LDV-FORUM

Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV

LDV-Forum **11.2 (1994)** Forum der Gesellschaft für Linguistische Datenverarbeitung e.V.

Herausgeber
Prof. Dr. Gerhard Knorz; Gesellschaft für Linguistische Datenverarbeitung e.V. (GLDV)
Anschrift: Fachhochschule Darmstadt, Fachbereich Information und Dokumentation (luD), Schöfferstr. 1-3, D-64295 Darmstadt Tel.: (06151)16-8490; Fax: (06151)16-8980; Email: knorz@fhdacom2.fhrz.fhdarmstadt.de

Redaktion
Gerhard Knorz, Ute Hauck

Wissenschaftlicher Beirat
Dr. Karin Haenelt, Prof. Dr. Christa Hauenschild, Prof. Dr. Gerhard Knorz, Prof. Dr. Jürgen Krause, Prof. Dr. Burghard Rieger, Dr. Dietmar Rösner, Prof. Dr. Burkhard Schaeder

Erscheinungsweise
Zwei Hefte im Jahr, halbjährlich zum 30. Juni und 30. Dezember

Bezugsbedingungen
Für Mitglieder der G LDV ist der Bezugspreis des LDV-Forum im Jahresbeitrag mit eingeschlossen. Jahresabonnements können zum Preis von DM 40,(incl. Versand), Einzelexemplare zum Preis von DM 20,- (zuzügl. Versandkosten bei der Redaktion bestellt werden.

Editorial

Positiv - *soll man den Tag beginnen..1,* und warum nicht auch ein Heft des LDV-Forum? Also, ich gewinne den Eindruck, daß sich die Angebotssituation bei den Fachbeiträgen stetig bessert. Das betrifft nicht nur die Anzahl der entsprechenden Seiten im vorliegenden Heft, sondern auch die Tatsache, daß Anfragen und Einreichungen für kommende Ausgaben vorliegen. Natürlich handelt es sich bei dieser Einschätzung um Schlußfolgerungen und Beobachtungen, für die eine signifikante statistische Grundlage (noch) nicht existiert.

Auch was die Rezensionen und Tagungsberichte betrifft, sieht es ganz nach einer Besserung aus: zum einen werden mittlerweile aktiv Beiträge dieser Art angeboten, zum anderen hat mich der Vorstand der GLDV bei der Akquisition effektiv unterstützt. Ich danke herzlich allen, die zu einem informationsreichen LDVForum beitragen und beigetragen haben.

Als jemandem, der die Kunst des Zauberns nie völlig aufgegeben hat, hätte mir eigentlich bewußt sein müssen, daß nichts unzweckmäßiger ist, als eine Absicht vorher anzukündigen. Nicht nur, daß die Zauberei durch Überraschung verblüfft - ein im Ansatz mißglückter Versuch läßt sich immer noch retten, solange man sich nicht auf ein Ergebnis festgelegt hat. In diesem Sinn will ich alte "Fehler" an dieser Stelle nicht wiederholen und auch die nicht eingelösten Ankündigungen nicht in der Sache kommentieren. Nur soviel für diejenigen, die tatsächlich auf in Aussicht gestellte Beiträge warten: Nichts davon ist aufgegeben - aber es hätte deutlich mehr an Engagement meinerseits bedurft, um in diesem Heft etwas Substantielles vorweisen zu können.

So, und nun wird sich zeigen, daß auch eine positiv begonnene Sache nicht unkritisch zu Ende gehen muß. "Was ist los mit der Computerlinguistik?" ist nämlich die Frage, die sich mir aufdrängt. Man muß nicht den Finger in die Wunde der geplatzten GLDV-Herbstschule legen - man muß sich nur etwa als Teilnehmer der KONVENS fragen, wieso denn eine Tagung dieser Art mit der Rückendeckung mehrerer einschlägiger wissenschaftlicher Verbände nicht eine Teilnehmerzahl erreicht, die über der früherer GLDV-Jahrestagungen liegt? Wenn es am Programm nicht gelegen hat (siehe dazu die Tagungsberichte in diesem Heft), den Tagungsort kann man schon gar nicht

verdächtigen: Wien ist allemal eine Reise wert, wenn etwa Graz wenige Wochen später bei der informationswissenschaft-

1 Unter dieser Überschrift jedenfalls erinnere ich mich an regelmäßige gute Ratschläge zu morgendlicher Stunde im Radio.

lichen Tagung ISI '94 *mit* ca. 250 Teilnehmern in historischer Umgebung geradezu unzeitgemäße2 Pracht entfaltete. Um es vorsichtig zu formulieren: Ein Signal des Aufbruchs ging jedenfalls von der Atmosphäre der KONVENS nicht aus.

Und wenn nun im Frühjahr 95 die GLDV nach Regensburg einlädt, um unter anderem 20 Jahre GLDV zu reflektieren? Ich wünsche den Veranstaltern und uns allen, daß nicht der Eindruck entstehen muß, daß heute eine Idee nur verwaltet wird, die vor 20 Jahren junge WissenschaftlerInnen *mit* Begeisterung erfüllt hat! Denn wenn es so wäre, wer sollte das verstehen in einer Zeit, in der Texte und Rechner mittlerweile untrennbar zusammengehören? In der man Laien umfangreiche Informationssysteme verkaufen will, *mit* denen sie ohne Recherchekenntnisse alles und jedes effektiv auffinden sollen? In der jede bessere Textverarbeitung grammatische Fehler auffinden lassen will? In der man den automatischen "Übersetzer" in der Westentasche *mit* sich herumtragen kann? In einer Zeit also, in der Computerlinguistik greifbar ist, ohne daß man "befürchten" müßte, die Probleme seien gelöst? Denn daß wir *mit* den genannten Werkzeugen und Produkten erst am Anfang stehen, kann nun wirklich jeder unvoreingenommene Nutzer ohne jede Mühe sehen.

Nun, ich sehe durchaus die thematische Konkurrenz der Multimedia-Welt der virtuellen Realität und der grenzenlosen Vernetzung. Und ich verstehe durchaus auch deren Attraktivität. Aber daß es *mit* der Faszination der natürlichen Sprache vorbei sein sollte, kann ich nicht wirklich glauben. Zumal das World Wide Web doch bestens demonstriert, daß bis auf das zufällig Gefundene so gut wie alles in Texte verpackte Wissen nur scheinbar zugänglich ist: ohne intellektuelle Aufbereitung oder aber Werkzeuge zum Durchsuchen von Texten versteckt sich jede Information in unübersehbarer Quantität.

Sehen wir uns in Regensburg?

G.K.

PS. Zu dem im letzten Forum thematisierten *Ligaturen-Problem* kann ich gegenwärtig nichts beitragen außer einen Hinweis auf den KONVENS-Beitrag von Steiner/Barth, der m.E. auf dieses Problem übertragbar sein sollte.

2 Natiirlich ist es gegenwärtig eine ungünstige Zeit für Reisekosten.

Titelgestaltung
Markus Allgayer, Saarbrücken

Fachbeiträge
Unaufgefordert eingesandte Fachbeiträge werden vor Veröffentlichung von mindestens zwei ReferentInnen begutachtet. Manuskripte (dreifach) sollten daher möglichst frühzeitig eingereicht werden und bei Annahme zur Veröffentlichung in jedem Fall zusätzlich auch noch auf Diskette (5 ¼“ bzw. 3 ½“) als ASCII oder LATEX-Datei übermittelt werden. Formatierungshilfen *(LDVforum.sty)* werden auf Wunsch zugesandt.

Rubriken
Die namentlich gezeichneten Beiträge geben ausschließlich die Meinung der Autoren wider. Einreichungen sind - wie bei Fachbeiträgen - an die Redaktion zu übermitteln.

Redaktionsschluß
Für alle Rubriken mit Ausnahme der als Fachbeiträge eingereichten Manuskripte:
für Heft 12.1/95: 30. Apr. 1995; für Heft 12.2/95: 31. Okt. 1995

Herstellung
IAI, Saarbrücken

Druck
reha GmbH, Saarbrücken

Auflage
400 Exemplare

Anzeigen
Preisliste und Informationen: Prof. Dr. Johann Haller, Institut für Angewandte Informationsforschung (IAI), Martin-Luther-Straße 14, D-66111 Saarbrücken; Tel.: (0681) 39313; Fax: (0681) 397482; Email: hans@iai.unisb.de

Bankverbindung LDV-Forum
(Prof. Haller): SaarLB Saarbrücken (BLZ 590 500 00) KtoNr. 20 00 21 43

GLDV-Anschrift
Prof. Dr. Winfried Lenders, Institut für Kommunikationsforschung und Phonetik (IKP), Poppelsdorfer Allee 47, D-53115 Bonn Tel.: (0228) 735638, Fax: (0228) 735639; Email: lenders@uni-bonn.de

GRUNDLAGEN DER GENERIERUNG DEUTSCHER VERBFORMEN MIT DEM COMPUTER

Rudolf Rausch & Horst Rothe

Zusammenfassung:

Vorgestellt wird ein auf IBM-PCs realisierter Algorithmus zur Bildung aller deutschen Verbformen aus ihrem Infinitiv. Der Algorithmus ist nicht an maschinelle Wörterbücher gebunden, reflektiert aber bei Ausnahmen auf Listen mit Ablautreihen etc. Er berücksichtigt sowohl starke als auch schwache Verben, trennbare und untrennbare Verben, Neologismen ebenso wie Onomatopoetika oder Entlehnungen aus anderen Sprachen. Infinitivphrasen werden unter Berücksichtigung der zu verändernden Wortstellung konjugiert.
Realisiert ist der Algorithmus als Unit in Borland Pascal 7.0 (OOP) und arbeitet äußerst schnell. Eingesetzt wird er in Lehrsoftware auf dem Gebiet Deutsch als Fremdsprache.

An algorithm for the generation of any German verbal form derived from its infinitive is presented. The algorithm, based on rules as far as possible, uses minimalized lists to handle the numerous exceptions and works without a computerized dictionary. It also deals with regular and irregular verbs, separable and unseparable prefixes, as well as neologisrns, onomatopoetica and borrowings from other languages. The inflection of infinitival phrases considers the necessary changes of word positions. The algorithm is implemented as a unit in Borland Pascal 7.0 (OOP) and is very efficient. It can be, and in fact already is, employed in teachware for German as a foreign language.

1 Zielsetzung

Bei der Vermittlung der Verben im Fach Deutsch als Fremdsprache (DaF) stehen Ausländer einer Fülle von Problemen gegenüber. Neben der richtigen Bildung der Formen spielen Fragen der unterschiedlichen morphosyntaktischen Eigenschaften, Reflexivität, Wortstellung, Abtrennbarkeit von Konstituenten und die Funktionalität eine wichtige Rolle. Die vorliegende Arbeit beschäftigt sich nur mit einem Teil dieser Probleme - nicht aus didaktischer Sicht, sondern mit der Absicht, den Computer als Lehr- und Übungspartner nutzbar zu machen. Damit der Rechner sich als geeigneter Partner erweist, muß er über die entsprechenden Regeln verfügen.

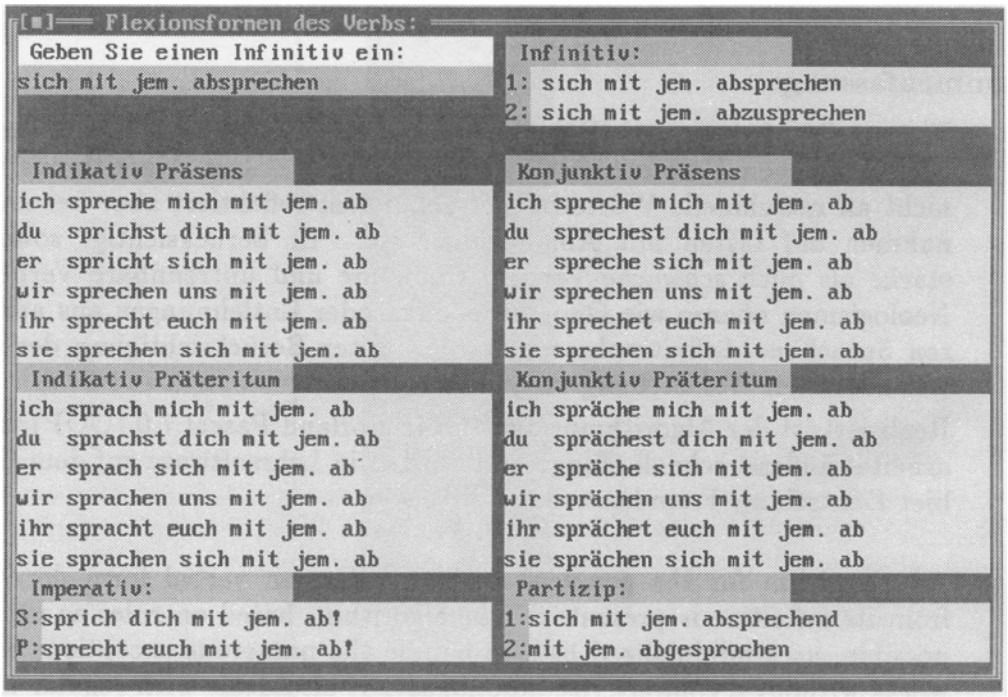
Ziel des vorgestellten Programmes ist es, zu allen deutschen Verben alle bildbaren Formen präsentieren zu können. Es erhebt sich die Frage, auf welche Weise die Menge aller

deutschen Verben überhaupt zu ermitteln ist, ob Lehnwörter und von Fremdwörtern abgeleitete Verben gleichberechtigt aufgenommen werden sollen, und ob die Vielfalt der durch Präfigierung und Zusammensetzung bildbaren Verben überhaupt erfaßbar sein kann. Tatsächlich behandelt der Algorithmus alle genannten Gruppen und besitzt die Fähigkeit, alle Formen von Onomatopoetica oder gar Neubildungen nach den Regeln der Konjugation deutscher Verben zu generieren.

Das Üben der Formen unter der Kontrolle des Computers kann nur dann zu befriedigenden Ergebnissen führen, wenn der Algorithmus

> den gesamten Formenbestand beherrscht und t> alle

(geläufigen) Formenvarianten akzeptiert.



Grammatiken, Regeln und Lernwille genügen leider nicht, den gesamten Formenbestand regelhaft ableiten zu können, da die in Nachschlagewerken angegebenen Beispiele nicht immer generalisierbar sind. Beispielsweise gibt der Duden keine Auskunft über die 1. Person Singular Präsens von" wissen". Die Vielzahl der Regeln, die Veränderungen des Basismorphems verlangen oder Elisionen erzwingen, erlauben oder verbieten, ist für den Lerner nur schwer zu überblicken *(messen, du mißt, er maß, gemessen, mißt, wandeln: ich wandle).*

Die Fähigkeit des Algorithmus, alle Verbformen erzeugen zu können, bildet die Basis für ein Nachschlagewerk sämtlicher Formen des Verbs im Präsens und im Präteritum sowohl im Indikativ als auch im Konjunktiv. Auch die Infinitive, Imperative und die beiden Partizipien fehlen nicht. Die physisch im Computer vorhandenen Listen sind auf ein Minimum beschränkt.

2 Algorithmus der Formenbildung

Der Algorithmus generiert zu einem gegebenen Infinitiv bzw. zu einer Infinitivphrase alle Flexionsformen. Die Lösung dieser Aufgabe gliedert sich wie folgt:

1. die syntaktische Analyse der Infinitivphrase

2. die morphologische Analyse der Infinitivform

3. die Generierung der jeweiligen Formen.

Während die syntaktische Analyse zu Beginn der Prozedur vergleichsweise einfach ist, gestaltet sich die morphologische Analyse der Infinitivform weit vielschichtiger. Die Ergebnisse beider Analyseschritte gestatten die Auswahl der zutreffenden Regeln bei der Generierung der einzelnen Formen.

2.1 Analyse des Infinitivs

Die Analyse von Infinitivphrasen wird im Punkt 4. behandelt, so daß zu Beginn der Betrachtungen die Infinitive im Zentrum stehen. Auch die Reflexivität soll zunächst unberücksichtigt bleiben (siehe 3.).

Eine der wesentlichen Aufgaben der Analyse ist es, die Basis des Verbs zu ermitteln. Zu diesem Zweck beginnt der Algorithmus mit einer Analyse des Infinitivs von rechts.

2.1.1 Analyse von rechts

Alle Infinitive des Deutschen enden auf *n.* Wird vom Algorithmus ein Wort erkannt, das diese Bedingung nicht erfüllt, wird es nicht weiter analysiert. Andernfalls ist es gerechtfertigt, anzunehmen, daß dieses *n* nicht der Basis angehört. In vielen Infinitiven geht diesem nein e voraus, das ebenfalls abgetrennt werden kann. Allerdings ist dabei zu berücksichtigen, daß es auch Infinitive wie *knien* gibt, bei denen das zum ie gehörige e nicht abgetrennt werden darf, während bei auf *-eien* ausgehenden Infinitiven (z. B. *schreien)* offensichtlich kein ie vorliegt und folglich das e zur Endung *-en* und nicht zur Basis gehört.

Zahlreiche Verben enden mit *-ern* oder *-eln.* Da bei diesen Verben in bestimmten Formen das e ausfallen kann bzw. muß, prüft der Algorithmus, ob es sich um ein Verb dieser beiden Gruppen handelt. Das e fällt nicht aus, wenn ihm ein Dehnungs-h, d. h. ein *h,* vor dem ein Vokal steht, vorangeht *(nähern).* Es erweist sich als zweckmäßig, diese Verben in der Behandlung nicht den Verben auf *-ern* oder *-eln* zuzuordnen.

Eine weitere Gruppe von Verben endet auf *-igen* oder *-lichen,* wobei *-ig* und *-lieh* nicht Bestandteil der Basis sind: *beschönigen} reinigen} verdeutlichen.* Zahlreiche Verben scheinen auf die Morphemgruppe *-igen* zu enden: *geigen} neigen} schweigen* usw., wobei das i dieser Verben als Bestandteil des Nukleus ei durchaus zur Basis gehört, so daß Verben, auf *-eigen* nicht zu dieser Gruppe gehören.

Die Gruppe der Verben auf *-ieren* ist für die Formenbildung ebenfalls von Bedeutung, kann jedoch zu diesem Zeitpunkt der Analyse vom Algorithmus noch nicht mit Sicherheit erkannt werden, man denke an *schmieren} verlieren,...,* die nicht zu dieser Gruppe gehören und bei denen ein Abtrennen von *-ieren* die Basis beschädigen würde!

2.1.2 Analyse von links

Derjenige Teil des Infinitivs, der nach dem Abtrennen der o. g. Buchstabengruppen von rechts übrigbleibt, enthält die Basis, der noch Folgen weiterer Wortkonstituenten vorangehen können. Dem Erkennen dieser vorangestellten Wortteile dient die anschließende Analyse des Wortrests von links.

2.1.2.1. *Abtrennbare Konstituenten*

Da viele deutsche Verbalpräfixe und andere vor der Basis des Infinitivs auftretende Konstituenten (im folgenden nur noch "Präfixe" genannt) reihenbildend sein können *(herkommen) hinausgehen} übereinkommen} umeinanderbinden,...*), erweist es sich als sinnvoll, diese iterativ abzutrennen. Eine Liste aller möglichen komplexen Präfixe für diese Analyse bereitzustellen, scheitert aufgrund der Vielzahl der möglichen Kombinationen. Obwohl

nicht alle Kombinationen von Präfixen auftreten, trennt der Algorithmus ohne Rücksicht auf die Zulässigkeit der Kombination zuverlässig an den Morphemgrenzen und isoliert auf diese Weise die Basis. Ein Hinweis auf die Sinnfälligkeit der Konstituentenkombination wird nicht gegeben. Auf die Bildung der Flexionsformen hat dies keinen Einfluß.

Dennoch sind bei der Abtrennung von Präfixen einige Probleme zu bewältigen:

> die Bestimmung aller möglichen Verbalpräfixe und die Klassifizierung ihrer Abtrennbarkeit
ablegen - er' legt ab, besitzen - er *besitzt,*
übersetzen - er *übersetzt* (einen Text), *übersetzen* - er *setzt über* (= überquert den Fluß)

> das Erkennen des Beginns der Basis
angeln (nicht: *an/geln), beten* (nicht: *be-ten),*
annullieren (nicht: er *nulliert an), abonnieren* (nicht: er *onniert ab)* usw.

Das erste dieser Probleme läßt sich anhand einer Liste lösen, die aus der Durchsicht der einschlägigen Nachschlagewerke zu erstellen war.

Schwieriger zu lösen ist das zweite Problem, da es nicht sinnvoll wäre, alle Wörter, die nur scheinbar mit einem Präfix beginnen, in Ausnahmelisten zu erfassen, zumal diese ihrerseits präfigiert sein könnten und im ungünstigsten Fall nur in präfigierter Form auftreten und deshalb weder in normalen noch in rückläufigen Wörterbüchern mit vertretbarem Aufwand aufzufinden wären - von der großen (endlichen?) Anzahl ganz zu schweigen!

Glücklicherweise ließ sich eine morphologische Regel formulieren, die dieses Problem erheblich vereinfacht, wenngleich auch sie nicht ohne Ausnahmen gilt.

Einsilbigkeitsregel:

1. Jede Basis eines deutschen Verbs enthält einen Nukleus: einen Vokal, einen Doppel vokal, ie oder einen Diphthong.

2. Jede Basis eines deutschen Verbs enthält **genau einen** silbentragenden Vokal, Doppelvokal, ie oder Diphthong, d. h. die Basen deutscher Verben sind einsilbig. (Das in *-ern* oder *-eln* auftretende e zählt dabei nicht zur Basis - es kann ja sogar ausfallen.)

Während 1. trivial zu sein scheint, ist 2. außerordentlich bemerkenswert, jedoch bisher kaum als Bestandteil einer Regel in Erscheinung getreten. Dabei dürfen einige Ausnahmen nicht unberücksichtigt bleiben. Es handelt sich dabei vorwiegend um Wörter fremder Herkunft wie *trompeten, posaunen, baldowern, orakeln, kalauern* u. a. Die deutschen Verben *arbeiten, antworten* und *heiraten* gehören ebenfalls in diese Gruppe. Diese Wörter müssen in Listen geführt werden, weil für die Bildung des Partizip II die Akzentstelle von Bedeutung ist:

'arbeiten - ge' arbeitet, 'kalauern - ge 'kalauert
trom 'peten - trom 'petet, 0 'rakeln - 0 'rakelt

Auf der Grundlage der **Einsilbigkeitsregel** läßt sich der überwiegende Teil der deutschen Verben in Präfixe und Basis aufspalten, indem während des Analyseprozesses iterativ Präfixe abgetrennt werden, solange solche an der linken Seite des verbleibenden Restwortes vorhanden sind und wenigstens eine Silbe als Basis verbleibt. (Wir benutzen hier nur den trivialen Teil 1 der Regel. Bei den zur Zeit bei uns anstehenden Arbeiten zur algorithmischen phonetischen Transkription von Texten erweist sich Teil 2 der Regel als bedeutsames Werkzeug zur Bestimmung der Akzentstellen längerer Wörter). Diese Vorgehensweise verhindert bei Verben wie den folgenden, daß falsche Morphemgrenzen ermittelt werden könnten:

Infinitiv	Analyse von rechts	Analyse von links	Bemerkung
dampfen	*dampf*	**dampf**	nicht: *da-**mpf***
einigen			weil *-igen* schon ab!
hinabwandern	ein	**ein**	
hinaufgehen	*hinabwand(e)r*	*hin-ab-**wand(*** **e)r**	
	hinaufgeh	*hin-auf-**geh***	nicht: *hin-auf-ge-**h***

(Die Basis ist durch Fettschrift hervorgehoben.)

Beim Abtrennen der Präfixe ist auch darauf zu achten, daß von einem Wort wie *hinterfragen* nicht *hin-* sondern tatsächlich *hinter-* abgetrennt wird. Verallgemeinert ausgedrückt: Stehen aufgrund der Zeichenfolge am linken Wortrand mehrere Präfixe für das Abspalten zur Wahl, muß der Algorithmus das zutreffende auswählen.

2.1.2.2. *Präfixe mit schwankendem Akzent*

Einen weiteren Problemfall stellen die Konstituenten *durch-, hinter-, über-, um-, unter, voll, wider-* und *wieder-* dar: Sie können sowohl akzentuiert als auch akzentlos auftreten. In Abhängigkeit davon sind sie abtrennbar oder nichtabtrennbar und das Partizip 11 wird (wenn nicht andere Bedingungen dies verhindern) mit bzw. ohne *-ge* gebildet. Die Verben mit *wieder-* sind alle auf dem Präfix betont, also unfest präfigiert, bis auf *wiederholen,* das (mit Bedeutungsunterschied) in beiden Akzentuierungen auftritt. Bei den übrigen sieben dieser Präfixe gibt es große Gruppen von Verben, die

1. stets auf dem Präfix akzentuiert sind
2. stets auf der Basis akzentuiert sind und
3. in beiden Akzentuierungen (mit Bedeutungsunterschied) auftreten.

Die ersten beiden Gruppen wurden mit entsprechender Kennzeichnung der Akzentstelle in einer Liste erfaßt, anhand derer der Algorithmus entscheiden kann,

> ob das Präfix abgetrennt werden muß oder nicht

> wie das Partizip 11 zu bilden ist

> ob das *zu* des Infinitiv 11 einzufügen ist oder ob es als Einzelwort vor den Infinitiv I tritt.

Manche dieser Präfixe werden eindeutig klassifizierbar, wenn sie sich mit bestimmten anderen Präfixen reihen wie z. B. *hindurch, herüber, umeinander.* Die zusätzliche Aufnahme solcher zusammengesetzter Präfixe in die Präfixliste hat die Liste zur Betonungskennzeichnung bedeutend verkürzt.

Gruppe 3 läßt wegen der Homographie beider Verben nicht erkennen, um welche Akzentvariante es sich handelt. In einem solchen Fall hilft nur eine Rückfrage an den Benutzer, der Eindeutigkeit dadurch herstellt, indem er die Akzentsilbe markiert:

über ´setzen oder *´übersetzen, um Jahren* oder *´umfahren*

Neben den abtrennbaren Präfixen müssen auch die nicht abtrennbaren Präfixe *(be-, ent-, emp-, er- ge-J ver-, zer-)* erkannt werden, weil Verben mit diesen Präfixen das Partizip 11 ohne *-ge* bilden. Außerdem kann nur dann, wenn diese erkannt werden, nach der Einsilbigkeitsregel festgestellt werden, ob *-ier* Basisbestandteil oder Suffix ist.

Bei einigen der mit einem Präfix mit schwankendem Akzent gebildeten Verben wird die Abtrennbarkeit durch Reihung mit einem nachfolgenden nichtabtrennbaren Präfix aufgehoben: *überbeanspruchen: du überbeanspruchst.* Der Algorithmus trägt diesem Sachverhalt auf der Basis entsprechender Einträge in den Ausnahmelisten Rechnung.

Bei Reihungen von Konstituenten, die *wieder* enthalten, bestehen zusätzliche Probleme der Getrennt- bzw. Zusammenschreibung. Im Infinitiv ist Zusammenschreibung möglich für

wiedereinsetzen, wiederaufführen, wiederherrichten, ...,

wenn der Ton nicht verteilt ist. In den finiten Formen ist in diesen Fällen die Basis abzuspalten und zu konjugieren, während die Reihe der übrigen Konstituenten nachgestellt werden muß. Die Besonderheit ist, daß diese an der Fuge zwischen *wieder* und dem Rest der Reihe zu trennen sind:

er *setzt sie wieder ein, sie führten es wieder auf, richte es wieder her!*

Für *wieder* bestätigt das Wörterverzeichnis des Dudens diese Regel.

Für *einander* liegen u. E. ähnliche Verhältnisse vor, wobei im Duden keine entsprechenden Wörter enthalten sind. Es ist jedoch anzunehmen, daß für *aneinandervorbeireden, auseinanderhervorgehen, ...,* Zusammenschreibung möglich sein sollte, auch wenn beispielsweise unter *aneinander* im Duden zu lesen ist "Getrenntschreibung, wenn... *aneinander* zu einem bereits zusammengesetzten Verb tritt" [Drosdowski 1991, S. 107]. Die finiten Formen weisen dieselbe Besonderheit wie Reihungen mit *wieder* auf:

redet nicht aneinander vorbei!, sie gingen auseinander hervor

Weder die Regelverzeichnisse verschiedener Ausgaben des Duden noch die durchgesehenen Grammatiken enthalten Regeln, die eine eindeutige Entscheidung der Schreibung erlauben. Der Algorithmus akzeptiert Zusammenschreibungen und fügt ggf. notwendige Leerzeichen in die Folge der abgetrennten Konstituenten ein.

Außerdem ergibt sich bei *einander* eine Einschränkung des Formenumfangs bei intransitiven Verben, denn reziproke Verben benötigen mindestens zwei Handlungsträger und bilden deshalb bei fehlendem Objekt keine Singularformen

auseinanderdriften: sie driften auseinander, nicht: *ich drifte auseinander.* hingegen:

etwas auseinandernehmen: ich nehme etwas auseinander

2.1.3 Zusammengesetzte Verben

Zusammengesetzte Verben, die nicht entsprechend 2.1.2. behandelt werden können, bereiten bei der Bestimmung der Morphemgrenzen erhebliche Schwierigkeiten. Mit Hilfe der *Einsilbigkeitsregel* gelingt es relativ leicht, sie zumindest als Zusammensetzungen zu erkennen. Handelt es sich um trennbar zusammengesetzte Verben, kann der Benutzer die Morphemgrenze vor der Basis markieren *(kalt/walzen).* Bei untrennbar zusammengesetzten Verben kann die Akzentstelle angeben werden. Dies ist einem Deutsch lernenden Ausländer nicht zuzumuten. Die folgenden Überlegungen dienen dem Ziel, dem Nutzer diese Entscheidungen durch den Algorithmus abzunehmen.

Bei einem Teil der zusammengesetzten Verben folgt der ersten Konstituente ein nicht abtrennbares Präfix *(vakuumbedampfen).* In solchen Fällen kann die Wortfuge mit einiger Sicherheit erkannt werden und aus dem Vorhandensein des nicht abtrennbaren Präfix folgt, daß das Partizip II ohne *ge-* zu bilden ist. Die Sicherheit beim Erkennen der Fuge und des Präfixes wird noch dadurch erhöht, daß überprüft wird, ob auf das ermittelte Präfix ein für eine Basis zulässiges Onset folgt. Dadurch werden Fehlanalysen wie *autogenschweißen: auto-ge-nschweißen* unwahrscheinlich. Verben dieses Typs treten hauptsächlich als Infinitive und als Partizipien auf. Die finiten Formen werden selten verwendet. Es erhebt sich die Frage, ob es sinnvoll ist, sie vom Algorithmus bilden zu lassen.

Weitere zur Formalisierung geeignete Regeln waren bisher weder aus der Literatur noch aus der Beschäftigung mit dem Gegenstand abzuleiten. Aus diesem Grunde war eine Liste zu erstellen, die die Wortfuge bzw. die Akzentstelle aller derjenigen Verben nachweist, die sich der bisher erläuterten algorithmischen Behandlung entzogen. Diese Liste ist nicht vollständig, da es eine große Zahl von Verben (die auch der Duden nicht nachweist) gibt, die z. B. für technische Prozesse stehen und nur in der jeweiligen Fachsprache gebräuchlich sind.

2.1.4 Starke Verben

Nachdem aus dem Infinitiv des Verbs die Basis ermittelt worden ist, kann überprüft werden, ob es sich um die Basis eines starken Verbs handelt. Dazu bedient sich der Algorithmus einer Liste der Basen aller starken Verben, die die zugehörigen Ablautreihen enthält. Außerdem weist diese Liste aus, ob das betreffende Verb zusätzlich eine schwache Form hat und ob diese der starken Form vorzuziehen ist. Besonderheiten der Formenbildung wie abweichende Endungen, abweichende Bildungen des Partizip II oder irreguläre Veränderungen der Basis sind enthalten. Die Existenz von Bedeutungs- oder Verwendungsunterschieden zwischen starken und schwachen Formen ist in dieser Liste ebenso vermerkt.

Beim Vergleichen der Basis mit den Listeneinträgen ist zu berücksichtigen, daß eine Reihe von Verben aufgrund ihrer Präfigierung schwach zu flektieren ist, obwohl die Liste der starken Verben die Basis enthält.

stark: *gleiten - glitt,* ebenso: *ausgleiten - glitt aus*

schwach: *begleiten - begleitete*

Damit erweist sich, daß nur das konsequente Abtrennen der Präfixe während der vorangehenden morphologischen Analyse zu korrekten Ergebnissen bei der Bestimmung der Basismorpheme führt, wohingegen das bedingungslose Aufsuchen der Basis eines starken Verbs in der Zeichenfolge des Infinitivs zu Fehlinterpretationen führen kann (stark: *leiden - litt* aber schwach: *kleiden - kleidete).* Ein Verb kann daher nur dann in der Liste der starken Verben als repräsentiert gelten, wenn sich die der Basis vorangehende Zeichenfolge vollständig in Präfixe oder Konstituenten von Zusammensetzungen aufgliedern läßt.

Wurde ein Verb als stark erkannt, ist anschließend der Nukleus zu bestimmen, damit dieser beim Bilden der flektierten Formen durch die entsprechenden Ablaute ersetzt werden kann. Als hilfreich erweist sich an dieser Stelle erneut die **Einsilbigkeitsregel,** die ausnahmslos für alle starken Verben gilt. Zu beachten ist lediglich, daß neben den Vokalen und Umlauten auch Diphthonge, Doppelvokale und ie erkannt werden müssen. Außerdem ist der Nukleus von rechts her zu suchen, um zu vermeiden, daß ein *u* , das Bestandteil von *qu* ist, fälschlich als Basisvokal betrachtet wird.

Bei einigen starken Verben ändert sich durch den Ablaut auch die Vokalqualität: *bitten* (ungespannt kurzes i) - *bat* (hinteres langes *a), gebeten* (gespanntes langes e)

Hier sind - wie die Beispiele zeigen - regelhafte Veränderungen des Basisauslauts erforderlich. Je nach Vokallänge müssen Doppelkonsonanten durch einfache ersetzt werden oder umgekehrt. Dazu ist es notwendig, die Qualität und Quantität des Basisvokals zu bestimmen. Der Nukleus gilt als lang, wenn

> es sich um einen Doppelvokal oder ie handelt

> es sich um einen Diphthong handelt

> dem Vokal ein Dehnungs-h folgt

> dem Vokal kein oder nur ein einzelner Konsonant folgt,

sonst kurz. Bei der späteren Synthese der Flexionsformen wird die so gewonnene Information benutzt, um die nötigen Anpassungen der Basisauslautkonsonanten entsprechend den Rechtschreibregeln auszuführen. Die zu verwendende Ablautreihe wird dem Listeneintrag zum Verb entnommen, ebenso abweichende Endungen dieses starken Verbs in bestimmten Formen.

2.1.5 Partizip 11 mit oder ohne *ge*

Ob das Partizip n mit oder ohne *ge* zu bilden ist, ermittelt der Algorithmus anhand des von Suffixen befreiten Infinitivs (vgl. 2.1.1). Betrachtet wird dabei

I. derjenige Teil, der rechts von allen abtrennbaren Präfixen und Wortkomponenten steht, falls solche vorhanden sind (Präfixe mit schwankender Betonung können zu diesem Zeitpunkt noch nicht als abtrennbar gewertet werden)

11. derjenige Teil rechts von der Einfügestelle für *ge* oder *zu,* falls eine solche bei der Analyse erkannt wurde. Zu diesen Verben gehören u.a. *notlanden - er notlandete notgelandet, mißbilden - mißgebildet, sonnenbaden – sie sonnenbadete - sonnengebadet.* Hier ist *ge-* an der im Eintrag der unter 2.1. 7.6. beschriebenen Liste angegebenen Stelle einzufügen.

In. der ganze von Suffixen befreite Infinitiv, wenn 1. und 2. nicht zutreffen

Das Partizip n ist mit *ge-* zu bilden, wenn dieser Teil (falls er allein zu sprechen wäre) auf der 1. Silbe akzentuiert ist, sonst ist das Partizip n ohne *ge-* zu bilden.

Ob ein abtrennbares Präfix vorliegt, wurde bereits bei der Analyse von links ermittelt. Abtrennbare Präfixe ziehen stets den Akzent auf sich, bis auf die oben genannten acht Konstituenten mit schwankender Akzentuierung. Wie in diesen Fällen zu verfahren ist, wurde bereits oben erläutert.

Zu klären bleibt, woraus der Algorithmus auf die Akzentuierung schließen kann. Folgende Fälle sind zu beobachten:

A Der zu überprüfende Wortteil enthält ein Akzentzeichen. Steht das Akzentzeichen vor dem Auftreten des ersten Vokals in der Zeichenfolge, ist die erste Silbe betont. Zu beachten ist, das *u* in *qu* nicht als Vokal gerechnet werden darf.

B Der zu überprüfende Wortteil ist in der Akzentliste (vgl. 2.1. 7.4.) enthalten, so daß verifiziert werden kann, ob der Ton auf der ersten Silbe liegt.

C Der zu überprüfende Wortteil ist einsilbig und damit auf der ersten (weil einzigen) Silbe akzentuiert.

D Ein nicht abtrennbares Präfix *(be-, ent-, emp-, er-, ge-, ver-, zer-)* ist vorhanden. Da diese Präfixe nie betont sind, ist der Wortteil nicht auf der 1. Silbe betont.

E Der zu überprüfende Wortteil ist mehrsilbig und endet mit *-ier* , das den Akzent auf sich zieht, folglich ist die erste Silbe unbetont.

F Der zu überprüfende Wortteil beginnt mit der Silbe *ur-* oder *un-*. Diese ziehen stets den Akzent auf sich *(urteilen – geurteilt),* die erste Silbe ist betont.

Liegt keiner dieser Fälle vor, was wenig wahrscheinlich ist, muß der Benutzer aufgefordert werden, ein Akzentzeichen zu setzen oder eine Abtrennstelle zu markieren, beim erneuten Versuch liegt dann entweder Fall A vor oder es wird entsprechend I eine Komponente abgetrennt.

Wenn *ge* anzufügen ist, erfolgt dies vor dem Wortteil, der entsprechend I bis In ermittelt wurde. Falls am Wortanfang eines der acht Präfixe mit schwankender Akzentuierung vorhanden ist, wird in der Fuge zwischen Präfixkette und Basis eingefügt, sofern die Präfixkette den Akzent hat.

2.1.6 Besonderheiten

Nach dem Aufgliedern in Basis und Präfixe (oder auch andere vorangestellte Wortteile) müssen noch einige Ausnahmen erkannt werden:

Wenn ein Konsonant bei der Bildung des Infinitivs elidiert wurde (z. B. *vollaufen voll-laufen),* ist, um die Formen korrekt bilden zu können, der entsprechende Konsonant im Onset der Basis zu ergänzen. Treffen später bei der Formenbildung wieder drei gleiche Konsonanten mit nachgestelltem Vokal zusammen, ist einer davon entsprechend der Rechtschreibregel wieder zu entfernen.

Besondere Ausnahmen bilden Verben wie *radfahren* oder *kegelschieben* , die im Infinitiv klein und zusammen geschrieben werden und sich sonst wie trennbare Verben verhalten. Bei der Abtrennung zur Bildung der finiten Formen muß die Großschreibung berücksichtigt werden:

sie fährt Rad, er *schiebt Kegel*

Die Verbform *möchten* wird vom Lerner oft als Infinitiv betrachtet. Um diesen Fehler zu erkennen, wird dieser Eingabe der zugehörige Infinitiv *mögen* zugeordnet.

Wie oben erwähnt, akzeptiert der Algorithmus das Vorhandensein von Akzentzeichen in der Eingabe. Falls ein Akzentzeichen enthalten ist, wird es in allen gebildeten Formen mitgeführt. Falsch gesetzte Akzentzeichen im Infinitiv können jedoch auch zu falschen Betonungen der finiten Formen führen. Schlimmstenfalls können dadurch sogar falsche konjugierte Formen erzeugt werden (*beurlauben — ge-, beurlaubt).* Der Benutzer sollte Akzente nur dann setzen, wenn der Algorithmus dazu auffordert.

Die Fuge zwischen Präfix und Basis bzw. zwischen den Konstituenten zusammengesetzter Verben kann durch / markiert werden. Dies sollte jedoch in allen Fällen, in denen der Computer nicht explizit danach verlangt, unterbleiben, da falsche Kennzeichnungen der Abtrennstelle zu falschen Formen führen können.

Untrennbar zusammengesetzte Verben, die *ge* oder *zu* bei der Bildung des Partizips bzw. des Infinitiv II in der Wortfuge aufnehmen, werden weitestgehend vom Algorithmus erkannt. Sollte dies - insbesondere bei Verben aus dem Fachvokabular spezieller Bereiche - nicht zutreffen, ist eine Markierung der Fuge mit dem Unterstrich (_) möglich. Auch hier gilt jedoch, daß fehlerhafte Kennzeichnungen der Wortfuge zu fehlerhaften Formen führen können.

2.1.7 Die verwendeten Ausnahmelisten

Wie die Ausführungen zeigen, kommt die algorithmische Lösung der Analyse nicht ohne Ausnahmelisten aus. Im einzelnen verwendet der Algorithmus folgende Listen:

2.1.7.1. *Liste der unregelmäßigen Verbstämme*

Diese Liste enthält die Basen aller starken Verben des Deutschen (über 200 Einträge).

Ein Eintrag enthält ein Basismorphem (gelegentlich mit vorangestellten Konstituenten) und folgende Angaben bzw. Kürzel:

r	wenn neben der unregelmäßigen auch die regelmäßige Form auftritt
R	wie r, jedoch ist die regelmäßige vorzuziehen
r\	wie r, jedoch hat das Partizip II nur die unregelmäßige Form
\	es existiert nur die regelmäßige Form, die starke Form darf nicht verwendet werden (*beauftragen - beauftragte* nicht: *beauftrug*)
ablaut [endung]	Ablaut des Präteritums (alternative Ablaute sind durch „|" getrennt.), Modifikation der Coda einer Basis und Endung (Ziffern bedeuten das Löschen von Zeichen der abgelauteten und orthographisch angepaßten Basis nach links)
ablaut [endung]	wie oben, jedoch für Partizip II
ablaut [endung]	wie oben, jedoch für 2. u. 3. Person Präsens Singular Indikativ
ablaut [endung]	wie oben, jedoch für Konjunktiv II
!s	Es gibt semantische Unterschiede zwischen starker und schwacher Form
!i	Hilfsverb *sein* muß angewendet werden
!b	Hilfsverb kann *sein* oder *haben* sein

Eintragungen in der Liste der unregelmäßigen Verben

beauftrag \	schreib ie ie	lad u: a: ä:|a: ü:
beantrag \	treib ie ie	leid i -2tt i -2tten
trag u: a: ä: ü:	verderb a o i ü	meid ie ie
hab a -1tte a: -t a -1 ä -2tt	sterb a o i ü!i	send r a -te a -t e e -
grab u: a: ä: ü:	schnaub R o: o: au ö:	wend r a -te a -t e e -
geb a: e: i: ä:	brauch R au -te au -t au äu|au -t	find a u i ä
heb o: o: e: ö:	sied r o -2tt o -2tten ie ö -2tt	empfind a u i ä
schieb o: o: ie ö	scheid ie ie	schind R u u i ü
werb a o i ü	schneid i -2tt i -2tten	schwind a u i ä!i

Ausschnitt aus der Liste der unregelmäßigen Verben

2.1.7.2. Liste der abtrennbaren Präfixe

ab	*dahinter*	*drum*
acht	*darunter*	*dar*
anheim	*darüber*	*da*
an	*drunter*	*durcheinander*
aufrecht	*drüber*	*'durch*
auf	*darum*	*durch*
aus	*davon*	*einander*
beisammen	*drein*	*ein*
bei	*drauf*	*empor*
dagegen	*drin*	*entgegen*

Ausschnitt aus der Liste der Präfixe

Diese Liste enthält alle abtrennbaren Präfixe sowie die Präfixe mit schwankender Akzentuierung. Es handelt sich hauptsächlich um einfache Präfixe, jedoch sind auch einige wenige Kombinationen von Präfixen enthalten, wenn diese Besonderheiten gegenüber ihren Bestandteilen haben oder nur in fester Kombination auftreten. Auch einige Morpheme, die strenggenommen keine Präfixe sind, sich aber im betrachteten Zusammenhang wie solche verhalten, wurden aufgenommen. Diese Liste enthält über 100 Einträge.

2.1.7.3. *Liste der scheinbar präfigierten Verben*

** ab-*	*anglisier*	*antizipier*
abandonnier	*anzmzer*	*antwort*
abortier	*annektier*	**bei*
absorbier	*annonczer*	*beimpf*
abonnier	*annotier*	*beinhalt*
abbreviier	*annullier*	*beirr*
an	*antichambrier*	**da*
anathematisier	*antikisier*	*datier*

Ausschnitt aus der Liste der scheinbar präfigierten Verben

Diese Liste dient der Vermeidung irrtümlicher Abtrennung von Präfixen. **In** unserer Liste sind derzeit etwa 30 Einträge enthalten. Die hier zu vermerkenden Verben lassen sich schwer in Nachschlagewerken auffinden, insbesondere dann, wenn sie nur in ihrerseits präfigierter Form gebräuchlich sind. Diese Liste verhindert auch, daß Beispiele wie *beinhalten* falsch interpretiert werden (nicht: *bei-nhalten).*

2.1.7.4. *Akzentuierungsliste*

Antwort1	*durchacker* 1	*volladl*
arbeiti	*durcharbeiti*	*vollbring2*
Argwöhn1	*Durchatm 1*	*vollend2*
baldower2	*Durchback 1*	*vollführ2*
Bea 1i	*durchbeb2*	*Vollfüll 1*
beend2	*Durchbeiß 1*	*Vollgie 1l*
befürwort2	*durchbieg* 1	*Volllauf 1*
beinhalt2	*Durchbild 1*	*Vollauf 1*
berliner2	*Durchblas 1*	*Vollmach 1*

Ausschnitte aus der Betonungsliste

Diese Liste dient der eindeutigen Bestimmung der Akzentstelle von Verben, bei denen der Algorithmus ohne diese Liste zu falschen Ergebnissen käme. Die Zahl hinter der Basis gibt die Nummer der Akzentsilbe an. Die Liste umfaßt etwa 500 Einträge, die meisten davon sind den Präfixen mit schwankender Betonung geschuldet.

2.1.7.5. *Liste zusammengesetzter Verben*

bekanntmach4	*sauberhalt4*	*sitzenbleib5*
bekanntwerd4	*schlittenfahr4*	*spazierengeh3*
blaumach4	*schönfärb4*	*spazierenfahr* 4
bleibenlass4	*schönmach4*	*spazierenreit4*
brachleg3	*schönred3*	*spazierenführ* 4

Ausschnitte aus der Liste der zusammengesetzten Verben

Das algorithmische Erkennen der Fuge zusammengesetzter Verben, die nicht im o. g. weit gefaßten Sinne als Präfigierung betrachtet werden können, gelingt nur in den unter 2.1.2. und 2.1.3. genannten Fällen. Mit dieser Liste ist es dem Algorithmus möglich, die gebräuchlichsten der übrigen zusammengesetzten Verben bei der Formenbildung an der richtigen Stelle zu trennen. Die Zahl rechts von der Basis gibt an, wieviel Zeichen weiter links die Wortfuge liegt, an der zu trennen bzw. *-ge-* oder *-zu-* einzuschieben ist. Die Liste umfaßt derzeit über 100 Einträge.

2.1.7.6. *Liste von Verben mit Fügestellen*

danksag3 notschlacht8
mißbild4, nottauf4,
mißstimm5 notwasser6
notland4, sonnenbad3

Ausschnitte aus der Liste der Verben mit Fügestelle

In dieser Liste sind die seltenen Wörter vermerkt, die bei der Bildung der finiten Formen nicht getrennt werden, und dennoch *ge-* bei der Bildung des Partizip 11 bzw. *-zu-* beim Infinitiv 11 in der Wortfuge aufnehmen. Die Zahl hinter der Basis gibt an, wieviel Zeichen weiter links sich die Wortfuge befindet.

2.2 Synthese der Formen

2.2.1 Anpassungen der Basis wegen Ablauts

Starke Verben lauten in zahlreichen Formen ab, d. h. der Nukleus der Basis ändert sich. Entsprechend der Vokalqualität und -quantität sind daher nicht selten Änderungen im Basisauslaut erforderlich, die den Rechtschreibregeln Rechnung tragen. Einschlägige Grammatiklehrbücher enthalten dazu leider kaum explizite Hinweise.

Lautet ein langer Vokal oder ein Diphthong auf einen kurzen ab, muß ein evtl. vorhandenes Dehnungs-h getilgt werden:

nehmen: ich nehme, du nimmst

Folgt dem Nukleus genau ein Konsonant in der Basis, ist dieser zu verdoppeln:

treten: ich trete, du trittst, er tritt

pfeifen: ich pfiff Lautet ein kurzer Vokal auf einen langen Vokal oder einen Diphthong ab, sind doppelte Konsonanten durch einfache zu ersetzen.

schaffen: ich schuf Besonderheiten ergeben sich bei ck, das durch *k* zu ersetzen ist *(backen - buk)* und bei *ss* bzw. *ß:*

Ein *ß* ist durch *ss* zu ersetzen, wenn der vorangehende Ablaut kurz wird und die nachfolgende Endung mit einem Vokal (also *e)* beginnt *(schließen - geschlossen, reißen - wir rissen).* Umgekehrt ist *ss* durch *ß* zu ersetzen, wenn ein konsonantisch beginnendes Flexionsmorphem folgt *(fassen - er faßt)* oder wenn die leere Endung folgt *(fassen - faß!, essen - iß-0* oder wenn ein kurzer Vokal auf einen langen Vokal ablautet *(essen - ich aß ich äße).*

2.2.2 Formen des Indikativ Präsens

Endungen im Indikativ Präsens	
ich ...-e	Wir ...-en
du . ..-st	ihr ...-t
er, sie, es ...-t	sie.. .-en

Um die konjugierten Formen des Indikativ Präsens zu generieren, geht der Algorithmus vom Infinitiv des Verbs aus. Dieser war während der Analyse bereits von der Endung und von abtrennbaren Konstituenten befreit worden. Im allgemeinen sind die Personalendungen anzufügen. Als letztes ist - falls das Verb als trennbar klassifiziert wurde - die abgetrennten Komponenten nach einem Leerzeichen rechts anzufügen. Es sind jedoch einige Besonderheiten zu berücksichtigen:

2.2.2.1. *Verben* auf *-ern und -eln*

Im Indikativ der 1. Person Singular fällt bei den Verben auf *-ein* das *e* aus.

wandeln: ich wandle nicht: *ich wandele*

Bei den Verben auf *-ern* ist sowohl die Form mit oder ohne dieses *e* möglich.:

wandern: ich wandre oder *wandere*

In der 1. und 3. Person Plural entfällt bei den Verben auf *-ein* und *-ern* das *e* der Endung *-en:*

wir handeln nicht: *handelen wir*
wandern nicht: *wanderen*

Geht der Infinitiv auf *-ssern* oder *-sseln* aus, erhebt sich die Frage, ob das *ss* in *ß* umzuwandeln ist, und ob überhaupt eine Elision des nachfolgenden *e* gestattet ist:

wassern – *ich wassre* oder *ich waßre* oder nur *ich wassere?*
fusseln – *ich fussle* oder *ich fußle* (zu fußeln?!) oder nur *ich fussele?*

Bedauerlicherweise geben ausgewiesene Nachschlagewerke darüber keine Auskunft, da wie oben bereits erwähnt - fast alle Nachschlagewerke nur einige Formen angeben, die 1. Person Singular findet man nur bei wenigen ausgewählten Beispielen.

Klar ist hingegen

wzr wassern sze wassern.

2.2.2.2. *Einschub eines e*

Bei der 2. und 3. Person Singular sowie in der 2. Person Plural wird der Einschub eines *e* erforderlich, wenn die Basis auf *d* oder tauslautet *(reden redest redet).* Endet die Basis auf m oder *n,* wird auch ein *e* eingeschoben, wenn davor ein Konsonant außer *I* oder r steht:

trocknen: wappnen:	*du trocknest, er trocknft*
rechnen:	*du wappnest, er wappnft*
filmen: du filmst	*du rechnest, er rechn§.t*
lernen: du lernst	nicht: *du filmest* (Konjunktiv!) nicht: *du lernest* (Konjunktiv!).

Ein Dehnungs-h darf an dieser Stelle nicht als vorangehender Konsonant betrachtet werden, sondern muß als Längenmarkierung des Vokals gelten:

rahmen du rahmst, er rahmt nicht: *rahmest, rahmet* (Konjunktiv)

Bei den Doppelkonsonanten mm und *nn* am Ende einer Basis erfolgt kein e-Einschub, da es sich nicht um zwei (verschiedene) Konsonanten handelt, sondern um eine orthographische Regelung, die den vorangehenden Vokal als kurz und ungespannt markiert.

kennen - du kennst, kommen - du kommst.

Diese Regeln des e-Einschubs gelten sowohl für schwache als auch für starke Verben.

Verben wiederum, deren 2. und 3. Person Singular ablauten, weisen nur in der 2. Person Plural diesen e-Einschub auf:

braten: - du brätst nicht: *du brätest,* jedoch: *ihr bratet*

2.2.2.3. *S-Laute im Basisauslaut*

Die Personalendung *-st* der 2. Person Singular muß in denjenigen Fällen, in denen die Basis auf einen S-Laut endet (dazu zählen neben *s* auch *ß, x* und *z)* auf *-t* geändert werden:

du nutzt, du hext, du reist, du reißt

2.2.2.4. *Anpassung der Basis* an *den Ablaut*

Bei starken Verben, für die die Ablautreihe bereits während der Analyse ermittelt wurde, ist ggf. in der 2. und 3. Person der Basisvokal durch den entsprechenden Ablaut zu ersetzen. Ändert sich dabei die Vokalqualität, sind weitere Anpassungen im Basisauslaut erforderlich (vgl. 2.2.1.).

2.2.2.5. *Ausnahmen*

Eine besondere Behandlung in den angegebenen Personen erfahren folgende Verben:

se1,n:	*ich bin,*	*du bist,*	er *ist,*	*wir sind,*	*ihr seid,*	*sie sind*
werden:	*du wirst,*	er *wird*				
tun:	*wir tun,*	*ihr tut,*	*sie tun*			

Desgleichen gilt dies für die Präteritopräsentia *(dürfen, können, mögen, müssen, sollen, wissen, wollen)* in den folgenden Formen im Singular:

Bei der 1. Person muß der Ablaut, der bei den übrigen starken Verben nur für die 2. und 3. Person zutrifft, eingesetzt werden:

ich darf, ich kann,...

Die 1. und 3. Person der Präteritopräsentia sind außerdem endungslos:

ich mag, er *darf, sie kann*

Schließlich ist in Abhängigkeit vom Ergebnis der vorausgegangenen Analyse zu berücksichtigen, welcher der vom Algorithmus generierten Formen der Vorzug zu geben ist, da hin und wieder konkurrierende Formen (mit oder ohne Elision, stark oder schwach) auftreten. Danach richtet sich, welche an erster und welche an zweiter Stelle zu präsentieren sind.

2.2.3 Formen des Indikativ Präteritum

Um die Formen des Indikativ Präteritum zu generieren, wird wiederum vom Infinitiv ausgegangen, der während der vorausgegangenen Analyse bereits von Endung und ggf. abtrennbaren Konstituenten befreit worden ist.

Entsprechend der Person sind die Endungen anzufügen:

Endungen im Indikativ Präteritum (starke Verben)		Endungen im Indikativ Präteritum (schwache Verben)	
ich...-	*wir ...-en*	*ich ...-te*	*wir ...-ten*
du ...-st	*ihr ...-t*	*du .. .-test*	*ihr... -tet*
er, Sie, es ...-	*Sie ...-en*	*er, sie, es .. .-te*	*sie ...-ten*

Nachdem die konjugierten Formen gebildet worden sind, müssen die abtrennbaren Konstituenten nach einem Leerzeichen rechts wieder hinzugefügt werden. Auch im Präteritum sind einige Besonderheiten zu berücksichtigen.

Bei schwachen Verben ist in allen Personen der Einschub eines *e* erforderlich, wenn die Basis auf *d* oder *t* auslautet *(reden redetest, redete).* Endet die Basis auf m oder *n,* wird auch ein e eingeschoben, wenn davor ein Konsonant steht (nicht aber *1* oder r: vgl. 2.2.2.2.):

trocknen:	*ich trocknete, du trocknetest, er trocknete, wir trockneten, ihr trocknetet, sie trockneten,*
wappnen:	*du wappnetest, er wappnete,..*
rechnen:	*du rechnetest, er rechnete*
filmen:	*du filmtest* nicht: *du filmetest*
lernen:	*du lerntest* nicht: *du lernetest*

In der 1. und 3. Person Plural verliert bei auf *ie* endenden Basen die Endung *-en* fakultativ das *e:*

wir schrie(e)nJ sie knie(e)n

Auch bei den starken Verben kann der Einschub eines *e* zwischen Basis und Endung erforderlich werden, im Indikativ Präteritum allerdings nur in der 2. Person sowohl im Singular als auch im Plural. Die Bedingungen dafür sind die gleichen wie bei schwachen Verben, jedoch wird außerdem auch dann ein *e* in der 2. Person Singular eingeschoben, wenn die Basis auf *8J ß* oder *z* ausgeht *(x* kommt im Basisauslaut starker Verben nicht vor): *du lasest, du maßest, du schmolzest*

Bei den starken Verben muß der entsprechende Ablaut eingesetzt werden. Das funktioniert analog zu dem Verfahren, das bereits für Präsens Indikativ beschrieben wurde. Die jeweiligen Anpassungen der Basis sind entsprechend 2.2.1. auszuführen:

saufen - er soff J greifen - er griff,
triefen - es troff, essen - sie aß, gießen - wir gossen,
kommen - er kam

Einige weitere unregelmäßige Veränderungen, die den Eintragungen der Liste starker Verben entnommen werden, sind zu berücksichtigen: z. B. *stehen – stand, gehen – ging, schneiden – schnitt, leiden - litt, denken, dachte, sein - war.*

Starke Verben mit Ablaut im Präteritum, deren Basen im Präteritum auf *d* oder *t* enden, zeigen in der 2. Person Singular gelegentlich entweder fakultatives oder obligatorisches *e* zwischen Basis und Endung [Drosdowski 1991].

wandest, tat(e)st, ritt (e)st, fochtest, flochtest, sottest tat(e)stJ bat(e)st

aber: *fandst, tratst, littst, schnittst, miedst*

Einige Verben können sowohl schwach als auch stark konjugiert werden, entweder mit oder ohne Bedeutungsunterschied. In diesen Fällen muß der Algorithmus beide Formen bilden. Die Präfigierung eines solchen Verbs kann jedoch bewirken, daß die eine oder die andere Form unzulässig wird:

schaffen:	*schuf* oder *schaffte*
erschaffen:	nur stark: *er erschuf'* nur
anschaffen:	schwach: *er schaffte an*

Die Liste der starken Verben trägt dieser Erkenntnis Rechnung und enthält für zusammengesetzte Verben mit dieser Eigenschaft zusätzliche Einträge. Um die Anzahl dieser zusätzlich benötigten Einträge gering zu halten, wurden verallgemeinerte Vergleichsmuster in die Liste aufgenommen, die es gestatten, ganze Klassen von Zusammensetzungen mit einer Basis in einem einzigen Eintrag zu behandeln.

Für alle diejenigen starken Verben, die auf verschiedene Weise ablauten können, sind die verschiedenen Ablautvarianten in der Liste der starken Verben vermerkt, so daß alle alternativen Varianten vom Algorithmus gebildet werden:

dreschen: er drasch oder *drosch*

2.2.4 Formen des Konjunktivs I

Der prinzipielle Ablauf der Synthese der Konjunktivformen gleicht dem der Synthese der Indikativformen.

Endungen im Konjunktiv I	
ich ...-e	*wir .. .-en*
du .. .-est	*ihr ...-et*
er, sie, es .. .-e	*sie.. .-en*

Generell könnten nun dem Basismorphem die einzelnen Personalendungen angefügt werden, Alle Formen des Konjunktiv I werden ohne Ablaut gebildet. Insgesamt müssen jedoch einige Besonderheiten berücksichtigt werden:

Auch im Konjunktiv sind für Verben auf *-ern* und *-eIn* ausfallende *e* zu berücksichtigen. Die Literatur erweist sich als unergiebig, gesicherte Erkenntnisse zu vermitteln, welche der vorliegenden Formen

du wandelest, du wandlest oder *du wandelst*

als standardsprachliche Konjunktive gelten.

Eine Ausnahme macht *sein* im Singular: *ich sei, du sei(e)st, er sez.*

Im Gegensatz zu den Formen im Indikativ darf weder vom Basisauslaut noch von der Endung ein *e* (wie bei *ich kniee)* entfernt werden.

2.2.5 Formen des Konjunktivs 11

Der Ablauf der Synthese konjugierter Formen im Konjunktiv II gleicht prinzipiell der Synthese der Indikativformen.

Endungen im Konjunktiv II (schwache Verben)		Endungen im Konjunktiv II (starke Verben)	
ich ...-te	wir ...-ten	ich ...-e	wir ...-en
du ...-test	ihr ...-tet	du . ..-est	ihr.. . -et
er, sie, es ...-te	sie ...-ten	er, sie" es ...-e	sie ...-en

Dem für die Bildung des Konjunktivs II gewonnenen Basismorphem werden zu diesem Zeitpunkt die für den Konjunktiv II geltenden Personalendungen angefügt. Danach sind die im Vorfeld abgetrennten Konstituenten rechts nach einem Leerzeichen wieder zu ergänzen.

Wie bei der Synthese des Indikativ bereits beschrieben, kann vor der Endung das Einschieben eines *e* (in allen Personen) nach den gleichen Regeln (vgl. 2.2.2.2.) erforderlich sein.

Die Ablaute der starken Verben sind der Liste der starken Verben zu entnehmen. Ggf. muß die Basis entsprechend den unter 2.2.1. angegebenen Regeln angepaßt werden *(kommen: er käme, saufen: er söffe, sprießen: es sprösse* usw.). Ebenso wie beim Indikativ Präteritum können in den Formen des Konjunktiv II alternative Ablaute auftreten (*er schwämme* oder *er schwömme)* und ebenso, daß starke und schwache Form miteinander konkurrieren *(glimmen: es glömme* oder *es glimmte).*

2.2.6. Formen des Imperativs

Verben, die nur im Plural auftreten (z. B. *auseinanderlaufen),* bilden keinen Imperativ Singular und unpersönliche Verben, die nur in der 3. Person gebräuchlich sind (z. B. *es regnet),* bilden keinen Imperativ.

Eine der Hauptschwierigkeiten bei der Bildung des Imperativ Singular zeigt sich im Auffinden von Regeln, die entscheiden helfen, ob ein Endungs-e unzulässig, fakultativ oder obligatorisch ist. Der Algorithmus benutzt folgende Regeln:

Verben, die im Indikativ Präsens in der 2. und 3. Person Singular von e auf i oder ie ablauten, lauten auch im Imperativ Singular ab und haben in dieser Form kein Endungs-e (z. B. *treffen:* er *trifft, triff!, befehlen:* er *befiehlt, befiehl!).*

Die ss-ß-Regelung (vgl. 2.2.1.) ist auch hier anzuwenden *fressen* - *friß!.*

Ausnahmen von dieser Gruppe sind *werden:* er *wird, werde!* und *bersten:* *es birst,* *berste!* sowie die spezielle Form *siehe* in Wendungen wie *siehe da!* oder *siehe* *Seite ...,* *siehe oben* etc.

Gesondert behandelt wird ebenfalls der Imperativ von *sein,* der stets ohne e zu bilden ist: *sei!.*

Verben auf *-eln* verlieren das e des Infinitivs und erhalten ein Endungs-e: *wandeln* *wandle!*

Verben auf *-ern* können sowohl mit oder ohne e den Imperativ bilden: *wandern* *wand(e)re!*

Gesicherte Erkenntnisse darüber, ob *wander,* oder *wander'* !, *wandel!* oder *wandel'!* in der Standardsprache akzeptabel sind, konnte die Literatur nicht nachweisen, in der Umgangssprache jedoch treten sie offensichtlich nicht selten auf.

Die Verben *dürfen, können, sollen* und *mögen* haben keinen (sinnvollen) Imperativ.

Alle verbleibenden Verben bilden den Imperativ Singular mit der Endung *-e,* basierend auf der in der Analyse gewonnenen Form der Infinitivbasis (ohne abtrennbare Konstituenten). Zusätzlich ist die Imperativform ohne die Endung -e erlaubt, wenn

> das Grundverb nicht zu den mit *-igen* gebildeten Verben gehört
beschönigen: nur *beschönige!* aber: *geig(e)!*

> das Grundverb nicht *müssen, wollen,* oder *wissen* ist.

> die Basis des Verbs nicht auf *d* oder *t* endet *(rede! schneide!)* t> der

Imperativ ohne das Endungs- e nicht schwer sprechbar sind.

Dieses "schwer sprechbar" bedarf einer genaueren Betrachtung, da der Computer nicht ohne weiteres in der Lage ist, dies zu verifizieren. Ob ein Imperativ ohne e als schwer sprechbar gilt, wird ermittelt, indem der Algorithmus die Coda der Basis in eine Folge verallgemeinerter phonetischer Transkriptionszeichen umwandelt. Der Vergleich mit den Eintragungen einer Liste, die die "gut sprechbaren" Folgen (etwa 80 verallgemeinerte Konsonantenmuster) enthält, entscheidet, ob die vorliegende Folge als "schwer sprechbar" einzustufen ist.

Bildet ein Verb den Imperativ auf e und die Basis endet mit *ss,* dann muß eine Anpassung von *ss* zu *ß* erfolgen: *faß!*

Die Form des Imperativs Plural ist mit der 2. Person Indikativ Präsens im Plural identisch und erfordert deshalb keine zusätzlichen Betrachtungen.

Abgetrennte Konstituenten werden sowohl beim Imperativ Singular als auch beim Imperativ Plural rechts nach einem Leerzeichen (einer Spaties) angefügt: *aufpassen: paß auf! paßt auf!*

2.2.7 Partizipien

Das Partizip I wird aus dem Infinitiv abgeleitet:

1. *-en* bzw. *-n* wird von rechts abgetrennt
2. bei Verben auf *-eIn* sowie *-ern* wird *-nd* angefügt, bei allen anderen *-end*

Das Partizip II wird aus folgenden Teilen zusammengesetzt:

> abtrennbare Konstituente (falls existent),
> *ge-*, wenn nötig (s. 2.1.5.)
> dem von abtrennbaren Konstituenten befreiten Infinitiv und
> einer Endung

Die Endung des Partizips II schwacher Verben ist *-t* und es muß nach den oben beschriebenen Regeln ein vorangestelltes *e* ergänzt werden. Starke Verben hingegen enden im Partizip II in der Regel auf *-en* **,** mit Ausnahme derjenigen Fälle, für die die Liste der starken Verben andere Endungen bereitstellt. Im Partizip II müssen die Regeln des Ablautes ebenso befolgt werden, wie dies bei der Generierung der konjugierten Formen nötig ist. Das gilt sowohl für den Wechsel von ss mit *ß* als auch für die Verdoppelung oder Vereinzelung von Konsonanten im finalen Bereich des Basismorphems. Bei auf *ie* endenden Partizipstämmen starker Verben erweist sich das *e* in der Endung *-en* als fakultativ: *schreien - geschrie(e)n.* Falls konkurrierende Formen des Partizip II auftreten, wird vom Algorithmus sowohl die starke als auch die schwache Form generiert. Bei *werden* sind in Abhängigkeit vom Vorhandensein eines weiteren Partizips II im Satz *geworden* bzw. *worden* möglich.

"Die Verben *brauchen* (im Sinne von: *müssen), dürfen, können, lassen, mögen, müssen, sollen, wollen* und *heißen, hören, sehen,* beschränkt auch *helfen* und *machen,* stehen im Perfekt und Plusquamperfekt als Infinitiv statt als Partizip II, wenn sie mit einem anderen Infinitiv zusammentreffen:" [Baer 87, S. 653]

Das hättest du nicht zu tun brauchen (nicht: *gebraucht).*

Er hat nicht kommen können / dürfen / sollen.

Ich hatte dich kommen sehen.

Unter syntaktisch anderen Bedingungen bilden diese Verben durchaus das Partizip II. Der Verbformengenerator trägt diesem Sachverhalt Rechnung und bildet in Abhängigkeit vom Kontext das übliche Partizip II oder den in dieser Funktion stehenden Infinitiv.

2.2.8 Infinitiv mit "zu"

Der Infinitiv mit *zu* ist relativ problemlos zu bilden, da der Algorithmus auf die Prozedur zur Generierung des Partizips II zurückgreifen kann. In denjenigen Fällen jedoch, in denen abtrennbare Präfixe oder Wortkonstituenten vorliegen oder Einfügestellen für *ge* beim Partizip II bzw. *zu* beim Infinitiv II vorliegen, muß das *zu* an der entsprechenden Stelle in den Infinitiv I eingesetzt werden, sonst ist es voranzustellen und mit einem Leerzeichen vom Infinitiv zu trennen:

abzulegen, kopfzustehen, notzutaufen, zu *rennen,* zu *bekommen*

Einige Verben zeichnen sich dadurch aus, daß die Fügestelle für *zu* nicht gleichzeitig die Fügestelle für *ge* ist. Zu dieser Gruppe gehören Verben, die mit einem der Präfixe mit schwankender Akzentuierung gebildet sind und sich zusätzlich mit einem nicht abtrennbaren Präfix verbinden:

überbeanspruchen: du überbeanspruchst, überbeansprucht, überzu beanspruchen, ebenso *überbelasten, überbelichten, überbewerten* etc.

Ferner sind Wörter zu beachten, bei denen durch Zusammensetzung einer von drei aufeinandertreffenden Konsonanten vor einem Vokal im Schriftbild nicht mehr vorhanden ist und nun durch das Einfügen von *zu* bzw. *ge* wieder geschrieben werden muß:

vollaufen - vollzulaufen.

3 Reflexive Verben

Das Reflexivpronomen reflexiver Verben ist in Abhängigkeit von der jeweiligen Person zu flektieren. Wenn ein Reflexivum in der Infinitivphrase enthalten ist, dann wird es zunächst abgetrennt, damit die Verbformen generiert werden können und erst dann, wenn dieser Prozess abgeschlossen ist, fügt der Generator das für die jeweilige Person gültige Reflexivpronomen zu. Als problematisch erweist sich die vom Algorithmus zu treffende Unterscheidung zwischen Akkusativreflexivität und Dativreflexivität, die erst mit Hilfe einer Liste aller reflexiven Verben oder bei Verwendung eines entsprechenden maschinenlesbaren Wörterbuchs gelöst werden könnte. In der gegenwärtigen Version stellt der Algorithmus für diese Differenzierung eine Lösungsmöglichkeit zur Verfügung, die eine noch etwas unbefriedigende Regelung darstellt:

Akkusativreflexivität wird angenommen, wenn *sich* vor dem Infinitiv angegeben ist;

sich wundern:	*ich wundere mich, du wunderst dich, sie wundert sich, wir wundern uns, ihr wundert euch, sie wundern sich*
sich aufregen:	*ich rege mich auf, du regst dich auf, er regt sich auf wir regen uns auf, ihr regt euch auf, sie regen sich auf*

Um Dativreflexivität zu kennzeichnen, ist vor dem Infinitiv sich3 einzugeben:

sich3 etwas vorstellen:	*ich stelle mir etwas vor} du stellst dir etwas vor, sie stellt sich etwas vor,* *wir stellen uns etwas vor, ihr stellt euch etwas vor, sie stellen sich etwas vor*

Das in *sich3 etwas vorstellen* enthaltene *etwas* stellt einen Vorgriff auf die im nächsten Abschnitt zu diskutierende Behandlung phraseologischer Einheiten dar.

Die Angabe von *sich* und *sich3* vor dem Infinitiv kann auch in reflexiven Konstruktionen benutzt werden, in denen das Reflexivpronomen nicht obligatorisch steht, sondern durch ein nicht mit dem Subjekt identisches Objekt ersetzt werden kann.

4 Konjugation von Infinitivphrasen

Der Algorithmus ist in der Lage, Infinitivphrasen zu konjugieren. Das zugrundeliegende Modell der Infinitivphrase ist

[sich/sich3 Leerzeichen] [Satzglieder Leerzeichen] Verbinfinitiv

Zu lesen ist dies wie folgt:

> Eckige Klammern geben an, daß die innerhalb stehenden Angaben fakultativ sind.

> Der Senkrechtstrich trennt alternative Angaben.

> Angaben in kursiver Fettschrift sind so wie angegeben zu verwenden,

> Wörter in der Schriftart Arial zeigen an, daß entsprechende Zeichen, Wörter oder Wortsequenzen substituiert werden können.

In der Synthese werden durch eine entsprechende Transformation alle erforderlichen Teile wieder aneinandergereiht. Für die finiten Formen geschieht dies beispielsweise nach folgendem Muster:
[Personalpron] finite_Form [flektiertes...Reflexivpron] [Satzglieder]
[Abgetrenntes]

Die Generierung von finite_Form und Abgetrenntes wurde unter 2.2. bereits beschrieben. Die Formen der Personalpronomina und die der flektierten Reflexivpronomina werden programminternen Tabellen entnommen. Der Teil Satzglieder wird unverändert aus der Eingabe übernommen. Da das der Konjugation unterliegende Verb in einer Infinitivphrase stets ungetrennt am Ende steht, braucht der Algorithmus von rechts her lediglich das davorstehende Leerzeichen zu suchen, um den Infinitiv aus der Infinitivphrase extrahieren zu können. Reflexivpronomen andererseits treten nur am Anfang einer Infinitivphrase auf, so daß das Leerzeichen rechts neben dem Pronomen das Abspalten von links her erlaubt. Eventuell dazwischenstehende Satzglieder (Objekte oder Adverbialbestimmungen) werden im allgemeinen durch die Konjugation weder verändert noch umgeordnet, so daß auf eine detaillierte Strukturanalyse verzichtet werden kann.

Beispiele:

Eingabe	eine der generierten Phrasen
sich an etwas erinnern: jemandem	*ich erinnere mich an etwas,...*
einen Gefallen tun: sich im *Urlaub*	*er tat jemandem einen Gefallen ihr ließt*
verwöhnen lassen:	*euch* im *Urlaub verwöhnen*

5 Bewertung und Einsatz

Zunächst einige kritische Bemerkungen: Da der Algorithmus über kein vollständiges morphologisches und semantisch konnotiertes Wörterbuch verfügt, können fehlerhafte Eingaben nur in beschränkten Maße als solche erkannt werden, was zur Folge hat, daß falsch eingegebene Infinitive bzw. Infinitivphrasen zu falschen Formen führen. Die gegenwärtige Version enthält keine Rechtschreibkontrolle, weil auch dazu ein Wörterbuch nötig wäre. Aus dem gleichen Grund können Akkusativreflexivität und Dativreflexivität nur über den Umweg der Eingabemarkierung differenziert werden.

Eine bestimmte Gruppe von Verben bildet nur den unpersönlichen Teil des Formenumfangs wie zum Beispiel die Verben: *regnen: es regnet, es regnete, es regne, geregnet.* Zu beobachten ist ferner, daß zahlreiche Verben aus dem technischen Fachwortschatz nur im Infinitiv oder als Partizipien gebräuchlich sind: *vakuumbedampfen, vakuumbedampft.* Bis auf die im Text genannten Ausnahmen präsentiert der Anzeigealgorithmus sonst stets den gesamten Formenumfang, ungeachtet ihrer Verwendbarkeit.

Zusätzliche Angaben innerhalb der Eingabe oder beim Aufruf des Algorithmus erlauben, den Umfang der zu präsentierenden Formen zu bestimmen, sowie Reflexivität und deren Kasus vorzugeben. Dies setzt im ersten Fall Kenntnisse beim Benutzer voraus. Im 2. Fall können Mechanismen in den Programmen, die den Verbformgenerator benutzen, diese Aufgabe übernehmen, sofern diese Programme ihrerseits über geeignete Mittel verfügen, die Einschränkungen und Besonderheiten zu erkennen, wozu umfangreiche Wörterbücher mit detaillierten Angaben zur Orthographie, Morphosyntax, Valenz und zur semantischen Konnotation erforderlich sind.

Dem Algorithmus könnten durchaus Wörterbücher zur Rechtschreibkontrolle vorgelagert werden, jedoch schränkte dies die Möglichkeiten ein, Neologismen oder ad-hoc-Bildungen sofort mit ihren Flexionsformen präsentieren zu können, ohne daß Nachfragen in Hinblick auf die Korrektheit der Schreibung beantwortet werden müssen.

Von Verben mit homographen Infinitiven und verschiedenen Bedeutungen *(schaffen: schuf/schaffte, schleifen: geschliffen/geschleift* usw.) werden alle alternativen Formen generiert und gezeigt. Es erfolgt ein Hinweis auf den Bedeutungsunterschied. Bei zusammengesetzten Verben, die solche homographen Basen enthalten, werden nur dann konkurrierende Formen generiert, wenn beim vorliegenden Kompositum beide Formen denkbar sind.

Der Verbformengenerator bildet deutsche Verbformen weitaus seltener fehlerhaft als ein durchschnittlich gebildeter deutscher Muttersprachler, wie die internen Testergebnisse zeigen. Diese Bewertung ist so vorsichtig gewählt, weil es außerordentlich schwierig, ist (wenn nicht gar unmöglich), die Aussage "Der Algorithmus macht keine Fehler" zu verifizieren. Die Bemühungen bei der Erarbeitung des vorliegenden Algorithmus waren jedoch durchaus von diesem Anspruch geprägt, dennoch wird die Praxis zeigen, daß sich gelegentlich einzelne, meist wenig gebräuchliche zusammengesetzte Verben finden, die Ergänzungen in den Ausnahmelisten erforderlich machen werden. Bei widersprüchlichen oder unterschiedlich interpretierbaren Aussagen in der Literatur bezüglich der Formenbildung wurde der Dudenausgabe [Drosdowski 1991] der Vorrang gegeben, sofern dieses Werk eine Entscheidung zuließ.

Von nicht zu unterschätzender Bedeutung erweist sich, daß vom Verbformengenerator die Formen der starken Verben generell richtig gebildet werden, was vom durchschnittlichen Muttersprachler nicht behauptet werden kann; es sei auf die Konjunktivformen oder an die Formen dieser Verben im Präteritum verwiesen.

Für Ausländer erweist sich die Nutzung des Programms ENDUNG, das diesen Algorithmus enthält, als eine große Hilfe beim Erwerb des Wissens über die Bildung der deutschen Verbformen, zumal im Gegensatz zu Lehr- und Nachschlagewerken nicht nur einzelne typische, sondern alle Formen präsentiert werden. Wie wichtig das ist, erkennt man, wenn man die zahlreichen Anpassungen der Basis oder die möglichen, obligatorischen oder verbotenen Elisionen betrachtet, ganz abgesehen davon, daß das Eintippen eines Infinitivs am Computer wesentlich schneller geht, als ihn in einem Nachschlagewerk aufzufinden.

Die Wortstellung im Deutschen ist für Ausländer in der Regel problematisch. Durch die Möglichkeit der Konjugation ganzer Phrasen wird dem Lernenden ein Werkzeug in die Hand gegeben, das wegen seiner Schnelligkeit und Vollständigkeit jedes gedruckte Lehrbuch oder Nachschlagewerk in den Schatten stellt.

Ständig werden auf den verschiedensten Fachgebieten, im Journalismus oder in der Alltagssprache neue Wörter gebildet, die in den Medien zwar schnell wiederzufinden sind, in für Ausländer gemachten Lehrwerken jedoch erst nach Jahren erscheinen. Der Verbformengenerator hingegen präsentiert die Formen der Neubildungen in der überwiegenden Anzahl der Fälle, ohne daß der Algorithmus oder die Ausnahmelisten geändert werden müßten.

Wie universal der Algorithmus arbeitet, zeigt sich im Umgang mit neuen Wortbildungen, die auf mannigfaltige Weise zustande kommen können. Zu nennen wären für die Bildung von Verben zumindest

> die Derivation von deutschen Wörtern sowie von Lehn- oder Fremdwörtern anderer Wortarten durch Anhängen der Endungen (für die Bildung des Infinitivs *-en* oder *n),* häufig in Verbindung mit Suffixen wie bei *-ieren, -lichen, -ern, -eIn* und schon selten *-igen. (faxen, grillen,* compilieren, palatalisieren, lenisieren, emulieren, verstimmlichen, spinnakern, umrubeln)

> die im Deutschen außerordentlich produktive Möglichkeit der Präfigierung mit abtrennbaren oder nichtabtrennbaren Präfixen oder ganzen Präfixreihen. Es fällt deshalb so schwer, Beispiele dafür zu nennen, weil die meisten so gebildeten Wörter völlig "gängige" Wörter sind. (umeinanderwickeln, rüberfaxen, verzurren)

> der Import aus anderen Sprachen, heute oft aus dem Englischen, wobei diese Wörter nach den morphologischen Regeln des Deutschen behandelt werden *(managen,* canceln, outen, *toasten,* mailen, einscannen, *abchecken, sich einchecken)*

> die Bildung von Onomatopoetika (schrapseln, brabbern)

Die fettgedruckten Verben - Jargon oder nicht - sind in der angegebenen Form im aktuellsten Duden [Drosdowski 1991] nicht enthalten. Dies unterstreicht die Vorzüge einer

algorithmischen Lösung, die problemlos zu jedem dieser Verben sämtliche konjugierten Formen liefert, obwohl keines dieser Verben in einer der vom Algorithmus verwendeten Listen enthalten ist.

Die klare Struktur des Algorithmus und die Einfachheit der verwendeten Ausnahmelisten macht es zu jeder Zeit möglich, aktuelle Erkenntnisse einfließen zu lassen, z. B. wenn doch noch ein Verb gefunden wird, das durch die Maschen geschlüpft war.

Der Algorithmus arbeitet außerordentlich schnell- die benötigte Zeit für die Auflistung der Formen auf dem Bildschirm (wenige Millisekunden) ist bei weitem größer als die Zeit zur Analyse und Synthese der Formen. Die hohe Geschwindigkeit wird durch ausgefeilte Teilalgorithmen, insbesondere bei der Erkennung der Ausnahmen, erreicht. Der Algorithmus kommt gänzlich ohne Backtracking aus, was sicher der Hauptgrund für seine hohe Geschwindigkeit ist. Auch die Wahl der Sprache Borland-Pascal 7.0 beeinflußte die Geschwindigkeit positiv, zumal die Verwendung von Inline-Assembler-Sequenzen in dieser Sprache auf besonders günstige Weise unterstützt wird.

Neben den außerordentlichen Vorteilen für den Lernenden, ein derart flexibel gestaltetes Nachschlagewerk zur Verfügung zu haben, fiel bei der Erstellung des Algorithmus eine tiefgehende Prüfung der in Grammatiklehrwerken niedergelegten Regeln nebenher mit ab bzw. war notwendige Voraussetzung, den vorliegenden Algorithmus überhaupt erstellen
zu können. An zahlreichen Stellen des Programmquelltextes finden sich Bemerkungen, wie außerordentlich schwierig es war, Hinweise oder gar Regeln in der Fachliteratur zu finden. Einige Fragen, insbesondere die potentielle Elision oder Apokopierung (hauptsächlich in Konjunktiv- bzw. Imperativformen), blieben gänzlich offen und konnten bis dato nicht befriedigend belegt werden. Die Anwendung linguistischen Wissens am Computer fördert schonungslos zutage, wo die aufgestellten Regeln tatsächlich Entscheidungen zulassen und wo deren Schärfe zu wünschen übrigläßt, ganz abgesehen davon, daß das Überprüfen der Prämissen semantisch motivierter Regeln bei der Umsetzung mit dem Computer immer noch große Schwierigkeiten bereitet. Bemerkenswert scheint zu guter Letzt, das mit dem `Verbformengenerator` der Beweis angetreten worden ist, daß Regeln selbst mit Prämissen wie "wenn es schwer sprechbar ist", durchaus formalisierbar sind, wenn der Betrachter tief genug in den Gegenstand und seine Nachbarwissenschaften (hier die Phonetik) eingedrungen ist und eine computergerechte Darstellungsform dafür gesucht hat.

Für die Computerlinguistik erwies sich einmal mehr, daß eine auf fundiertem linguistischen Wissen basierende morphologische Analyse unerläßlich ist und den Schlüssel zum Erfolg liefern kann und daß andererseits gewisse Unzulänglichkeiten bleiben, solange der Computer nicht "versteht", was inhaltlich gemeint ist. Man denke nur an die Präsentation verschiedener möglicher Flexionsformen zu homographen Infinitiven mit Bedeutungsdifferenzierungen, wo im gegebenen Kontext nur eine der alternativen Formen korrekt ist, der vorgestellte Algorithmus jedoch nicht in der Lage ist, eine gültige Entscheidung zu treffen.

Verwendet wird der vorgestellte Algorithmus in dem mit dem Deutsch-Österreichischen Hochschul-Softwarepreis ausgezeichneten Programm ENDUNG, das Aussprache und Grammatik des Deutschen für Ausländer, Übersiedler und Dialektsprecher vermittelt. Er dient sowohl als Nachschlagewerk, indem alle Formen zu einem vom Nutzer vorgegebenen

Infinitiv präsentiert werden als auch zur Kontrolle der Nutzerantworten bei Übungen, so

daß der Autor der Übungsaufgaben von der zeitraubenden Arbeit des Zusammenstellens der richtigen Lösungen befreit ist.

Literatur

Baer, D. et. al. (Hrsg.) (1987): Der Große Duden. Rechtschreibung Leipzig: VEB Bilbliographisches Institut

Drosdowski, G. et.al. (Hrsg.) (1991): Duden. Die deutsche Rechtschreibung Mannheim, Leipzig, Wien, Zürich: Dudenverlag

Dückert, J. u. Kempke, G. (Hrsg) (1984): Wörterbuch der Sprachschwierigkeiten. Bibliographisches Institut Leipzig

Helbig, G. , Buscha, J. (1991): Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht Leipzig, Berlin, München: Verlag Enzyklopädie Langenscheidt

Helbig, G., Buscha, J. (1989): Deutsche Übungsgrammatik. Verlag Enzyklopädie, 5. unveränderte Auflage

Jung, W. (1980): Grammatik der deutschen Sprache Leipzig: VEB Bibliographisches Institut

Muthmann, G. (1991): Rückläufiges deutsches Wörterbuch Tübingen: Max Niemeyer Verlag

Rausch, R., (1993): Die graphische Silbentrennung und ihre Anwendung auf die Aussprache von Vokalen und Konsonanten im Deutschen. - Sprechen: Z.f. Sprechwiss., H:2 / 91

Rausch, R., Rausch, I. (1992): Deutsche Phonetik für Ausländer Leipzig, München, Basel, New York: Verlag Enzyklopädie Langenscheidt

Rausch, R., Rothe, H. (1993): Das Aussprachetrainingsprogramm „ENDUNG" in Paechter, M. (Hrsg): Tagungsband des 6. Expertentreffens des Arbeitskreises „Pädagogische Software mit digitaler Sprachverarbeitung" TU Braunschweig

Rausch, R., Rothe, H. (1994): Bemerkungen zum Autorenprogramm „ENDUNG" in Gadler (Hrsg.): Sprache-Therapie-Computer, Tagungsband des 1. Internationalen Kongresses Graz, Karl-Franzens-Universität

HANGUL, THE KOREAN WRITING SYSTEM, AND ITS COMPUTATIONAL TREATMENT

Kiyong Lee Professor of Linguistics
Korea University and DAAD Visiting Scholar
Abteilung Computerlinguistik
Universität Erlangen-Nürnberg
klee@ling.korea.ac.kr klee@linguistik.uni-erlangen.de *

Hangul, the writing system of Korean, was invented in the mid-15th century. In one respect, Hangul may be viewed as a syllabic system like the Japanese Kana; in another as an alphabetic system like that of most Western languages. Because of these dual characteristics it is a non-trivial task to represent Hangul on the computer.

For most applications, Hangul characters may be viewed holistically as syllable units by default. For purposes of automatic text processing and word-form recognition, however, each syllable should be analyzed as consisting of consonants and vowels. The question is how these two different ways of analyzing Hangul characters should best be coded.

In this paper I will describe how the efficiency of the Hangul writing system can be transferred to its coding on computers. I will also review the encodings of Hangul by the precomposed Hangul code set of KSC 5601, a combinational Hangul code set, and ISO 2022, and their applications to the development of Hangul pro grams for sorting, editing, and processing Korean text. I then conclude with remarks on the merger efforts of the Unicode standard and ISO DIS 10646 on Hangul.

1 Introduction

Written language used to exist only in the traditional media of engraved stone, or handwritten or printed paper. Today it exists increasingly in the electronic medium on the computer.

Representing written language on the computer goes beyond creating the letters or symbols on the screen. In addition, the coding of written language should allow its electronic processing, consisting of the editing, formatting, searching, ordering, etc. of text.

For Hangul, several different coding systems have been proposed in recent years, each with its own merits and drawbacks. In this paper I will discuss which properties should be coded for the purpose of linguistic analysis, specifically automatic word-form recognition.

In Section 2, the origin of Hangul is described. Sections 3-6 explains the letters, syllable structures, and syllable representation as well as the key-in and display of Hangul. Section 7 presents possible ways of encoding Hangul. Section 8 shows how the encoded Hangul can be applied to some computing tasks like sorting, editing, and word-form recognition or generation in an efficient way. Section 9 discusses Unicode, a newly emerging venture to promote a universal encoding scheme for ail worldwide characters in a uniform way. Section 10 concludes the paper.

*This work was partially supported by Korea Research Foundation and Deutscher Akademischer Austauschdienst. I am grateful to these organizations for their financial support and to the sponsoring organization, Universität Erlangen-Nürnberg, and the director of the Program for Computational Linguistics, Prof. Roland Hausser. I am also grateful to Korea University, which granted me a six-month sabbaticalleave for research during the academic year of 1994.

2 Origin of Hangul

Different cultures have produced various types of writing systems. Some are ideographic like the Chinese Hanzi, some are syllabic like the Japanese Kana, and some are alphabetic like Latin. Hangul shares properties of all three types of writing systems.

In the mid-15th century, King Sejong[1] of the Yi Dynasty recognized that the Chinese writing system was too complicated and too time consuming to learn by ordinary people. Because Chinese writing had been the only one available in Korea for over a millenium, the king initiated work on a new writing system for Korean suitable for daily use. Through some long years of hard work and collaboration, his entourage of brilliant scholars succeeded in inventing a truly remarkable system of writing, originally called Hunminjŏngŭm the "Correct Sounds for Teaching People" but now called Hangul, the "Great Writing".[2]

Despite the royal promulgation of Hunminjŏngŭm in 1446, its use was restricted. Only at the founding of the Republic after World War II, Hangul was accepted as the official writing system of Korean. In Korean writing of legal documents or scholarly works, however, Chinese characters may still be found frequently. Even daily newspapers printed in Hangul are heavily mixed with Chinese characters. The trend is, however, rapidly changing towards the pure use of Hangul in everyday life and business in Korea.

In its appearance, Hangul syllable units resemble Chinese or Japanese Kana characters.[3] These writing systems represent syllable structures. For example, the word 한글 'Hangul' consists of two syllables, 한 'Han' and 글 'gul'. Hangul differs from Chinese and Japanese syllables, however, in that Hangul syllable units, say these two syallables 한 'Han' and 글 'gul', may be decomposed into phonemic units ㅎ 'H', ㅏ 'a', ㄴ 'n', and ㄱ 'g', ㅡ 'u', ㄹ 'l', respectively. Thus Hangul is a phonemic writing system consisting of consonant or vowel letters that represent distinct phonemes. In this respect, Hangul partially resembles the alphabetic writing system of Western languages.

3 Hangul Letters

There are three layers to the structure of Hangul characters: atomic (basic letters), molecular (e.g. double consonant letters), and syllabic (characters). There are 24 basic letters, which divide into 14 consonants and 10 vowels, constituting the atomic level of the Hangul alphabet.

The Hangul 'alphabet' begins with 14 atomic consonant letters:

Atomic Consonants:

	1	2	3	4	5	6	7	8
Hangul:	ㄱ	ㄴ	ㄷ	ㄹ	ㅁ	ㅂ	ㅅ	ㅇ
Phonemic:	k	n	t	l	m	p	s	Ø, ŋ
Romanized:	k	n	t	r, l	m	p	s	x, ng

	9	10	11	12	13	14
Hangul:	ㅈ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
Phonemic:	c	c^h	k^h	t^h	p^h	h
Romanized:	c	ch	kh	th	ph	h

[1] In romanizing Hangul letters, I follow the Yale Romanization with some modifications. However, for romanizing proper names like Sejong and Hangul or names of encoded Hangul letters like IEUNG (instead of Se.cong, Han.kul, or .I.UNG), I just follow the conventional ways of romanizing them.

[2] The term "Hangul" was introduced in 1910 by Chu, Shi-Kyŏng, the founding father of Korean modern linguistics.

[3] *Kana* collectively refers to the two orthographic styles of the Japanese syllabic writing system, Katakana and Hirakana.

There are the five aspirated consonants ㅊ /c^h/, ㅋ /k^h/, ㅌ /t^h/, ㅍ /p^h/, and ㅎ /h/. To the first four correspond the unaspirated consonants ㄱ /k/, ㄷ /t/, ㅂ /p/, and ㅈ /c/. These two classes are differentiated with an extra stroke or a slight variation of shape. For example, the aspirated consonant letter ㅊ /c^h/ has a stroke over the unaspirated ㅈ /c/. Besides these nine consonants, the basic set of consonant letters includes one liquid letter ㄹ /l/, one sibilant letter ㅅ /s/, and three nasal letters ㄴ /n/, ㅁ /m/, and ㅇ /ŋ/. The final letter ㅇ /ŋ/ called IEUNG represents a null sound syllable-initially.[4]

In addition, there are 10 atomic vowel letters, which are ordered immediately after the 14 consonant letters.

Atomic Vowel Letters:

	15	16	17	18	19	20	21	22	23	24
Hangul:	ㅏ	ㅑ	ㅓ	ㅕ	ㅗ	ㅛ	ㅜ	ㅠ	ㅡ	ㅣ
Phonemic:	a	ya	ə	yə	o	yo	u	yu	ɨ	i
Romanized:	a	ya	e	ye	o	yo	u	yu	w	i

The 10 atomic vowel letters represent both simple vowels and diphthongs. The letters ㅏ, ㅓ, ㅗ, ㅜ, ㅡ, and ㅣ represent simple vowels /a/, /ə/, /o/, /u/, /ɨ/, and /i/, respectively. The letters ㅑ, ㅕ, ㅛ, and ㅠ each represent a diphthong formed with an on-glide /y/ and one of the simple vowels /a/, /ə/, /o/, and /u/. The diphthongs are marked with an extra dot-like stroke. The diphthong ㅑ /ya/, for instance, is formed by adding an extra stroke to the vowel ㅏ /a/.[5]

On the basis of these 24 atomic letters, a set of 16 additional letters is derived. The set consists of 5 double consonant letters and 11 compound vowel letters.

Double Consonants:

	25	26	27	28	29
Hangul:	ㄲ	ㄸ	ㅃ	ㅆ	ㅉ
Phonemic:	k'	t'	p'	s'	c'
Romanized:	kk	tt	pp	ss	cc

The double consonants are formed by geminating the atomic simple obstruents ㄱ [k], ㄷ [t], ㅂ [p], ㅅ [s], and ㅈ [c]. Although these letters are not atomic, they each represent a single phoneme with the common phonetic feature *tense*, marked with an apostrophe " ' ": /k'/, /t'/, /p'/, /s'/, and /c'/, respectively.

On the basis of the 10 atomic vowel letters, 11 compound vowel letters are formed. These letters are written as orthographic compounds. They are formed by combining or recombining atomic vowels with each other.

Compound Vowels:

	30	31	32	33	34	35	36	37	38	39	40
Hangul:	ㅐ	ㅒ	ㅔ	ㅖ	ㅘ	ㅙ	ㅚ	ㅝ	ㅞ	ㅟ	ㅢ
Phone:	æ	yæ	e	ye	wa	wæ	œ	wə	we	wi	ɨi
Roman:	ay	yay	ey	yey	wa	way	oy	we	wey	uy	wy

Among the 11 compound letters, the three compound letters ㅐ, ㅖ, and ㅚ represent the simple vowels /æ/, /e/, and /œ/, respectively. In contrast, the other 8 compound letters ㅒ, ㅔ, ㅘ, ㅝ, ㅞ, ㅟ, and ㅢ represent the diphthongs /yæ/, /ye/, /wa/, /wə/, /we/, /wi/, and the sequence of two vowels /ɨi/, respectively.

[4] As in the ISO Romanization system, the initial RIEUL ㄹ is romanized into 'r', and the final ㄹ into 'l' in this paper. The initial IEUNG ㅇ is romanized into 'x', while the final is romanized into 'ng'.

[5] The Yale system is followed in romanizing vowel letters except that the rounded back vowel ㅜ is romanized into 'u' and the unrounded vowel ㅡ into 'w'.

Both the atomic and non-atomic letters form the molecular layer of Hangul characters. At the atomic layer, the basic letters are orthographically atomic, while the derived letters are non-atomic. At the molecular layer, however, they are both elementary units, from which syllable characters are formed.

The composition of Hangul letters may be analyzed from an orthographic and a phonological point of view. Orthographically, of the 40 Hangul consonant and vowel letters, 24 are treated as atomic, whereas the double consonants and compound vowels are composed of more than one atomic letter. The possibility of a single letter representing a sequence of two phonemes and of a compound letter representing a single phoneme result in the following possible types of correlating the orthographic and the phonological level.

	Type 1	Type 2	Type 3	Type 4	Type 5
Letters:	ㅏ	ㅑ	ㄱㄱ	ㅓㅣ	ㅕㅣ
	ㅣ	/\	\/	\/	/\/
Sounds:	/a/	/ya/	/k'/	/e/	/y e/

Type 1 is the most basic, representing a single sound with a single letter. This class comprises 20 atomic letters: the 14 atomic consonant letters ㄱ /k/, ㄴ /n/, ㄷ /t/, ㄹ /l/, ㅁ /m/, ㅂ /p/, ㅅ /s/, ㅇ /null, ŋ/, ㅈ /c/, ㅊ /c^h/, ㅋ /k^h/, ㅌ /t^h/, ㅍ /p^h/, ㅎ /h/, and the six atomic vowel letters ㅏ /a/, ㅓ /ə/, ㅗ /o/, ㅜ /u/, ㅡ /ɨ/, and ㅣ /i/.

Type 2 consists of the four atomic vowel letters ㅑ /ya/, ㅕ /yə/, ㅛ /yo/, and ㅠ /yu/. While being orthographic atoms, they represent phonologically a diphthong consisting of a glide /y/ and a vowel.

Type 3 is represented by the five double consonant letters: ㄲ /k'/, ㄸ /t'/, ㅃ /p'/, ㅆ /s'/, and ㅉ /c'/. Orthographically, they are composites. Phonologically, they each represent a simple phoneme with a common *tense* feature, marked with the apostrophe " ' ".

Type 4 is an orthographic composite which represents a single phoneme. Of this type are the three compound vowels: ㅐ /æ/, ㅔ /e/, and ㅚ /œ/.

Type 5 is represented by the seven compound vowel letters: ㅒ /yæ/, ㅖ /ye/, ㅘ /wa/, ㅝ /wə/, ㅞ /we/, ㅟ /wi/, and ㅢ /ɨi/. These letters are composites both orthographically and phonologically. They each consist of two vowel letters and represent a diphthong or a sequence of two vowels.

Referring to both, the orthographic and the phonological levels of Hangul, is important for morphological descriptions. For instance, there are the syllable reduction and combination phenomena in Korean:

(1) 이것이(this-Subject)	/ikəsi/	:	이게	/ike/
(2) 아이 (child)	/ai/	:	애	/æ/
(3) 먹이어(feed-Imperative)	/məkiəla/	:	먹여라	/məkyəla/
(4) 오(come)+ 았다(Suffix)	/o+as'ta/	:	왔다	/was'ta/

The first and second examples illustrate the reduction of the two syllables 것 /kəs/ and 이 /i/ into a single syllable /ke/, and the two syllables 아 /a/ and 이 /i/ into a single syllable /æ/. These reductions result from the deletion of the intervocalic /s/ or the null consonant represented by the IEUNG ㅇ and can best be described by analyzing the orthographic compositions of the vowels ㅔ /e/ and ㅐ /æ/, for these vowels are written as compounds, consisting orthographically of ㅓ and ㅣ, and ㅏ and ㅣ, respectively.

A phonological analysis of the composition of the atomic vowel letter ㅕ /yə/ is also needed to account for the syllable reduction in the third example. Here the two syllables 이어 /iə/ is reduced to a single syllable 여 /yə/. This results from the diphthongization of the high front

vowel /i/ occurring before a vowel and can best be correlated to the phonological composition of the basic vowel letter ㅖ which represents a diphthong consisting of a glide /y/ and a vowel /ə/. Its orthography could not explain why the reduction of ㅇ|어 into 여 occurs because it is written as an atom.

The final example illustrates how the verb 오(Stem come) combines with a past tense ending ㅏㅆ다(Suffix) into the finite form 왔다 /was'ta/(come-past). The vowel letter ㅘ is orthographically composed of ㅗ and ㅏ. Phonologically, however, it represents a single diphthong /wa/.

These four examples show that, for linguistic study, both the orthographic and the phonological composition of Hangul letters must be analyzed. The same holds for a computational treatment of such phenomena.

4 Syllable Structures

We turn now to the third layer of the Korean writing system, the syllable. A syllable is represented as a character; a word is written as a string of characters.[6] A syllable may consist of one of the following structures: CV, CVC, and CVCC. Thus, each syllable must begin with a consonant C followed by a vowel V. A syllable of the form CV is called an *open* syllable. A syllable of the form CVC or CVCC is called a *closed* syllable.

In modern Korean, the initial position of a syllable is occupied by exactly one of the 19 consonant letters. Even if a syllable starts phonetically with a vowel sound without a consonant, the syllable initial must be marked with the null consonant IEUNG ㅇ. For example, the syllable ㅇㅏ pronounced as [a] has the null consonant letter at its initial. Hence, all syllable characters written in Hangul begin with a consonant letter.

The vowel forms the nucleus of a syllable.[7] A syllable may be closed with one or two consonants. All of the 14 atomic consonants can occur syllable-finally. Of the five double consonants, only the two ㄲ 'kk' and ㅆ 'ss' can occur syllable-finally. Then there are 11 clusters of consonants that can occur syllable-finally as molecular units of a syllable.

Admissible Final Consonant Clusters

Hangul:	ㄳ	ㄵ	ㄶ	ㄺ	ㄻ	ㄼ
Romanized:	ks	nc	nh	lk	lm	lp
Hangul:	ㄽ	ㄾ	ㄿ ㅍ	ㅀ	ㅄ	
Romanized:	ls	lth	lph	lh	ps	

Given the restrictions on syllable structure and letters in certain syllable positions, how many syllables may be formed with the 40 letters of Hangul? For centuries, this question has been relevant to printing shops trying to set up an adquate set of character blocks for all the possible Hangul syllables. Today it also arises in the context of encoding Hangul on the computer. The answer depends on how syllables are counted.

By one count, we get 10,773 syllables. We obtain this number by calculating $19 \times 21 \times (14 + 2 + 11)$, where the intial 19 is the possible number of inital consonants, the second 21 the possible

[6] I will use the term *letter* to refer to a consonant or vowel symbol in Hangul that stands for a phoneme, but the term *character* to a syllable structure in a square block composed of consonant and vowel symbols, or letters. When used collectively, the term *letter* thus specified corresponds to the Korean term JAMO, which refers to the set of consonant(JA.UM) and vowel(MO.UM) signs. In computing, however, letters are treated as part of characters. See **character** in the Glossary.

[7] In treating the tripartite structure of Hangul syllables, the terms 'cho.seng' (initial sound), 'chwung.seng'(medial sound), and 'cong.seng' (final sound) are often used. The term 'onset', 'nucleus', and 'coda' are also used to refer to the initial, medial, and final of a syllable, respectively. They are phonological terms referring to phonological objects. Martin (1992) and S.S. Kang (1993) extended the use of these phonological terms to describe orthographic properties of Hangul.

number of vowels or diphthongs, and the final number (14+2+11) the addition of the 14 atomic consonants, the two double consonants, and the 11 consonant clusters that can be admitted as finals. By another count[8] the number of possible syllables is 11,172. This is so because the so-called FILLER position is added to the list of admissible finals: $11,172 = 19 \times 21 \times (27 + 1)$.[9]

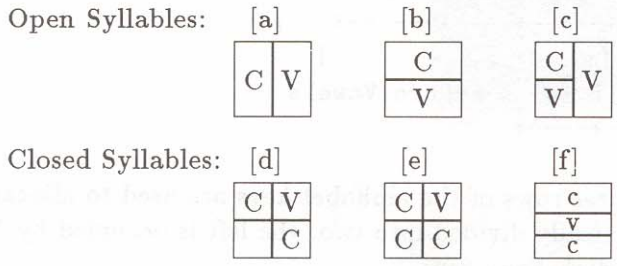
A more sophisticated way of counting reduces the number to 8,873 characters, because consonant clusters cannot serve as syllable finals for certain diphthongized vowel letters.[10] Scanning through modern Korean text, however, the number of actually used Hangul syllables has been estimated at 2,350 only. The Korean Industrial Standards Association took up this number and coded each of the 2,350 syllables as precomposed like Chinese characters.[11]

The restriction on admissible combinations of consonant letters is due to the phonological properties of Korean. In representing languages other than modern Korean, Hangul may allow any number of consonant letters at the initial or final position of a syllable structure. Hangul only lays down the general tripartite structure of a syllable without any specific constraints on the doubling or composition of individual consonant or vowel letters.[12]

5 Syllable Representation

Syllables in Hangul resemble Chinese characters in their layout. Words are printed as a string of two-dimensional syllable characters in a strict sequence of normally invisible equal-sized square frames. Each square block framing a Hangul syllable character then splits into two, three, or four smaller rectangles to house character components.

The internal placement of character components within the square frame of a syllable depends on the syllable structure as well as on the shape of a vowel. Depending on the shape of a vowel, an open syllable CV has an evenly divided square block [a], [b] or [c], a closed syllable CVC or CVCC a square block [d], [e] or [f], as shown:



For example, the syllable 가 'ka' is of shape [a], because the vowel ㅏ 'a' is of a vertical column shape. The syllable 구 'ku' is of shape [b], because the vowel ㅜ 'u' is of a horizontal bar shape. The syllable 과 'kwa' is of shape [c], because the diphthong ㅘ 'wa' is both horizontally and vertically shaped. The syllables 감 'kam' and 삶 'salm' are of shape [d] or [e], because they are closed syllables and their vowel ㅏ 'a' is vertically shaped. The syllable 꿈 'kkum' is of shape [f], because the vowel ㅜ 'u' is horizontally shaped and also because it is a syllable closed with the final ㅁ 'm' supporting the medial vowel from underneath.

[8] See S.S. Kang (1993: 49-50), for instance.

[9] The notion of FILLER was devised to treat both open and closed syllables in an uniform manner by assuming that every open syllable has a FILLER as its final.

[10] See Hangul and Computer Co.'s Report (1992:13).

[11] This number can be reduced further for some practical applications, for S.S. Kang (1993: 50), for instance, claims that only 1,996 syllable characters are needed for morphological analysis.

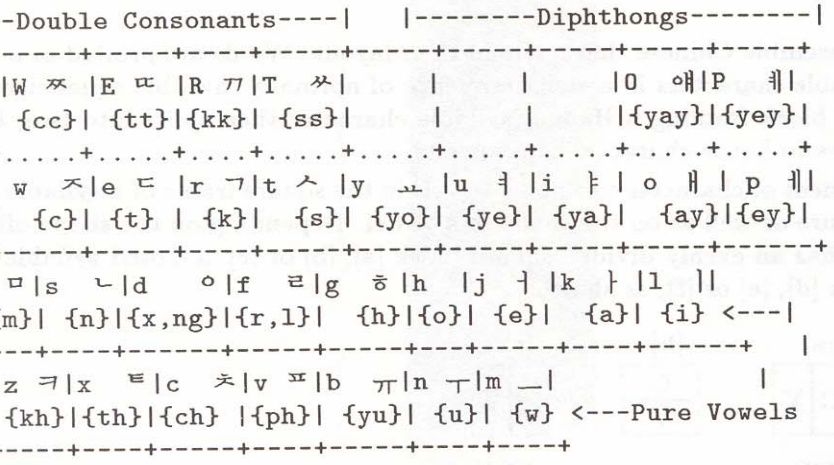
[12] In Middle or Late Middle Korean texts, one can find a series of two or three consonant letters occurring at the initial position of a syllable. Examples are {ᄇᆞ뒤} 'ptoy'(dirt) and {ᄇᆞᄉᆞ대} 'pstay'(time). See Martin (1992:28). Hence, the notion of *possible* combinations of letters must be distinguished from that of *admissible* or *actual* combinations. For some computing purposes, one need not consider every possible combination of Hangul letters. For linguistic or orthographic discussions, however, all possible and even impossible combinations should be displayable in Hangul.

Syllable squares or their component rectangles are not represented in Hangul. They just provide a conceptual frame for forming syllable units.[13] Since each consonant or vowel letter must be squeezed into some space inside an equal-sized square, its size and shape must vary accordingly; otherwise, syllable units look imbalanced. Because of such variations, Hangul actually uses a much larger number of fonts than its 24 basic letters. Also, for stylistic or artistic reasons these structures may undergo a great deal of variations.

6 Key-in and Display

On Korean typewriters and keyboards, there are two keyboard layouts in use: the two-set and the the three-set modes. The so-called *two-set* mode is more generally accepted. The toggle key for changing from the English mode to the Hangul mode or vice versa is "Shift+Bar" or some other designated key. The allocation of keys on the two-set mode keyboard is shown in the following table:

Two-Set Mode Hangul Keyboard Layout



According to the two-set mode layout, the three rows of the alphabet keys are used to allocate 33 Hangul letters.[14] These rows are then vertically divided into two: the left is occupied by 19 consonant keys and the right by 14 vowel or diphthong keys.

The specific allocation of Hangul keys is based on phonetic groupings. The five left-most keys at the third row are for obstruent letters: the lowercase keys are for simple obstruents and the uppercase keys for their respective double obstruents. The right-half of the keys at the third rows are for non-atomic vowels: the right-most keys in the upppercase are for two complex diphthongs, ㅒ 'yay' and ㅖ 'yey'. In the second row, the five keys on the left-half are occupied by three nasals ㅁ 'm', ㄴ 'n', ㅇ 'ng', a liquid ㄹ 'l', and a fricative ㅎ 'h' and the four keys on the right-half by four pure vowels ㅗ 'o', ㅓ 'e', ㅏ 'a', ㅣ 'i'. The left-side keys at the bottom row are for aspirated consonants ㅋ 'kh', ㅌ 'th', ㅊ 'ch', ㅍ 'ph' and the right-side keys for a diphthong ㅠ 'yu' and two pure vowels ㅜ 'u' and ㅡ 'w'. This keyboard layout well reflects the natural grouping of phonemes that the Hangul letters represent. It differs conceptually from the layout of Western keyboards, which are arranged according to statistical frequency.

While the two-set mode of keyboard layout is phonetically based, the three-set mode is statistically oriented. The allocation of keys on the three-set keyboard also seems to be based on

[13] While practicing penmanship or calligraphy, however, one may wish to draw these squares and try to be exact in putting each letter in a right place.

[14] On the keyboard layout, for the convenience of readers of this paper, Hangul letters are each romanized and braced by a pair of curly brackets.

the frequency of use and presumably on the efficacy of finger uses. Even though this seems to support the three-set mode, it appears to be less favored in its use. The public prefers to have a smaller set of objects to learn and thus a smaller number of Hangul keys to memorize.

7 Hangul Code Sets

On the computer, all symbols such as the letters of the Latin alphabet or the characters of Hangul, are represented as binary numbers. If each character is represented by a binary number of length 7, then there are $128(=2^7)$ binary numbers available.

With 7-bit numbers, ASCII, the American Standard Code for Information Interchange, coded exactly 128 characters, consisting of the Latin alphabet letters, numerals, special graphic characters, *sp*(for spacing), and *del*(for deletion) as well as 32 control characters. The following ASCII table, which will be referred to below, represents the hexadecimal numbering system.

The ASCII Table

```
Hexadecimal Numbering:
C0 Area: control characters
|00 nul|01 soh|02 stx|03 etx|04 eot|05 enq|06 ack|07 bel|
|08 bs |09 ht |0a nl |0b vt |0c np |0d cr |0e so |0f si |
|10 dle|11 dc1|12 dc2|13 dc3|14 dc4|15 nak|16 syn|17 etb|
|18 can|19 em |1a sub|1b esc|1c fs |1d gs |1e rs |1f us |
GL Area: graphic characters
|20 sp |21  ! |22  " |23  # |24  $ |25  % |26  & |27  ' |
|28  ( |29  ) |2a  * |2b  + |2c  , |2d  - |2e  . |2f  / |
|30  0 |31  1 |32  2 |33  3 |34  4 |35  5 |36  6 |37  7 |
|38  8 |39  9 |3a  : |3b  ; |3c  < |3d  = |3e  > |3f  ? |
|40  @ |41  A |42  B |43  C |44  D |45  E |46  F |47  G |
|48  H |49  I |4a  J |4b  K |4c  L |4d  M |4e  N |4f  O |
|50  P |51  Q |52  R |53  S |54  T |55  U |56  V |57  W |
|58  X |59  Y |5a  Z |5b  [ |5c  \ |5d  ] |5e  ^ |5f  _ |
|60  ` |61  a |62  b |63  c |64  d |65  e |66  f |67  g |
|68  h |69  i |6a  j |6b  k |6c  l |6d  m |6e  n |6f  o |
|70  p |71  q |72  r |73  s |74  t |75  u |76  v |77  w |
|78  x |79  y |7a  z |7b  { |7c  | |7d  } |7e  ~ |7f del|
```

This table consists of two areas: the C0 Area for control characters and the GL Area for graphic letters or characters. Note that *sp* and *del* are treated as genuine characters and coded among the graphic characters. For several different sets of Latin alphabet letters including diacritics, ASCII had to extend the set of binary numbers and introduced 8-bit or 1-byte numbers for coding. Furthermore, even for 7-bit characters, an extra bit is used for so-called *parity checks*. Since 1-byte binary numbers are 8 digit long, the octal or hexadecimal numbering system is normally used to represent the coding of characters in ASCII.

Because of its peculiar syllable compositions and representations, the coding of Hangul is more complicated than that of the Latin alphabet. There are two major approaches to the coding of Hangul: one is to treat syllables as precomposed units and the other to treat them as composed of initial, medial, and final. The former is represented by a coding called KSC 5601-1987 and the other by a variety of other combinational code sets. In order to have enough room for coding all possible Hangul characters, both approaches use two bytes in contrast to ASCII, which uses only one byte.

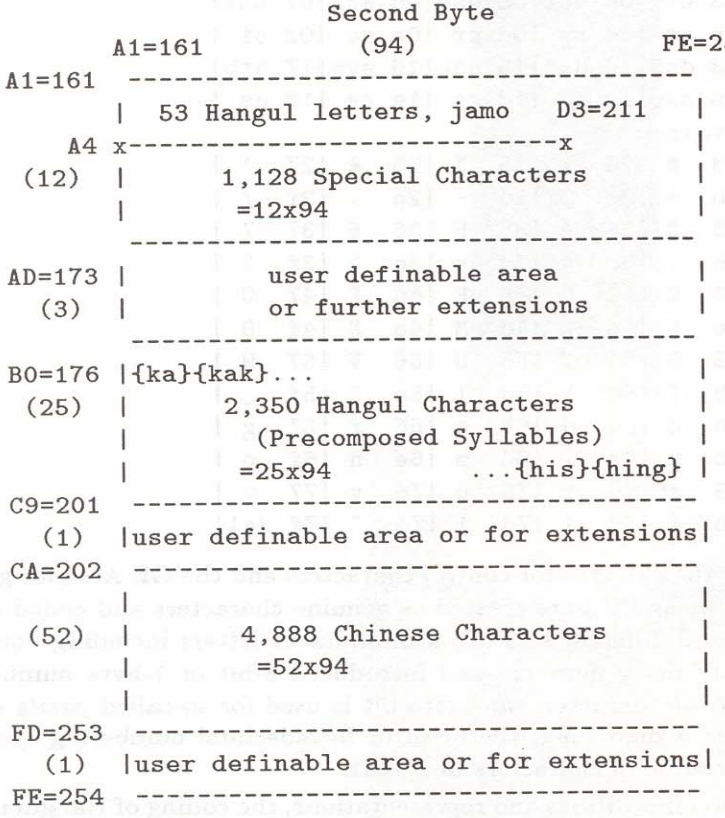
KSC 5601-1987 was the standard Hangul code proposed for industrial use by the Korean Industrial Standards Association in 1987. Based on an 8-bit 2-byte allocation of codes, it assigns

a pair of two non-ASCII code points to each precomposed character in Hangul. Assignment values are restricted to the non-ASCII area ranging from 161 to 254 (or from hexadecimal A1 to FE). By placing Hangul code points in the 2-byte area of A1_FE×A1_FE beyond the ASCII area 127 of 7f, KSC successfuly distinguishes Hangul characters from Latin alphabet letters. This is shown in the sketch of the KSC Code Allocation Table given on the following page.

On the subarea B0_C8×A1_FE of the table, 2,350 Hangul syllables are coded as unit characters. Isolated Hangul consonants and vowels are treated as special characters and coded in the subarea on the top of the table with other special characters. 4,888 Chinese Hanja characters are also coded. The remaining 470 code points are allocated for any future extensions or private coding.[15]

For academic uses and linguistic purposes, however, we need a wider code zone for Chinese characters[16] and a theoretically infinite number of code spots for Hangul characters. The current KSC 5601-1987/1989 should thus be augmented by a more expressive code set to accommodate a potential variety of the rapidly growing research and industrial activities in the area of Hangul engineering.

The KSC 5601 Code Allocation Table for Hangul



Before KSC 5601 appeared, a variety of 8-bit 2-byte combination code sets were devised by the IBM Korea and a host of other PC makers. The so-called *sang-yong* 2-byte combination code set is still popular among PC makers and users in Korea. The 2-byte approach accepts the basic tripartite syllable structure of Hangul consisting of initial, medial, and final, and classifies

[15] In 1989, KSC 5601 has been expanded to accommodate a new list of letters in archaic Hangul as well as additional modern and archaic syllables.

[16] Chinese characters should have been treated as belonging to a separate code set rather than a genuine part of the Korean coding system. For such efforts, see **Unicode** in the concluding section.

Hangul letters accordingly. It then allocates 5 bits out of 16 bits (= 8bits×2bytes) to each of the 3 letter sets and uses the remaining 1 bit as MOB, the most significant bit, for distinguishing Hangul letters from Latin letters: 0 is for Latin letters and 1 for Hangul letters.

One of the most recent efforts has been to develop an extended unix code, EUC, by encoding KSC 5601 into ISO 2022 which specifies various character sets, or charsets. Thereby, the original specification of Hangul and other characters by KSC 5601 is retained through various modes of conversion and reconversion to the orginal form. ISO 2022 was primarily devised to incorporate code sets for many different languages, thus allowing the use of multilingual characters in a text. The goal is to use Hangul letters, English alphabet letters, and Japanese characters, and even Arabic characters together in the same text without much difficulty in shifting from one character mode to other modes.

The technical basis is the assignment of a leading character, LC, in ISO 2022 to each character set, thus identifying it uniquely. It then characterizes each charset with specifying (1) the byte-length of its code, (2) columns that it occupies on a screen, (3) the number of characters that each set contains, (4) its graphic set, (5) its final character, (6) its displaying direction, and so on.

8 Applications

8.1 Sorting Order

In order to facilitate the retrieval of a desired item from a list, the items in the list are compiled in a certain order. Using this order, items may be looked up in personal address-books, telephone directories, indexes in publications or dictionaries. The operation in itself is very simple: it operates on strings of characters depending on the order assigned to each character in a string.

Problems may arise, however, if the ordering principle is subject to uncertainties, caused by special letters like the umlaut ü in German. Suppose a user wants to look up the postal code of "Nürnberg". Luckily he happens to know that the umlaut ü in German is often written as "ue" and, going through a long list of names in a column, successfully finds "Nürnberg" shortly below "Nudesdorf" but notes that it is listed above "Nunsdorf". Hence, "Nürnberg" is listed between "Nudesdorf" and "Nunsdorf".

However, this is not the only possible order in German. For example, in Hermann Paul's *Deutsches Wörterbuch*, "Küche" comes before "Kuchen". Hermann Paul's convention is also used in the popular Duden's German dictionaries. The choice is thus either treating 'ü' as a single vowel that comes after 'u' or treating it as a combination of 'u' and 'e'.

A similar but more complex problem occurs in dealing with Hangul, for it contains not only a set of compound letters, but also a large set of syllables. Sorting depends on how these compound letters and syllables are coded. Here I will just discuss the double consonants to illustrate different ways of sorting in Hangul.

There are two ways of ordering the double consonant letters in Hangul. In South Korea, they are placed immediately after their correponding base letters. Hence, the double consonant letter ㄲ 'kk' immediately follows its base ㄱ 'k', the sorting order ㄱ-ㄲ-ㄴ. In North Korea, however, they are ordered at the end of the 14 atomic consonant letters. Hence, the double consonant letter ㄲ 'kk', for instance, follows the letter ㅎ , the last of the 14 atomic consonant letters. In this ordering, we thus have: ㄱ-ㄴ-ㄷ-...-ㅎ-ㄲ-ㄸ-ㅃ-ㅆ-ㅉ.

Suppose now we follow the South Korean way of ordering the double consonant letters. There are still two different versions of sorting, for sorting in Hangul may depend on the type of syllable structure. Version One sorts out open syllables before closed syllables. For example, the open syllable 까 'kka' is ordered after 가 'ka' but before 각 'kak', although the letter ㄱ 'k' is ordered before the double consonant ㄲ 'kk'.

Version Two, on the other hand, sorts strings of letters without analyzing the syllable structures of strings. It simply scans each letter sequentially and sorts according to the fixed order of each letter. It thus sorts 까 'kka' after 깊 'kiph', after sorting out every sequence starting with ㄱ 'k'.

Sorting Order Illustrated

```
Version One                Version Two
     "ka"                       "ka"
     "kka"-------               "kak"
     "kak"       |              "kan"
     "kan"       |              "kas"
     "kas"       |              "kiph"
     "kiph"       -------->     "kka"
     "na"                       "na"
```

Version One has often been adopted for compiling Korean dictionaries.[17] From a computational point of view, this complicates the process of sorting without adding any obvious merits. In order to follow this version, the computer must first scan the entire string, analyzing its syllable structure, and then return to the beginning to carry on.[18]

Version Two operates on a string of letters in a strict sequence of how each letter is assigned a code point. The following table shows how a code point is assigned by KSC 5601 to each of the letters listed before.

가	'ka'	Ox3021
각	'kak'	Ox3022
간	'kan'	Ox3023
갓	'kas'	Ox302b
깊	'kiph'	Ox316d
까	'kka'	Ox316e
나	'na'	Ox332a

Here, sorting simply applies to each syllable character by checking each code point. The sorting of these syllables may be analyzed as checking each letter in each syllable character. It checks the initial, the medial, and the final of each syllable. Since ㄱ 'k' precedes ㄲ 'kk' in coding order, every string beginning with ㄱ 'k' must be sorted before scanning any other strings including the ones beginning with ㄲ 'kk'.

In 1933, the ordering of the 24 atomic letters was formally established by the Unified Spelling *Thong.il.an* of Hangul. It was then incorporated into the Hangul Orthography, *Hangul Mac.chum.pep*, and officially promulgated by the Ministry of Education in 1989.[19]

Till then, the intrinsic ordering of non-basic letters had been considered obvious. It had been assumed that the orthographic compositions of non-atomic letters would decide on their ordering unequivocally. Unexpectedly, there arose varying views on the compositions of letters, resulting in different orderings. Therefore, the 1989 *Hangul Mac.chum.pep* explicitly laid out the sorting order for the 40 letters of Hangul which include the 16 composite letters: the five double consonant letters and the 11 compound vowels, as in the table given on the following page.

[17] See Si-sa's *Si-sa Elite Concise Korean-English Dictionary* 2nd ed., 1992.

[18] This argument may not be valid, however, if there is a precomposed code set of Hangul in which the syllable 까 'kka' is assigned a smaller code point than the syllable 각 'kak'. Such an encoding is, however, unacceptable, for it operates rather arbitrarily.

[19] For a brief survey, see Lee and Ahn (1989: 23-27).

This ordering of Hangul letters is based on their orthographic composition. The five double consonant letters are ordered immediately after their respective atomic letters, for each of them is formed by doubling an atomic consonant letter. The compound vowel letters are ordered in the same manner. For instance, the compound letter ㅐ is immediately ordered after the atomic vowel letter ㅏ, because its first componet is ㅏ.[20]

Sorting Order of Hangul Letters

	1	2	3	4	5	6	7	8	9	10	
Consonants:	ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	ㅁ	ㅂ	ㅃ	ㅅ	
	11	12	13	14	15	16	17	18	19		
Consonants:	ㅆ	ㅇ	ㅈ	ㅉ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ		
	20	21	22	23	24	25	26	27	28	29	30
Vowels:	ㅏ	ㅐ	ㅑ	ㅒ	ㅓ	ㅔ	ㅕ	ㅖ	ㅗ	ㅘ	ㅙ
	31	32	33	34	35	36	37	38	39	40	
Vowels:	ㅚ	ㅛ	ㅜ	ㅝ	ㅞ	ㅟ	ㅠ	ㅡ	ㅢ	ㅣ	

Although the numbering is given explicitly, the ordering of Hangul letters specified in the table is still relative. The numbering may change, for example, when other composite letters like final consonant clusters or archaic composite letters are added to the list of 40 letters of Hangul.

As the character set of Hangul increases, an algorithm for Hangul sorting may become complex. In KSC 5601-1989, a large number of Hangul letters and syllables, both modern and archaic, were added and allocated a separate code zone without reshuffling. As a consequence, the problem of sorting through archaic text has arisen and awaits a solution.

8.2 Hangul on X

X is a network-transparent window system which allows multiple applications to run simultaneously in windows. Because it is network transparent, applications can run machines scattered through the network. As a virtual screen, it generates multi-font text and graphics in monochrome, color, or a bitmap display.

Controling all access to input and output devices,[21] the X Window System as a server allows various client programs to run on it. One type of its major clients is a terminal emulator program like the *xterm* or hpterm[22] which creates a virtual terminal and allows terminal-based programs to run on it. In fact, the xterm program is the official terminal emulator for the X Window System itself, providing DEC VT102 and Tektronix 4014 compatible terminals for programs that cannot use the window system directly.

All these ideas of X have been incorporated into the design and implementation of a Hangul xterm called *hanterm*.[23] It emulates all terminals, except for Tektronix, for Hangul input and display. It uses EUC, an Extended Unix Code, which encodes the KSC 5601-1987 code set, and runs on X under the platforms like SunOs 4.1.1.-KLE1.1 or LINUX on 386 or higher PC.

The upgraded version *han3term*, Hanterm Version 3.0, now supports a 3-byte Hangul code[24] as well a combinational code set. It also allows the option of choosing the standard two-set mode keyboard and the three-set mode keyboard.[25]

[20] In North Korea, the double consonant letters are immediately ordered after the 14 atomic consonant letters because they are treated not as geminates, but as atoms.

[21] Typically, input devices refer to the mouse and keyboard, and output devices to the display screen.

[22] Terminal emulator for Term0 terminals.

[23] The program *hanterm* was produced by the GUI Consortium at KAIST under the supervision of Prof. Kilnam Chon, director of HCI Lab, Korea Advanced Institute of Science and Technology.

[24] For further information on this newer version, contact Chang, Hyeong-Kyu at chk@ssp.etri.re.kr.

[25] The new version has been successfully installed on HP Series 700 at the Computational Linguistics Lab of the

Hanterm is still at the preliminary stage of testing its support of editors and document preparation systems. The Hangul vi, *helvis*, seems to work very nicely on it, but Hangul Emacs, *NEmacs*, is rather clumsy.[26] Although Mule, as a multi-lingual enhancement of Emacs, runs independently of any language specific environment, it supplements hanterm effectively without causing any incompatibilities. Mule provides a powerful editor, while hanterm provides an X environment for various supporting programs like *hscreen* which create virtual terminals or *hcode* which handle code conversions or romanization of Hangul.

8.3 Text Editor: Mule

The coding of Hangul characters is the basis for writing and editing Hangul on the computer. Only now editors like vi or Emacs, which are normally used for texts written in the Latin alphabet, have become available to users of Hangul. A test version of Hangul Emacs, which was a modified Japanese NEmacs, appeared in 1991 or so.

Mule is a new version of the GNU[27] Emacs that can handle multilingual texts. The name is an acronym derived from a **Mul**tilingual **E**hancement to GNU Emacs. It can handle both ASCII and non-ASCII characters, both 7 bits (ASCII) and 8 bits (Latin-1) characters as well as 2-byte (Japanese, Chinese, Korean) characters coded in the ISO 2022 standard or its variants like EUC. It accommodates all these characters in a single buffer, thus providing an ideal environment for developing multi-lingual programs like interlingual translation systems.

Mule is the basis for running Hangul LaTeX. K. Un's *jhtex* version 0.7 running on Mule 2.0 or enhanced Emacs 19.25 is a Hangul formatter based on LaTeX2ε.[28] With this Hangul LaTeX, all the vowels and consonants, whether atomic or compound, can be displayed and printed, indicating the final consonant clusters, except for ㄹㅍ 'lph',[29] can also be printed.

Despite its continuous upgrading, Mule still requires many additional improvements for Hangul. Currently, it can only display or print a restricted number of Hangul letters and syllables. These component letters are dealt with only as non-combining elements. Besides its inability to display many of the desirable Hangul characters, it cannot, for instance, close a syllable without hitting a spacebar. For example, to type "서울"(Seoul), one has to type "tj< *sp* >dnf"[30] in the Hangul mode. If the spacebar is omitted after the open syllable "서", there results an ill-formed string of letters ㅅㅓㅇㄹ. In the Hangul mode, the two keys for ", <" and ". <" are not active.

8.4 Romanization and Round-Trip Compatibility

For Western readers, it is often useful to translate Hangul text into the Latin alphabet. For Korean readers, the romanized text must then be translated back into the original Hangul. These two-way processes of transliteration can be carried out in an automatic way with an appropriate code conversion program. A program satisfies the so-called *round-trip compatibility*, if it succeeds in mapping Hangul letters into the Latin alphabet and then back into the orginal Hangul letters.[31]

University of Erlangen-Nürnberg, too. Some time-consuming work was needed for its compilation. For obtaining installation patches, contact Rolf Haberrecker, rolf@linguistik.uni-erlangen.de.

[26] For further information, contact Woohyung Choi, HCI Lab, KAIST, Taejon, Korea, whchoi@cosmos.kaist.ac.kr.

[27] Pronounced as [gnyu], an acronym derived from 'GNU's Not UNIX,' standing for the name of Free Software Foundation which circulates various UNIX-compatible programs free.

[28] This new version was officially released on June 1, 1994, replacing LaTeX2.09.

[29] The ㄹㅍ 'lph' as a cluster unit is coded in KSC 5601, but cannot be printed because no font has been made available. No one seems to have been interested in making a special font for it because it is rarely used, neither by itself nor in combinations other than the syllable combination of 읊 'wlph'.

[30] The < *sp* > is not part of the string, but just indicates the hitting of the spacebar.

[31] In recent years, a couple of Hangul romanization programs have been developed. One is Beom-mo Kang's (1993) program *h2kr.pas* and another June-Yub Lee's (1992) *hcode*. The *h2kr.pas* is written in PASCAL and runs

In Hangul text, syllable boundaries are clearly recognizable because Hangul syllables are each represented by a clearly demarcated character block. Romanized Hangul, written as a string of Latin alphabet letters, does not automatically indicate those syllable boundaries. The following examples illustrate ambiguities resulting from the loss of syllable structure in a simple-minded romanization.

	Hangul	Ambi.		Hangul	Ambi.
(1)	가짜(fake)	kacca	(3)	바름(correctness)	palum
	갖자(Let's have)	kacca		발음(pronunciation)	palum
(2)	안짜(not salty)	ancca	(4)	고기(meat)	koki
	앉자(Let's sit)	ancca		곡이(song+Subject)	koki

Using a hyphen to mark syllable breaks, the romanization may be disambiguated, resulting in 'ka-cca' vs. 'kac-ca', 'an-cca' vs. 'anc-ca', 'pa-lum' vs 'pal-um', and 'ko-ki' vs. 'kok-i'.

The ISO Romanization and the Yale Romanization disambiguate syllables in different ways, both meeting round-trip compatibility. Compare the two systems with respect to romanizing Hangul consonant letters.

Romanization of Consonants

	1	2	3	4	5	6	7	8	9	10
Hangul:	ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	ㅁ	ㅂ	ㅃ	ㅅ
ISO										
Initial:	K	Kk	N	T	Tt	R	M	P	Pp	S
Final:	k	kk	n	t		l	m	p		s
Yale										
Initial:	k	kk	n	t	tt	l	m	p	pp	s
Final:	k.	kk.	n.	t.		l.	m.	p.		s.

	11	12	13	14	15	16	17	18	19
Hangul:	ㅆ	ㅇ	ㅈ	ㅉ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
ISO									
Initial:	Ss		C	Cc	Ch	Kh	Th	Ph	H
Final:	ss	ng	c		ch	kh	th	ph	h
Yale									
Initial:	ss		c	cc	ch	kh	th	ph	h
Final:	ss.	ng.	c.		ch.	kh.	th.	ph.	h.

The 19 Hangul consonant letters romanized in this table consist of the 14 atomic and the 5 double consonant letters. In the ISO system, the initial IEUNG ㅇ is not represented. With this exception, each of the remaining 18 initial Hangul consonant letters is romanized in uppercase. For example, the initial Hangul consonant letter ㄱ of number 1 is romanized into 'K' and the initial Hangul double consonant letter ㄲ of number 2 into 'Kk'. On the other hand, the final Hangul consonant letters are romanized in lowercase. Hence, the final ㄱ and ㄲ are simply transliterated into the lowercase Latin alphabet letters 'k' and 'kk'. The initial Hangul liquid letter ㄹ of number 6 is romanized into the uppercase 'R', although this Latin alphabet letter is distinct from the Latin letter 'l' into which the final ㄹ is transliterated.

By transliterating initial Hangul consonant letters into uppercase Latin letters, the ISO Romanization system aims at disambiguating syllable boundaries in romanized Hangul text. The following table shows how the ISO system succeedes in disambiguating syllables.

on a DOS-operated PC. With some options on the choice of a romanization system, it can process both Hangul and romanized textual materials.

ISO's Disambiguation of Syllables

(1)	가짜(fake)	KaCca	(3)	바름(correctness)	PaRum
	갖자(Let's have)	KacCa		발음(pronunciation)	Palum
(2)	안짜(not salty)	anCca	(4)	고기(meat)	KoKi
	앉자(Let's sit)	ancCa		곡이(song+Subject)	Koki

The use of uppercase in the ISO Romanization system has a few drawbacks, however. First, the ISO system fails to operate in a computer system which is case insensitive. Secondly, the convention of writing the first letter of proper names in uppercase cannot be retained in romanized Hangul. Lastly, the conspicuous proliferation of uppercase letters may reduce readability and lose textual elegance.

The Yale Romanization system, which seems to have been accepted as the standard in linguistic circles, uses a different convention to disambiguate romanized Hangul syllables. Instead of using uppercase, the Yale system adopts a set of conventions for using the dot '.'. By one convention, every final consonant may be marked with a dot '.'. Hence, we obtain the following:

Yale's Disambiguation of Syllables

(1)	가짜(fake)	kacca	(3)	바름(correctness)	palum
	갖자(Let's have)	kac.ca		발음(pronunciation)	pal.um
(2)	안짜(not salty)	an.cca	(4)	고기(meat)	koki
	앉자(Let's sit)	anc.ca		곡이(song+Subject)	kok.i

The dot marking each final consonant letter automatically marks the end of a syllable, thus differentiating 'kacca'(가짜) from 'kac.ca'(갖자), 'an.cca' (안짜) from 'anc.ca'(앉자), 'palum' (being right 바름) from 'pal.um'(발음), and 'koki'(고기) from 'koki' (곡이).

The problem of syllabification also occurs in converting romanized Hangul vowel letters back into its original. For illustration, we select a few vowel letters, say ㅏ, ㅑ, ㅖ, ㅐ, and ㅒ, and discuss their romanization.

Hangul:	ㅏ	ㅑ	ㅖ	ㅐ	ㅒ
ISO:	a	ya	e	ae	yae
Yale:	a	ya	ey	ay	yay

As shown in the table, the ISO system romanizes these characters into 'a', 'ya', 'e', 'ae', and 'yae', respectively, and the Yale system into 'a', 'ya', 'ey', 'ay', and 'yay', respectively. Syllable ambiguities occur in the ISO system, for the romanized letter 'ae' can be interpreted as representing either 애 or 아에, and 'yae' as representing either 얘 or 야에.

In the Yale system, however, no such ambiguities occur, due to the convention of representing the initial IEUNG ㅇ with the dot '.'. In this system, for instance, the romanized string 'ayay' must be either '.a.yay' 아얘 or '.ay.ay' 애애. Hence, if every occurrence of the initial IEUNG ㅇ is represented by the dot in romanized Hangul, there arises no syallable ambiguity and round-trip compatibility is ensured.

9 Remarks on Unicode

On the international scene, there is a gradual, but revolutionary movement towards unifying all character sets into a unified universal character set, often identified with *Unicode*. In 1988, the first serious endeavors were made to unify all the pre-exiting coding systems into a single consistent system for encoding all languages in a uniform manner. For the purpose of efficiently processing multilingual text, the pre-existing character codes were found inadequate and often inconsistent due to the basic architecture of many of the present character codes as well as

conflicting national standards. Despite its simplicity and elegance, the ASCII's 7-bit character size could not accommodate several thousand Hangul syllables or over 20,000 Han characters. As was seen in implementing Mule by adopting ISO 2022, currently encoded multingual text contained too many ESC sequences, thus slowing down efficiency in processing.

The International Organization for Standardization thus commissioned an *ad hoc* committee to study the possibility of adopting multibyte encoding as a worldwide standard. By the beginning of 1991, the Committee developed a 32-bit character encoding known as ISO DIS 10646.[32] It was, however, voted down by a majority of ISO member countries. Instead, resolutions constituting the merging of DIS 10646 with the Unicode standard were approved.

The Unicode standard was modeled on the general layout of the ASCII character set. However, it had to expand the size of a character from ASCII's 7-bit to a 16-bit size. This made it possible to assign numerical codes of the same size to both ASCII and non-ASCII characters in a uniform manner. Besides retaining uniformity in charcter size, Unicode aimed at completeness and efficiency in its encoding scheme. To make its encoding complete, the Unicode Consortium formed *Unicode, Inc.*, an international non-profit organization promoting the uniform encoding of all possible characters in one and the same codespace. With uniform encoding, text is to be produced in a plain format with sequences of fixed-width characters and with no escape sequences for differentiating character sets. As a consequence, text processing can be carried out more efficiently.

As for Hangul, Unicode Version 1.0 simply copied the KSC 5601 character set as part of the CJK[33] Auxiliary character zone into its general codespace. The selection and ordering of characters are preserved in the Unicode standard as in KSC. Each of the Hangul letters, both modern and archaic, is assigned a code point in the ranges U+3131 → U+3163 and U+3164 → U+318E, and all the 2,350 Hangul syllables in the range U+3400 → U+3D2D. A special set of circled or parenthesized Hangul characters is also encoded.[34] As in KSC, the published version 1.0 of the Unicode standard basically treats Hangul not as a phonetic or phonemic, but as a syllabic writing system. To reflect the alphabetic features of Hangul, it should be able to treat Hangul letters as combinational elements in a later version of the Unicode standard.

Under the merger efforts with ISO 10646, the upcoming Version 1.1 of the Unicode standard plans to incorporate 93 combinining elements as well as 1,930 modern and 1,754 old Korean syllables of Hangul into its enlarged sets of character codes. In this version, the CJK ideographics zone is pushed further away from the Hangul or CJK Auxiliary zone by leaving behind a wider gap available for encoding larger Hangul sets: the range of encoding CJK ideographs moves from the original range of 4000 → 8BFF to the new range of 4E00 → 9FFF, thereby increasing its own size from 19,456 to 20,992 code points. The maximal unassigned area for Hangul now increases from 722 to 4,050 code points, for the vacant space is expanded from 3D2E to 4DFF.

One remarkable improvement over KSC 5601 is that Chinese Han characters were taken out of the Korean character zone and placed under the general heading of CJK ideographs. This should certainly reduce the work load of encoding characters in Korean text, for KSC 5601-1987 allocated a sizable area of codespace for Han characters. By properly excluding their treatment, KSC can better concentrate on encoding Hangul letters and syllables.

10 Conclusion

This paper is another illustration of the close connections among language, orthography, and its computational treatment. I have shown that the linguististic consideration of traditional

[32] DIS stands for "Draft International Standard."

[33] CJK is an acronym for "Chinese, Japanese, and Korean".

[34] However, many of the other characters which are considered to be necessary for Hangul text processing are left for future additions.

concepts of Hangul structure and letters is not only useful but necessary for an efficient implementation on the computer.

Explaining to the Western reader both the phonemic and the syllabic nature of the writing system of Korean, I have discussed problems of sorting, coding, romanization, and implementing computational tools for Hangul. The requirements of representing letters and syllable characters on the computer, especially when applied to a strange and exotic writing system, provided a new glance and a better understanding of the nature of language. Similarities between the coding of different writing systems find their formal expression the new development of Unicode.

Acknowledgments

Portions of this paper have been read and commented on by Roland Hausser, Gerald Schüller, and Marco Zierl at a series of colloquia. I thank them all. I am also grateful to the help of Hung-Gyu Kim and Beom-mo Kang of Korea University with their material on Hangul codes, Koaunghi Un of Universität Tübingen with his Hangul LaTeXjhtex and advice in program installation, and Woohyung Choi, HCI Lab, KAIST, for his prompt e-mail replies to my frequently asked questions. I also thank Rolf Haberrecker for installing various Hangul programs on our HP workstations, Jörg Schreiber and Markus Schulze for kindly responding to tons of trivial questions, and Jochen Leidner and Martin Scherbaum who proofread and commented on the earlier draft.

References

[1] Chang, Hyeong-Kyu, et al. (1992). 3-Byte Hangul Coding System. Unpublished. KAIST, Taejon, Korea.

[2] Cho, See-Young (1992). Scannen von Koreanischen Texten mit OCTOPUS. *MAKROLOG Newsletter* 1, 16-20.

[3] Crystal, David (1990). *A Dictionary of Linguistics and Phonetics.* 2nd ed. Basil Blackwell, Oxford.

[4] Hangul and Computer Co. (1992). Report of A Basic Research on Hangul Codes and Keyboards [written in Korean], sumbitted to the Ministry of Culture, Seoul.

[5] Hewlett-Packard, Interface Technology Operation. (1991). *Using the X Window System.* 4th ed. Hewlett-Packard Co., Covallis, Oregon.

[6] Hong, Yun-Pyo, et al. (1991). A Research on the Computational Treatment of Archaic Letters in Hangul [written in Korean]. *Hankuke Censanhak (Korean Computational Linguistics)* 1, 1-84.

[7] Kang, Beom-mo (1993). Document and Source Code for h2kr.pas. Research Institute for Language and Information, Korea University, Seoul. Unpublished.

[8] Kang, Seung Shik (1993). *Korean Moorphological Analysis Using Syllable Information and Multiword Unit Information* [writen in Korean]. Unpublished doctoral dissertation, Seoul National University.

[9] Lamport, Leslie (1994). *LaTeX User's Guide and Reference Manual.* Updated for LaTeX2ε. Addison-Wesley, Reading, Massachusetts.

[10] Lee, June-Yub (1992). Document and Source for HCODE Version 2.0. Unpublished.

[11] Lee, Hi-Sung and Ahn, Byung-Hee (1989). *Lectures on Hangul Orthography* (*Hangul Machumpep Kanguy*)[written in Korean]. Shin-Ku Munwha-Sa, Seoul.

[12] Maebashi, Takahiro (1993). Frequently Asked Questions and Answers for Mule. Translated by Naoto Takahashi. FTP from etlprt.etl.go.jp.

[13] Martin, Samuel E. (1992). *A Reference Grammar of Korean.* Charles E. Turtle, Tokyo.

[14] Rosenberg, Jerry M. (1986). *Dictionary of Artificial Intelligence and Robotics.* John Wiley and Sons, New York.

[15] Scheifler, Robert W., James Gettys, et al. (1992). *X Window System: The Complete Reference to Xlib, X Protocol, ICCCM, XLFD.* 3rd ed. Digital Press, Burlington, Massachusetts.

[16] Schoonover, Michael A. et al. (1992). *GNU Emacs: UNIX Text Editing and Programming.* Addison-Wesley, Reading, Massachusetts.

[17] Shin, Jungshik (1993). Hangul and Internet in Korea.FAQ. Mailed from jshin@minerva.cis.yale.edu.

[18] Un, Kouanghi (1994). How to Use Hangul LaTeX: Based on jhtex Version 0.7 [written in Korean]. Unpublished. Universität Tübingen.

[19] Unicode Consortium, The (1991). *The Unicode Standard: Worldwide Character Encoding Version 1.0*, Vol. 1. Addison-Wesley, Reading, Massachusetts.

[20] Unicode Consortium, The (1992). *The Unicode Standard: Worldwide Character Encoding Version 1.0*, Vol. 2. Addison-Wesley, Reading, Massachusetts.

[21] Unicode Consortium, The (1993). *The Unicode Standard, Version 1.1.* Prepublication Edition.

[22] Unicode, Inc. (1994). UnicodeData-1.1.2.tx. Unpublished. Anonymous FTP from unicode.org.

[23] Unicode, Inc. (1994). KSC 5601 to Unicode Table. Unpublished. Anonymous FTP from unicode.org.

Aufnahmeantrag zur GLDV – Mitgliedschaft

Name: ..... Vorname: .....

Straße: ..... PLZ: ..... Ort: .....

Telefon: ..... email: .....

Akad. Grade und/oder Berufsbezeichnung: ..... Firma oder Hochschule: .....

Student: NEIN / JA Semesterzahl: ..... Fächer: .....

Arbeitsschwerpunkt bzw. Interessengebiete: .....

Ort: ..... Datum: ..... Unterschrift: .....

Referenz– Erklärung: Ich bin (nicht–studentisches) Mitglied der *GLDV* seit 19 und befürworte diesen Antrag

Name: ..... Vorname: ..... email: .....

Ort: ..... Datum: ..... Unterschrift: .....

Die *Gesellschaft für Linguistische Datenverarbeitung e.V. (GLDV)* wurde 1975 - zunächst unter dem Namen "LDV-Fittings e.V." - als Verein zur Förderung der wissenschaftlichen linguistischen Datenverarbeitung gegründet. Sie ist der wissenschaftliche Fachverband für die maschinelle Sprachverarbeitung in Forschung, Lehre und Beruf. Sie vertritt die Interessen ihrer Mitglieder nach außen und fördert die Kooperation zwischen den Mitgliedern und ihren verschiedenen Arbeitsbereichen. Zu diesen Arbeitsbereichen gehören die computerlinguistische Grundlagen- und Anwendungsforschung, die sprachorientierte Forschung zur Künstlichen Intelligenz, die sprachliche Informations- und Wissensverarbeitung sowie die philologische Datenverarbeitung. Die *GLDV* fördert diese Arbeitsbereiche durch Veranstaltungen, Arbeitskreise und Publikationen.

Die *GLDV* unterstützt die Zusammenarbeit mit Nachbardisziplinen (wie z.B. Linguistik und Semiotik, Informatik und Mathematik, Psychologie und Kognitionswissenschaft, Informations- und Dokumentationswissenschaft) und unterhält Kontakte zu den entsprechenden Fachverbänden. International kooperiert die *GLDV* mit Organisationen wie der *Association for Computational Linguistics (ACL)*, der *Association for Literary and Linguistic Computing (ALLC)* und der *Association for Terminology and Knowledge Engineering (TKE)*.

— Veranstaltungen —

Die Jahrestagungen der *GLDV* bieten einen Überblick über die aktuellen Entwicklungen im Bereich der maschinellen Sprachverarbeitung:

1985: *Sprachverarbeitung in Information und Dokumentation*
1986: *LDV und philologische Datenverarbeitung*
1987: *Analyse und Synthese gesprochener Sprache*
1988: *Computerlinguistik und ihre theoretischen Grundlagen*
1989: *Interaktion und Kommunikation mit dem Computer*
1990: *Lexikon und Lexikographie*
1991: *Quantitative Linguistik*
1993: *Sprachtechnologie: Methoden, Werkzeuge, Perspektiven.*

Seit 1992 alternieren diese *GLDV*-Tagungen mit der *Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*, die gemeinsam von den verschiedenen deutschsprachigen LDV/CL–Vereinigungen veranstaltet wird. Dies bedeutet in der Zukunft einen Wechsel zwischen besonderen Schwerpunktthemen aus der Arbeit der *GLDV* für ihre Tagungen und Themen eher gemeinsamen Interesses für die *KONVENS*. Darüber hinaus organisiert die *GLDV* mit ihren Arbeitskreisen Kleintagungen, Workshops, Tutorials und Herbstschulen.

— Arbeitskreise —

Gegenwärtig bestehen folgende Arbeitskreise, in denen neben Wissenschaftlern auch Studierende mitarbeiten:

AK **Kodierung/Normierung maschinenlesbarer Texte (TEI)**
Leiter: Peter Scherber, email: pscherb@ibm.gwdg.de

AK **Hypertext** (im Aufbau)
Leiter: N.N.

AK **Ausbildung und Berufsperspektiven**
Leiter: N.N.

AK **Lexikographie**
Leiter: Nico Weber, email: ncw@cl2.ikp.uni-bonn.de

AK **Linguistische Software (ASK–Linguistik)**
Leiter: N.N.

AK **Maschinelle Übersetzung**
Leiter: Johann Haller, email: hans@iaisun.iai.uni-sb.de

AK **Quantitative Linguistik**
Leiter: Reinhard Köhler, Trier

AK **Parsing in Morphologie und Syntax**
Leiter: Roland Hausser,
email: rrh@linguistik.uni-erlangen.de

AK **Korpora**
Leiter: Robert Neumann,
email: rbneu@mx300b.coling.ids-mannheim.de

— Publikationen —

Publikationsorgan der *GLDV* ist das *LDV-Forum*. Die Zeitschrift veröffentlicht Fachbeiträge und Berichte, Diskussionen und Rezensionen aus dem gesamten Spektrum der Linguistischen Datenverarbeitung und Computerlinguistik (LDV/CL); einzelne Hefte erscheinen dabei zu besonderen Themenschwerpunkten. Über die Annahme von *Fachbeiträgen* entscheidet der Herausgeber nach (mindestens zwei) Gutachten von Mitgliedern des wissenschaftlichen Beirats (Herausgebergremium). Das *LDV-Forum* erscheint zweimal im Jahr, der Bezug ist für *GLDV-Mitglieder* im Jahresbeitrag enthalten. Der *GLDV-Newsletter* (gldv-nl) ergänzt das *LDV-Forum* als ein schneller, elektronischer Info-Dienst. Beiträge mit Bezug zur Computerlinguistik, Fachinformationen, Kurzrezensionen von Büchern und Software, Veranstaltungsinformationen, Anfragen, Diskussionsanregungen, Stellungnahmen, Adressen, etc. geben über den Moderator (via *electronic mail*) an alle Mitglieder.
Die *GLDV* und der Georg Olms Verlag Hildesheim haben eine eigene Buchreihe *Sprache und Computer* (Hrsg.: P. Hellwig, J. Krause), die Mitglieder preisermäßigt beziehen können. Besondere *GLDV-Publikationen* zur Computerlinguistik in Deutschland – wie den *Studienführer LDV/CL für die deutschen Universitäten* (2. Auflage: 1991, 95 S., DM 10,–) und die *Ausbildungsprofile von CL-Lehrangeboten* (1991, 132 S., DM 15,–) – können über den 2. Vorsitzenden der *GLDV* (Prof. Dr. Haller, IAI, Postfach 1150, 66041 Saarbrücken) bezogen werden.

*Aufnahmeanträge senden Sie bitte an Herrn Lenders, die genaue Anschrift finden Sie in diesem Band auf Seite 2 des Editorials

EINFÜHRUNG IN DIE COMPUTERLINGUISTIK
Vermittlung computerlinguistischer Grundlagen im Rahmen eines sprachwissenschaftlichen Studiengangs

Uta Seewald
Romanisches Seminar
Universität Hannover

Der Lehrveranstaltungstyp der Einführung kennzeichnet im Rahmen zahlreicher Studiengänge Veranstaltungen, die sich in der Regel an Studienanfänger richten und Grundlagen des jeweiligen Studienfaches vermitteln. Es erstaunt deshalb nicht, daß computerlinguistische Einführungen vor allem in Lehrprogrammen solcher Hochschuleinrichtungen angeboten werden, bei denen das Fach Computerlinguistik als eigenständiges Studienfach etabliert ist oder zumindest im Rahmen eines N ebenfachstudiums betrieben werden kann.

Ohne Zweifel besteht Interesse an computerlinguistischen Fragestellungen jedoch auch über diese Studiengänge hinaus, insbesondere im Rahmen linguistischer Studiengänge sowie der sprachwissenschaftlichen Zweige der Philologien. Verschiedentlich geführte Diskussionen um Ziele und Zukunft der Computerlinguistik führten bei zahlreichen Diskutanten auch zu der Erkenntnis, daß eine linguistische Ausbildung langfristig nur schwer ohne computerlinguistische Anteile vorstellbar ist.

Diese Situation war im Rahmen der Romanistik an der Universität Hannover vor vier Jahren Anlaß, das Angebot an sprachwissenschaftlichen Seminaren um computerlinguistische Veranstaltungen zu erweitern. Die angebotenen Veranstaltungen, die von Anfang an auch Studierenden anderer Philologien offenstanden, waren schließlich auch Ausgangspunkt für die Einrichtung eines sogenannten Anwendungsfaches "Romanistische Linguistik" für Informatiker.

Ein Ausweis für den Mangel an derartigen Veranstaltungen auch im Rahmen anderer Studienfächer ist die Tatsache, daßdie an der Universität Hannover angebotenen computerlinguistischen Einführungsveranstaltungen von einem sehr heterogenen Publikum besucht werden, das sich sowohl aus Romanisten, zum Teil auch Anglisten und Germanisten, als auch aus Informatikern zusammensetzt. Während die Studenten der philologischen Fächer den Computer - wenn überhaupt - nur als Schreibinstrument kennen und in der Regel nicht über die Kenntnis einer Programmiersprache verfügen, operieren die Informatikstudenten, die diese Veranstaltung meist im 3. Studiensemester besuchen, seit Studienbeginn mit dem Computer und haben bereits Kenntnisse in mindestens zwei Programmiersprachen. Dafür verfügen sie zum Zeitpunkt des Besuchs dieser Veranstaltung nur über geringe oder gar keine linguistischen Kenntnisse.

Da der fehlende Umgang mit dem Instrumentarium Computer und Programmiersprache die Formulierung der für alle Teilnehmer gleich gestellten Aufgaben wesentlich erschwert hat, werden die Studenten der philologischen Fächer in einer der eigentlich computerlinguistischen Einführungsveranstaltung vorausgehenden Veranstaltung mit informatischen bzw. programmiertechnischen Grundlagen ver-

traut gemacht. Bei den programmiertechnischen Hilfsmitteln handelt es sich um die Vermittlung der Programmiersprache LISP, die auch im Rahmen des Informatikstudiums von den Informatikstudenten erlernt wird. 1

Im folgenden werden Konzeption und Inhalte der Einführung in die Computerlinguistik skizziert, wie sie derzeit an der Universität Hannover für Romanisten und Informatiker angeboten wird:

1. Aufgaben und Anwendungsbereiche der Computerlinguistik
2. Segmentierung - "vom Text zum Graphem"
3. Lexikon: Modelle und Implementierung
4. Morphologische Analyse und Generierung
5. Parsingverfahren
6. Syntaktische Analyse mit Netzwerkgrammatiken (RTN und ATN)

(1) Da die Teilnehmer der Veranstaltung meistens nur sehr schemenhafte Vorstellungen davon haben, was Computerlinguistik überhaupt ist bzw. mit welchen Aufgabenstellungen man sich in der Computerlinguistik beschäftigt und welche möglichen Anwendungen für Ergebnisse computerlinguistischer Arbeit bestehen, bildet ein Überblick über Forschungsrichtungen und Anwendungsgebiete der Computerlinguistik den Einstieg in die Einführungsveranstaltung. Dieser Überblick wird eingeleitet mit einem Videofilm, der über Forschungszentren und computerlinguistische Projekte informiert.2 Ausgehend von dieser Darstellung werden einzelne Aufgabenstellungen und Anwendungsgebiete, wie *Information Retrieval* und Maschinelle Übersetzung, skizziert. Die Teilnehmer werden auf die Zusammenstellungen in Handke (1988), Hausser (1994), Klavans (1989), Schmitz (1992) und Smith (1991) hingewiesen.

(2) Den ersten Schwerpunkt der Veranstaltung bildet das Thema Segmentierung, daß es sich hierbei um einen der zentralen Prozesse der linguistischen Datenverarbeitung handelt, der den Kern jeder Analyse bildet und Voraussetzung für jede weitergehende Verarbeitung sprachlicher Information ist, sei sie geschrieben oder gesprochen. Hierbei ist von Bedeutung, dass die Studenten ein Problembewußtsein dafür entwickeln, daß jede Information, die sie aus einem Text zur weiteren Verarbeitung extrahieren wollen, mittels Segmentierung isoliert werden kann und daß sich dieser Prozeß auf verschiedenen sprachlichen Ebenen bis hinunter zu einzelnen Phonemen oder Graphemen fortsetzen läßt. Um das Problembewußtsein hierfür zu schärfen, wird zur Übersicht der Artikel "Segmentierung in der Computerlinguistik" , Lenders (1989), behandelt.

Da möglichst viele Überlegungen von den Teilnehmern selbst in gemeinsam erarbeitete oder eigene Lösungen programmiertechnisch umgesetzt werden sollen und das verwendete Programmierwerkzeug hierfür die Programmiersprache LISP ist, wird als begleitende Lektüre und als Arbeitsgrundlage das Buch "Natürliche Sprache: Theorie und Implementierung in LISP" von Handke (1988) herangezogen.

(3) Bei der Verarbeitung natürlicher Sprache ist in der Regel der Zugriff auf ein Lexikon erforderlich, das deshalb einen weiteren Themenschwerpunkt der Einführung bildet. Zunächst wird mit den Teilnehmern die Funktion des Lexikons bei der Verarbeitung natürlicher Sprache erörtert sowie die daraus resultierenden Anforderungen an die Art der Information, die das Lexikon enthalten muß. Da eine der Aufgabenstellungen der Computerlinguistik, die im Grenzbereich der Künstlichen Intelligenz angesiedelt ist, die Simulation menschlicher Sprachverstehensprozesse ist, wird an dieser Stelle zunächst eine Parallele zwischen dem mentalen Lexikon und dem Computerlexikon gezogen. Einzelne Modelle, die die Psycholinguistik für die

1 Als Lehrbuch wird den Teilnehmern das bereits als Standardwerk geltende Buch von Winston/Horn (1987) empfohlen.
Die Entscheidung für LISP oder PROLOG ist arbiträr. Bei einer Wiederholung der Veranstaltung soll anstelle von LISP die Programmiersprache PROLOG eingesetzt werden.

2 „Computer und Sprache". IBM, Videothek der Informationsverarbeitung, 1988.

die Repräsentation des mentalen Lexikons entwickelt hat, werden in Handke (1988) dargestellt. Da dort die Frage der Übertragbarkeit der psycholinguistischen Modelle auf den Computer unmittelbar an die Darstellung der verschiedenen Modelle des mentalen Lexikons anschließt, wird diese Fragestellung anhand des Kapitels "Lexikon", Handke (1988:21-43), mit den Teilnehmern bearbeitet. Weil ein wesentlicher Gesichtspunkt bei der Entscheidung für ein bestimmtes Lexikonmodell in der Computerlinguistik die Effizienz der Darstellung und des Zugriffs ist, wird die Repräsentation des Lexikons als Diskriminationsnetzwerk (*trie structure*) vertieft und auch programmiertechnisch umgesetzt. 3 Ein weiteres Kriterium für die Güte computerlinguistischer Anwendungen ist die Forderung nach einfacher Modifizierbarkeit bestehender Datensammlungen. Um zu einem späteren Zeitpunkt Einträge aus dem Lexikon zu entfernen, Angaben zu verändern oder neue Einträge hinzuzufügen, ist es wichtig, daß das Lexikon bzw. die darin enthaltenen Einträge für den Benutzer (Lexikographen) leicht lesbar sind. Aus diesem Grunde erstellen die Teilnehmer zunächst ein Lexikon, das alle vorgesehenen Einträge listenartig und übersichtlich aufführt. Unter Zuhilfenahme der in Handke (1988) erläuterten LISP-Funktionen wird das Lexikon schließlich in ein Diskriminationsnetzwerk überführt. Um auch von vornher
ein die Modularität als wichtiges Gütekriterium hervorzuheben, wird den Teilnehmern die Aufgabe gestellt, das Lexikon im ersten Schritt als Textdatei anzulegen, diese von LISP in eine Liste einlesen zu lassen und schließlich in eine *trie structure* zu überführen.

(4) Den nächsten thematischen Schwerpunkt des Seminars, bei dem das Lexikon dann erstmalig als Datenbasis eingesetzt wird, bildet die morphologische Analyse sowie die Generierung von Wortformen. Vor der Erstellung eines Analyse und eines Generierungsmoduls, mit dem französische - oder je nach der Zusammensetzung der Teilnehmer auch deutsche, englische oder italienische - Verbformen analysiert bzw. generiert werden sollen, erarbeiten die Teilnehmer in der Diskussion auf der Grundlage vorbereitender Lektüre4 zunächst die Besonderheiten, Vorzüge und Nachteile eines auf einem Vollformenlexikon basierten Verfahrens auf der einen sowie eines auf einem Stammlexikon basierten Verfahrens auf der anderen Seite. Der einfachen programmiertechnischen Umsetzung wegen werden die Teilnehmer zur Implementierung eines paradigmaorientierten Ansatzes5 angeleitet, den jeder Teilnehmer dann eigenständig für eine vorgegebene Liste von Verben erweitert.6 In diesem Zusammenhang wird auch das Phänomen der Allomorphie 7 behandelt und die Alternativen zur Behandlung dieses Phänomens bei der morphologischen Analyse erörtert.

(5) Den letzten Themenbereich des Einführungsseminars bildet die syntaktische Analyse. Obschon das Parsen sprachlicher Einheiten im Sinne von Segmentieren bereits Gegenstand der morphologischen Analyse ist, werden verschiedene Parsingstrategien, wie *bottom-up, topdown, breadth-first, depth-first* , erst an dieser Stelle behandelt.8

(6) Als ein in der linguistischen Datenverarbeitung sehr verbreiteter Formalismus zur syntaktischen Analyse werden schließlich Netzwerkgrammatiken bzw. Netzwerkparser vorgestellt.9 Es wird zunächst

3 Vgl. Smith (1991:111), Sproat (1992:111-113), Handke (1988:43-64) sowie Dei et al. (1990:97108).

4 Guzman et al. (1989), Sproat (1992:1-123), Schaeder/Willee (1989).

5 Bauer (1988:151-163).

6 Als Hilfsmittel zur Systematisierung der Verbformen des Französischen eignet sich *Le nouveau Bescherelle, L 'art de conjuguer* sowie die Flexionstabellen im *Grand Robert,* für die Verbformen des Italienischen *Verbi italiani* von Buratti (1993) sowie die von Cappelletti (1990) in der Collection Bescherelle erschienenen *8000 verbes italiens* und die Übersicht in Dardano/Trifone (1985).

7 Bauer (1988:13-16).

8 Zur Lektüre werden Hellwig (1989a), Klavans (1989), Lenders/Willee (1986), Sabah (1989) sowie Winograd (1983) herangezogen.

9 Auch an dieser Stelle kann zur Einführung in die Thematik und gleichzeitig zur Wiederholung bisher behandelter Themen ein Videofilm eingesetzt werden, der unter dem Titel "Natürlichsprachliche Systeme" neben Mustererkennung auch eine Präsentation von ATN-Grammatiken bietet. ("Einführung in die Künst-

ein einfaches Netzwerk behandelt, das von den Teilnehmern dann in ein LISPProgramm zur Analyse ausgewählter einfacher syntaktischer Phänomene umgesetzt wird. Im nächsten Schritt werden dann rekursive Netzwerkgrammatiken (RTN) vorgestellt. Die Teilnehmer erhalten die Aufgabe, das von ihnen bisher erarbeitete Programm zu modifizieren und daraus eil}.e rekursive Netzwerkgrammatik zu erstellen.

Sofern ausreichend Zeit zur Verfügung steht, wird die Aufgabe gestellt, ein syntaktisches Problem zu bearbeiten, dessen Lösung mit Hilfe eines RTN nicht möglich ist. Im Französischen bieten sich hier u. a. Sätze im *Passe compose* an, bei denen sich das *participe passe* (Partizip Perfekt) in Genus und Numerus nach dem ihm vorausgehenden pronominalisierten oder durch *que* Anschluß aufgenommenen direkten Objekt richtet. Die Einsicht in die Notwendigkeit eines" Gedächtnisses", das die kategoriale Information über das Subjekt aufnimmt, leitet dann die Darstellung erweiterter Übergangsnetzwerke (ATN) ein, deren Behandlung den Abschluß der Veranstaltung bildet. 10

Bei den bisher durchgeführten Veranstaltungen hat sich gezeigt, daß der parallele Besuch einer einstündigen Programmierübung in LISP eine nützliche Ergänzung für die informatisch nicht oder nur wenig vorgebildeten Teilnehmer ist. Diese Programmierübung hat besonderen Effekt, wenn für die Betreuung der Teilnehmer zusätzlich eine studentische Hilfskraft zur Verfügung steht, die den Studenten bei der Programmerstellung, Fehlersuche und Problemlösung Hilfestellung gibt.

Die hier skizzierte Einführungsveranstaltung in die Computerlinguistik hat sich ins- . besondere für die Teilnehmer der sprachwissenschaftlich-philologischen Fächer als von besonderer Bedeutung erwiesen, daß sie durch die programmiertechnische Umsetzung der gestellten Probleme nach entsprechender Programmierarbeit zum einen sichtbare Ergebnisse erzielen und sich zum anderen ein Instrumentarium aneignen, das auch für andere Fragestellungen nutzbar ist. Darüber hinaus ermöglicht eine Einführung in die Computerlinguistik an Hochschulen, an denen trotz ansonsten ausgebauter Philologien Angewandte Sprachwissenschaft nicht vertreten ist, mit der Computerlinguistik jedenfalls in einen Bereich der Angewandten Linguistik Einblicke zu gewinnen.

liehe Intelligenz 3. Natürlichsprachliche Systeme, Teil 1 und 2, von Robert Trappl, Wien. Spektrum Videothek, 1991).

10 Winograd (1983), Winston (1987), Handke (1988), Habert (1983), Charniak/MeDermott (1985), Smith (1991), Hellwig (1989a, 1989b), Sabah (1989).

Literatur

Bei der hier aufgeführten Literatur handelt es sich um Werke, auf die zuvor im Text verwiesen wird, sowie um weitere Literatur, die bei der oben beschriebenen Lehrveranstaltung als Lektüre empfohlen wird.

Art de conjuguer. (Le nouveau BeschereHe) Bearb. v. D. Langendorf. Frankfurt, Berlin, München: Diesterweg.

Barr, Avron/Feigenbaum, Edward A., (eds.) (1981): The Handbook of Artificial Intelligence, Vol. 1-4, Los Altos: Kaufmann.

Blitori, Istvan S. / Lenders, Winfried / Puschke, Wolf gang (1989): Computational Linguistics. Computerlinguistik. Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen. Berlin: Walter de Gruyter.

Bauer, Laurie (1988): Introducing Linguistic Morphology. Edinburgh: Edinburgh University Press.

Buratti, Rosalia (1993): Verbi italiani. Mailand: Garzanti Editore

CappeHetti, Luciano (1990): 8000 verbes italiens. Collection Bescherelle. Paris: Hatier.

Charniak, Eugene/McDermott, Drew (1985): Introduction to Artificial Intelligence. Addison- Wesley.

Dardano, Maurizio /Trifone, Pietro (1985): La lingua italiana. Bologna: Zanichelli.

Deiss, Klaus/Handke, Jürgen/Meyer, Bodo (1990): Professionelles Programmieren mit LISP. Hamburg: McGrawHili.

Görz, Günther (1988): Strukturanalyse natürlicher Sprache. Bonn: AddisonWesley.

Gregor, Bernd/Krifka, Manfred (1986): Computerfibel für Geisteswissenschaften. München: C. H. Beck.

Guzman, V. P. De/O'Grady, W./ Aronoff, M. (1989): Morphology: The Study of Word Structure. In: O'Grady et al. (Hrsg.): Contemporary Linguistics. New York: St. Martin's Press, S. 89-122.

Habert, Benoit (1983): Un regard sur les ATN. Manchester, CCL/UMIST Report No. 83/6.

Handke, Jürgen (1987): Sprachverarbeitung mit LISP und PROLOG auf dem PC. Wiesbaden: Vieweg.

Handke, Jürgen (1989): Natürliche Sprache: Theorie und Implementierung in LISP. Hamburg: McGraw-Hill.

Handke, Jürgen (1994): Zugriffsmechanismen im mentalen und maschinellen Lexikon. In: Börner, Wolfgang/Vogel, Klaus (Hrg.): Kognitive Linguistik und Fremdsprachenerwerb. Das mentale Lexikon. Tübingen: Narr, S. 89106.

Hellwig, Peter (1989a): Parsing natürlicher Sprachen: Grundlagen. In: Batori/Lenders/Puschke (1989), S. 348-377.

Hellwig, Peter (1989b): Parsing natürlicher Sprachen: Realisierungen. In: Batori/Lenders/Puschke (1989), S. 378430.

Hausser, Roland (1993): Aufgaben der Computerlinguistik. In: LDV-Forum 10/2 S. 63-77.

Hausser, Roland (in Vorbereitung): Grundlagen der Computerlinguistik. Erlangen.

Klavans, Judith (1989): Computational Linguistics. In: O'Grady et al. (Hrsg.): Contemporary Linguistics. New York: St. Martin's Press, S. 413-445.

Krifka, Manfred (1988): Die Sprache der Computer und die Sprache der Menschen. In: Forum für interdisziplinäre Forschung 1/1988, S. 22-27.

Lenders, Winfried (1989): Segmentierung in der Computerlinguistik. In: Batori/Lenders/-Puschke (1989), S. 159-166.

Lenders, Winfried /WilIee, Gerd (1986): Linguistische Datenverarbeitung. Ein Lehrbuch. Opladen: Westdeutscher VIg.

Reyle, Uwe/Rohrer, Christian, Hrg. (1987): Natural language parsing and linguistic theories. (Studies in Linguistic and Philosophy). Dordrecht: D. Reidel Publishing Company.

Rolshoven, Jürgen/Seelbach, Dieter, Hrg. (1991): Romanistische Computerlinguistik. Tübingen: Niemeyer.

Sabah, Gerard (1989): L'intelligence artificielle et le langage. Processus de comprehension. Paris: Hermes.

Schaeder, Burkhard/Willee, Gerd (1989): Computergestützte Verfahren morphologischer Beschreibung. In: Batori/Lenders/Puschke (1989), S. 188203.

Schmitz, Ulrich (1992): Computerlinguistik. Eine Einführung. Opladen: Westdeutscher VIg.

Smith, George, W. (1991): Computers and Human Language. New York/Oxford: Oxford Univ. Press.

Sproat, Richard (1992): Morphology and Computation. Cambridge/Landon: MIT Press.

Winograd, Terry (1983): Language as a Cagnitive Pracess, Val 1: Syntax. New York: Addison-Wesley.

Winston, Patrick Henry (1987): Künstliche Intelligenz. Bonn: AddisonWesley.

Winston, Patrick Henry/Horn, Berthold Klaus (1987): LISP. Bann: Addisan- Wesley.


Gesellschaft für
Linguistische
Datenverarbeitung
9. GLDV-Jahrestagung
30. 3. – 31. 3. 1995
UNIVERSITÄT REGENSBURG
20 Jahre GLDV
Angewandte
Computerlinguistik

Jorna/van Heusden/Posner (Hrsg.): Signs, Search and Communication. Semiotic Aspects of Artificial Intelligence. Walter de Gruyter. Berlin/New York 1993, 378 S.

In den letzten Jahrzehnten zeichnen sich Konturen eines neuen multidisziplinären Forschungsfeldes ab, das sich u. a. mit solchen Themen wie Informationsverarbeitung, Symbolmanipulation, Repräsentation, Komputation, Problemlösen, Kommunikation und Wissen beschäftigt. Anliegen dieser Forschungsrichtung, die als Kognitionswissenschaft bezeichnet wird, ist es, die natürliche Intelligenz biologischer Organismen, insbesondere von Menschen, und die künstliche Intelligenz menschengemachter Maschinen, darunter vor allem von Computern, zu erforschen. Zur Realisierung der Aufgabenstellung der Kognitionswissenschaft werden mindestens zwei Modelle angeboten. Das eine basiert auf der Auffassung, daß sowohl der menschliche Geist als auch Computer gleichermaßen als informationsverarbeitende Systeme zu verstehen sind. Dabei wird von der Voraussetzung ausgegangen, daß Informationsverarbeitungsprozesse an Symbole als Repräsentationen kognitiver Prozesse gebunden sind. Man könnte dieses Modell auch als semiotisches Paradigma der Kognitionswissenschaft bezeichnen. Ein diesem Modell entgegengesetzter Ansatz ist der Konnektionismus. Er geht von der Annahme aus, daß Gehirne neuronale Netzwerke sind, die auf der Grundlage zahlloser weitverzweigter Verknüpfungen arbeiten. Die Beziehungen zwischen Neuronengruppen verändern sich durch Erfahrungen und Lernprozesse. In dieser Konzeption spielen Konzepte der Symbolverarbeitung keine wesentliche Rolle mehr. Der Konnektionismus wird manchmal auch als subsymbolisches Paradigma bezeichnet.

Unterschiedliche wissenschaftliche Disziplinen haben mit ihren Untersuchungen zur Herausbildung der Konturen der Kognitionswissenschaft beigetragen, indem sie wichtige Einsichten in das Verständnis kognitiver Leistungen geliefert haben. Zu erwähnen sind hier die Kognitive Psychologie, Kognitive Linguistik, Neurowissenschaften, Computerwissenschaft, Künstliche- Intelligenz- Forschung, Philosophie und Logik. Im Kanon dieser Wissenschaften hat man lange Zeit eine Disziplin vermißt, die sich seit ihrer Entstehung explizit mit der Analyse von Zeichen- oder Symbolprozessen beschäftigt. Es handelt sich um die Semiotik. Erst in den letzten Jahren haben sich Semiotiker mit Aufsätzen und vereinzelt auch mit größeren Arbeiten in die kognitionswissenschaftliche Debatte eingebracht. Einen wichtigen Durchbruch leistet in diesem Kontext die zu besprechende Publikation, die im Ergebnis einer Konferenz in Groningen mit dem Thema Expert Systems} Culture and Semiotics entstanden ist. Sie verfolgt das Ziel, die semiotischen Aspekte der Kognitionsforschung im allgemeinen und der Künstlichen Intelligenz im besonderen zu explizieren. Darüber hinaus geht es den Autoren darum, "Schnittstellen" aufzuzeigen, an denen die mehr empirisch und technisch orientierte community der Künstlichen Intelligenz mit den mehr philosophisch orientierten Semiotikern effektiv zusammenarbeiten kann. (S. 19)

Die Beiträge des Buches sind in drei Abschnitte gegliedert, denen einführende Bemerkungen von R. Jorna und B. van Heusden (Groningen) unter dem Titel Signs} search and communication: Towards an empirical future vorangestellt sind. Die beiden Autoren beginnen ihre Ausführungen mit einer Definition der Semiotik, die Bezug nimmt auf neueste Untersuchungen

über menschliche Kognition und Computersysteme. So besteht nach ihrer Auffassung die Aufgabe der Semiotik als Wissenschaft von den Zeichen- und Symbolprozessen darin, alle Arten der Kommunikation und des Austauschs von Wissen zwischen und innerhalb von Informationsverarbeitungssystemen, wie Menschen, anderen Organismen und Maschinen zu untersuchen. (S. 11) Dabei wollen sie in Analogie zur empirisch orientierten Computerwissenschaft die Semiotik ebenfalls als eine empirische Wissenschaft von Zeichen verstanden wissen. (S. 20) Jorna und van Heusden favorisieren unter dem Blickwinkel der von ihnen gegebenen Definition der Semiotik verständlicherweise den informations- oder symboltheoretischen Ansatz in der Kognitionswissenschaft. Sie machen aber gleichzeitig auch darauf aufmerksam, daß einige Autoren des Bandes den Beitrag der Semiotik zur Kognitionswissenschaft aus der Perspektive des Konnektionismus analysieren.

Der erste Abschnitt der Publikation, in dem fünf Autoren zu Wort kommen, trägt den Titel *Signs and representations.*

J. Pelc (Warschau) diskutiert in seinem Aufsatz *Semiosis, cognition, interpretation* unter Berücksichtigung von Arbeiten der polnischen Schule der Logik und Epistemologie einen pragmatischen Zugang zu semiotischen Problemen. Dieser Zugang impliziert, daß Semiose, Kognition und Interpretation als Handlungen verstanden werden. Interessant sind die Überlegungen von Pelc bezüglich der von ihm unterstellten Analogie zwischen Sprachstrukturen und Strukturen des Geistes. Sie münden in die Behauptung, daß der Computer menschliches Denken auf der Grundlage des Wissens über die Organisation semiosischer Aktivitäten und die Struktur semiotischer Systeme simuliert.

L. Santanella Braga (Sao Paolo) untersucht in ihrem Beitrag *A triadic theory of perception* die Wahrnehmungstheorie von Peirce. Sie weist in diesem Zusammenhang ausdrücklich darauf hin, daß Peirces Theorie insofern eine besondere Bedeutung zukommt, als in ihr die dyadische Wahrnehmungstheorie durch eine triadische Theorie ersetzt wird. Dabei scheint ihr das *percipuum,* das von Peirce neben dem Wahrnehmungsinhalt (dem *percept)* und dem Wahrnehmungsurteil als dritter Begriff in seine Wahrnehmungstheorie eingeführt wird, von besonderer Bedeutung zu sein. Anknüpfend an die triadischen Untersuchungen von Peirce schlägt die Autorin eine dreistufige Gliederung des *percipuum* vor.

E. M. Barth (Groningen) befaßt sich in ihrem Aufsatz *Systems of logical representation and inference: An empiricist approach to cognitive science* mit der Vielfalt von Systemen logischer Formen in der menschlichen Kognition und Sprache. Ganz im Sinne der einleitenden Bemerkungen von Jorna und van Heusden versucht sie, sowohl die Aufgaben der Semiotik wie diejenigen der Logik unter empirischen Aspekten zu analysieren. Ausgehend von Problemen, die im Kontext asymmetrischer binärer Relationen entstehen, entwickelt Barth ein neues Schema für eine adäquate Analyse solch wichtiger Kategorien wie *Nichtübereinstimmung, Mißverständnis* und *Konflikt.*

Der Artikel von J .G. Meunier (Montreal) trägt den Titel *Semiotic primitives and conceptual representation of knowledge.* Meunier geht von der sehr provokanten, aber äußerst interessanten Behauptung aus, daß sowohl der symboltheoretische wie auch der konnektionistische Ansatz in der Kognitionswissenschaft hinsichtlich der Wissensrepräsentation auf das semiotische Paradigma nicht verzichten können. Diese Behauptung findet seiner Meinung nach ihren Niederschlag in der Annahme, daß eine Repräsentation zum Ausdruck bringt, daß "etwas für etwas anderes steht" (S. 85) oder daß etwas auf etwas anderes bezogen ist. Nicht zu übersehen ist, daß der Begriff *Repräsentation* eng mit dem Begriff der Konstruktion verknüpft ist. Abschließend diskutiert der Autor die Frage, ob ein semiotisches System nicht-konzeptuell und dennoch adäquat für die Wissensrepräsentation sein kann.

M. F. Peschi (Wien) macht in seinem Beitrag *Semiotic aspects of neurally based representation of knowledge* auf zahlreiche Probleme aufmerksam, die im Zusam-

menhang mit dem symboltheoretischen Zugang zur Wissensrepräsentation auftreten. Das schwierigste Problem, auf das er sich im weiteren vor allem konzentriert, scheint ihm dasjenige der Bedeutung zu sein. Zur Behandlung dieses Phänomens schlägt PeschI einen interessanten alternativen Weg vor, der seinen Ausgang in der Überlegung nimmt, daß die Bedeutung eines Symbols das Resultat eines Konstruktionsprozesses ist. Dabei wird "die Bedeutung eines Symbols nicht als ein Netzwerk von anderen Symbolen verstanden, sondern vielmehr als Etablierung konstruierter Strukturen, als eine Korrelation zwischen einem externen Objekt, einem externen Symbol und dem internen Zustand des Nervensystems einer nicht-sprachlichen ... Stufe neuronaler Aktivitäten". (S. 103)

Der zweite Abschnitt, der den Titel *Abduction and reasoning in expert systems* trägt, behandelt solche für die Kognitionswissenschaft wichtigen Begriffe wie *Schluß folgern, Suche* und *Problemlösen.* Die fünf Autoren, die in diesem Abschnitt zu Wort kommen, knüpfen an die von dem amerikanischen Philosophen und Semiotiker Peirce getroffene Unterscheidung zwischen drei Arten des Schließens an, die als *Deduktion, Induktion* und *Abduktion* bezeichnet werden. Gleichzeitig weisen sie übereinstimmend auf die besondere Rolle hin, die die Abduktion in alltäglichen Schlußprozessen spielt. Zudem betonen sie zu Recht die wichtige Funktion der Abduktion in Expertensystemen, d. h. in Computerprogrammen, die mit dem Ziel entworfen werden, Prozesse des Schließens menschlicher Experten so gut wie möglich zu imitieren.

J. C. A. van der Lubbe und E. Backer (Delft) untersuchen in ihrem Aufsatz *Human-like reasoning under uncertainty in expert systems* die Beziehungen zwischen den oben erwähnten Schlußarten insbesondere unter den Bedingungen unbestimmter (vager) Informationen. Dabei versuchen sie herauszufinden, wie sowohl diese Typen des Schließens als auch die verschiedenen Methoden zur Regelerzeugung in gegenwärtigen Expertensystemen Anwendung finden können. Überzeugend weisen sie nach, daß die Implementation vor allem des abduktiven Schließens in Expertensystemen ein äußerst kompliziertes Problem ist, seine Lösung aber sehen sie gegenwärtig als eine der größten Herausforderungen bei der Entwicklung von Expertensystemen an.

A. H. Marostica (Los Angeles) führt in ihrem Beitrag *Abduction: The creative process* die Überlegungen von van der Lubbe und Backer über Abduktion weiter. Sie konzentriert sich dabei vor allem auf die Analyse der Beziehung zwischen Abduktion und Methoden zur Problemlösung. Die Abduktion ist ihrer Meinung nach im Unterschied zum Problemlösen ein kreativer Prozeß. Aus dieser Annahme zieht sie nun folgerichtig die Konsequenz, daß die Abduktion qualitative Verfahren verwendet und daß sie demzufolge keinesfalls nur ein mechanisches Verfahren ist, wie es für Problemlösungsmodelle charakteristisch ist.

G. Luger und C. Stern (Albuquerque) beginnen ihren Aufsatz *Expert systems and the abductive circle* mit der Behauptung, daß Expertensysteme zunehmend an Bedeutung gewinnen bei der Unterstützung menschlichen abduktiven Problemlösens. Die Grundidee ihrer Überlegungen ist die Ansicht, daß Expertensysteme symbolverarbeitende Architekturen verwenden. Damit sind natürlich diese Architekturen semiotischer Natur. Ihre Darlegungen münden in den Vorschlag, einen zeichenbasierten Zugang zur abduktiven Erklärung zu wählen, der ganz offensichtlich einen signifikanten Vorteil gegenüber dem gegenwärtigen entailmentbasierten Zugang liefert.

B. G. Silverman (Washington, D.C.) betont in seinem Beitrag *Gan machines compensate for biased human reasoning? Gase studies in medical decision making* eine interessante Dimension der allgemeinen Praxis von Expertensystemen. Normalerweise dienen Expertensysteme dazu, den Benutzer zu unterstützen. Silverman aber fordert, daß diese Systeme auch eine Funktion als Kritiker ausüben sollten. Damit Expertensysteme die geforderte kritische Funktion erfüllen können, ist es seiner Ansicht nach notwendig, das *knowledge engineering* durch das "critic engineering" zu

ergänzen.

Mit H. Vissers (Tilburg) Aufsatz *Procrustes, or the future 01 ftexibility* wird der zweite Teil der vorliegenden Publikation beendet. Visser widmet sich dem in der Künstlichen Intelligenz viel diskutierten Problem der Formalisierung des commonsense-Schließens. Er kann sich mit einer derartigen Formalisierung nicht anfreunden. Die Begründung für seine skeptische Haltung leitet er daraus ab, daß seiner Meinung nach die Befürworter der Formalisierung ihre Überlegungen nur auf die Linguistik und Logik gründen, aber dabei die Kombination von natürlicher Sprache und Psychologie außer acht lassen. Ferner betont er, daß die einzig richtige Position in diesem Kontext diejenige ist, nach der Intelligenz aus Wissen plus Handlung, Flexibilität und *common sense* besteht.

Die Beiträge der Autoren des dritten Abschnittes der vorliegenden Publikation, sind - wie schon der Titel *Communication with expertsystems* anzeigt - um die Untersuchung der Beziehung zwischen Kommunikation und Expertensysteme gruppiert. Dabei wird die Rekonstruktion der grundlegenden Prozesse in der menschlichen Kommunikation unter dem Blickwinkel der Konstruktion solcher Prozesse in der Mensch-Maschine-Kommunikation und der Maschine-Maschine-Kommunikation durchgeführt (vgl. S. 262).

R. Posner (Berlin) verfolgt mit seinem Beitrag *Believing, causing, intending: A hierarchy 01 sign concepts* die Absicht, eine Antwort auf die Frage zu geben, was sich in der Kommunikation zwischen Menschen ereignet. Die Beantwortung dieser Frage ist eine unabdingbare Voraussetzung für die Analyse der Kommunikation zwischen Informationsverarbeitungssystemen in der Künstlichen Intelligenz. Da Posner zufolge Kommunikation an Zeichenprozesse gebunden ist, steht eine Untersuchung der letzteren im Mittelpunkt seines Aufsatzes. Diese Zeichenanalyse verdient nun insofern besondere Aufmerksamkeit, als sie auf der Basis von nur drei Grundbegriffen der intensionalen Logik - *glauben, verursachen, intendieren* - eine Hierarchie von Zeichentypen einführt. Diese Hierarchie umfaßt die Zeichen *Signal, Anzeichen, Ausdruck* und *Geste.* Auf dieser Zeichenhierarchie konstruiert der Verfasser in Analogie zu Searles fünf Sprechakttypen Kommunikationstypen, denen die Formulierung entsprechender Kommunikationsbedingungen folgt. Posner betrachtet "die Implementation der gegenwärtigen Hierarchie von Zeichenprozessen in einen Computer als ein Testverfahren für die Konsistenz" (S. 263) seiner Überlegungen.

A. Müller (Göttingen) unternimmt in seinem Aufsatz *On knowledge representing interacting systems* den interessanten Versuch, zu zeigen, inwiefern die beiden folgenden Entwicklungen der Kognitionswissenschaft mit wichtigen semiotischen Fragen verbunden sind: es handelt sich zum einen um das Problem der Adaption und zum anderen um den Begriff der Koordination (distributiertes Schließen). Müller beginnt seine Betrachtungen mit der nicht unbedeutenden Frage, wie zwei fremde Personen in einer fremden Umgebung miteinander kommunizieren können. In semiotischen Begriffen ausgedrückt handelt es sich um das Problem der Semiogenese, d.h. um die Frage, wie Zeichen entstehen und wie eine semiotische Relation zwischen zwei intelligent Handelnden erzeugt wird. Zur Beantwortung der aufgeworfenen Frage bedient sich der Autor der Theorie der Parallelverarbeitung sowie der Theorie neuronaler Netzwerke.

J. A. Michon (Groningen) erörtert in seinem Beitrag *Implementing a sense 0time in intelligent systems* ein scheinbar nebensächliches Problem in bezug auf Expertensysteme - und zwar das Problem der Zeit. Ihm scheint diese Frage nun aber überhaupt nicht nebensächlich zu sein, denn - so argumentiert er überzeugend -, "wenn wir nicht wissen, welcher Teil der Informationsverarbeitungsarchitektur in der Verarbeitungszeit enthalten ist, dann sind wir nicht in der Lage, eine realistische Kommunikation zwischen natürlichen und künstlichen intelligenten Systemen zu erreichen" (S. 318). Empirische Befunde würden deutlich zeigen, daß der Zeitablauf der Interaktion sehr entscheidend für eine erfolgreiche Kommunikation ist. Aus diesem Grunde erachtet es Michon für erfor-

derlich, einen Sinn von Zeit in künstliche Systeme zu implementieren.

Y. Vogelenzang und J. de Vuyst (Groningen) widmen sich mit ihren Überlegungen, die unter dem Titel *Towards an evaluation tool for natural language interfaces* stehen, zwar einer der interessantesten Fragen der Kognitionswissenschaft, aber auch - wie sie selbst betonen - einer der schwierigsten Fragen. Es handelt sich um die Entwicklung von natürlichsprachlichen Schnittstellen, in deren Ergebnis eine gewisse "natürliche" Kommunikation zwischen Menschen und Computern möglich ist. Im Verlaufe ihrer Darlegungen betrachten sie verschiedene Versuche zur Entwicklung eben solcher Schnittstellen und geben eine Klassifikation der dabei verwendeten linguistischen Formalismen an.

F. Rastier (Paris) konzentriert sich in seinem Beitrag *The linguistic analysis of expert texts* auf zwei Themen: erstens erläutert er das Ziel einer "linguistic ergonomics" (linguistischen Ergonomik). Dabei verweist er darauf, daß die Hauptströmungen der Linguistik sich zu sehr auf idealisierte Vorstellungen der Sprache konzentriert hätten. Er hingegen favorisiert einen ergonomischen Zugang insbesondere zu Texten, weil dieser kognitive Aspekte menschlicher Experten einschließt. Zweitens testet Rastieranhand von praktischen Beispielen seine theoretischen Überlegungen.

M. H. Chignell (Toronto) diskutiert in seinem Aufsatz *Cooperative human machine reasoning: Communication through the user interface* ähnlich wie bereits Vogelenzang und de Vuyst Kommunikationsprozesse zwischen Menschen und Maschinen bei Software-Anwendungen. In diesem Kontext fordert er bezüglich des Entwurfs von Benutzerschnittstellen ein konzeptuelles Modell, ein Modell also, das mehr aufgaben-orientiert und aufgabenabhängig ist. Ausführlich beschreibt Chignell ein von ihm entworfenes Modell einer Schnittstelle. Aus semiotischer Sicht im speziellen und aus kognitionswissenschaftlicher Sicht im allgemeinen ist von besonderem Interesse seine Diskussion über visuelle Semiotik und den naturalistischen Entwurf von Ikons.

Hatte sich die Kognitionswissenschaft viele Jahre nahezu hinter dem Rücken der Semiotik entwickelt, so haben die Autoren des vorliegenden Bandes mit ihren Aufsätzen einen Beitrag zur Überwindung dieses Defizits geleistet. Dies ist ihnen zum einen durch die Einbeziehung und Weiterführung von Ideen solcher Klassiker der Semiotik wie Locke, Leibniz, Peirce, Morris, Hjemslev und Goodman gelungen. Deren zeichentheoretische und philosophische Ansätze waren im kognitionswissenschaftlichen Diskurs bislang unverständlicherweise völlig unterbelichtet. Zum anderen haben sie durch die Aufhellung eines solch wichtigen Zeichenprozesses wie dem der Kommunikation (im weitesten Sinne) auch neue Zugänge zur Lösung solcher Probleme aufgezeigt wie denen der semiotischen Wahrnehmung, der Erzeugung von Zeichen, der Repräsentation von Symbolen als Resultat von Konstruktionsprozessen, des abduktiven Schließens in Expertensystemen, der Analyse des *common sense,* der Untersuchung von *Zeit* in Expertensystemen sowie der Rolle von Ikons in der visuellen Kommunikation. In diesem Kontext ist wohltuend zu bemerken, daß sich nahezu alle Autoren dessen bewußt sind, daß diese Probleme nur unter Einbeziehung von Resultaten anderer kognitionswissenschaftlicher Disziplinen zu lösen sind. Mit dieser Ausrichtung leistet der Band auch einen wertvollen Beitrag zu der erfolgreich geführten Reihe *Grundlagen der Kommunikation und Kognition,* die sich der Publikation interdisziplinär orientierter Arbeiten verpflichtet fühlt .

Prof. Dr. Evelyn Dölling
Arbeitsstelle für Semiotik
Technische Universität Berlin
Sekr. TEL 6 Ernst-Reuter-Platz 7
10587 Berlin

Kolloquium

„Theorie der Semantik und Theorie der Lexikographie — Angewandte Semantik und Praxis der Lexikographie"

AK – Lexikographie

27. und 28. Januar 1995
im IKP der Universität Bonn

Leitung: Nico Weber

Interessentinnen / Interessenten, die an dem Kolloquium teilnehmen möchten, bitten wir, sich unter Angabe von Namen, Adresse und Telefonnummer per E-Mail oder Brief anzumelden. Sie erhalten dann weitere Informationen.

Hotelinformationen schicken wir auf Anfrage. Wir sind zu erreichen:

Poppelsdorfer Allee 47
53225 Bonn

Telefonisch unter:
(0228) 73 56 44 (Jens Ostermann / Nico Weber)
73 56 21 (Monika Braun / Jens Ostermann)
73 56 38 (Sekretariat des IKP, Frau von Neffe)

über Fax: (0228) 73 56 39

über E-Mail: upk004@ibm.rhrz.uni-bonn.de

Vorläufiges Programm im Überblick:

⇒ Henning Bergenholtz (Aarhus): Verteilung der enzyklopädischen Informationen in einem erklärenden Fachwörterbuch.
⇒ Manfred Bierwisch (Berlin): Lexikon und Universalien.
⇒ Gregor Buechel (Köln): Können Verben semantische Relationen markieren?
⇒ Ulrich Engel (Heppenheim): über semantische Relatoren und anderes. Ein Entwurf für künftige Valenzwörterbücher.
⇒ Udo L. Figge (Bochum): Kognitiv orientierte Lexikographie.
⇒ Ulrich Heid (Stuttgart): Lexikalische Beschreibungen und ihre Überprüfung in Textcorpora.
⇒ Manfred W. Hellmann (Mannheim): Lexikographische Erschließung des Wendekorpus.
⇒ Erhard Hinrichs / Heike Winhart (Tübingen): Praktische Semantik in der Computerlinguistik - ein Erfahrungsbericht.
⇒ Ursula Klenk (Göttingen): Zur Beschreibung der Interdependenz von Syntax und Semantik in der HPSG.
⇒ Herbert Kuestner (Berlin): From formula to natural language - automatic paraphrase generation from lexical items.
⇒ Petra Ludewig (Osnabrück): Inkrementelle wissensbasierte Wörterbuchanalysen.
⇒ Stephan Mehl (Duisburg): Semantische Relationen - Akquisition und Repräsentation.
⇒ Burkhard Schaeder (Siegen): Ansichten von Bedeutung: fachsprachliche vs. gemeinsprachliche Semantik.
⇒ Uta Seewald (Hannnover): Wortbedeutungen in Wörterbüchern - Wortbedeutungen in Texten.
⇒ Bernhard Schröder (Bonn): Zum Kompositionalitätsprinzip in der Semantik.
⇒ Astrid Steiner-Weber (Bonn): Zur Beschreibung von Komposita im "Lexikon der byzantinischen Gräzität".
⇒ Rainer Stuhlmann-Laeisz (Bonn): Was ist philosophische Logik der Zeit?
⇒ Nico Weber (Bonn): Thematische Einführung.
⇒ Werner Wolski (Marburg): Konzeption eines Bedeutungswörterbuchs zum Werk von Paul Celan.
⇒ Suzanne Wolting (Düsseldorf): Semantische Repräsentation mehrwertiger Verben in einem vererbungsbasierten Lexikon.
⇒ Gerd Wotjak (Leipzig): Zur Beziehung von lexikalischem Bedeutungswissen und enzyklopädischem Weltwissen.

KONVENS 94 IN WIEN

Vor einigen Jahren beschlossen die Gesellschaften DGfS, GI, GLDV, ITG/DEGA und ÖGAI zweijährlich eine gemeinsame Tagung zum Thema Verarbeitung natürlicher Sprache zu veranstalten. Ziel der KONVENS soll es sein, jeweils einen Querschnitt der aktuellen Forschung auf allen Gebieten der Verarbeitung natürlicher Sprache am Computer zu bieten. Damit ist sie ein Forum etwa der Begegnung zwischen Forschern auf dem Gebiet gesprochener Sprache und der Computerlinguistik, aber auch der Künstlichen Intelligenz und der Kognitionswissenschaft geworden.

Die erste KONVENS, die im September 1992 in Nürnberg stattfand und von der GI organisiert wurde, fand erfreulich großen Anklang. Damit waren die Erwartungen für die Folgetagung recht .hoch. Die KONVENS 94 wurde von der OGAI organisiert und fand vom 27. bis 30. September in Wien statt. Das Niveau war auch diesmal erfreulich hoch, wohl eine Folge davon, daß die insgesamt 38 Beiträge aus mehr als 90 Einreichungen ausgewählt wurden. Die Bandbreite war groß: sowohl von den behandelten Themengebieten als auch von der geographischen Verteilung. Überraschend der relativ hohe Anteil von Teilnehmern aus nicht deutschsprachigen Ländern: insgesamt zehn Beiträge aus Irland, Frankreich, Kanada, den Niederlanden, den USA und der Türkei.

Von den Teilnehmerzahlen her gab es mit insgesamt etwa 120 Teilnehmern einen leichten Rückgang gegenüber der letzten KONVENS. Offensichtlich wird es immer schwieriger, eine Tagung zu besuchen, ohne einen veröffentlichten Beitrag dazu zu leisten. Auf diese geänderten Rahmenbedingungen werden die Veranstalter - auch der KONVENS - in Zukunft reagieren müssen.

Jede KONVENS steht unter einem Schwerpunktthema. In Wien war es "Das Lexikon in der Sprachverarbeitung". Diesem Schwerpunktthema werden insbesondere auch die eingeladenen Vorträge und Tutorials untergeordnet.

Traditionellerweise beginnt die KONVENS mit einem Tag für Tutorials. Der Besuch der Tutorials ist im Tagungspreis enthalten und soll einen Überblick und eine Einführung in ein zum Schwerpunktthema gehöriges Gebiet bieten. Ein Tutorial mit dem Titel "Der Aufbau integrierter Lexika" wurde von Dafydd Gibbon (Univ. Bielefeld) abgehalten, das zweite" Aussprachelexika in der Sprachsignalanalyse und synthese" von Klaus Kohler (Univ. Kiel). Beide Tutorials waren gut besucht.

Am zweiten Tag begann die eigentliche Konferenz mit dem wissenschaftlichen Programm. Am Anfang jedes der drei Tage gab es einen eingeladenen Vortrag, danach die referierten Beiträge in zwei parallelen Sitzungen. Im folgenden ein - notwendigerweise subjektiver und lückenhafter - Überblick über das wissenschaftliche Programm. Ich möchte an dieser Stelle darauf hinweisen, daß es möglich ist, den Tagungsband direkt von der OGAI um AS 300,- zu beziehen (Postadresse: ÖGAI, Postfach 177, A-1014, Österreich).

Der erste eingeladene Vortrag kam von Bran Boguraev (Apple, USA). Unter dem Titel "Computational Lexicography far Natural Language" stellte er eine Methode vor, einen automatischen Index für die Hilfefunktion eines Betriebssystems zu erstellen. Durch einen gelungenen Mix von

Methoden auch aus der Computerlinguistik kann die Funktionalität gegenüber rein strukturellen Methoden stark verbessert werden. Ein gutes Beispiel für ein~ auf den ersten Blick vielleicht wenig spektakuläre, aber sehr nützliche Applikation. Meiner Meinung nach sind solche Anwendungen das beste Rezept, bei Anwendern und Herstellern die Akzeptanz für natürlichsprachige Produkte und Produktfeatures zu erhöhen.

Am Donnerstag sprach Hermann Ney über "Die Funktion des Lexikons in der maschinellen Spracherkennung" . Dabei gab er im wesentlichen einen Überblick über den derzeitigen Stand der Kunst in der Spracherkennung. Zum Abschluß seines Vortrags gab es eine Vorführung des bei Philips entwickelten Speechwriters zur Erfassung von medizinischen Befunden. Ein weiteres Beispiel für eine bereits wirtschaftlich nutzbare Entwicklung.

Beim letzten eingeladenen Vortrag von Antje Meyer (MPI Nijmegen) stand der kognitionswissenschaftliche Aspekt im Vordergrund. Sie beschäftigte sich mit dem "Zugriff zum mentalen Lexikon in der Sprachproduktion" . Für mich immer wieder faszinierend, wie in dieser Disziplin durch geschickte Experimente Einblicke in die Sprachverarbeitung beim Menschen gewonnen werden können, die langfristig auch die Methoden der Computerlinguistik beeinflussen und befruchten.

Von den insgesamt 37 begutachteten Vorträgen konnte ich selbst nur einen kleinen Teil besuchen. Mein Eindruck ist daher notwendigerweise lückenhaft und sicher auch subjektiv. Wenn man Trends feststellen möchte: Die "klassischen" Themen wie Parsing und Syntax waren nicht so dominant wie noch vor zwei Jahren. Es gab eine Reihe von Beiträgen zu Morphologie und Lexikon, wohl aufgrund des Schwerpunktthemas. Die Beschäftigung mit Semantik und Diskurs im weitesten Sinne gewinnt an Bedeutung, wobei der Schwerpunkt auf praktischen Lösungen für gut abgegrenzte Gebiete liegt - etwa der Vortrag von Egg und Herweg "A Type Hierarchy for Aspectual Classification". Die international in Mode gekommene Verwendung statistischer Methoden und Korpora hat bei dieser KONVENS noch nicht durchgeschlagen, entsprechende Beiträge kamen meistens aus dem Bereich der Verarbeitung gesprochener Sprache, wo sie traditionell dominieren. Eine Ausnahme war der Vortrag von Sutcliffe und O'Sullivan, der eine Verwendung von WordNet zur Nomenklassifikation präsentierte.

Inwieweit konnte der Anspruch der Begegnung zwischen den unterschiedlichen Disziplinen erfüllt werden? Hier bietet sich ein differenziertes Bild. Der Bereich Gesprochene Sprache war mit acht Beiträgen gut vertreten. Aus der kognitionswissenschaftlichen Ecke dagegen nur ein - allerdings sehr interessanter Vortrag - von Hemforth, Konieczny und Scheepers. Alle anderen aus Computerlinguistik und NLP, wobei die Grenze zwischen diesen Bereichen mittlerweile – erfreulicherweise - sehr unscharf geworden ist.

Neben den Vorträgen gab es noch Posters und eine ganze Reihe von Systemvorführungen. Leider standen nur zwei Rechner zur Verfügung, die sich die verschiedenen Gruppen teilen mußten, was auch zu leichtem Unmut führte. Insgesamt gelang es aber doch, alle geplanten Vorführungen einem interessierten Publikum zu präsentieren. Auch hier war die Bandbreite groß: von der Morphologiekomponente über die Grammatik-Testumgebung zum Merkmalsformalismus und der Verarbeitung von Phraseologismen. Wahrscheinlich sollten künftige KONVENS-Veranstalter mehr Zeit und Ressourcen für diese interessante Art der Präsentation bereitstellen.

Abgerundet wurde das Programm durch zwei Gelegenheiten zum geselligen Beisammensein: am Abend des ersten Tages gab es einen kleinen Willkommens empfang in den Räumen des Österreichischen Forschungsinstituts für Artificial Intelligence und am Donnerstag abends den für Wien obligaten Heurigenbesuch.

Harald Trost
Österreichisches Forschungsinstitut für
Artificial Intelligence
harald@ai.univie.ac.at

KONVENS 94 IN WIEN

"Verarbeitung natürlicher Sprache" aus der Sicht einer Studentin

Wissenschaft und Wien? Was sollte an den "Hidden-Markov-Modellen" reizen, wenn gleichzeitig Peymann an der Burg das "Käthchen von Heilbronn" inszeniert? Würde ich mich wirklich für "die kompositionelle Bildung von Diskursrepräsentationsstrukturen über einer Chart" erwärmen können, während nur wenige hundert Meter weiter Madame Butterfly auf ihren amerikanischen Marineoffizier wartet?

All' diese Fragen stellte ich mir, während ich im Zug nach Wien zu der von der Österreichischen Gesellschaft für Artificial Intelligen ce ausgerichteten Tagung "Konvens '94 - Verarbeitung natürlicher Sprache" saß. Ich, das ist eine 24jährige Studentin und Hilfskraft am Institut für Kommunikationsforschung und Phonetik der Universität Bonn unter Leitung von Prof. Winfried Lenders. Gespannt war ich aber letztlich trotzdem - schließlich war es die erste internationale Tagung meines Studienfaches, an der ich Gelegenheit hatte teilzunehmen: eine seltene und auch einmalige Chance für eine Studentin, "Forschung hautnah" zu erleben.

Mein Ziel war anfangs natürlich bewußt niedrig gesteckt: Ich wollte einfach einmal über den Tellerrand des eigenen Institutgeschehens sehen, was auf der ganzen Welt in dem Fach passiert, das man selber studiert, eventuell mehr über praktische Anwendungsmöglichkeiten erfahren und Ideen für die Zukunft, das Berufsleben, bekommen.

Durch die Organisation im Vorfeld der "Konvens 94" jedoch wurden meine Erwartungen bei weitem übertroffen. Am ersten Tag hatten wir die Wahl zwischen zwei Tutorials als Einstimmung, ich entschied mich für den "Entwurf integrierter Lexika". Hier wurden meine bisher nur theoretischen Kenntnisse über maschinelle Lexika gesprochener Sprache durch die Berichte von Prof. Gibbon über die praktische Arbeit und den Aufbau eines solchen Lexikons erweitert, so daß sich meine vage Vorstellung zu einem echten Bild von der Entwicklung, den Problemen und Grenzen eines solchen Lexikons formierten.

Sodann sollte ich Einblick in das rege Treiben eines internationalen Kongresses bekommen. Nach persönlicher Programmplanung - immerhin standen in den drei Tagen ungefähr 40 Vorträge zur Wahl, die meisten in englischer Sprache gehalten - saß ich gespannt im Auditorium und mußte bald merken, daß hier ein anderer Stil vorherrschte als der an der Uni praktizierte. Dr. Boguraev von Apple Computers stellte seine Erfahrungen aus der maschinellen Lexikographie dar, unterstrichen von vielleicht zu vielen bunten Bildchen, aber immerhin konnte ich sogar die englische Sprache ohne größere Probleme verstehen, was auch während der weiteren Vorträge glücklicherweise so blieb. Und so sollte es auch weiterhin ein Vergnügen sein, den Vortragenden zu folgen, denn die Länge von 1/2 Stunde, die Diskussionen nach den einzelnen Vorträgen, die Cafepausen zwischendurch und vor allem der schon während der Tagung zur Verfügung stehende Tagungsband, in dem alle Vorträge referiert sind, so daß ein Mitschreiben überflüssig war, ermöglichten den Teilnehmern, sich ganz und gar den Re-

feraten zu widmen. Zeitweise irritierte mich zwar das an die Wand projizierte Formelwirrwarr einzelner Vorträge, wahrscheinlich muß man schon sehr lange selbst wissenschaftlich geforscht haben, um die Darstellung natürlicher Sprache in kryptischen Formeln auf Anhieb verstehen zu können, auf der anderen Seite jedoch wurde zum Beispiel mein unklares Bild von einem Interlingua-Modell der maschinellen Sprachübersetzung durch den Vortrag von Bianka Buschbeck-Wolf konkretisiert; zumindest weiß ich jetzt, wie interlinguale Konzepte für räumliche Präpositionen konkret auszusehen haben. Interessant war auch eine praktische Anwendung computerlinguistischer Kenntnisse: Ein maschinelles Diktiersystem von Phillips ermöglicht das automatische Diktieren von Krankenberichten, so daß die Arzthelferin oder Sekretärin nur noch kurz posteditieren muß. Zwar werden hier einmal mehr die Grenzen der maschinellen Spracherkennung sichtbar, denn auch bei trainiertem Sprecher kommt es immer wieder zu Fehlern, aber dieses System kann die praktische Anwendung der Forschung aufzeigen.
Am erstaunlichsten aber für mich war, wie ertragreich das Treffen am Rande der Veranstaltung mit Wissenschaftlern ist, deren N amen stellvertretend für die aktuelle Forschung stehen, Namen, die mir als Studentin bisher nur aus den Literaturangaben bekannt waren. Ich habe durch sie neue Ideen und Ratschläge erhalten, von denen ich einige schon realisieren konnte. So tauschte man Erfahrungen entweder zwischen den Vorträgen oder aber bei den gemeinsamen Abenden aus und als Studentin lauschte ich den langjährigen Erfahrungsberichten, die sehr häufig auch ihre unterhaltsamen Seiten hatten.

Am Freitag stand ich dann, wer hätte das vorher erwartet, eher traurig als erlöst wieder am Bahnhof. Auf der Rückfahrt ordnete ich meine Eindrücke und hatte den Kopf voller neuer computerlinguistischer Ideen für das folgende Semester und meine Magisterarbeit. Natürlich habe ich durch die Vorträge viel hinzugelernt, jedoch war es mir besonders wichtig, Ansprechpartner gefunden zu haben, die mir Anregungen vermittelten und mir Perspektiven für die Zukunft als Computerlinguistin gaben.

Daneben trugen das abendliche Ausklingen beim Heurigen, der entspannte fachfremde Plausch und auch die wienerische Atmosphäre dazu bei, die Konvens aus der Sicht einer Studentin für rundum gelungen anzusehen.

Monika Braun, Univ. Bonn

SIGIR '94

Die diesjährige Konferenz der Special Interest Group Information Retrieval (SIGIR) der ACM fand vom 4. bis 6. Juli 1994 an der Dublin City University unter der Leitung von Alan Smeaton statt. Insgesamt kamen ca. 270 Teilnehmer aus 20 Ländern (inkl. Australien, Südkorea, Singapur, Japan, Slovenien, USA, Canada und den meisten europäischen Staaten). Ca. ein Drittel erhielt Mittel aus dem HCM-Förderfond der EU. SIGIR '94 wurde im Tagungsband von dessen Herausgebern Bruce Croft und Keith van Rijsbergen als die bisher größte SIGIR Konferenz angekündigt, da 35 reguläre Vorträge, zwei eingeladene Vorträge, zwei Panels, acht Tutorials, ein ACM-SIGIR Preisvortrag und mehrere Systemdemonstrationen stattfanden. Trotz dieser angekündigten Größe konnte man allerdings den Eindruck gewinnen, daß die inhaltliche Breite der Vorträge relativ gering war. Vor allem spiegelten sich neue Ansätze und IR-Paradigmen mit wenigen Ausnahmen (z.B. die Vorträge von Peter Ingwersen oder Matthias Hemmje) nur in den beiden Panel-Sitzungen zur Integration von IR und Datenbanken und zur Evaluierung Interaktiver Retrieval Systeme wider. Sowohl in vielen Vorträgen als auch Diskussionsbeiträgen wurde die Dominanz eines sehr traditionellen, engen Verständnisses von IR deutlich.

Enttäuschend war vor allem der eingeladene Vortrag von Jaime Carbonell, der sich im Gegensatz zum Vortragstitel nur sehr rudimentär mit IR und noch rudimentärer mit der Verzahnung IR und Natural Language Processing befaßte und im wesentlichen einen State-of-the-Art zum NLP im Bereich der Maschinellen Übersetzung behandelte. Der eingeladene Vortrag von Dennis Tsichritzis stellte vor dem Hintergrund der aktuellen Hardware-Entwicklung den Wandel des Informationsbegriffs innerhalb des Informationsprozesses von Zahlen, über Daten und Texte zur Vielschichtigkeit multimedialer Information dar und zeigte Bereiche auf, mit welchen sich IR-Leute aufgrund dieser Veränderungen werden auseinandersetzen müssen (z.B. die Erschließung dieser neuen Informationstypen). Insofern hätte dieser Vortrag als richtungsweisend für die Konferenz angesehen werden können, wären ihm weitere, mit ähnlich aktueller Thematik gefolgt.

Nun kurz zu den regulären Vorträgen der Tagung, die sich auf zwölf Sektionen verteilten, wovon einige parallel zueinander oder den Panel-Sitzungen stattfanden.

In der ersten Sektion zur Textkategorisierung fanden vier Vorträge statt. David Lewis und William Gale befaßten sich mit einem Trainingsalgorithmus zum Klassifizieren von Texten, wobei der Schwerpunkt auf dessen Effizienz lag. Yiming Yang stellte mit Expert Network einen Ansatz vor, der auf der Basis menschlicher Entscheidungen ein Netzwerk erstellte, welches die Links zwischen den Knoten aufgrund von Wort- und Kategorienverteilungen im Trainingsset berechnet. In der durchgeführten Evaluierung zeigte sich bzgl. recall und precision eine gegenüber anderen Methoden signifikante Verbesserung. Der Vortrag von Apt, Damerau und Weiss befaßte sich mit einem Textkategorisierungsmodell, das sprachunabhängig arbeitet. Die Experimente hierzu fanden auf der Basis deutscher und englischer Reuters- Texte statt. Rainer Hoch präsentierte das System INFOC

LAS, das im Bereich von Geschäftsbriefen unter Rückgriff auf verschiedene Wissensquellen (z.B. Wortfrequenzstatistik, morphologisches Wissen, typ-spezifische Wortlisten etc.) eine automatische Briefarchivierung vornimmt. In der Indexierungssektion stellte Chris Paice eine Evaluierungsmethode für Wortstammalgorithmen vor, die auch für die Optimierung der Algorithmen von Wert ist. David Eilis, Jonathan Furner-Hines und Peter Willett befaßten sich mit der Fragestellung, wie konsistent manuell vergebene Links in Hypertextbasen ausfallen, wobei besonders die Relation dieser Konsistenz zur Retrievaleffektivität herausgestellt wurde.

Im Vortrag von Ellen Voorhees ging es um die Erweiterung von Anfragen durch lexikalisch-semantisch ähnliche Wörter. Bzgl. Effektivität gemessen auf der Basis der TREC-Kollektion ergab sich, dass die Komplexität der Ausgangsanfrage eine entscheidende Rolle spielte. Die vier Vorträge in der Sektion Benutzermodellierung repräsentierten sehr unterschiedliche Sichtweisen und Ansatzpunkte. Bryce Allen untersuchte experimentell den Einfluß kognitiver Fähigkeiten wie der Wahrnehmungsgeschwindigkeit auf das Lernverhalten und die Suchperformanz von Endbenutzern und die Wirkungsweise dieser Mechanismen. Der Vortrag von Amanda Spink befaßte sich mit möglichen Quellen für Frageerweiterungen im Sinne einer Effektivitätssteigerung und zeigte Möglichkeiten auf, diese innerhalb von Relevance Feedback- Techniken zu integrieren. Brian Logan, Steven Reece und Karen Sparck Jones beschrieben ein theoretisches Modell über Informationsvermittlung basierend auf der KI - Technik der "belief revision". Der Vermittlungsprozeß und das Verhalten von menschlichen Informationsvermittlern wurde in Experimenten untersucht mit der Zielrichtung, Grundstrategien für eine Simulation menschlicher Informationsvermittler herauszuarbeiten. Peter Ingwersen schloß die Sitzung und damit den ersten Konferenztag mit seinem Vortrag über Polyrepräsentation von Benutzerbedürfnis und semantischen Einheiten innerhalb seiner kognitiven Theorie der IR-Interaktion, die durch das Zusammentreffen verschiedener kognitiver Strukturen und Transformationen bestimmt wird, wobei sich der individuelle Benutzer mit bestimmten kognitiven Ausprägungen im Einflußbereich verschiedener Modellelemente befindet.

Die erste Sitzung des 2. Konferenztages mit zwei Vorträgen galt der Theorie und Logik von IR. P.D. Bruza und T.W.C. Huibers stellten ein Verfahren zum Vergleich von Retrievalmechanismen vor, das nicht experimentell, sondern induktiv vorgeht und dabei die Axiome ausfiltert, die den Retrievalprozeß steuern. In der folgenden Präsentation von Fabrizio Sebastiani ging es um eine terminologische Logik zur Modellierung des IR-Prozesses, wobei durch die Einführung eines probabilistischen Elements die Relevanzbeziehung zwischen einem Dokument und einem Benutzerbedürfnis durch Wahrscheinlichkeitsgrade ausdrückbar ist.

Innerhalb der Sektion "Natural Language Processing" war der erste Vortrag von Christian Jacquemin und Jean Royaute der Termidentifikation gewidmet, wobei ein Verfahren beschrieben wurde, das durch partielle Parsing-Techniken aufgrund logischer Regeln und Metaregeln Basisterme und deren Wortformen auffindet. Mark Sanderson präsentierte Ergebnisse aus Studien, die untersuchten, ob korrekte Wortdisambiguierung eine Performanzsteigerung im IR-Prozeß bewirkt. Seiji Miike, Etsuo Itoh, Kenji Ono und Kazuo Sumita stellten ein japanisches Volltext-Retrievalsystem vor, das über eine dynamische Abstract Generierungsfunktion verfügt, wobei der Benutzer die Möglichkeit hat, die Textbereiche selbst festzulegen, die zusammengefaßt werden sollen. Parallel zur NLP-Sektion fanden drei Vorträge aus dem Bereich der statistischen Modelle statt. Der erste von Ijsbrand Jan Aalbersberg befaßte sich mit einem auf der Termfrequenz basierenden Retrievalmodell, wobei der Neuheitswert darin bestand, daß mehrere lokale Ähnlichkeiten zwischen Termrangnummer in der Anfrage und im Dokument in eine globale Ähnlichkeit transformiert werden.

Brian Bartell, Garrison Cottrell und Rik Belew zeigten, daß durch Kombination der Ergebnisse verschiedener Retrievalal-

gorithmen die Retrievalperformanz gesteigert werden kann, was daran liegt, dass unterschiedliche Methoden unterschiedliche Eigenschaften bei der Relevanzbestimmung favorisieren und damit unterschiedliche Schwerpunkte setzen. Weiterhin wurde gezeigt, wie die verschiedenen Methoden kombiniert werden können. Im Vortrag von Joon Ho Lee werden verschiedene, das Boolesche Retrievalmodell erweiternde Verfahren am Beispiel der AND- und 0R- Verknüpfung und der Anfragegewichtung untersucht und durch mathematische Eigenschaften angereichert, um die Retrievaleffektivität zu steigern. Zu Beginn der Evaluierungssektion stellten William Hersh, Chris Buckley und David Hickam eine Reihe von Retrievalexperimenten vor, die zeigten, daß unerfahrene Benutzer aus dem medizinischen Bereich mit einem auf dem Vektor-Raum-Modell basierenden Retrievalsystem gleich gute Ergebnisse erzielten wie erfahrenere Benutzer, die nur über ein Boolesches System verfügten. Aus den Ergebnissen konnte außerdem neues Testmaterial gewonnen werden. Kazem Taghva, Julie Borsack und Allen Condit verglichen in ihrer Studie, wie sich die Retrievaleffektivität verhält, wenn OCR-Texte ohne oder mit manueller Bearbeitung als Dokumentgrundlage verwendet werden. Die Autoren konnten keine signifikante Unterschiede im Hinblick auf recall und precision nachweisen. In dem Vortrag von Howard Turtle konnte unter Hinweis auf die methodischen Probleme gezeigt werden, daß natürlichsprachliche Anfragen im Vergleich zu Booleschen im Bereich juristischer Volltext-Dokumente eine Performanzsteigerung bewirken.

Die bei den regulären Vorträge aus dem Bereich probabilistischer Modelle befaßten sich zum einen mit der Verwendung der logistischen Regression als Dokumentrankingfunktion und deren Evaluierung (Fredric Gey), zum anderen mit Erweiterungen des 2-Poisson Modells durch zusätzliche Variablen und wiederum Modellevaluierung auf der Basis des TREC Testmaterials (S. Robertson und S. Walker). W.S.
Cooper erhielt den Triennial ACM SIGIR Award. Er stellte in seinem Vortrag die kritische Frage, ob der Aufwand, der eingesetzt wurde, um die probabilistische Theorie auf gesunde Beine zu stellen, nicht besser hätte in theoretisch weniger anspruchsvolle Untersuchungen explorativer Art investiert werden sollen. "Time will tell whether the theoretical baggage that accompanies the probabilistic method is more a benefit or an encumbrance", schloß Cooper.

Der 3. und letzte Konferenztag begann mit der Sektion Benutzerschnittstellen. Der erste Vortrag von Matthias Hemmje, Clemens Kunkel und Alexander Willett präsentierte mit LyberWorld ein 3D-System zur Visualisierung verschiedener Elemente des IR-Prozesses (räumliches Navigieren, Aufbau der Suchanfrage, Relevanzbestimmung etc.). Jack Conrad und Mary Hunter Utt berichteten über ein Zugriffs verfahren auf große Textdatenbanken mittels automatisch extrahierter domänenspezifischer Merkmale und deren Ähnlichkeitsbeziehungen, wobei jedoch der Schwerpunkt des Vortrags nicht auf der Gestaltung der Oberfläche lag.

Die Routing-Sektion beinhaltete drei Vorträge. Der Vortrag von Masahiro Morita und Yoichi Shinoda stellte die Ergebnisse verschiedener Experimente und einer Feldstudie dar, in welchen es darum ging, Benutzerinteressen zum Filtern neuer Information zu benutzen. Gegenstand des Vortrags von David Hull war das sog. Latent Semantic Indexing (LSI) als neuer IR-Ansatz, der versucht, die zugrundeliegende Termassoziationsstruktur durch eine Repräsentation der semantischen Faktoren zu modellieren. Experimente zeigen eine extreme Performanzsteigerung, wenn LSI in Verbindung mit statistischer Klassifikation eingesetzt wird. Chris Buckley, Gerard Salton und James Allan befaßten sich mit Relevanz Feedback im Kontext der TREC-Studien. Sie weisen nach, daß ein direkter Zusammenhang besteht zwischen der Effektivität und der Anzahl der aus relevanten Dokumenten hinzugefügten Termen.

In der Sektion Passage Retrieval ging James P. Callan auf eine differenzierte Definition und Rolle der Passagen ein. Experimentelle Ergebnisse wurden vorgestellt, die verschiedene Textsegmente (Paragra-

phen und unterschiedlich große Textfenster) als Passagen definierten sowohl für homogene als auch für heterogene Dokumentkollektionen. Ross Wilkinson ging in seinem Vortrag in die gleiche Richtung und zeigte, daß sich eine Effektivitätssteigerung erzielen läßt, wenn Wissen über die Dokumentstruktur vorhanden ist und verwendet wird. Elke Mittendorf und Peter Schäuble stellten abschließend einen neuen Ansatz vor, der auf der Basis des Hidden Markov Models für Passage Retrieval relevante Textfragmente generiert. Die letzte Sektion der Tagung war Systemimplementierungen gewidmet. Dabei ging es um die Update-Problematik bei IRS (Kurt Shoens, Anthony Tomasic, Hector Garcia-Molina), eine Methode für effizientes Ranking (Michael Persin) und um ein Volltext-IRS aus dem Toubleshooting-Bereich (Peter G. Anick).

Neben den angeführten Vorträgen fanden zwei interessante Panels über die Integration von IR und Datenbanken (Moderation: Norbert Fuhr) und die Evaluierung Interaktiver Retrieval Systeme (Moderation: Susan Dumais) statt. Die Tagungsbeiträge sind im Tagungsband (Croft, W.B., van Rijsbergen, C.J., SIGIR 94. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Springer-Verlag) enthalten. Unbedingt erwähnt werden muß die hervorragende lokale Organisation von Alan Smeaton und seinen Mitarbeitern, das Conference Dinner im Royal Hospital Kilmainharn, wo die Tagungsteilnehmer bei Harfenklängen und Irish Folk zum Mitsingen und -tanzen animiert wurden, und nicht zuletzt die Fußballweltmeisterschaft, welche die Tagung bis zum Ausscheiden der Irischen Nationalmannschaft intensiv begleitete. Wenn es nach uns gegangen wäre, hätten die Iren den World Cup gewonnen!

Christa Womser-Hacker, Regensburg

MEHRWERT INFORMATION - **ISI** '94
Informationswissenschaftliches Symposium an der Universität Graz

Anstelle eines geplanten ausführlichen Tagungsberichtes können hier aus Zeit gründen nur einige wenige Stichpunkte zum Tagungsverlauf gegeben werden.

Diese vom Hochschulverband für Informationswissenschaft (HI) veranstaltete und vor Ort von Prof. Rauch organisierte Tagung war aus mehreren Gründen ein voller Erfolg für alle Beteiligten: Die besondere Atmosphäre in einem historischen Universitätsgebäude und einer adäquaten Teilnehmeranzahl von ca. 250 Personen, die reibungslose Organisation und insgesamt qualitativ gute Beiträge (Proceedings sind im Universitätsverlag Konstanz erschienen).

Einen guten Einstieg hatte die Tagung mit dem Beitrag von Prof. Rauch, der die naive Erwartung problematisierte, daß ein Mehr an Informationstechnik auch ein Mehr an Wirtschaftlichkeit und Erfolg produzieren müsse und der die Frage aufwarf, wie wir als Nutzer von unseren Informationssystemen gestaltet werden. Leider wurde dieser Aspekt, für den er den Begriff der *Informationsdynamik* benutzte, im weiteren Verlauf der Tagung nicht erkennbar aufgegriffen (soweit ich dies sagen kann, denn es gab durchweg zwei parallele Sessions) .

Der Beitrag von Prof. Picot, der die Problematik von Mehrwert Information aus der Sicht der Wirtschaftsinformatik diskutierte, brachte für mich zwei Erkenntnisse: Erstens sollte ich diesen breit angelegten und viele Facetten streifenden Beitrag nochmals nachlesen, wenn er denn schriftlich vorliegen sollte (er fehlt im Tagungsband und ist versprochen) und zum anderen hatte er das gleiche Thema wie der sich an schließende Vortrag von Tegethoff, zu dem er in bemerkenswertem Gegensatz stand. Beide Vortragende beschäftigten sich mit dem Problem von "Verständigung": Picot im Verständnis eines Universitätsprofessors, der eine differenziert ausgearbeitete Vorlesung hält - ohne jeden Versuch, Verständigung durch die Art der Präsentation zu erreichen. Tegethoff als Märchenerzähler mit einer beeindruckenden Übereinstimmung von Inhalt und Form, der das Märchen als alte Kunst der Informationswissenschaft als junge Disziplin gegenüberstellte.

Bei den vielen Beiträgen der Tagung fiel auf, daß sich die meisten Vortragenden bemühten, das zentrale Thema der Tagung, *Mehrwert Information* explizit mit zu berücksichtigen, allerdings ohne daß damit eine tiefere Beschäftigung mit diesem Aspekt verbunden war.

In der Abschlußdiskussion versuchte Prof. Kuhlen als Vorsitzender des HI und als Protagonist des Begriffs vom "informationellen Mehrwert" diese Begrifflichkeit kontrovers diskutieren zu lassen, was ihm nicht zuletzt wegen seines Talentes als offensiver Talkmaster im Publikum und einiger engagierter Kontrahenten mit Erfolg gelang.

Fazit: Die Meßlatte für die nächste ISI '96 liegt hoch!

G.K.

3. INTERNATIONALE FACHTAGUNG FÜR COMPUTEREINSATZ IN DER HISTORISCHEN SPRACHWISSENSCHAFT
Dresden, 7.-8. Oktober 1994

Methoden der linguistischen Datenverarbeitung haben in der historischen Sprachwissenschaft bisher noch kaum Anwendung gefunden, obwohl viele der hier anstehenden Probleme der Archivierung, Manipulation und Analyse von Texten den Computereinsatz geradezu herausfordern. Dies liegt daran, daß die speziellen Probleme der historischen Sprachwissenschaft auch spezielle Anwendungsprogramme erfordern, die jedoch von der marktorientierten Softwareindustrie wegen des beschränkten Abnehmerkreises nicht zur Verfügung gestellt werden, und da die Erstellung maßgeschneiderter Programme zur Lösung von Detailproblernen recht kostspielig ist. Aus diesem Grund hat sich 1991 ein Kreis interessierter Fachvertreter zusammengeschlossen, um in regelmäßigen Abständen einen

Gedankenaustausch zu pflegen. Ein erstes Treffen fand 1992 an der Universität Bam

berg statt, das 1993 an der Universität Prag wiederholt wurde. Der gute Erfolg und die rege internationale Beteiligung haben den Veranstalter bewogen, die diesjährige Veranstaltung an die Technische Universität Dresden zu holen. Bei dieser Gelegenheit etablierte sich dieser bislang informelle Interessenverband als Arbeitskreis der GLDV, wobei als Leiter Jost GIPPERT (Frankfurt) und als sein Stellvertreter Johann TISCHLER (Dresden) gewählt wurden (s. dazu die Kurzbeschreibung von Jost GIPPERT im vorliegenden Heft).

Zum Vortrag kamen folgende Referate:

DURKIN, Desmond (Münster): Verwendung einer Datenbank zur Textanalyse (DataPerfect - Westmitteliranisch)

DVOliA.K, Jan (Prag): Advantage of a nongraphic mode of screen (using VGA character font editor and printing the results)

GANTER, Bernhard (Dresden): Einsatz von TEX zur Wiedergabe von Texten in nichtlateinischen Alphabetschriften.

GIPPERT, Jost (Frankfurt): Zum Stand der Arbeit an der indogermanischen Textdatenbank.

HALLER, Johann (Saarbrücken): Maschinelle Übersetzung - Linguistischer Hintergrund und Beispiele aus kommerziellen und aus Forschungssystemen.

MARATSCHNIGER, Martina (Klagenfurt): Statistische und kartographische Aufarbeitung der Infinitivprominenz in Norditalien und im angrenzenden slawischen Sprachgebiet.

RAHMAN, Furat (Prag): Sumerische Zeichenliste im Computer. Einige Erwägungen zur Umschrift und Bearbeitung.

SCHANZE, Helmut (Siegen): Aufbau und Handhabung einer (Bild-) Datenbank

TISCHLER, Johann (Dresden): CD-ROM und Historische Sprachwissenschaft: Griechisch, Lateinisch, Hethitisch.

VAVROUSEK, Petr (Prag): Einige Erwägungen zu einer künftigen Computer-Datenbank zur Erforschung des Hethitischen.

ZEMANEK, Petr (Prag): Hypertext und die Bearbeitung und Analyse der Bibel.

ZEILFELDER, Susanne (Jena): Juristische Aspekte der Indogermanischen Textdatenbank.

Jost Gippert
Johann Wolfgang Goethe-Universität
Frankfurt am Main

9. GLDV Jahrestagung 1995 30.-31.3.1995, Univ. Regensburg

Die GLDV wurde vor genau zwanzig Jahren unter dem Namen ldv-fittings gegründet. Die Ziele des Vereins waren damals die Förderung der Linguistischen Datenverarbeitung v. a. durch den Austausch von EDV-know-how und Programmen. Mit der Etablierung der Computerlinguistik als Wissenschaftsgebiet wurde die GLDV zum Fachverband der Computerlinguisten.

Jubiläen sollten auch dazu dienen, sich kritisch damit auseinanderzusetzen, ob die bei der Gründung des Vereins gesteckten Ziele sinnvoll waren, ob und wie sie erreicht wurden und wie weit eine Revision notwendig und wünschenswert ist. Im Rahmen einer Podiumsdiskussion mit Gründungsmitgliedern sollen dieser kritische Rückblick diskutiert und Perspektiven für die zukünftige Arbeit entwickelt werden.

Ein zweiter Schwerpunkt der Tagung soll auf die Angewandte Computerlinguistik gelegt werden. Dabei geht es um die Bedeutung, die NLP-Systeme im Kontext graphischer und multimedialer Anwendungen haben kann. Beispiele dafür sind neuere Entwicklungen im Information Retrieval incl. Hypertext, Hilfsysteme, Anwendungsperspektiven von NLP-Systemen, Evaluierung u. a.

Diese Thematik wird in drei Sektionen vertieft werden:

Sektion:	Fuzzy Linguistik (Organisation Ch. Womser-Hacker)
Sektion:	gesprochene Sprache v. a. Anwendungen und Oberflächen (Organisation E. Noeth)
Sektion:	NLP - Anwendungen (Organisation G. Thurmair) Grammatik
Sektion:	und Implementation (Organisation: Hausser)

Zeitplan:

9.94	Call for Papers
14.1.95	Deadline: Einreichen von extended abstracts
7.2.95	Benachrichtigung über Annahme der Vorträge
30.-31.3.95	Tagung in Regensburg;
8.4.95	Abgabe der druckfertigen Fassung der Beiträge für den Tagungsband. Der Tagungsband erscheint nach der Tagung im Verlag Georg Olffis.

Programmkommitee:

Haller (Saarbrücken)
Hausser (Erlangen)
Heyer (Leipzig) Hitzenberger (Regensburg) Krause (Regensburg) Lenders (Bonn)
Lutz (Koblenz)
Pütz (Kiel)
Seewald (Hannover)
Thurmair (München)

Organisationskomitee:

J. Krause
L. Hitzenberger Ch. Womser-Hacker

Tagungsort:

Universität Regensburg
Inst. für Allg. und Indogermanische Sprachwissenschaft
FG: Informationswissenschaft

Auskünfte:

L. Hitzenberger
Universität Regensburg
Phil. Fak. IV
Universitätsstraße 31
93040 Regensburg
e-mail:
Ludwig .Hitzenberger@sprachlit.uni-regensburg.de
Tel.: 0941/943-4195
Fax: 0941/943-3585

Dr. Christa Womser-Hacker University of Regensburg Information Science
D-93040 Regensburg
e-mail:
christa.womser-hacker@sprachlit.uniregensburg.de
Tel.: 0941-943-3600
Fax: 0941-943-3585

Morpholympics 96

In den letzten beiden Ausgaben des LDVForums nahm die Morphologie und besonders die erste Morpholympics einen breiten Raum ein. Die vom Arbeitskreis "Morphologie und Parsing" (R. Hausser, Erlangen) initiierte Veranstaltung erzielte ein lebhaftes Echo in der Fachwelt, erntete eine Reihe positiver und negativer Kritiken; wie bereits geplant, soll eine zweite Morpholympics voraussichtlich am 4. und 5. März 1996 in Saarbrücken stattfinden. Ihr soll wieder ein entsprechender kleiner Workshop mit den potentiellen Teilnehmern vorausgehen; als neue Hauptsprache ist Französisch vorgesehen, eine zweite Runde für Deutsch soll auch durchgeführt werden.

Die GLDV hat Kontakt mit ihrer französischen Schwesterorganisation ATALA aufgenommen, die ein großes Interesse an einer solchen Veranstaltung gezeigt hat; gleichzeitig läuft zu diesem Thema ein französisches Projekt, das vom IN ALF in Nancy geleitet wird. Mit beiden Stellen werden Gespräche geführt, die eine gemeinsame Trägerschaft der Morpholympics 1996 zum Ziel haben.

J. Haller, Saarbrücken
Koordinator der 2. Morpholympics 1996

KONVENS 96 und 98

Kurzbericht des GLDV- Vertreters im KONVENS- Vorbereitungskomitee

Die KONVENS 1994 in Wien war die zweite gemeinsame Konferenz der Gesellschaften, die sich im deutschsprachigen Raum mit maschineller Sprachverarbeitung beschäftigen. Sie brachte vor allem qualitativ einen Sprung nach vorne; eine strenge Auswahl aus den knapp 100 eingereichten Vorträgen gewährleistete ein durchweg hohes Niveau (siehe auch den Bericht von Monika Braun, Univ. Bonn in diesem Heft).

Das Komitee traf sich kurz am Rande der KONVENS und nahm das Angebot der DGFS-Sektion Computerlinguistik an, die nächste KONVENS 1996 in Bielefeld zu organisieren; für 1998 wird diese Konferenz voraussichtlich von der GLDV in Saarbrücken veranstaltet.

J. Haller, Saarbrücken

ESSIR'95

Preliminary Announcement 2nd European Summer School in Information Retrieval, Ilmenau/Thuringia, Germany September 3-8, 1995

The European Summer School in Information Retrieval aims at providing academic education in the field of IR for students, young scientists and engineers. The first summer school of this kind took place in Bressanone, Italy in 1990.

ESSIR'95 will be held at the C2.mpus of the Technical University of Ilmenau. Ilmenau is located in Thuringia forest near Weimar, which will be the Cultural Capital of Europe in 1995.

Please inform your colleagues, assistants and students of ESSIR'95.

Courses:

. *Introduction*
(C.J. van Rijsbergen, U. Glasgow, GB)

. Natural *Language Processing*
(G. Ruge, Sietec, Munich, D)

. *Models*
(N. Fuhr, U. Dortmund, D)
. *Architecture* of IR *systems*
(P. Willett, U. Sheffield, GB)
. *User interfaces*
(P. Ingwersen, Royal School of Librarianship, DK)
. IR and Databases
(Y. Chiaramella, U. Grenoble, F)
. *Evaluation*
(P. Bollmann, TU Berlin, D)
. *Intelligent Retrieval*
(U. Thiel, GMD, Darmstadt, D)
. *Multimedia Retrieval*
(P. Schaeuble, ETH Zurich, CH)
. IR and *the Network*
(N.N.)
. IR and *Hypermedia*
(M. Agosti, U. Padova, I) ESSIR'95 is organized by the IR group of GI, the German association for computer science.

Conference Chair:	Norbert Fuhr (U. Dortmund, D)
Local Organization:	Reinhard Schramm (TU Ilmenau, D)
Treasurer:	Christa Womser-Hacker (U. Regensburg, D); Ulrich Thiel (GMD, Darmstadt, D)
Publicity:	Gerda Ruge (SIETEC, Munich, D)
Program:	Norbert Fuhr (U. Dortmund, D); Ulrich Thiel (GMD, Darmstadt, D)

For getting more information about ESSIR'95, look at the ESSIR page in World Wide Web at

http: //ls6-www.informatik.uni-dortmund.de /essir/announce.html

Via e-mail, you can subscribe to the ESSIR mail server by sending the following message:

To: mailserv@ls6.informatik.uni-dortmund.de
Subject: Command
@SH subscribe ESSIR95 <first name> <last name>

In case you have no access to any of these services, send a letter to the following address:

ESSIR95
Informatik VI
University of Dortmund D-44221 Dortmund Germany

Kolloquium" Theorie der Semantik und Theorie der Lexikographie
ıdte Semantik und Praxis der Lexikographie"
AK - Lexikographie
27. *und* 28. Januar 1995 *im IKP der Universität Bonn*

Leitung: Nico Weber

Interessentinnen / Interessenten, die an dem Kolloquium teilnehmen möchten, bitten wir, sich unter Angabe von Namen, Adresse und Telefonnummer per E-Mail oder Brief anzumelden. Sie erhalten dann weitere Informationen.

Hotelinformationen schicken wir auf An frage. Wir sind zu erreichen:

. Über E-Mail:
upk004@ibm.rhrz.uni-bonn.de.
. In der Poppelsdorfer Allee 47, 53225 Bonn.
. Telefonisch unter:
(0228) 73 56 44
(Jens Ostermann / Nico Weber)
73 56 21
(Monika Braun / Jens Ostermann) 73 56 38
(Sekretariat des IKP, Frau von Neffe)
. Über Fax:
(0228) 73 56 39

Vorläufiges Programm *im Überblick:*

Henning Bergenholtz (Aarhus): Verteilung der enzyklopädischen Informationen in einem erklärenden Fachwörterbuch.

Manfred Bierwisch (Berlin): Lexikon und Universalien.

Gregor Buechel (Köln): Können Verben semantische Relationen markieren?

Ulrich Engel (Heppenheim): Über semantische Relatoren und anderes. Ein Entwurf für künftige Valenzwörterbücher.

Udo L. Figge (Bochum): Kognitiv orientierte Lexikographie.

Ulrich Heid (Stuttgart): Lexikalische Beschreibungen und ihre Überprüfung in Textkorpora.

Manfred W. Hellmann (Mannheim): Lexikographische Erschließung des Wendekorpus.

Erhard Hinrichs / Heike Winhart (Tübingen): Praktische Semantik in der Computerlinguistik - ein Erfahrungsbericht.

Ursula Klenk (Göttingen): Zur Beschreibung der Interdependenz von Syntax und Semantik in der HPSG.

Herbert Kuestner (Berlin): From formula to natural language - automatic paraphrase generation from lexical items.

Petra Ludewig (Osnabrück): Inkrementelle wissensbasierte Wörterbuchanalysen.

Stephan Mehl (Duisburg): Semantische Relationen - Akquisition und Repräsentation.

Burkhard Schaeder (Siegen): Ansichten von Bedeutung: fachsprachliche vs. gemeinsprachliche Semantik.

Uta Seewald (Hannover): Wortbedeutungen in Wörterbüchern - Wortbedeutungen in Texten.

Bernhard Schröder (Bonn): Zum Kompositionalitätsprinzip in der Semantik.

Astrid Steiner- Weber (Bonn): Zur Beschreibung von Komposita im Lexikon der byzantinnschen Gräzität".

Rainer Stuhlmann-Laeisz (Bonn): Was ist philosophische Logik der Zeit?

Nico Weber (Bonn): Thematische Einführung.

Werner Wolski (Marburg): Konzeption eines Bedeutungswörterbuchs zum Werk von Paul Celan.

Suzanne Wolting (Düsseldorf): Semantische Repräsentation mehrwert iger Verben in einem vererbungsbasierten Lexikon.

Gerd Wotjak (Leipzig): Zur Beziehung von lexikalischem Bedeutungswissen und enzyklopädischem Weltwissen.

Protokoll
der Mitgliederversammlung der
GLDV vom 29.9.1994 in Wien

Beginn: 16.30 Uhr Ende: 18.20 Uhr Sitzungsleitung: W. Lenders

Tagesordnung

TOP	1:	Eröffnung der Sitzung, Regularien
TOP	2:	Endgültige Festlegung der Tagesordnung
TOP	3:	Bericht des Vorstands mit Kassenbericht und Bericht der Kassenprüfer
TOP	4:	Haushaltspläne 1994/95
TOP	5:	Entlastung des Vorstands
TOP	6:	Wahl von Kassenprüfern
TOP	7:	Bericht des Beirats
TOP	8:	Berichte aus den Arbeitskreisen
TOP	9:	Verabschiedung der Satzung in ihrer fortlaufenden Form
TOP	10:	Diskussion einer möglichen Änderung von §13 der Satzung bezüglich der Wahl und Stellung von Beiratsmitgliedern
TOP	11:	Jahrestagung 1995 und KONVENS 1996
TOP	12:	Arbeitsprogramm 1994/95
TOP	13:	Verschiedenes

TOP 1: Eröffnung der Sitzung, Regularien

Der Vorstandsvorsitzende, W. Lenders, stellt fest, daß die Tagesordnung mit der Einladung an die Mitglieder versandt und im LDV-Forum veröffentlicht worden ist. Es sind 21 Mitglieder anwesend. Die Öffentlichkeit wird von der Mitgliederversammlung zugelassen. Dem Antrag auf Stimmübertragung eines nicht anwesenden Mitglieds wird stattgegeben. Das Protokoll der letzten Mitgliederversammlung in Kiel, das den Mitgliedern durch das LDV-Forum zugegangen ist, wird genehmigt.

TOP 2: Endgültige Festsetzung der Tagesordnung

Anträge zur Tagesordnung liegen nicht vor. Die Tagesordnung wird in der vorgelegten Form angenommen.

TOP 3: Bericht des Vorstands mit Kassenbericht und Bericht der Kassenprüfer

W. Lenders berichtet, daß die GLDV derzeit 321 Mitglieder zählt. Die Mitgliederzahl ist etwa konstant geblieben. 15 Beitritten stehen 21 Austritte (darunter zuvor geführte Mitglieder, die zu Postrückläufen geführt hatten) gegenüber. Die Zahl der Postrückläufer konnte erheblich reduziert werden. Zu den Gegenständen der Vorstandsarbeit zählten in der bisherigen Amtszeit des Vorstands die Arbeitskreise, das LDV-Forum, die Vorbereitung der Tagungen, die Mitgliederwerbung sowie die Kontaktaufnahme mit anderen

Gesellschaften. Die einzelnen Punkte werden in den jeweils dafür vorgesehenen Tagesordnungspunkten gesondert behandelt. Der Vorsitzende berichtet, daß eine der vornehmsten Aufgaben des Vorstands in der Neubelebung der Arbeitskreise besteht. Ungelöst ist in diesem Zusammenhang nach wie vor die Frage der Fortführung des Arbeitskreises "Berufsperspektiven" . W. Lenders teilt mit, daß die Herbstschule in Leipzig, deren Organisation vor Ort Prof. Heyer übernommen hatte, aufgrund zu geringer Teilnehmerzahl abgesagt werden mußte. Es besteht die Vermutung, daß die Unterkunftsprobleme für Studenten eine der Ursachen für die geringe Zahl an Anmeldungen war. Die Herbstschule ist nun für das Jahr 1995 geplant. Der Schatzmeister R. Hausser legt den Kassenbericht des Haushaltsjahres 1993 vor. Er berichtet, daß die Beiträge für das laufende Haushaltsjahr im August eingezogen wurden. Die nicht per Einzugsermächtigung durchgeführten Beitragszahlungen sind noch nicht abgeschlossen. Die Kasse wurde von Herrn Lutz und Herrn Schröder, der das Amt stellvertretend für Herrn Willee ausübte, geprüft. Die Mitgliederversammlung stimmt dem Einspringen Herrn Schröders als Kassenprüfer nachträglich zu. Herr Schröder trägt vor, daß die Prüfung den Zeitraum vom 1.1.93 bis 31.12.93 umfaßte. Es lag eine Ein- und Ausgabenrechnung sowie das Journal '93 zur Prüfung vor. Herr Schröder beantragt die Entlastung des Vorstands.

TOP 4: Haushaltspläne 1994/95

Für das Jahr 1994 werden mit ca. 22.000 bis 25.000 DM Einnahmen gerechnet. Aus diesen Einnahmen werden die Kosten für das LDV-Forum, die Arbeitskreise, den Vorstand und einzustellende Hilfskräfte finanziert.

TOP 5: Entlastung des Vorstands

Nach erfolgtem Bericht der Kassenprüfer (vgl. TOP 3) stimmt die Mitgliederversammlung einstimmig für die Entlastung des Vorstands.

TOP 6: Wahl von Kassenprüfern

Herr Lutz wird in Abwesenheit zur Fortsetzung seines Amtes als Kassenprüfer vorgeschlagen. Als zweiter Kassenprüfer wird Herr Schröder nominiert. Die Mitgliederversammlung stimmt der Nominierung von Herrn Lutz und Herrn Schröder einstimmig zu.

TOP 7: Bericht des Beirats

Der Vorstandsvorsitzende stellt fest, daß nur ein Beiratsmitglied anwesend ist. Aus dem Beirat gibt es nichts zu berichten, das nicht auch im Bericht des Vorstands enthalten ist. Nach einem Beschluß des Vorstands vom 27.9.94 soll der Beirat künftig stärker zur Arbeit in der Gesellschaft, etwa zur Akquirierung und Auswahl von Beiträgen für das LDVForum, herangezogen werden. Desweiteren wird der Beirat aufgefordert, einen Sprecher aus seinen Reihen zu bestellen. Beiratsmitglied G. Knorz berichtet in seiner Funktion als Herausgeber des LDV-Forums, daß sich die Beitragslage für das Forum entspannt hat. Fachbeiträge für das anstehende Heft liegen vor. Herr Knorz nimmt die Anregung auf, eine Rubrik im LDV-Forum einzurichten, die wichtige Email-Adressen der Gesellschaft verzeichnet.

TOP 8: Berichte aus den Arbeitskreisen

W. Lenders gibt zunächst einen Überblick über die Aktivitäten der Arbeitskreise (1. und 2.), die nicht durch Arbeitskreisleiter auf der Mitgliederversammlung vertreten sind:

1. *AK "Kodierung und Normung"*
   Leitung: Peter Scherber

Der Arbeitskreis hat im vergangenen Jahr zwei Treffen veranstaltet. Er verfolgt das Ziel, einen Leitfaden für deutsche Benutzer zu den Vorschlägen der TEI zu erarbeiten und eine Evaluierung der TEI-Standards vorzunehmen. Die Zahl der aktiven Arbeitskreismitglieder liegt bei 13.

2. *AK "Lexikographie"*
Leitung: Nico Weber

Der Arbeitskreis hat zur Mitarbeit an einer Bibliographie zur maschinellen Lexikographie aufgerufen. Ein Arbeitskreistreffen soll voraussichtlich im Dezember stattfinden. Ein Themenvorschlag ("Repräsentation von Wörterbuchtexten") für dieses Treffen liegt bereits vor.

3. *AK "Korpora"*
Leitung: Robert Neumann

Herr Neumann berichtet über die Treffen des Arbeitskreises. Bei dem letzten zweitägigen Treffen in Münster wurde über Projekte aus dem Bereich der Europäischen Union sowie über Teilprojekte aus dem "Verbmobil"-Projekt referiert. In diesem Jahr findet noch eine Tagung zum Thema "Korpora und Annotationen" statt. Eine Folgeveranstaltung dieser Tagung ist bereits für das Frühjahr 1995 in Stuttgart geplant.
Derzeit erfolgt der Aufbau eines Listservers, der Informationen des Arbeitskreises bereithalten soll. Die Zahl der aktiven Arbeitskreisteilnehmer beträgt ca. 15.

4. *AK "Maschinelle Übersetzung"*
Leitung: Hans Haller

Ein Nachrichtenaustausch der Arbeitskreisinteressenten über Email ist erfolgt. H. Haller berichtet, daß der Teilnehmerkreis klein ist, und stellt fest, daß das akademische Interesse an Maschineller Übersetzung insgesamt aufgrund der derzeitigen Förderpolitik nur gering ist.

Es ist geplant, gemeinsame Veranstaltungen des Arbeitskreises mit dem "MTAnwenderkreis" durchzuführen. Die Tagung des "MT-Anwenderkreises" im Frühjahr 1995 soll voraussichtlich auch als Treffen des AK "Maschinelle Übersetzung" stattfinden.

5. *AK "Parsing in Morphologie und Syntax"*
Leitung: Roland Hausser

R. Hausser berichtet über die 1. MORPHOLYMPICS, bei der sich im März '94 acht Systeme dem Wettbewerb stellten. In Heft 1, 1994 des LDV-Forums sind die Ergebnisse dieser Veranstaltung dokumentiert. Eine Dokumentation, die eine Darstellung aller Systeme, die sich auf der MORPHOLYMPICS präsentierten, umfaßt, wird in Kürze erscheinen.

Die 1. MORPHOLYMPICS war der Auftakt einer Reihe von Veranstaltungen (MORPHOLYMPICS und PARSOLYMPICS), die künftig in mehr oder weniger regelmäßigen zeitlichen Abständen durchgeführt werden sollen. Die zweite MORPHOLYMPICS soll voraussichtlich 1996 in Saarbrücken mit Französisch als Haupttestsprache stattfinden.

Der Arbeitskreis wird auch als Sektion auf der Jahrestagung der GLDV in Regensburg aktiv werden.

6. *AK "CL-Studiengänge/Berufsperspektiven"*

W. Lenders unterstreicht, daß die Bemühungen um die Wiederaufnahme der Arbeit des Arbeitskreises "Berufsperspektiven" sowie die Suche nach einer Nachfolge der Leitung

des Arbeitskreises fortgesetzt werden müssen, da eine Neuauflage der Broschüren erforderlich ist. Hierzu sollen Mittel für eine Hilfskraft gestellt werden. Der Vorsitzende ruft die Mitglieder auf, über eine Aktivität in diesem Arbeitskreis nachzudenken. W. Lenders berichtet, daß die Einrichtung eines Arbeitskreises "Historisch-Vergleichende Sprachwissenschaft" geplant ist.

TOP 9: Verabschiedung der Satzung in ihrer fortlaufenden Form

Die Mitgliederversammlung nimmt die Satzung in ihrer fortlaufenden Form einstimmig an.

TOP 10: Diskussion einer möglichen Änderung von §13 der Satzung bezüglich der Wahl und Stellung von Beiratsmitgliedern

W. Lenders ruft die Mitgliederversammlung zur Stellungnahme bezüglich der vom Vorstand vorgeschlagenen Satzungsänderung auf, die eine Berufung von Mitgliedern in den Beirat ermöglichen soll. Der Vorschlag sieht vor, neben 4 gewählten Beiratsmitgliedern maximal weitere vier Beiratsmitglieder zu kooptieren. Die Mitgliederversammlung unterstützt den Vorschlag des Vorstands mit einer Gegenstimme, eine diesbezügliche Satzungsänderung auszuarbeiten. Der Vorstand wird der Mitgliederversammlung bei ihrer nächsten Zusammenkunft einen Vorschlag zur Änderung des §13, Abs. 2 unterbreiten. Ein Formulierungsvorschlag soll vorab per Email und im Newsletter veröffentlicht werden, um Reaktionen einer größeren Mitgliederzahl zu ermöglichen. Eine Verankerung der AK-Leiter im Beirat soll im Vorstand noch einmal diskutiert werden.

TOP 11: Jahrestagung 1995 und KONVENS 1996

L. Hitzenberger berichtet, daß der "Call for papers" für die Jahrestagung 1995 in Regensburg erfolgt ist und fordert die Mitglieder zur Teilnahme auf. W. Lenders erinnert daran, daß auf der nächsten Mitgliederversammlung während der Jahrestagung in Regensburg (30./31.3.95) ein Wahlvorstand gewählt werden muß, da Wahlen zum Vorstand und Beirat anstehen. H. Haller berichtet über die Sitzung der Vorbereitungsgruppe der KONVENS vom 28.9.94. Die KONVENS '96 wird von der Sektion Computerlinguistik der DGfS organisiert werden und in Bielefeld stattfinden. Die GLDV wird die KONVENS '98 ausrichten.

TOP 12: Arbeitsprogramm 1994/95

Das Arbeitsprogramm für 1995 sieht die Organisation der Herbstschule sowie eine weitere Belebung der Arbeitskreise und die Weiterführung des Newsletters vor.

TOP 13: Verschiedenes

Der Vorstand wird aufgrund der geringen Zahl der Tagungsteilnehmer um eine Einschätzung der KONVENS gebeten. In erster Linie werden für die geringe Teilnahme die zahlreichen Konkurrenzveranstaltungen verantwortlich gemacht.

Uta Seewald (Schriftführung)

Winfried Lenders (Sitzungsleitung)

Entwurf einer Satzungsänderung

Wie bei verschiedenen Gelegenheiten angekündigt und bei der letzten Mitgliederversammlung beschlossen, wird der Vorstand der nächsten Mitgliederversammlung in Regensburg eine Änderung der Satzung der GLDV vorschlagen und zur Abstimmung stellen, die auf eine Modifikation der Wahl des Beirats hinausläuft. Es handelt sich um die Änderung von Par. 13, Abs. (2) der Satzung, der folgende neue Fassung erhalten soll (die neuen Passagen sind in,,[ ]" eingefügt):

> "Der Beirat besteht aus 7 Mitgliedern. [Vier seiner Mitglieder werden] auf die Dauer von zwei Jahren, vom Tag der Wahl an gerechnet, von den Mitgliedern gewählt. [Bis zu drei weitere Mitglieder können durch den Vorstand kooptiert werden. Der Beirat] bleibt bis zur Neuwahl im Amt. Wählbar sind auch natürliche Personen, die der Vereinigung nicht angehören. Vorstandsmitglieder können nicht zugleich Mitglieder des Beirats sein. Der Beirat wählt einen Beiratssprecher" .

Erläuterungen:

Durch diese Satzungsänderung soll dem Vorstand die Möglichkeit eingeräumt werden, Persönlichkeiten, die er für die GLDV für wichtig hält, in den Beirat zu berufen. Von früheren Vorständen war das Fehlen einer solchen Möglichkeit immer bedauert worden.

Die Änderung der Satzung, die der Vorstand vorschlägt, ist minimal. Sie schränkt die Rechte der Mitglieder nicht ein, erweitert aber den Handlungsspielraum des Vorstands und würde auch die bisher wenig effektive Beiratsarbeit verbessern.

Der Vorstand bittet alle Mitglieder, die vorgeschlagene Satzungsänderung zu bedenken und auch schon vor der Mitgliederversammlung zu diskutieren, z.B. über electronic mail. Bitte teilen Sie ihre Meinung dem Vorsitzungen oder einem anderen Vorstandsmitglied mit! Die E-mail-Adressen sind wie folgt:

Lenders:	Lenders@uni-bonn.de
Haller:	hans@iai.uni-sb.de
Hausser:	rrh@linguistik.uni-erlangen.de
Seewald:	nhtcseew@rrzn-user.uni-hannover.de
Hitzenberger:	Lud wig.Hitzenberger@sprachlit.uni-regensburg.de

GLDV-Mitgliederversammlung 1995

Während der GLDV-Jahrestagung 1995, die vom 30. - 31.3.1995 in Regensburg stattfindet, wird auch die diesjährige Mitgliederversammlung der GLDV abgehalten:

am Donnerstag, dem 30.3.1995 16.30 Uhr
Universität Regensburg
(genauer Ort wird noch bekannt gegeben)

Zu der Mitgliederversammlung sind alle Mitglieder herzlich eingeladen.

Tagesordnung:

1. Regularien
2. Bericht des Vorstandes und Kassenbericht 1994
3. Entlastung des Vorstandes
4. Wahl von Kassenprüfern
5. Bericht des Beirats
6. Berichte der Arbeitsgruppen/-kreise
7. Zusammensetzung und Wahl des Beirats.
   Diskussion und Beschlußfassung zu einer vorgeschlagenen Änderung von Paragraph 13, Abs. 2 der Satzung
8. Neuwahlen
   - 8.1 Wahl eines Wahlvorstandes
   - 8.2 Kandidatenliste: Vorstand
   - 8.3 Kandidatenliste: Beirat
9. Arbeitsprogramm 1995/96
10. Nächste Jahrestagungen
11. Verschiedenes

Die offizielle Einladung zur Mitgliederversammlung 1995 sowie nähere Erläuterungen zu TOP 7 und die Kandidatenlisten zu TOP 8 werden mit Rundbrief an alle Mitglieder verschickt.

Winfried Lenders, 1. Vorsitzender

ARBEITSKREISE DER GLDV

Kurzbeschreibung der Zielsetzungen und Aktivitäten des Arbeitskreises "Historisch-Vergleichende Sprachwissenschaft"

Der Arbeitskreis "Historisch-Vergleichende Sprachwissenschaft" etablierte sich während der ,,3. Internationalen Fachtagung für Computereinsatz in der historischen Sprachwissenschaft" in Dresden am 8. Oktober 1994; als sein erster Leiter wurde Jost GIPPERT, Frankfurt, gewählt, als Stellvertreter Johann TISCHLER, Dresden (vgl. den Tagungsbericht von Johann TISCHLER im vorliegenden Heft).

Hervorgegangen ist der Arbeitskreis aus einer zunächst nicht organisierten Gruppe von Sprachwissenschaftlern, die sich erstmals im Jahre 1992 an der Universität Bamberg (Veranstalter Jost GIPPERT), zum zweiten Male dann 1993 an der Karls- Universität Prag trafen (Veranstalter Petr VAVROUSEK). Derzeit gehören dem Arbeitskreis rund 30 Sprachwissenschaftler an, die an verschiedenen europäischen Universitäten arbeiten und dabei durchaus divergierende Spezialgebiete vertreten, die aber sämtlich unter den Überbegriff "Historisch-Vergleichende Sprachwissenschaft" fallen (z.B. Indogermanistik, Indo-Iranistik, Hethitologie, Keilschriftkunde) und damit auch die gemeinsamen Zielsetzungen des Arbeitskreises zu umreißen gestatten: Wie bei den genannten Tagungen deutlich wurde, eröffnet die Anwendung elektronischer Verfahren bei der wissenschaftlichen Beschäftigung mit historischen Sprachstufen und in solchen Sprachstufen überlieferten Zeugnissen zahlreiche zukunftsweisende Perspektiven. Diese beschränken sich nicht etwa nur auf den (inzwischen auch in historischen Fächern schon recht weit verbreiteten) Computereinsatz bei der Erstellung von Druckvorlagen, sondern erstrecken sich zunächst auf alle Bereiche der Analyse von sprachlichen Zeugnissen (Textanalyse: Erstellung von Konkordanzen und Indizes, grammatische Auswertung, morphologisches und syntaktisches Parsing etc.), darüber hinaus mehr und mehr auch auf die nicht-sprachliche Grundlage solcher Zeugnisse (Bildverarbeitung: Epigraphik, Paläographie usw.).

Als eine typische Anwendung in diesem Sinne kann z.B. das zunächst als einfache Textdatenbank konzipierte, inzwischen aber unter dem umfassenderen Titel "Thesaurus indogermanischer Text- und Sprachmaterialien (TITUS)" stehende Projekt gelten, das 1987 von Jost GIPPERT ins Leben gerufen wurde und an dem sich inzwischen rund 40 Wissenschaftler europäischer und außereuropäischer Universitäten beteiligen; Ziel dieses Projektes ist zunächst die Sammlung eines möglichst vollständigen Corpus von Textmaterialien in den für den indogermanistischen Sprachvergleich relevanten Sprachen, im weiteren dann die Entwicklung und Erprobung elektronischer Analyseverfahren, die es gestatten, das in den Texten enthaltene sprachliche Material historisch und vergleichend auszuwerten.

Der neu gegründete Arbeitskreis "Historisch - vergleichende Sprachwissenschaft" soll eine Plattform darstellen, auf der Verfahrens- und Lösungsvorschläge für derartige Anwendungen vorgestellt und diskutiert werden können. Eine nächste Fachtagung des Arbeitskreises ist für 1995 an der Universität Frankfurt geplant; besonderes Augenmerk soll dabei u. a. auf juristischen

Aspekten des TITUS-Projekts sowie auf der Anwendung von Verfahren zur Bildverarbeitung in der historisch-vergleichenden Sprachwissenschaft liegen.

Kontaktadresse:
Prof. Dr. Jost GIPPERT
Vergleichende Sprachwissenschaft Georg-Voigt-Str. 6
Postfach 11 19 32
D-60054 Frankfurt
Telefon +49-69-7988591
Fax +49-69-7982873
e-mail Gippert@em.uni-frankfurt.d400.de

AK - Maschinelle Übersetzung 10. Europäische Tagung

Der "Anwender-Arbeitskreis Maschinelle Übersetzung" ist ein seit mehreren Jahren bestehender loser Zusammenschluß von aktuellen Anwendern und interessierten Personen aus Industrie, Behörden und Universitäten, der sich etwa zweimal im Jahr zum Erfahrungsaustausch trifft. Die Organisation liegt derzeit in den Händen von Hans Billing und Ursula Bernhard von der GMD; als ideale Tagungsorte werden Institutionen angesehen, die selbst Softwarewerkzeuge in Sprachabteilungen einsetzen. Das 10. Treffen dieses Arbeitskreises fand am 7. und 8. 11. 1994 bei der Fa. Böhringer in Ingelheim statt und erfreute sich regen Besuchs.

Zu Anfang begrüßten die Organisatoren und Alain Paillet von der Fa. Böhringer die Gäste; ein Vertreter des Bereichs Öffentlichkeitsarbeit der Fa. Böhringer (Arzneimittel, ca. 24.000 Beschäftigte weltweit) stellte die Firma und ihre Geschichte kurz vor und veranschaulichte seine Darstellung mit vier Kurzfilmen über Medikamentenproduktion, Sicherheit, Umweltschutz und Krankheiten, die mit den hergestellten Medikamenten erfolgreich bekämpft werden können. Alain Paillet schilderte dann die Arbeit im Sprachendienst der Firma, der international knapp zwanzig Mitarbeiter hat und mehr als 30.000 Seiten pro Jahr in den größeren europäischen Sprachen bearbeitet. Er erläuterte die aktuelle und geplante Verwendung der zu diesem Zweck angeschafften Softwarewerkzeuge METAL, MemCat (Translation Memory) und TermBase.

METAL war auch das Thema der beiden anderen Vorträge an diesem Nachmittag, diesmal jedoch von der Entwicklungs- und Vertriebsseite. Gerda Klimonow und Andreas Küstner zeigten in einer Vorführung die Arbeiten der Gesellschaft für Multilinguale Systeme in Berlin an einem Russisch-Deutschen METAL-System mit ca. 12.000 Lexikoneinträgen. Chris Pyne von Sietec und Gregor Thurmair stellten im Anschluß daran Entwicklung und Chancen der linguistischen Softwaretechnik sowie die Bedeutung einer Integration der einzelnen Werkzeuge dar. Der Nachmittag endete mit einigen Kurzberichten von Tagungen sowie dem Austausch von Neuigkeiten; Tom Gerhardt von der Universität Saarbrücken wies auf die Möglichkeit hin, das traditionsreiche MÜ-System SUSY per Internet für experimentelle Zwecke zu nutzen (Auskünfte bei Dirk Luckhardt unter dlu@rz.uni-sb.de). Man hatte sich jedoch noch viel zu sagen; jedenfalls kamen die letzten Ingelheimer Kneipenbesucher erst beim Morgengrauen nach Hause.

Der Morgen des 8.11. begann dann mit einem Bericht von Hans Billing über eine Umfrage zur Fördersituation im Bereich Maschinelle Übersetzung geschriebener Sprache sowie über eine vom BMFT veranstaltete Expertenversammlung zu diesem Thema; erst bei einer Ausschreibung "intelligente Systeme" Ende diesen Jahres taucht das Thema wieder auf (unter vielen anderen Themen aus der Künstlichen Intelligenz).

Seit dem Ende von EUROTRA und seiner Begleitforschungsprojekte waren alle Förderaktivitäten des BMFT auf VERBMOBIL konzentriert, das jedoch nach Meinung aller Teilnehmer nur einen Teil der Probleme behandelt, die für die Übersetzung von Texten relevant sind. Johann Haller gab in Ergänzung hierzu einen kurzen Bericht über die letzte Projektlenkungssitzung dieses großen Verbundprojektes in München sowie über die Leistungsfähigkeit des dort vorgestellten Mini-Demo-Systems.

Der zweite Schwerpunkt des Arbeits-

kreistreffens war LOGOS gewidmet; Friederike Bruckert, Eschborn, berichtete über den gegenwärtigen Status des Systems und der gleichnamigen Firma, die 1994 ihr 25jähriges Jubiläum feiert. Zwei Drittel der derzeit rund 50 Mitarbeiter sind in den USA beschäftigt; weltweit werden ca. 50 Kunden betreut.

Harald Zimmermann, Softex Saarbrücken, stellte in Ergänzung dazu Arbeiten zur leichteren Kodierung von LOGOS-Wörterbüchern vor; dieses auf PC vorliegende Werkzeug ermögliche eine Eingabe von 1000 (ALEX-) Wörtern pro Stunde. Auch an der Rationalisierung der Kodierung von Mehrwortausdrücken wird gearbeitet.

Die Tagung des Arbeitskreises wurde dann mit einer allgemeinen Diskussion abgeschlossen; Ursula Bernhard und Hans Billing werden die nächste Tagung in der zweiten Novemberwoche 1995 voraussichtlich bei der Fa. SAP in Walldorf organisieren. Unter den verschiedenen Themenvorschlägen wurde auch die Möglichkeit erwogen, wieder einmal die Frage der Ausbildungsproblematik zu behandeln: Softwarewerkzeuge in der Ausbildung von Übersetzern. Die anwesenden Vorsitzenden der GLDV, Winfried Lenders und Johann Haller boten die Mitwirkung der GLDV bzw. des entsprechenden Arbeitskreises in diesem Punkte an. Auch de~. Vorschlag, die Vorträge zu konkreten MU-Systemen für einen weiteren Teilnehmerkreis, z.B. Studenten, zu öffnen, wurde positiv diskutiert.

Alle Mitglieder zeigten Interesse an dem vom 11.-13. Juli 1995 in Luxemburg stattfindenden "MT Summit", der die nächste Gelegenheit zu einem Wiedersehen bieten dürfte. Organisiert wird diese Veranstaltung von der European Association for Machine Translation, Genf.

Die lockere Atmosphäre des Arbeitskreises und die Vielzahl der aktuellen Informationen ließen alle Teilnehmer zufrieden nach Hause reisen.

P.S: Informationen zum AK - Lexikographie finden Sie unter der Rubrik "Veranstaltungen" auf *Seite* 69