis 1 : 1 : 1 besteht, daß es, anders ausgedrückt, homographe und polyseme Wortformen gibt:

(2) $(g_1, L_{1,2,\ldots,n}, h_{1,2,\ldots,o})$

Die im Text auftretende Wortform g_1 ist nur über eine Kontextanalyse zu lemmatisieren. Innerhalb von ST müssen demnach Regeln formuliert werden, die die Bedingungen angeben, nach denen die Zuordnung von g, L und h eindeutig wird.

Da neben der Morphologie die Syntax das Gebiet ist, über das sich relativ gesicherte Aussagen machen lassen, beschränkte sich die maschinelle Kontextanalyse bislang auf die syntaktische Einheit Satz. In einigen Fällen kann zur Lösung von Homographen eine Wahrscheinlichkeitsschätzung oder Berechnung der relativen Häufigkeit herangezogen werden. Schätzen läßt sich die Wahrscheinlichkeit nach dem Alter des Textes, seinem Inhalt, der Sprachschicht, der er angehört, und anderen Faktoren.

Für eine mathematisch genaue Bestimmung dagegen kann man von den Zahlenverhältnissen ausgehen, die sich für den gleichen Homographen aus einem dem zu bearbeitenden Text vergleichbaren Sample ergeben. Die zuverlässigsten Orientierungsgrößen gewinnt man natürlich aus der Berechnung der relativen Häufigkeit aus dem vorliegenden Text. Allerdings setzt diese Ermittlung schon die Angaben voraus, für deren Gewinnung sie angesetzt wird und ist daher sinnlos.

Beide Methoden, die Schätzung nach philologischen Kriterien sowie die Berechnung, kann man nicht als automatische Homographenprogramme in dem genannten Sinn bezeichnen, denn die Mehrdeutigkeiten werden vor Beginn des Programmdurchlaufs nach den geschätzten oder errechnete Maßgaben reduziert oder aufgehoben. Ein Beispiel dafür berichtet Busa: Tritt in den Texten von Thomas von Aquin die Graphemfolge *Ibis* auf, so wird die Möglichkeit, daß es sich um den Namen des exotischen Vogels handeln könnte, als unwahrscheinlich ausgeschaltet. *Ibis* wird nur als 2. FUT Sg IND AKT zu *Ire* lemmatisiert[1]. Die Zahl der Fälle, die so eindeutig entschieden werden können, ist freilich sehr klein. Hinzu kommt, daß sich aus der Wahrscheinlichkeitsschätzung keine allgemeinen Regeln ableiten lassen, da sie auf textabhängigen Variablen basiert. Für die lateinische Abhandlung eines Ornithologen würde sich die Wahrscheinlichkeit für *Ibis* eventuell zugunsten des Vogelnamens verschieben.

Allein aus diesen knapp skizzierten Überlegungen ist ersichtlich, daß automatische Lemmatisierung durch bloßen Wörterbuchvergleich nicht möglich ist.

---

1 Sinngemäß zitiert nach Busa, R., 'Un lexique latin électronique', S. 258.

3.2.2 Regelintensive Verfahren

Das Prinzip regelintensiver Verfahren ist dem in Kap. 3.2.1 geschilderten sozusagen entgegengesetzt. Es besteht darin, möglichst alle Graphemfolgen, ausgehend von Form und Satzkontext, einem Lemma zuzuordnen, ohne auf direkte Informationen eines Wörterbuchs angewiesen zu sein[1]. Daß dieser Weg ebenso wenig erfolgreich sein kann, wie der zuvor beschriebene, ist offenbar. Wenn wir ihm dennoch einige Überlegungen widmen, dann deshalb, weil er verschiedene interessante Einsichten vermittelt, die für die Entwicklung einer sinnvollen und erfolgversprechenden Kombination zwischen ST-Wörterbuch und 'regelmäßiger' Textbearbeitung von Bedeutung sind. Wäre ein Sprachsystem gegeben, das sich ohne Wörterbuch vollständig und eindeutig durch Regeln analysieren ließe[2], so bestünde ein Lemmatisierungsprogramm im wesentlichen aus den folgenden Teilprozessen:

- Segmentierung in Graphemfolgen zwischen Leerzeichen,
- Klassifizierung der einzelnen Wortformen nach morphologischen Kriterien (Endungs- und Präfixanalysen),
- Ausstattung mit allen von der Struktur der Graphemfolge her möglichen Informationen,
- Verifizierung bzw. Falsifizierung der im zweiten Schritt als mehrdeutig erkannten Wortformen anhand syntaktischer Gesetze, wie Stellungsregeln, Satzzeichen, zulässige Gruppenbildung und andere.

(1) Beispiel: Für die Wortform *Verse* des in Kap. 3.1 gewählten Beispielsatzes (1), S. 10, ergäbe sich aus der morphologischen Untersuchung das redundante Ergebnis, daß sie allen Wortklassen angehören könnte, denn das Graphem E kann in jeder Wortklasse als letztes auftreten; es sei denn, die Einteilung ist so getroffen, daß Infi-

---

1 Ein derartiges Verfahren erwähnt Krallmann in seiner Dissertation, geht aber nicht näher darauf ein; 'Statistische Methoden...', S. 95 - 96.

2 Das Russische weist zum Beispiel homographe Wortformen in weit geringerem Maße auf als etwa das Deutsche.

nitive, Partizipien und Kardinalzahlwörter je eigene Klassen bilden[1]. Wie aber innerhalb der Wortklassen die rein morphologische Analyse bestimmte Formen des Paradigmas ermittelt, zeigt die folgende Aufstellung.

Funktionswortklassen[2] sind nicht berücksichtigt, was hinsichtlich des Programms bedeutet, daß schon ein ST-WOBU vorausgesetzt ist.

(2)

I*	*Verse*	IMP Sg	zu *versen*	(VRB)[3]
II*	*Verse*	1. Sg PRÄS IND AKT	zu *versen*	(VRB)
III*	*Verse*	1. Sg PRÄS KONJ AKT	zu *versen*	(VRB)
IV*	*Verse*	3. Sg PRÄS KONJ AKT	zu *versen*	(VRB)
V*	*Verse*	nom Sg	zu *Verse*	(SUB FEM)
VI*	*Verse*	gen Sg	zu *Verse*	(SUB FEM)
VII*	*Verse*	dat Sg	zu *Verse*	(SUB FEM)
VIII*	*Verse*	akk Sg	zu *Verse*	(SUB FEM)
IX*	*Verse*	nom Sg	zu *Verse*	(SUB MASK)
X	*Verse*	dat Sg	zu *Vers*	(SUB MASK)
XI	*Verse*	nom Pl	zu *Vers*	(SUB MASK)
XII	*Verse*	gen Pl	zu *Vers*	(SUB MASK)
XIII	*Verse*	akk Pl	zu *Vers*	(SUB MASK)
XIV*	*Verse*	dat Sg	zu *Vers*	(SUB NEUT)
XV*	*Verse*	nom Pl	zu *Vers*	(SUB NEUT)
XVI*	*Verse*	gen Pl	zu *Vers*	(SUB NEUT)
XVII*	*Verse*	akk Pl	zu *Vers*	(SUB NEUT)
XVIII*	*Verse*	nom Sg	zu *verse*	(ADJ)
XIX*	*Verse*	akk Sg	zu *verse*	(ADJ)
XX*	*Verse*	nom Pl	zu *verse*	(ADJ)
XXI*	*Verse*	akk Pl	zu *verse*	(ADJ)
XXII*	*Verse*		zu *verse*	(ADV)

Gehen wir einen Schritt weiter und nehmen an, daß die im ST-WOBU verzeichneten Formen (*die, mehr, als*) mit Informationen versehen wären, so ergäbe sich als Input für die Falsifikationsphase:

---

1 wie es etwa das Saarbrücker Projekt vorsieht; vgl.: Eggers, H., 'Elektronische Syntaxanalyse...'. S. 57 - 59

2 Siehe zu diesem Begriff Anm.1 , S.23.

3 Siehe Abkürzungsverzeichnis

(3)	*Verse*	*entstehen,*	*die*	*mehr*	*tönen*
	SUB	SUB	DEM	ADJ	SUB
	VRB	VRB	REL	ADV	VRB
	ADJ	ADJ			ADJ
	ADV	ADV			ADV
		INF			
		PART			

als	*sagen*	*wollen*[1].
KON	SUB	SUB
	VRB	VRB
	ADJ	ADJ
	ADV	ADV
	INF	INF
	PART	PART

Jede Homographie potenziert natürlich die Möglichkeiten der Gruppenbildung und erschwert die Prozedur der automatischen Zusammenfassung. Über einen Satz, dessen Wortformen wie in (3) mehrdeutig sind, läßt sich durch syntaktische Regeln nur die eine Klarheit gewinnen:

(4) (*Verse entstehen*), (*die mehr tönen als sagen wollen*)[2].

Aber nicht eine Wortklassenangabe ist damit zu falsifizieren; alle in (3) mehrdeutigen Wortformen bleiben es in der gleichen Weise.

Die in (4) dargestellte Gruppierung läßt in Hinsicht auf die Satzart die folgenden Möglichkeiten zu:

(5) 1 (Hauptsatz), (rel. Nebensatz)
2 (Hauptsatz), (Hauptsatz)
3 (Nebensatz)[3], (Hauptsatz)

Mehr ist über den Satz nicht zu erfahren, und die einzige Möglichkeit, ohne menschliche Zwischenkorrektur zu lemmatisieren, besteht darin, das ST-WOBU erheblich zu erweitern. Je mehr Graphemfolgen eines Satzes durch die Suche und den Vergleich im ST-WOBU

---

1 Die in (2) verzeichneten detaillierten Angaben sind mitzudenken; sie werden nicht noch einmal aufgeführt, um die Darstellung überschaubar zu halten.

2 Die Schreibweise soll andeuten, daß die syntaktischen Bindungen der Elemente innerhalb der Klammer enger sind als diejenigen über die Klammern hinaus.

3 hier ein uneingeleiteter Nebensatz, etwa: *fiele Regen, ...*