DK 001.8 : 801.55 : 025.4 : 681.3.053

Morphologische Relationen durch Reduktionsalgorithmen

Von *Rainer Kuhlen*, Frankfurt/Main

Summary

Es werden drei Algorithmen vorgestellt, mit denen englische Textwörter wörterbuchunabhängig und vollautomatisch auf lexikographische bzw. formale Grundformen oder Stammformen reduziert werden können. Die Verfahren sind im Rahmen des Forschungsprogramms der Zentralstelle für maschinelle Dokumentation (ZMD) entwickelt worden, das sich mit dem Aufbau eines Wörterbuchsystems für das automatische Indexing und Retrieval beschäftigt; sie sind auch für andere Dokumentationsprozesse einsatzbereit, z. B. für die Herstellung von komfortableren KWOC-Registern. Zu allen Algorithmen werden quantitative Aussagen bezüglich der Reduktionsquoten gemacht. Die Verfahren werden unter Anwendung der im Information Retrieval gebräuchlichen Parameter Recall und Precision bewertet.

This paper precents three algorithms which can be used to reduce English text words automatically and without a dictionary to their lexical or formal standard form or to their stem. The methods were developed at the Zentralstelle für maschinelle Dokumentation (ZMD) within a research program on dictionary construction for automatic indexing and retrieval. The algorithms can also be applied to documentation problems such as the production of convenient KWOC indexes. A quantitative analysis of the effectiveness of reduction is given for each of the algorithms. For the evaluation of the reduction quality the well-known retrieval measures of precision and recall are adapted.

1. Einleitung

Die automatische morphologische Analyse zum Zwecke der Vereinheitlichung von semantisch weitgehend gleichen und graphematisch nicht sehr unterschiedlichen Wörtern hilft eines der Grundprobleme der Dokumentation in automatisierten Informationssystemen zu lösen. Insofern die natürliche Sprache und damit ein nicht durchgängig normiertes Vokabular Grundlage von Textherstellung im weiten Sinne (Schreiben eines Textes, Abstracting, Indexing) und Recherche ist, werden von verschiedenen Menschen gleiche Sachverhalte verschieden dargestellt bzw. verschieden befragt. Eine Interkonsistenz ist nur annähernd zu erreichen. Eine EDVA, deren Stärke gerade diese Interkonsistenz ist, braucht hier Hilfestellung. Die Dokumentation sollte deshalb zunehmend mit der Formel arbeiten: *Anpassung durch Relationen*. Die morphologische Analyse stellt nur einen Sonderfall zur Relationengewinnung dar.

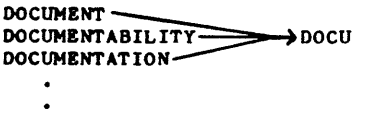
Aus dem Problembereich der morphologischen Analyse werden hier ausgeklammert die Präfixe und die zusammengesetzten Ausdrücke, ebenso syntaktische und semantische Klassifizierungen, die mit Hilfe der Identifizierung von Endzeichenketten von Wörtern in Annäherungsverfahren möglich sind[1]. Hier wird lediglich das Problem der Anpassung behandelt.

Unterschiedlich flektierte und derivierte Wörter können vom Computer durch die Rückführung auf Grund- oder Stammformen oder andere identische Wortfragmente als zusammengehörig erkannt werden. Die Flexionsendungen, *Flexive* genannt, und die Derivationsendungen, *Derivative* genannt[2], werden hierbei abgetrennt:

DOCUMENTS → DOCUMENT → DOCU

Durch die Pfeile sei vereinfacht angedeutet, daß zwischen DOCUMENTS und DOCUMENT eine bestimmte Relation besteht — die Grundformenrelation — und daß zwischen DOCUMENT und DOCU eine Stammformenrelation entstanden ist[3]. Das Moment der Anpassung wird dann verständlich, wenn man sieht, daß durch die Reduktion von DOCUMENTING und DOCUMENTED ebenfalls Grundformenrelationen entstehen, die alle zusammengefaßt werden können. Der Vollständigkeit halber sollte die Relation DOCUMENT → DOCUMENT mit aufgeführt sein:

Eine Stammformenrelation baut sich dann ähnlich auf:



Solche Relationen und natürlich noch andere (semantische, statistische, fachspezifische usf.) werden sowohl auf der Input- als auch auf der Output-Seite von Do-

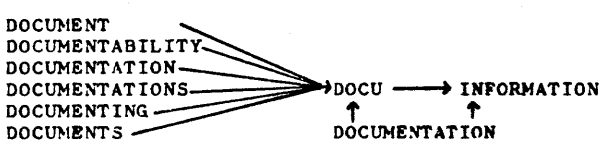
kumentationsprozessen gebraucht, das soll hier speziell heißen: beim automatischen Indexing und beim automatischen Retrieval[4].

Die dafür nötigen maschinellen Verfahren sind auf Grund linguistischer experimenteller Untersuchungen im Rahmen des Forschungsprogramms der ZMD entwickelt worden, mit dessen Hilfe ein maschineninternes Wörterbuch für das automatische Indexing und Retrieval aufgebaut werden soll[5]. Alle im folgenden vorgestellten Reduktionsalgorithmen sind in PL/1 in Unterprogrammtechnik programmiert und auf einer IBM 370/145 ausgetestet, so daß sie in der dokumentarischen Praxis routinemäßig eingesetzt werden können. Wie schon aus den Beispielen ersichtlich, wurde die Untersuchung am *Englischen* durchgeführt; als Analyse- und Textcorpora fungierten Texte des Referateorgans FOOD SCIENCE AND TECHNOLOGY ABSTRACTS (FSTA), für das die ZMD die EDV-Arbeiten leistet, so daß große Mengen maschinenlesbaren Materials zur Verfügung standen.

2. Morphologische Relationen beim automatischen Indexing.

Die folgende Beschreibung stellt lediglich eine modellhafte Idealisierung und damit Vereinfachung der Vorgänge beim automatischen Indexieren dar. Das Prinzip ist jedoch auch so zu erkennen. Mit Hilfe von Relationen, die in ein maschineninternes Wörterbuchsystem eingetragen sind, wird von Textwörtern auf Deskriptoren verwiesen, die so den Inhalt eines zu indexierenden Dokuments beschreiben. Falls im Wörterbuch der Deskriptor DOCUMENT vorgesehen ist, kann einem Text, in dem DOCUMENT vorkommt, auf Grund der Relation DOCUMENT → DOCUMENT dieser Deskriptor zugewiesen werden, aber schon nicht mehr, wenn die Pluralform DOCUMENTS vorkommt. Die Reduktion DOCUMENTS → DOCUMENT müßte vorab geleistet sein, denn eine EDVA kann nur die eindeutige Zeichenkette DOCUMENT als Einstieg für die angegebene Relation benutzen. Ohne morphologische Analyse müßten alle flektierten und derivierten Formen Bestandteile besonderer Relationen sein. Dies würde nicht nur das Wörterbuch unnötig aufblähen, sondern auch eine Unvollständigkeit im Relationengefüge bewirken; denn wenn man noch bei den Flexionen alle Formen vorhersehen kann, so ist dies bei der sehr produktiven Derivation des Englischen kaum möglich.

Mit Hilfe der morphologischen Analyse kann man dann auch längere Relationenketten aufbauen. Wenn eine Relation DOCUMENTATION → INFORMATION vorgesehen ist, dann kann das Einstiegswort der Relation, DOCUMENTATION, der morphologischen Analyse unterzogen werden, so daß mehrere Textwörter, die der gleichen Analyse unterworfen werden, für die Relation verwendet werden können:



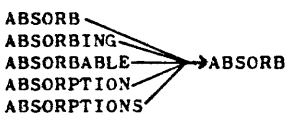
DOCU fungiert sozusagen als *formaler Oberbegriff* für die verschiedenen Flexionen und Derivationen.

3. Morphologische Relationen beim Retrieval.

Da Indexing und Retrieval reziproke Vorgänge sind, stellt sich das Problem ähnlich: Die Formulierung der Suche muß mit der Systemsprache kompatibel gemacht werden[6]; dabei bleibt es gleich, ob man mit einem deskriptororientierten System arbeitet oder die free-text-seach anwendet. Im einen Fall müssen Frageformulierungen und die im System vorhandenen Deskriptoren, im anderen Fall Einstiegswörter und Textwörter angepaßt, d. h. vereinheitlicht werden. Hierbei erweist sich eine algorithmisierte, linguistischen Prinzipien folgende morphologische Analyse der schrittweisen right-hand-truncation als überlegen, da weniger Homographen entstehen und da den graphematischen Veränderungen der Grundformen bzw. Stammformen bei Flexion bzw. Derivation besser Rechnung getragen werden kann, z. B. ABSORB → ABSORPTION.

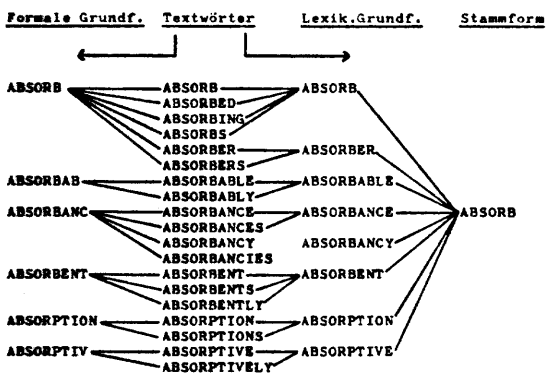
4. Beschreibung der Algorithmen

Die Vereinheitlichung des Vokabulars kann nun unterschiedlich *stark* erreicht werden. Es ist unmittelbar einleuchtend, daß der Informationsverlust bei Reduktionen



unterschiedlich hoch ist. Dieser Tatsache wird dadurch Rechnung getragen, daß drei unterschiedliche Reduktionsalgorithmen bis zur Produktionsreife[7] entwickelt wurden, so daß drei qualitativ verschiedene Relationsarten entstehen. Bei jeder Anwendung sollte man sich überlegen, welche Stärke angemessen ist, und man könnte die entstandenen Relationen unterschiedlich gewichten.

Das folgende Beispiel erläutert die Arbeitsweise der drei Algorithmen:



Es muß besonders darauf hingewiesen werden, daß alle Verfahren *wörterbuchunabhängig* arbeiten. Eine Reduktion kann hier also nicht dadurch geschehen, daß nach Abschneiden eines Derivativs, z. B. ABSORB-ABLE → ABSORB, das verbleibende Wort oder der Stamm, ABSORB, nach Identität mit einem Wörterbucheintrag überprüft werden kann. Die Algorithmen helfen daher, konventionelle oder maschineninterne Wörterbücher aufzubauen. In dem Beispiel entstehen durch die lexikographische Grundformenreduktion acht Grundformen, durch die formale Grundformenreduk-

tion sechs formale Grundformen und durch die Stammformenreduktion eine Stammform.

Lexikographische Grundformen entsprechen Einträgen in konventionellen Lexika. Die englischen Flexive werden entweder ersatzlos abgetrennt (FISHES → FISH) oder rekodiert (CITIES → CITY). Die graphematischen Konditionen hierfür sind zum Teil recht kompliziert, vor allem für die Verben, da ungefähr die gleiche Anzahl Verben auf Vokal (meist E) und auf Konsonant enden (INFORMING → INFORM; STRIKING → STRIKE). Es muß also genau geklärt sein, wann ein Verbalflexiv durch E ersetzt werden muß. Als ein Beispiel für graphematische Regeln seien die Regeln für die Abtrennung des Plural-S angeführt (vereinfacht). Im Algorithmus ist vorab die mögliche Abtrennung bzw. Rekodierung von IES bzw. ES untersucht worden. Hierbei gelten folgende ad hoc eingeführte Symbole:

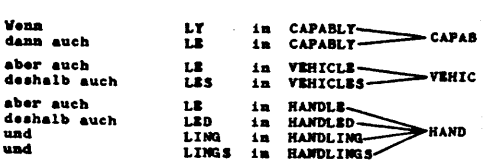
J	Anzahl der Zeichen eines Wortes
(J-2,2)	Zeichenkette des Wortes, die beim drittletzten Zeichen beginnt und die zwei Zeichen lang ist
β	alle Konsonanten
%	alle Vokale (auch Y)

S darf abgetrennt werden, wenn

(J-1,1)	= β (nicht S)	BIRDS →	BIRD
	= E	HOUSES →	HOUSE
(J-2,2)	= α Y	BOYS →	BOY
	= α O	RADIOS →	RADIO
	= OA	COCOAS →	COCOA
	= EA	FLEAS →	FLEA

Computerorientierte linguistische Verfahren arbeiten selten fehlerfrei, jedoch halten sich die Fehler hier in tolerablen Grenzen. Mit Hilfe von Zufallsstichproben aus einem Corpus von 72 000 verschiedenen Wörtern wurde errechnet, daß mit 95 Prozent Wahrscheinlichkeit die Fehlerquote nicht höher als 0,5 Prozent sein wird. Ein typischer Fehler, der allerdings durch Ausnahmelisten vermieden werden könnte: CARIES → CARY.

Formale Grundformen sind Wortfragmente, bei denen die ‚normalen' englischen Flexive (IES, ES, S, ING, IED, ED, ER, EST, IER, IEST) und die in wissenschaftlichen Texten häufig vorkommenden lateinischen Flexive (EUM, IUM, UM, US, EA, IA, A, I, AE, EX, IX, ICES) abgetrennt werden, ohne daß die entstehenden Wortfragmente rekodiert würden CITIES → CIT). Um das für die Zwecke der Dokumentation angestrebte Ziel zu erreichen, zusammenzuführen, was auf der jeweiligen Ebene der Wortanalyse zusammenführbar ist, müssen zuweilen zusätzlich gewisse Endzeichenketten abgetrennt werden, die keine echten Flexive sind, z. B. E in BELIEVE wegen BELIEV ING. Es müssen also sozusagen Folgeerscheinungen der unrekodierten Deflexion wieder in Ordnung gebracht werden. Das kann zum Teil über mehrere Stufen gehen:



Die ganze Angelegenheit ufert aber nicht aus. Die formale Grundformenreduktion benötigt 43 Flexive, dazu einige Restriktionsregeln (z. B. Y nicht abtrennen, wenn (J-1,1 = a) und einige Regeln zur Doppelrekodierung (STOPPING → STOPP → STOP).

Stammformen entstehen durch Abtrennen[8] der Derivative von vorab deflektierten Wörtern. Die lexikographische Grundformenreduktion wird also der Stammformenreduktion vorgeschaltet. Es werden z. B. nicht mehr ATE und ATES und ATED und ATING als Derivative benötigt, sondern lediglich ATE. Dies war in bisherigen Verfahren nicht möglich[9], da die Regeln für die lexikographische Grundformenreduktion noch nicht wörterbuchunabhängig formuliert waren. Die entstehenden Stammformen brauchen nicht identisch mit linguistischen Stammformen zu sein, und ebenso braucht die Bedeutung nicht mehr für einen kompetenten Sprecher der Sprache erkennbar zu sein, da lediglich der Computer das Wort identifizieren soll. Zwei ‚dokumentationsfunktionale' Bedingungen müssen erfüllt sein:

a) Die Zusammenführung von Wörtern verwandter Bedeutung und gleicher graphematischer Herkunft muß gewährleistet sein.

b) Es dürfen keine Homographen entstehen.

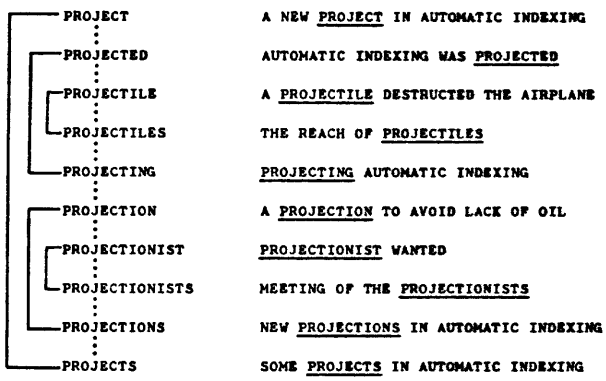
Beide Kriterien sind nur im Idealfall immer ganz erfüllt. (Bewertung s. unten)

5. Anwendungsmöglichkeiten

Allein der für das zu dokumentierende Wissenschaftsgebiet zuständige Dokumentar kann entscheiden, ob eine schwache oder starke Wortreduktion dem jeweiligen Zweck angemessen ist. Ein automatisiertes Informationssystem sollte jedoch nicht nur in der Experimentierphase mehrere Möglichkeiten bereitstellen, damit das parallele Benutzen verschiedener Verfahren je nach Bedarf möglich ist. Als *Zwischenstufen* zwischen der formalen Grundformen- und der Stammformenreduktion sind mehrere Verfahren denkbar, die weniger stark reduzieren, also in geringerem Maße semantische Information vernachlässigen als die Stammformenreduktion. Dabei könnten folgende Kriterien berücksichtigt werden: Häufigkeit des Vorkommens der Derivative; es könnten nur Derivative unterhalb einer gewissen Länge akzeptiert werden; es könnten nur solche Derivative berücksichtigt werden, die sich eindeutig als Substantiven zugehörig erkennen lassen; die Auswahl könnte semantische Gesichtspunkte geltend machen usf.

Mit etwas Phantasie wird der erfahrene Dokumentar erkennen, daß Wortreduktionsalgorithmen nicht nur für morphologische Relationen in Projekten für das automatische Indexing und Retrieval verwendbar sind. Hier sei nur auf die Registerherstellung, speziell das *KWOC-Register* verwiesen. In Registern, die meistens alphabetisch geordnet sind, stehen nicht alle semantisch zusammengehörigen und in ihrer phonologischen Gestalt weitgehend ähnlichen Wörter zusammen, da Flexion und Derivation eine Veränderung der Sortierreihenfolge bewirken. Wenn man akzeptiert, daß für die Zwecke der Dokumentation die wichtige Information in den Grund- oder Stammformen enthalten ist, dann ist es sinnvoll, unter Anwendung der verschiedenen Reduktionsverfahren Grund- oder Stammformenregi-

ster zu erstellen, so daß die phonologischen Veränderungen durch Flexion und Derivation unberücksichtigt bleiben. Der Nutzen solcher Register zeigt sich besonders bei den KWOC-Registern (Keyword-out-of-context), bei denen das herausgerückte Keyword als Sortierbegriff dient, so daß beim fertigen Register alle gleichen Keywords zusammenstehen, nicht aber immer die flektierten und derivierten Formen. Das kann vor allem bei großen Registern, bei denen die Keywords sehr viele Zeilen ausmachen, von Nachteil sein.



In der folgenden Zusammenstellung ist auf das herausgerückte Keyword die lexikographische Grundformenreduktion angewendet worden.

Beispiel für KWOC nach der Grundformenreduktion:

PROJECT	A new *project* in automatic indexing
	Automatic indexing was *projected*
	Projecting automatic indexing
	Some *projects* in automatic indexing
PROJECTILE	A *projectile* destructed the airplane
	The reach of *projectiles*
PROJECTION	A *projection* to avoid lack of oil
	New *projections* in automatic indexing
PROJECTIONIST	*Projectionist* wanted
	Meeting of the *projectionists*

Eine Stammformenreduktion wäre, wie das Beispiel zeigt, für diese Anwendung nicht so günstig, da die Dokumente, die den Filmvorführer angehen (PROJECTIONIST → PROJECT), mit denen, die allgemein Projekte angehen, zusammenkämen. Wohl wäre aber eine weitere Verbesserung zu erzielen, wenn über die *Stammform maschinenintern sortiert* und über die *Grundform ausgegeben* würde. PROJECTILE würde dann nicht mehr den Zusammenhang von PROJECT und PROJECTION stören, da ILE nicht als Derivativ vorgesehen ist.

6. Bewertung

Um die unterschiedliche Stärke der Reduktionsalgorithmen einschätzen zu können, sollen *quantitative* Aussagen gemacht werden. Für statistische Zwecke, u. a. auch um Aussagen über Reduktionsquoten bei größer werdenden Textmengen machen zu können, sind aus dem Material von FSTA fünf ineinanderverschachtelte Corpora erstellt worden. Die kleinere Textmenge ist in der nächst größeren vollständig enthalten. Aus den Texten sind Stopp- oder Funktionswörter entfernt[10]; sie sind in den folgenden Zahlen nicht enthalten.

FSTA-Hefte	1/72	1-3/72	1-6/72	1-12/72	7/71-6/73
Types[11]	11824	22263	32475	48445	72278
Tokens	80513	244283	484594	949839	1851480
LGF	2305	3912	5249	6988	9427
FGF	2754	4848	6630	9159	12816
STF	4259	7426	10030	14136	19743
LGF	19,5	15,6	16,1	14,4	13,0
FGF	23,2	21,7	20,4	18,9	17,7
STF	36,0	33,3	30,9	29,2	27,3

LGF: Lexikographische Grundformenreduktion
FGF: Formale Grundformenreduktion
STF: Stammformenreduktion

Durch die Stammformenreduktion fallen also z. B. bei einem Corpus von 72278 Types 19743 oder 27,3 Prozent aller Types fort, d. h., sie werden mit anderen zusammengeführt.

Diese quantitativen Aussagen zur Reduktionsquote reichen jedoch zur Beurteilung nicht aus. Es müssen zusätzlich Bewertungsverfahren angegeben werden, die die Richtigkeit und Vollständigkeit der Reduktionen überprüfen. Es muß *bewertet* werden, ob alles, was hätte zusammengeführt werden sollen, zusammengeführt worden ist und ob das, was zusammengeführt ist, auch richtig zusammengeführt worden ist. Hier bieten sich die im Information Retrieval gebräuchlichen Parameter ‚Recall' und ‚Precision' an, die wie folgt umdefiniert werden.

Der *Recall* für jedes *einzelne* Wort kann wie folgt berechnet werden:

$$R(W) = \frac{\text{Anzahl der mit W richtig zusammengeführten Wörter}}{\text{Anzahl aller (im Corpus vorhandenen) mit W zusammenzuführenden Wörter}}$$

Für die *Precision* gilt:

$$P(W) = \frac{\text{Anzahl der mit W richtig zusammengeführten Wörter}}{\text{Anzahl aller mit W zusammengeführten Wörter}}$$

Für die Bewertung von *gesamten* Corpora, im Idealfall der vollständigen englischen Sprache, werden die Durchschnitte R und P der Werte R(W) und P(W) eingeführt:

$$R = \frac{\text{Summe aller R(W)}}{\text{Anzahl aller W, für die R(W) definiert ist}[12]}$$

$$P = \frac{\text{Summe aller P(W)}}{\text{Anzahl aller W, für die P(W) definiert ist.}}$$

Die Berechnung und Einzelheiten des Bewertungsverfahrens werden hier nicht diskutiert, sondern lediglich die Ergebnisse mitgeteilt.

	STF	FGF ST	LGF ST	FGF	LGF
R	0.90	0.57	0.49	0.95	0.97
P	0.90	0.96	0.97	0.96	0.97

Die Zahlen in den Spalten 2 und 3 verstehen sich so, daß die formale und die lexikographische Grundformenreduktion unter dem Anspruch der Stammformenreduktion bewertet werden, d. h., es wird so getan, als ob die beiden Verfahren unzureichende Stammreduktionsverfahren wären. Dieses an sich ‚ungerechte' Verfahren liefert aber Vergleichswerte für die unterschiedliche Stärke der Verfahren. In den Spalten 4 und 5 werden die beiden Grundformenreduktionen nach ihrem eigenen Anspruch bewertet, inwieweit sie auf der jeweiligen Wortanalyseebene all das zusammenführen — und nicht mehr —, was zusammenzuführen ist.

Anschrift des Autors:
Rainer Kuhlen, Lehrinstitut für Dokumentation in der DGD (LID), 6000 Frankfurt/Main 1, Westendstraße 19.

Anmerkungen

(1) Vgl. hierzu *Kuhlen*, R.: Flexive und Derivative in der maschinellen Verarbeitung englischer Texte. Abschlußarbeit zur zweijährigen Nachuniversitären Ausbildung von Informationswissenschaftlern in der ZMD. Frankfurt/M. 1974. 287 S. (Als Manuskript vervielfältigt). Auf diese Arbeit wird auch im folgenden ständig verwiesen.

(2) Die linguistische Definitions- und Methodenproblematik bleibt hier ausgeklammert.

(3) An diesem Beispiel kann man schon erkennen, daß die für die Zwecke der Dokumentation gebildete Stammform nicht in jedem Fall mit der linguistischen übereinzustimmen braucht.

(4) Einen Überblick über den augenblicklichen Stand des automatischen Indexing gibt: *Sparck Jones*, K.: Automatic Indexing. A state of the art review Computer Laboratory, Univ. of Cambridge (England), April 1974.

(5) Vgl. *Lustig*, G.: Probleme der Wörterbuchentwicklung für das automatische Indexing und Retrieval. in Nachr. Dok.: 25 (1974) Nr. 2, S. 50–54.

(6) Vgl. *Holmberg-Dickson*, D.: Wortstammretrieval: Eine Studie zur gegenwärtigen Lage in Praxis und Forschung. (Berichte zur Anwendungsentwicklung 1.) Konstanz: Telefunken Computer GmbH, Dez. 1972.

(7) In der in Anm. (1) erwähnten Arbeit sind weitere Verfahren auf linguistischer und statistischer Basis vorgeschlagen, die der experimentellen Beschreibung der suffixalen Struktur des Englischen dienen.

(8) Für die Abtrennung sind zwei Verfahren gebräuchlich: Unter *„longest matching"* versteht man ein Derivativabtrennungsverfahren, bei dem durch Vergleich mit einer Derivativliste die längstmögliche Endzeichenkette bei Übereinstimmung mit einem Derivativ abgetrennt wird. Die Aufstellung der Derivative — in der in Anm. (1) genannten Arbeit sind 441 (deflektierte) Derivative klassifiziert worden — orientiert sich dabei mehr an strukturalistischen Prinzipien des amerikanischen Distributionalismus. Das Prinzip der *Iteration* beruht darauf, daß Derivative bzw. Derivativbestandteile in Klassen eingeteilt werden, entsprechend ihren Möglichkeiten, sich mit dem Stamm oder anderen Derivativen zu verbinden. Man trennt bei der Reduktion nun iterativ ab, zuerst wird das Verzeichnis terminaler Endzeichenketten abgesucht, ist eine gefunden, so wird zur nächsten Klasse übergegangen usf. Aus jeder Klasse darf nur eine Zeichenkette genommen werden. Das iterative Verfahren arbeitet also nicht allein mit tatsächlichen, sondern auch mit potentiell zu erwartenden Derivativen, es ist also generativ.

(9) So nicht bei *Salton* und *Lovins*. Vgl. *Lovins*, J. B.: Development of a stemming algorithm, in: Mechanical translation 11 (1968), Nr. 1/2, S. 22–31.

(10) Bei etwa 200 dieser Wörter, die alle sehr häufig vorkommen, verringert sich der Umfang der Corpora um ca. 40 Prozent.

(11) Die Begriffe „Type" und „Token" werden nicht ausdrücklich definiert. Die Anzahl der Types gibt die Anzahl der unterschiedlichen Wörter eines Textes, die Anzahl der Tokens die Anzahl aller Wörter dieses Textes an. Wenn in einem Text HOUSE hundertmal vorkommt, wird es als ein Type und als 100 Tokens gezählt.

(12) R(W) und analog P(W) sind dann nicht definiert, wenn ein Wort isoliert vorkommt, also ohne Flexions- oder Derivationsformen. Zähler und Nenner sind dann 0, da die Übereinstimmung des Wortes mit sich selbst nicht als besondere Leistung der Algorithmen angesehen werden sollte.

Literatur

Chapin, P. G.; *Norton*, L. M.: A procedure for morphological analysis. Information system language studies No. 18. Report MTP-101. Bedford, Mass.: Mitre Corp. (1968).

Chapin, P. G.: On affixation in English. in: Bierwisch, M.; Heidolph, K. E. (Eds.): Progress in linguistics. (Janua Linguarum Series Major 43.) The Hague, Paris: Mouton 1970. S. 51–63.

Dietrich R.: Automatische Textwörterbücher, Studien zur maschinellen Lemmatisierung verbaler Wortformen des Deutschen. Tübingen: Niemeyer 1973.

Dolby, J. L.; *Earl*, L. L.; *Resnikoff*, H. L.: The application of English word morphology to automatic indexing and extracting. (Report No. M-21–65–1.) Palo Alto, Calif.: Lockheed Missiles & Space Company 1963.

Harris, B.: „Polygram" grapho-morphology analyzer for English (Microf. Montreal 1971).

Kay, M.; *Martins*, G. R.: The MIND system: The morphological-analysis program. Santa Monica, Calif.: Rand Corp. (1970).

Lovins, J. B.: Vgl. Anm. (9).

Pfeiffer, M.: Automatische Suffixanalyse. (Diplomarbeit, Institut für Informatik der Univ. Stuttgart, Masch.schr. 1973).

Salton, G.; *Lesk*, M. E.: Information analysis and dictionary construction, in: *Salton*, G. (Ed.): Information storage and retrieval. Scientific report ISR-11 (1966).

Schott, G.: Linguistische Algorithmen zur Deflexion als Mittel zum automatischen Indexieren im Deutschen. 3. Jahrestagung, Gesellschaft für Informatik, Hamburg, 8.–10. Okt. 1973. (Lecture notes in computer science 1.) Berlin (u. a.): Springer 1973 S. 421–430.