LDV-FORUM

Forum der Gesellschaft für linguistische Datenverarbeitung GLDV

LDV-Forum 11.1 (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 (IuD), Schöfterstr. 1-3, D-64295 Darmstadt Tel.: (06151)168490j Fax: (06151)16-8980j Email: knorz@fhdacom2.fhrz.fh-darmstadt.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 Schäder

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

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

Denn erstens kommt es anders. .. Der Beginn dieses etwas albernen Zweizeilers drängt sich mir auf, wenn ich nun als den letzten von mir zu liefernden Baustein für das LDV-Forum 1/94 dieses Editorial anfertige.

Es beginnt beim letzten Heft, und - wie könnte es anders sein - es setzt sich in der vorliegenden Ausgabe 94/1 nahtlos fort, zum Glück nicht nur in negativer Hinsicht.

Eine Überraschung der bösen Art erlebte ich, als das letzte Heft des LDV -Forum mit dem Postversand bei mir eintraf und ich die einzelnen Seiten mit dem zufriedenen Gefühl desjenigen durchblätterte, der die Inhalte ja bereits kennt, aber nun erstmals das vollendete Werk in den Händen hält. Ich blieb gleich bei dem programmatischen Vorwort "Zur Entwicklung der GLDV" von Winfried Lenders hängen, dessen Einleitungssatz der einer Vorfassung zu sein schien. Ein weiterer Blick genügte, um die Panne zu bestätigen: Dieses Vorwort war im Prinzip eine Minute nach Zwölf erst angeliefert worden und in den Vorabdruck noch in einer nicht veröffentlichungsreifen Fassung übernommen worden. Die redaktionelle und inhaltliche Überarbeitung durch den ersten Vorsitzenden der GLDV lief parallel mit meinem Check des Heft-Vorabdrucks in Darmstadt - und im gedruckten Heft fand sich tatsächlich trotz aller hektischen Arbeit (oder gerade deswegen) die unbearbeitete Fassung. In der Fülle der E-Mails und der satztechnisch/ organisatorischen Arbeiten war in Saarbrücken diese wichtige Änderung nicht richtig "angekommen". Was die wohlgesetzten Worte betrifft, so mag diese Panne läßlich erscheinen. Daß aber mit diesem Vorwort der Eindruck erweckt werden kann, daß das LDV-Forum auch in der Zeit nach 1991 aus Darmstadt kam, kann so nicht stehen bleiben. Zumal Burghard Rieger in dieser Zeit die Redaktion des LDV-Forums übernommen hat, als er als Vorsitzender der GLDV schon ansonsten sehr belastet war (die mehr als verdiente Dankadresse für dieses "Doppelengagement" nachzulesen auch im Editorial zu 93/1). Für die technische Panne, die so in einem heiklen Aspekt zur inhaltlichen Panne wurde, eine Entschuldigung in Richtung Trier!

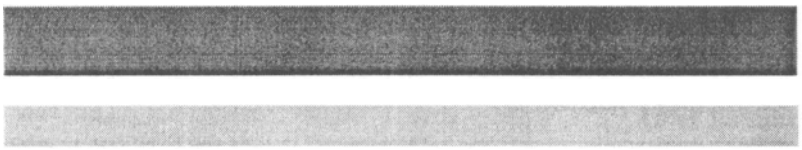
Wenn nicht *es erstens anders käme,* ginge es in diesem Heft schwerpunktmäßig um die Verantwortung der Hochschulen für "ihre" Absolventen der Computerlinguistik. In diesem Fall ist der Grund für die Planungsänderung ein erfreulicher: Es ergab sich die Gelegenheit, die *Ergebnisse der Morpholympics* zum Generalthema zu machen und damit das Schwerpunktthema *"Morphologie"* der letzten Ausgabe fortzusetzen. Ein Thema

im übrigen, das der Redaktion ihren ersten - konstruktiv kritischen - Leserbrief eingebracht hat - vielleicht sogar mit praktischen Konsequenzen? Wir werden sehen!

Auch die Produktion dieses Heftes hat wieder einen kräftezehrenden Endspurt erfordert - hoffentlich ohne Konsequenzen der oben geschilderten Art. Die Hauptlast wird dabei vom IAI, bzw. - um es genauer zu sagen - von Frau Ute Hauck getragen, ohne deren Engagement es gegenwärtig kein gedrucktes LDV-Forum geben könnte. Eine Verbesserung der Situation sehe ich nur, wenn sich die Substanz, aus der heraus das LDV-Forum lebt, erweitert, so daß Teile des Heftes schon weit vor Drucklegung bearbeitet werden können. In jedem Falle sollte die Leistung von Frau Hauck (und damit des IAI) angemessen zur Kenntnis genommen und gewürdigt werden. Herzlichen Dank! Und außerdem wäre mittel- bis ,langfristig in den Zeiten von Internet und World Wide Web die Publikationsform zu überdenken.

So, und nun hoffe ich, daß Sie das LDV -Forum mit Interesse zur Hand nehmen und sich als Mitglieder der GLDV auch selbst angesprochen fühlen, mit Beiträgen für eine angemessene Qualität dieses Publikationsorgans zu sorgen. Vielleicht treffen wir uns auf der Konvens? Es wäre kein gutes Zeichen, sollte die GLDV dort nur unter ferner liefen präsent sein!

G.K.



PS. Der Redaktion liegt ein interessantes und zum Thema passendes Rezensionsexemplar vor: *Uta Seewald: Maschinelle morpho-semantische Analyse des Französischen: >MORSE<. Eine Untersuchung am Beispiel des Wortschatzes der Datenverarbeitung. Tübingen: Niemeyer*, 1994 *(Sprache und Information; Bd.* 26)!

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 1/4 bzw. 3 1/2) als ASCII oder LA-TEX-Datei übermittelt werden. Formatierungshilfen *(LDVforum.sty)* werden auf Wunsch zugesandt.

Rubriken
Die namentlich gezeichneten Bei': träge 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 11.2/94: 31. Okt. 1994 für Heft 12.1/95: 30. Apr. 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

VORWORT

Liebe Mitglieder der GLDV,

in meiner" Grußadresse" im letzten LDV-Forum habe ich anläßlich meiner Amtsübernahme als Vorsitzender allen ausgeschiedenen Mitgliedern des vorhergehenden Vorstandes und des Beirats für ihre langjährige Arbeit sowie Gerhard Knorz für seine Tätigkeit als Redakteur des Forums gedankt. Unerwähnt blieb dabei leider, dass Burghard Rieger, zusätzlich zu seinen Belastungen als Vorsitzender, die Jahrgänge 1991 und 1992 des FORUMs redaktionell betreut hat (vgl. das Editorial in Bd. 10, Nr. 1). Dafür gebührt ihm ein ganz besonderes Dankeschön, hat er doch durch seinen Einsatz unserer Vereinszeitschrift über eine schwierige Zeit hinweggeholfen.

Es steht nun im September im Rahmen der Wiener KONVENS unsere nächste Mitgliederversammlung an; die Einladungen dazu sind Anfang Mai verschickt worden. Einigen von Ihnen wird aufgefallen sein, daß das Einladungsschreiben ein Datum aus dem vorigen Jahr trug. Für dieses Versehen, das auf allzu großes Vertrauen in die Selbständigkeit unserer Textverarbeitung zurückgeht, bitte ich Sie alle um Entschuldigung. Ich hoffe, daß Sie trotz dieses Fehlers Zeit, Ort und Tagesordnung der nächsten MV ordnungsgemäß notieren konnten und würde mich freuen, viele von Ihnen in Wien wiederzusehen.

In den Erläuterungen zur Tagesordnung der MV auf der zweiten Seite ist mir ein weiterer Fehler unterlaufen. Die Erläuterungen beziehen sich auf die TOPs 9 und 10 (nicht auf TOP 8 und TOP 9). Auch dieses Versehen bitte ich zu entschuldigen.

Die Tagesordnung der MV sieht vor, daß auf Anregung des Vorstandes die in der Satzung vorgesehene Zusammensetzung des Beirats sowie das Verfahren für die Wahl des Beirats erörtert werden sollen. Hintergrund dieser Anregung ist der Wunsch des Vorstandes, zusätzlich zu gewählten Beiratsmitgliedern selbst Personen für die Beiratsarbeit berufen zu können. Außerdem möchte der Vorstand die Frage zur Diskussion stellen, ob und wie man ggf. die Arbeitskreisleiter stärker an die satzungsgemäßen Organe der Gesellschaft binden kann.

Als weiteres wichtiges Thema der MV sollten wir diskutieren, wie die Arbeitskreise, die leider nur z. T. in befriedigender Weise aktiv sind, mobilisiert werden können. Besonders liegt mir hier die Fortsetzung der Tätigkeit unserer Gesellschaft auf dem Gebiet der Dokumentation von Ausbildungs- und Berufsperspektiven am Herzen. Ich bitte Sie herzlich, dem Vorstand ihre Anregungen mitzuteilen.

Wir haben also insgesamt genügend Themen für eine fruchtbare Diskussion, und ich ermuntere Sie daher nochmals dringend zur Teilnahme an unserer Mitgliederversammlung in Verbindung mit der KONVENS. Die KONVENS ist zwar nicht die Jahrestagung der GLDV, aber sie ist auch 'unsere' KONVENS, und wir sollten mit einer starken Beteiligung vertreten sein. Wien ist eine Reise wert, besonders wenn dort die GLDV tagt.

In diesem Sinne wünsche ich Ihnen, auch in der Hoffnung auf ein Treffen in Wien, einen schönen und sonnigen Sommer!

Winfried Lenders

LESERBRIEFE

Aus Ihrem Editorial, Herr Knorz, spricht Engagement für die Sache und Enttäuschung über das Ergebnis - Mangel an Beiträgen für das LDV-Forum, fragliche Akzeptanz der ganzen Unternehmung Computerlinguistik - recht deutlich.

Was den letzten Punkt betrifft - sind wir (ich spreche als Mitglied der GLDV) da nicht auch selbst mit schuld?

Morphologie ist das Thema des Heftes
das Heft selbst demonstriert, wie wenig Einfluß auf die Praxis die Beschäftigung mit diesen Themen hat. Mir jedenfalls gibt es - konventionell wie ich nun mal bin und durch das Lesen vieler mit TEX publizierter Texte noch immer nicht abgestumpft jedesmal noch einen kleinen Schlag, wenn mir die praktische Mißachtung so deutlich vorgeführt wird wie in diesem Heft.

Ich spreche von Ligaturen. Nach Duden, Rechtschreibung (ich habe die Auflage von 1980) S. 72 faßt die Ligatur Buchstaben zusammen, die im Wortstamm zusammengehören. Keine Ligatur steht zwischen Wortstamm und Endung und in der Wortfuge von Zusammensetzungen.

Wenn Sie sich das vorliegende Heft einmal durchsehen, verstehen Sie, was ich meine. Es fängt in Ihrem Editorial (ich weiß durchaus bzw. hoffe wenigstens, daß Sie den Satz nicht selbst auch noch besorgen müssen; und daß die Zeit zur Kontrolle knapp ist) in der letzten Zeile des ersten Abschnitts mit dem Wort (welches Omen!) *sträflich* an und hört (welcher Zufall auch hier) mit *berufliche Akzeptanz* auf. Noch peinlicher, weil auffälliger, wird es freilich auf den Seiten 44-54, wo in der Ankündigung *Begriffliche Informationsverarbeitung* die Mißachtung morphologischer Regeln auch noch im Fettdruck über Seiten hinweg ad oculos ~demonstriert wird.

Quisquilien, Kleinigkeiten, werden Sie vielleicht sagen. Ich sehe es nicht so. Hier geht es um die typographische Umsetzung von Morphologie.

. Und in einem Heft, das sich Morphologie zum Thema setzt, sollte sich eine Gesellschaft für Linguistische Datenverarbeitung so etwas nicht leisten - die Entschuldigung, daß TEX diese Ligaturen automatisch macht, lasse ich nicht gelten.

Diese Anmerkungen sind kein persönlicher Vorwurf an Sie - ich möchte mich für Ihren Einsatz vielmehr ausdrücklich bedanken. Ich halte es vielmehr für symptomatisch, daß auf die Praxis so wenig Sorgfalt verwandt wird. Wobei wir wieder bei der Verwunderung über die berufliche Akzeptanz von Leuten wären, die ihre Ausbildung in diesem Umfeld erfahren haben.

Wilhelm Ott,
Eberhard- Karls- Universität Tübingen

MORPHOLYMPICS - EIN UNTERNEHMEN DER GLDV

Winfried Lenders

Seit Bestehen der GLDV ist es eine ihrer wichtigsten satzungsgemäßen Aufgaben gewesen, die Kommunikation und Kooperation der Wissenschaftler auf diesem Gebiet der maschinellen Sprachverarbeitung zu fördern und zu einer Koordination der jeweiligen Projekte beizutragen. In dieses Aufgabenprofil paßt sich die von Roland Hausser im Rahmen des Arbeitskreises "Parsing in Morphologie und Syntax" eingebrachte Morpholympics-Initiative nahtlos ein. Vorbild sind, wie es der Name anzeigt, sportliche Wettkämpfe, und die Idee besteht darin, computerlinguistische Programme zur morphologischen Analyse und Synthese gegeneinander antreten zu lassen, mit dem Ziel, das momentan beste System zu ermitteln. Die Idee ist nicht unbedingt neu. Schon vor Jahren gab es Versuche, maschinelle Übersetzungssysteme im direkten Vergleich aufeinandertreffen zu lassen und zu bewerten. Doch noch nie wurde die Idee so konsequent angegangen und in die Tat umgesetzt, wie es zur Morpholympics geschah. Dafür sei den örtlichen Veranstaltern, den Teilnehmern, den Mitgliedern der Vorbereitungsgruppe und der Jury ganz herzlich gedankt. Daß sich acht Systeme dem Vergleich stellten, ist an sich schon ein Erfolg, nicht nur für Roland Hausser und die Erlanger Computerlinguistik, nicht nur für die GLDV als ausrichtender Gesellschaft, sondern vor allem auch für die Disziplin selbst. Morpholympics hat gezeigt, daß es über alle Unterschiede der Systeme im Einzelnen hinweg einen respektablen Forschungsstand auf einem begrenzten Gebiet der maschinellen Sprachverarbeitung gibt, daß es Lösungen für eine wissenschaftliche und praktische Problemstellung gibt, die nur darauf warten, in größere Aufgaben eingebunden zu werden. Wenn Morpholympics auch - wie in der Welt des Sports zwischen Siegern und Dabeigewesenen unterscheidet, so sollte doch klar sein, daß es Sieger und Besiegte nicht gibt. Es wurde eine Zwischenbilanz gezogen, die erkennen läßt, wo die Stärken und Schwächen der einzelnen Verfahren liegen, und das Urteil der Jury ist im Grunde nichts anderes als das Ergebnis einer Abwägung in einer insgesamt ausgewogenen Bilanz.

Im Namen der Veranstaltung, in der Art der Durchführung und in der Art der Bewertung wurde die Analogie zum sportlichen Wettkampf gewählt. Dies bringt für Teilnehmer und Publikum eine besondere Spannung mit sich. Klarheit sollte aber darüber bestehen, daß, anders als in der Welt des Sports, die Frage der Kriterien nicht so einfach zu beantworten ist. Es ist die Frage, ob das 'schneller, weiter, höher' vieler Sportarten, übertragen auf unseren Fall also das Zeitverhalten eines Systems und die Quote richtiger Lösungen, als Kriterien für die Güte ausreichen, ob es nicht - ähnlich wie in der Kür des Eislaufs - vielmehr auch um das Wie einer erreichten Leistung, im sprachwissenschaftlichen Kontext also um die Frage nach der Erklärungsadäquatheit eines Systems gehen muß. Vermutlich liegt, wie immer, die Antwort auf diese Fragen in der Mitte, jedenfalls haben die Juroren der diesjährigen Morpholympics versucht, einen solchen mittleren Weg zu gehen.

Die Erlanger Morpholympics sollte der Auftakt einer Reihe ähnlicher Veranstaltungen sein. Als nächstes ist für 1996 geplant, französische morphologische Systeme gegeneinander antreten zu lassen, desweiteren sollte zu einem späteren Zeitpunkt in einer 'Parseolympics' eine 'höhere' linguistische Ebene angegangen werden. Es bleibt zu hoffen, daß sich nach der diesjährigen gelungenen Erlanger Veranstaltung weitere 'Wettkämpfe' dieser Art verwirklichen lassen - Spannung wird Teilnehmern und Publikum garantiert.

STELLUNGNAHME DER JURY FÜR DIE MORPHOLYMPICS 94

Juroren: I. Batori, Koblenz/ G. Dogil, Stuttgart/ G. Görz, Erlangen/ W. Lenders, Bonn/ U. Seewald, Hannover

1 Grundlagen

Mit der ersten Morpholympics wurde hinsichtlich der Bewertung linguistischer Softwaretools sowohl bei den Testmaterialien als auch bei den Bewertungskriterien Neuland betreten. Wie bei einer Olympiade sollte sich die Jury für ein Siegersystem entscheiden, gegebenenfalls ein oder zwei weitere Systeme auf weiteren Siegerplätzen rangieren. Anders aber als im sportlichen Bereich stehen bei linguistischen Systemen standardisierte und quantifizierbare Verfahren, auf die man zurückgreifen kann, nur zur Messung der Schnelligkeit, mit der eine Analyse erreicht wird, zur Verfügung. Qualitative Kriterien sind nur bis zu einem gewissen Grad standardisierbar. So wird man den Grad der Korrektheit und Vollständigkeit morphologischer Analysen immer in Bezug setzen müssen zu der eingeschlagenen Analysemethode und ihrer theoretischen Fundierung. Man wird ferner auch gewisse Zufälligkeiten und äußere Bedingungen in Rechnung stellen müssen, aus denen heraus das eine System vollständiger oder weniger vollständig ist als das andere. Um in der sportlichen Analogie zu bleiben: Die Juroren waren sich einig, daß es eher darum ginge, eine Kür im Sinne des Eiskunstlaufs zu bewerten, als darum, den 'ersten' auf einer Zeit- oder Entfernungsskala zu ermitteln. Das Kollegium der Morpholympics-Juroren hatte sich in dieser Situation in einer ersten Besprechung am Vorabend des ersten Austragungstages mit Grundlagen und Methode der Bewertung zu befassen. Es bestand Einigkeit darüber, daß ein kollegiales Votum abgegeben werden sollte, das auf einer Anzahl von Bewertungskriterien und deren Gewichtigkeit beruhen sollte, auf die das Jurorenkollegium sich geeinigt hatte. Es bestand weiterhin Einigkeit darüber, daß immanente Kriterien, insbesondere das der linguistischen Fundiertheit eines Ansatzes und der Korrektheit der Ergebnisse, in erster Linie für die Rangierung eines Systems maßgebend sein sollten, an zweiter Stelle Kriterien des Laufzeitverhaltens und der Robustheit und an dritter Stelle Kriterien der Dokumentation und Präsentation.

2 Die Kriterien

1. *Immanente Systemeigenschaften:*
Korrektheit der Analyse bzw. der Formenbildung: Geltungsbereich des Verarbeitungssystems wie viele Wortformen werden als richtig erkannt, wie viele bleiben unanalysiert oder falsch gedeutet? Wird der Geltungsbereich des Systems durch Derivation und Komposition erweitert, bzw. wie werden abgeleitete und zusammengesetzte Wörter behandelt?

Linguistische Fundiertheit: Ist die morphologische Beschreibung (Repräsentation) ad hoc oder systematisch? Wird eine bestimmte linguistische Theorie. befolgt? Konsistenz der Analyse.

2. *Sekundäre Systemeigenschaften*
Integrierbarkeit in Anwendungsumgebungen, insbesondere in Analyse- und in Taggingverfahren, Text-to-Speech-Systemen u. a.

Robustheit, darunter auch die Handhabung der diakritischen Zeichen, Sonderzeichen, Großschreibung u.ä.

Benutzerfreundlichkeit: Bedienung des Systems, Transparenz der Analysen, Tools für Korrekturen und Systemerweiterungen.

Geschwindigkeit und Speicherbedarf.

3. Präsentationstechnik

Qualität des Vortrags, Präsentationstechnik, Vorführung, Erfüllung der technischen Bedingungen und Dokumentation, Vollständigkeit, Transparenz, Zuverlässigkeit der Beschreibungen.

Die genannten Kriterien repräsentieren mehrere Bewertungsdimensionen, die nicht gleichwertig sind und nicht isoliert betrachtet werden können. So sind die Juroren der Ansicht, daß die immanenten Kriterien, vor allem das der linguistischen Fundiertheit, unter den gegebenen Umständen und auch angesichts der Auswahl einer überwiegend linguistisch orientierten Jury vor anderen, z.B. denen des Zeitverhaltens, vorrangig zu bewerten ist. Besonders kritisch ist der Zusammenhang zwischen Umfang, Geschwindigkeit und Erweiterbarkeit eines Systems. Dies sollte hier verständlich gemacht werden.

3 Rangierung

Auf der Grundlage dieses Kriterienkatalogs gelangte die Jury zu folgender Rangierung:

Platz 1: GERTWOL (Koskenniemi)
Platz 2: MORPH (Hanrieder)
Platz 3: LA MORPH (Hausser)

Die übrigen Systeme blieben unrangiert; sie werden gleichwohl in der nachfolgenden Stellungnahme und Begründung der Entscheidung der Jury bezüglich ihrer jeweiligen Vor- und Nachteile gewürdigt.

4 Begründung

4.1 Präsentationstechnik, Dokumentation etc.

Alle vorgestellten Systeme, unterschiedlich wie sie waren, haben die Juroren positiv beeindruckt. Es war bei allen acht Systemen klar erkennbar, daß die Bewerber an den vorgestellten Systemen inspiriert gearbeitet haben. Alle Bewerber haben in Vortrag, Dokumentation und Vorführung ein profiliertes System auf hohem Niveau vorgestellt und die Ausschreibungsbedingungen erfüllt. Die Gutachter würdigen ausdrücklich den hohen Standard sowie die Sorgfalt, mit der alle Systeme die geforderte Dokumentation beigebracht und sich den angesetzten Wettkampfbedingungen unterworfen haben. Anerkennenswert ist, daß alle acht Systeme mit hohen Trefferquoten und korrekt arbeiten. Ablehnung und Fehlanalysen sind relativ selten; zu ihrer weiteren Interpretation bedarf es einer ausführlichen Sichtung und Klassifikation der Fehler, die von der Jury nicht geleistet werden konnte.

4.2 Zeitverhalten

Demgegenüber zeigten sich signifikante Abweichungen im Zeitverhalten, wie aus der nachfolgenden Tabelle der Erkennungsraten (Wortformen/Sekunde) hervorgeht (da das System MORPHY nur auf PC verfügbar ist, sind hier keine Laufzeiten aufgeführt).

Platform: HP9000/735, HP-UX 9.01			
	MORPHIX	LAMORPH	GERTWOL
Text1	4210	3413	722
Text2	5263	4900	983
List1	1820	1668	282
List2	1175	969	473

Platform: HP9000/735, HP-UX 9.01			
	MORPH	MPRO	PC-K
Text1	152	12	11
Text2	204	18	11
List1	106	5	9
List2	148	5	7

Platform: HP9000/735, HP-UX 9.01		
	MORPHY	PLAIN
Text1	-	3
Text2	-	3
List1	-	2
List2	-	1

Nach dieser Tabelle lassen sich leicht zwei Gruppen auseinanderhalten, 1. Systeme mit hohen Erkennungsraten, mit Geschwindigkeiten von 10^2 – 10^3 Wörtern/Sekunde (MORPHIX, LAMORPH, GERTWOL, und MORPH) und 2. die übrigen Systeme mit niedrigeren Erkennungsraten/Sekunde.

Obwohl die Juroren sich einig waren, daß die Verarbeitungsgeschwindigkeit kein vorrangiger Faktor der Bewertung sein darf, war sie als nachrangiges Kriterium für die Auswahl der drei ausgezeichneten Systeme,

die alle der ersten genannten Gruppe angehören, durchaus von Bedeutung.

Das vierte System dieser Gruppe stellt aufgrund seiner sehr hohen Geschwindigkeit nahezu alle Mitbewerber in den Schatten. MORPHIX benutzt besonders effiziente Techniken für die Beschleunigung der morphologischen Analyse. Sie beruhen auf der extensiven Nutzung von Vollformen, Hardwarecache und Vermeidung von Rekursionen. Letzteres erweist sich in der Segmentierung von Wortformen (Derivations- und Kompositionsanalyse) als Nachteil.

Hinsichtlich der weiteren sekundären, eher technischen Eigenschaften ist MORPHIX als gut und solide, aber nicht als herausragend einzuordnen. Das System ist robust und softwaretechnisch ausgereift. Portabilität und Integrierbarkeit sind als sehr gut zu bewerten. Zur Lexikonerweiterung gibt es einen (etwas hölzernen) Klärungsdialog.

MORPHIX leistet nur Flexionsmorphologie in Analyse und Generierung, keine Derivation und keine Komposition. Dies wird als zu einschränkend empfunden; die Argumente der Autoren, daß bzgl. Derivation halt keine Fehler gemacht werden könnten, denn es müsse eben jedes abgeleitete Wort explizit im Lexikon stehen, kann aus linguistischer Sicht nicht akzeptiert werden. Gerade aus linguistischer Sicht sollte ein System auch Derivation bieten. Was die Komposition betrifft, so ist deren Fehlen besonders im Deutschen ein großer Nachteil. So gut wie jeder Text enthält neue Komposita. Der Abdeckungsgrad ist mit ca. 10000 Wortstämmen eher bescheiden; er wird durch das Fehlen von Derivation und Komposition eher ungünstiger.

Zusammenfassend ist festzustellen, daß Morphix, wenn es auch unter Gesichtspunkten des Zeitverhaltens und der technischen Solidität zur Spitzengruppe gehört, aus linguistischer Sicht keine hohe Bewertung erhalten kann, da ihm ein rein phänomenologischer Klassifikationsansatz zugrunde liegt. Damit ist keinerlei linguistische Generalisierungsmöglichkeit gegeben. Das mag zwar gut sein für eine schnelle morphologische Black Box, und da liegt seine Stärke, fällt aber deutlich gegen seine sehr viel mehr linguistisch charakterisierten Mitbewerber zurück. Es ist klar, daß ein ingenieursmäßiger Ansatz mit dem obersten Ziel hoher Performanz auf der Seite der Flexibilität (s. Komposition) Schwierigkeiten hat. Für bestimmte Anwendungen spielt das keine Rolle, ist aber generell zu kritisieren. Bezüglich der Korrektheit liegt Morphix im Mittelfeld mit Tendenz nach unten.

4.3 Korrektheit

Was die Korrektheit der Analyseergebnisse anbetrifft, so muß zunächst festgestellt werden, daß den Juroren eine ausführliche Sichtung aller Ergebnisse in der Kürze der Zeit unmöglich war. Hier machte sich der Mangel einer gezielt erstellten Testsuite bemerkbar. Allenfalls konnten aus der Liste der kompletten Verbformen (List2) verschiedener Paradigmata vergleichende Bewertungen abgeleitet werden. Die Juroren waren also bezüglich der Korrektheit einerseits auf die von den Systemen gelieferten statistischen Angaben über erkannte Formen und Fehlerquoten, andererseits auf Stichproben angewiesen, wobei sich die Wortformen- und Paradigmenlisten für die Beurteilung der Leistungsfähigkeit von morphologischen Analysatoren als besonders aussagekräftig erwiesen.

Zur Beurteilung der Korrektheit wurden von den meisten Systemen automatisch Statistiken der erkannten Wortformen geliefert. PC-KIMMO und GERTWOL schrieben erkannte und nicht erkannte Wörter in verschiedene Dateien, wo sie gezählt und ihre Erkennungsrate ermittelt werden konnte. Die folgende Tabelle listet die jeweiligen Anteile an erkannten Wortformen auf:

Text1 (ca. 2100 Wortformen)			
MORPH	LA-MORPH	PC-K	PLAIN
95.5	93.6	93.5	92.9
Text2 (ca. 1620 Wortformen)			
MORPH	LA-MORPH	MORPHY	PC-K
98.6	96.2	95.9	94.6
List1 (3817 Wortformen aus dem LIMAS-Korpus)			
GERTWOL	MPRO	LA-MORPH	PC-K
99.9	96.8	95.3	95.0
List2 (282 Wortformen aus Flexionsparadigmata, davon ca. 5 nicht wohlgeformt)			
PC-K	MORPH	PLAIN	GERTWOL
96.8	89.7	88.0	86.9

Text1 (ca. 2100 Wortformen)			
GERTWOL	MPRO	MORPHY	MORPHIX
92.6	91.1	89.2	86.0

Text2 (ca. 1620 Wortformen)			
GERTWOL	MORPHIX	MPRO	PLAIN
93.2	92.1	90.1	89.6

List1 (3817 Wortformen aus dem LIMAS-Korpus)			
MORPH	PLAIN	MORPHY	MORPHIX
93.8	93.6	86.9	74.3

List2 (282 Wortformen aus Flexionsparadigmata, davon ca. 5 nicht wohlgeformt)			
MPRO	LA-MORPH	MORPHY	MORPHIX
84.6	83.3	75.8	73.4

Diese Erkennungsraten besagen nur, wieviel Prozent der Wortformen akzeptiert wurden, nicht aber, wieviele davon RICHTIG analysiert wurden. Dies ist für die Jury aber ein wichtiger Aspekt gewesen, auch wenn sie sich nur anhand von Stichproben einen Eindruck von der Korrektheit verschaffen konnte. Die Frage bleibt offen, wie hoch die Erkennungsraten sein würden, wenn man die fehlerhaften Wortformen im Text herausrechnen würde.

Weiterhin zeichnet sich für die Jury ab, daß bei einer Beurteilung der Systemleistung eines morphologischen Analysesystems bezüglich des Kriteriums der Korrektheit auch der Typus der linguistischen Phänomene, mit denen der Test stattfindet, in Rechnung zu stellen ist. Im besonderen ist der Zusammenhang zwischen Systemleistung und Vorkommens häufigkeit linguistischer Phänomene zu berücksichtigen. Wortformen kommen in Texten mit unterschiedlicher Häufigkeit vor. Bekanntlich reicht eine relativ kleine Anzahl von Wortformen (Types) für die Abdeckung der überwiegenden Mehrzahl der Textwörter (Tokens). Daher kann man mit einer kleinen gut ausgewählten Anzahl von Wortformen leicht etwa bis zu 3/4 der fortlaufenden Texte parsen. Am anderen Ende des Häufigkeitsspektrums liegen die Wörter, deren Häufigkeit extrem niedrig ist und deren Erfassung aufwendig und unwirtschaftlich ist. Daraus folgt, daß die Häufigkeitsverteilung der Wörter in Texten auch bei der Beurteilung der Systemleistung der morphologischen Analyseverfahren relevant ist: Es kann nicht gleichgültig sein, ob die Analyse einer Wortform mit hohem oder mit niedrigem Rang in der Häufigkeitsskala gelingt oder fehlschlägt. Es ist legitim und auch zu erwarten, daß in den verschiedenen Häufigkeitsbereichen unterschiedliche Analysetechniken benutzt werden. Während die dichten Häufigkeitsbereiche mit einfachen (lexikalischen) Listenverarbeitungsverfahren bewältigt werden können, verlangen die dünnen Häufigkeitsbereiche (Wortformen mit niedriger Vorkommenshäufigkeit) andere, regelbasierte (linguistische) Methoden. Die häufigen Wortformen können u.U. vollständig aufgelistet und kodiert als Vollformenlexikon gespeichert und effizient für die Ermittlung der morphologischen Beschreibungen eingesetzt werden. Die Benutzung von Vollformenlexika ist angesichts der heute verfügbaren Speicherkapazitäten attraktiv, sowohl schnell als auch robust. Allerdings wird die eigentliche morphologische Verarbeitung (Zerlegung des Wortes in seine Bestandsteile und die Identifizierung ihrer konstitutiven Elemente (=Morpheme)) umgangen. Die Wörter selbst bleiben unanalysiert, die Zuordnung der funktionalen Informationen zu den materialisierten Trägermorphemen findet nicht statt. Es mag sein, daß ein solches Vorgehen in manchen Fällen unumgänglich ist (wie etwa bei Suppletivformen gut-besser oder sein-ist u.ä.), wo aber eine normale Wortzerlegung möglich ist (holen: V, hol-e, hol-st, hol-t, usw.; grün: A, grün-e, grün-en, grün-er usw.) bleibt diese Form der 'Analyse' unterhalb der linguistischen Intuition. Der Intuition des Sprechers über den Aufbau dieser Wörter kommt in diesen Fällen klar die Zerlegung des Wortes in sinntragende Einheiten (Morpheme) besser entgegen. (Es ist dabei argumentativ nebensächlich, ob eine solche spontane Gliederung des Wortes linguistisch tatsächlich zutreffend ist, wesentlich ist das Erkennen der Zerlegbarkeit.) Die regelgesteuerte Zerlegung der Wörter ist allerdings nicht nur intuitionsgerecht, sondern sie ist die Grundlage der Erweiterbarkeit der Analyseverfahren, ohne welche kein System auskommen kann. Regelgesteuerte Wortzerlegung ist also mehr als eine akademische Forderung, sie macht erst breit ausgelegte Systeme möglich, die auch in den dünnbesetzten Bereichen des Wortschatzes effizient eingesetzt werden können. Ohne regelgesteuerte Wortzerlegung sind

die Systeme auf die individuelle Erfassung der Einzellexeme angewiesen und können mit dem ständig wachsenden Lexikon der natürlichen Sprachen nicht fertig werden.

Aus diesen Gründen waren sich die Juroren einig, daß für die Bewertung der linguistischen Korrektheit insgesamt eher Materialen heranzuziehen seien, wie sie in den Wortlisten (List1, List2) enthalten sind. In diesen Listen ist Fehlerfreiheit weitestgehend garantiert (es sei denn, daß aus Kontrollgründen Fehler gezielt eingebaut wurden, wie in List2), so daß vom Material her Gründe für Fehlinterpretationen entfallen. Die Testergebnisse für List1 und List2 bestätigen deutlich, daß Systeme mit reinem Tabellensuchverfahren und ohne primär regelgesteuerte Flexions-, Derivations- und Kompositionsanalyse, die zwar recht gute Ergebnisse bei Volltexten einspielen, in linguistischer Hinsicht unzureichend sind.

Insgesamt ist festzustellen, daß es aus den genannten Gründen problematisch, ja sinnlos ist, aus den von den Systemen insgesamt gelieferten Erkennungsraten eine qualitative Reihung abzuleiten.

4.4 Linguistische Fundiertheit

Alle Systeme, die sich auf der MORPHOLYMPICS präsentierten, behandeln die Flexion des Deutschen. In der Regel verfügen die Systeme auch über Mechanismen der Kompositaanalyse, die allerdings häufig auf bestimmte Kompositionstypen beschränkt sind. Derivationsprozesse werden hingegen nur von den wenigsten Systemen behandelt, und auch hier ist - mit der Ausnahme der von Hanrieder und Maas präsentierten Systeme - die Behandlung der suffixalen Derivation auf einige wenige Derivationssuffixe begrenzt. Dennoch ist die Berücksichtigung der Derivation, auch wenn sie sich auf einzelne Suffixe reduziert, bereits als besonderes Gütekriterium der Systeme bewertet worden, da hierdurch dem Phänomen der "Produktivität" Rechnung getragen und eine deutliche Steigerung der Erkennungsrate nichtlexikalisierter Formen erreicht wird, was insbesondere dann deutlich würde, setzte man die Größe des Lexikons der einzelnen Systeme in Relation zu den von ihnen jeweils korrekt erkannten Formen.

Die linguistische Fundiertheit der Systeme zeigte größere Schwankungen. Wie oben bereits begründet, wurden dabei die theoretische linguistische Fundiertheit im engeren Sinne des Wortes und die Beherrschung der sprachlichen Empirie jeweils gesondert betrachtet. Klar ist das Streben nach einer theoretischen Fundierung in GERTWOL, bei PC-KIMMO, LAMORPH, bei MORPH und bei PLAIN zu erkennen. Allerdings erweist sich die tragende Theorie (ein von der Chomskyschen Transformationstheorie abgeleitetes Pattern-Matching-Verfahren) bei dem letzteren als unzulänglich. GERTWOL und PC-KIMMO stützen sich auf dieselbe theoretische Grundlage der Two-Level-Morphologie. Ohne linguistische Orientierung bleiben die Systeme im Rahmen der einfachen Item-und-Arrangement-Morphologie im Sinne von Charles Hockett, und ihre Implementierungen sind konzeptuell auf der Ebene der Tabellensuchverfahren. So operiert das MORPHIX-System konsequent durchgehend mit einem Tripel: Präfix, Stamm und Suffix. Alternative (funktional gleichwertige) Formvariationen werden durch Allomorphe und das Problem der Komposition und Derivation mit Hilfe des Lexikons bewältigt. Die Größe des Lexikons und die schnelle Suche darin kompensieren allerdings nicht das Fehlen einer regelgesteuerten Derivations- und/oder Kompositionskomponente, das auch in den relativ hohen Fehlerraten von MORPHIX einen Niederschlag findet.

Den bezüglich der linguistischen Fundierung und auch insgesamt besten Eindruck hinterließ das GERTWOL-System der LINGSOFT, Inc. aus Helsinki. Das System stützt sich auf das Konzept der Two- Level- Morphologie von Kimmo Koskenniemi (TWOL) und füllt diesen theoretischen Rahmen mit der kompletten lexikographischen Informationsmenge des Deutsch-Englischen Wörterbuchs von Collins. Das TWOL-Sytem sichert Effizienz durch die systematische Benutzung von einfachen finite state Automaten und kombiniert sie mit der Verarbeitungstechnik

der iterativ konzipierten Fortsetzungsklassen. Hierdurch ist es möglich, Morpheme effizient und intuitionsgerecht zu identifizieren. Da die parallelisierten Automaten zwischen der lexikalischen und der textuellen Repräsentation richtungsneutral definiert werden, kann das TWOL-System mit der gleichen Effizienz sowohl in der Analyse- als auch in der Syntheserichtung benutzt werden. Das System operiert auf einem Wortschatz von ca. 100 000 Lexemen, deren Geltungsbereich durch Aufstellung von Fortsetzungsklassen für Derivation und Komposition weiter vergrößert wird und in der Dimension von 1 Million liegt. Dazu kommen noch die Formenmengen, die durch die Flexionsendungen bedingt sind und regelgesteuert erfaßt werden. Die Fortsetzungsklassen sichern auf natürliche Weise die Erweiterbarkeit des Systems. Die erfaßten Derivationsklassen - Ableitungen mit den Suffixen -ung, -heit, -keit, -lich - sind sicherlich inkomplett, aber decken zweifelsohne die häufigsten Typen ab. Die Erfassung der Komposition beruht auf einer eigenständigen Klassifizierung der Kompositionsglieder, ausgearbeitet von Mariikka Haapalainen. Für die Bedienung des Systems stehen professionelle Tools der Xerox Corporation zur Verfügung, die es erlauben, die morphologischen Regeln linguistisch transparent zu formulieren und in parallelgeschaltete einfache Automaten zu übersetzen. Die Ergebnisse der morphologischen Verarbeitung sind zuverlässig, schnell und beruhen auf einer feinmaschigen linguistischen Analyse. Die hohe Qualität des Endproduktes wird allerdings durch hohe Professionalität während der Systementwicklung erkauft.

GERTWOL ist (neben PC-KIMMO) das einzige System, das die morphophonologische Alternation während der Analyse mithilfe eines linguistisch fundierten und computerlinguistisch wohl verstandenen Formalismus erfaßt (Two-Level System). Aufgrund dieser Konzeption ist das System für mehrere Sprachen einsetzbar. Diese universelle Einsetzbarkeit des Systems beruht nicht auf der geschickten Implementierung, sondern auf seiner linguistischen Konzeption. Die linguistische Konzeption erlaubt es auch, das System GERTWOL sowohl für die Analyse, als auch für die Generierung ohne jeglichen zusätzlichen Implementierungsaufwand zu benutzen.

Umfang und Übersichtlichkeit der morphologischen Analyse - GERTWOL verfügt über das umfangreichste Wortformenlexikon des Deutschen (100.000 Formen, die andern Systeme beinhalten weniger als die Hälfte). Die Ergebnisse der Analyse werden in einer vom linguistischen Standpunkt aus sehr übersichtlichen Form dargestellt.

Mit dem zweiten Platz bedachte die Jury das System MORPH, präsentiert von seinem Autor Gerhard Hanrieder (der das System im Rahmen seiner Magisterarbeit an der Universität Trier erstellt hat). Dieses System zeigte nicht nur eine sehr hohe Akzeptanzrate (obwohl hier auch ungesicherte Vermutungen durchrutschten) und eine sehr günstige Verarbeitungszeit, die Platzierung ist vor allem mit der soliden linguistischen Fundiertheit und der Feinheit der morphologischen Analyse des Deutschen begründet. Insbesondere die linguistisch motivierten Merkmalverarbeitungstechniken sind zu würdigen. MORPH setzt sehr konsequent die Prinzipien der generativen Morphosyntax (besonders das Kopfprinzip) für die Beschreibung der Morphologie des Deutschen ein und weist eine systematisierte Unterscheidung zwischen der schwachen, starken und gemischten Deklination der Adjektiva auf. Positiv bewerteten die Juroren ferner die regelgesteuerte Behandlung der Derivation und Komposition. Der modulare Aufbau des Systems erlaubt es weiterhin, einige Feinheiten der deutschen Morphologie elegant zu erfassen. So werden z.B. lexikalisierte Komposita, die im Deutschen besonders zahlreich sind, nur der Flexionsanalyse unterzogen und von der Wortbildungsanalyse ausgeschlossen. Aufgrund der einheitlichen linguistischen Konzeption ist das System sehr übersichtlich und seine Funktionsweise ist leicht nachvollziehbar. Erwähnenswert ist auch, dass MORPH zu den Systemen gehört, die von einer Person entwickelt worden sind. Unter diesen Halbprofis ist das System eindeutiger Sieger.

An dritter Stelle plazierte die Jury

das System LA-MORPH der Arbeitsgruppe Hausser, Erlangen, das ebenfalls eine solide linguistische Fundiertheit bietet. LA-MORPH stützt sich auf die linksassoziative Grammatik von Roland Hausser. Interessant ist die Konzeption von verschiedenen Paradigmatypen (regulär, semi-regulär, semi- irregulär und irregulär) sowie die Idee der distinktiven und exhaustiven Nutzung dieser Merkmale bei der Formenbildung. Das System unterscheidet zwischen allomorpherzeugenden und allomorphkombinierenden Regeln. Die Regeln zur Erzeugung der Allomorphe sind mit der Konzeption der Varia-Regeln der Natürlichen Generativen Phonologie verwandt. Sie werden bei der Initialisierung des Systems, also vor der eigentlichen morphologischen Analyse, auf die Einträge des Lexikons angewendet und erzeugen das sogenannte Allomorphielexikon, dessen Elemente bei der Wortformenanalyse herangezogen werden. Die Anwendung der Konkatenationsregeln wird von der Links-Assoziativen Grammatik gesteuert. In LA-MORPH werden mögliche Fortsetzungen einer Morphemfolge nicht mithilfe von Fortsetzungsklassen für bestimmte Morphemtypen angegeben, sondern durch die Angabe potentiell anzuwendender Folgeregeln, die als Regelmenge jeder Kombinationsregel zugeordnet sind. Die Kombinationsregeln werden sowohl zur Analyse als auch zur Generierung verwendet. Bei der Generierung werden alle von einem Grundallomorph ableitbaren Formen erzeugt. An einer Komponente zur Generierung einzelner Wort formen wird gegenwärtig gearbeitet. LA-MORPH ist auch auf andere Sprachen anwendbar. Das System wird intensiv gepflegt. Eine Benutzeroberfläche wird zur Zeit implementiert. Es ist auf mehreren Plattformen verfügbar.

Die weiteren von den Juroren nicht plazierten Systeme weisen Besonderheiten auf, die, hätte es verschiedene 'Startgruppen' gegeben, das eine oder andere dieser Systeme sicherlich jeweils in seiner Gruppe 'preisverdächtig' hätte werden lassen.

So verdient das System MORPHY von Wolfgang Lezius Anerkennung. MORPHY ist ein tabellengesteuertes System, das mit einer Flexionstabelle für Nomina arbeitet, auch wenn (möglicherweise) diese substantivische Flexionstabelle von Standardbeschreibungen des Deutschen übernommen worden ist. Allerdings fehlt eine ähnlich transparente Darstellung der Verben, deren Komplexität ohne linguistische Professionalität nicht zu bewältigen ist. Bezüglich Benutzerfreundlichkeit ragt MORPHY heraus, da es nicht nur über eine leicht bedienbare Benutzeroberfläche, sondern auch über eine einfach zu handhabende Komponente zur Lexikonerweiterung verfügt. MORPHY, nur unter DOS verfügbar, ist das einzige der acht Systeme, das sich für den naiven Benutzer eignet, alle anderen Systeme verlangen zu ihrer Benutzung einen professionellen Computerlinguisten oder Informatiker, der sich entweder in LISP-Strukturen, EMACS oder sonstigen UNIX-Tools auskennen muß. Wörterbucherweiterung ist auch bei den anderen Systemen möglich, aber meist nur mithilfe dieser komplexeren Tools.

Ein weiteres nicht plaziertes System zeichnet sich hinsichtlich der Berücksichtigung der Derivation aus, nämlich das von Anne Schiller präsentierte System PCKIMMO. Das System verfügt derzeit über ein ca. 20 000 Stämme und 500 verschiedene Endungen umfassendes Lexikon und liegt, gemessen an der Erkennungsrate, die bei den Testdaten im Durchschnitt bei 95 liegt, auf dem 3. Platz. Die Einbeziehung der Derivation erfolgt bei PCKIMMO durch ein sogenanntes Hypothesenlexikon, das beliebige Zeichenketten als Wortstämme zuläßt und mit Hilfe von Derivationssuffixen, aber auch mit Flexionsendungen und Wortstämmen, Hypothesen über die vorliegende Wortform als Analyse liefert. Die Begrenzung der suffixalen Derivation auf die Ableitungssuffixe -lich, bar und -ung hat jedoch zur Folge, dass auch von PC-KIMMO ein Substantiv wie 'Unterschiedlichkeit' nicht analysiert werden kann, sofern es - wie das hier der Fall war - nicht im Stammlexikon aufgeführt ist.

Was die Komposition anbelangt, so werden von PC-KIMMO substantivische Komposita erkannt, wobei als Erstglied des Kompositums neben Nomina auch Adjektive auftreten können. Mit Blick auf die

Kompositaanalyse wurden in das Stammlexikon für Nomina auch Elemente aufgenommen, die selbst in Texten nicht frei vorkommen. Hierzu gehören Elemente wie Anti, Nicht-, und Pseudo-, deren Status unter linguistischen Gesichtspunkten wohl zutreffender mit der Einordnung in ein Präfix oder Präfixoidlexikon beschrieben worden wäre als mit der Zuordnung in ein Lexikon von Nominalstämmen. Bei einer Ausweitung der Kompositionsmechanismen auf die Verb-Nornen-Komposition ließe sich dann ebenfalls der Eintrag von nicht isoliert auftretenden Erstgliedern wie 'Sammel-' in das Lexikon der Nominalstämme vermeiden.

In bezug auf den Aufbau des Systems erfüllt das 2-Ebenen-System PC-KIMMO die Forderung nach Modularität: Lexikon, phonologisch/orthographische Regeln sowie das eigentliche Programm sind voneinander getrennte Einheiten. PC-KIMMO folgt der Philosophie der Two-Level-Morphologie von Kimmo Koskenniemi; indem seine Autorin dieses System übernahm, sparte sie die Entwicklungsarbeit für die Implementierung und konnte sofort auf einer höheren Ebene mit der effektiven Entwicklungsarbeit beginnen. Dies ist durchaus legitim und in der Forschung und Entwicklungspraxis auch üblich. Allerdings wird bei dem Einzelnen die Isolierung der eigenen Leistung gegenüber der Gruppe erschwert, da die individuelle Leistung konsequent in dem Rahmen eines bestehenden Systems erbracht wird.

Mit Hilfe der systematisierten Flexionsmuster (440 patterns) wird im PLAIN-System (vorgetragen von Fisser/Koch) eine breite und auch verläßliche Erfassung der Wortformen erzielt. Die nicht erkannten Formen sind typischerweise Abkürzungen. Es wird aber auch deutlich, daß sich die statischen Flexionsmuster für die Wortformenerkennung nicht direkt eignen und unvertretbar niedrige Erkennungsraten Wort/Sekunde erbringen. Was die Produktivität der Komposition im Deutschen anbelangt, so wurde ihr auch in PLAIN Rechnung getragen. Technisch wird dies dadurch realisiert, daß von dem als Übergangsnetzwerk realisierten Lexikon aus, in dem Stämme und Grundformen enthalten sind, eine Kante an den Anfangsknoten des Netzwerkes führt, so daß potentiell beliebig lange Komposita erkannt werden. Derivationsmechanismen werden von PLAIN nicht berücksichtigt. Die Flexionsanalyse erfolgt bei diesem System durch die Zuordnung von Flexionsparadigmen zu einzelnen Stämmen. Die zutreffenden Paradigmen eines Stammes werden auf der Grundlage eines Musterabgleichs mit repräsentativen Formen eines Lemmas, die im Fall der Substantive aus dem Artikel und der Form des Nominativs Singular, der Genitiv Singular- Form sowie der Form im N ominativ Plural bestehen, ermittelt.

Die Juroren bestätigen schließlich die hochgradige individuelle Professionalität von Heinz Dieter Maas, der mit MPRO eine eigens für die MORPHOLYMPICS aus den Saarbrücker Analyseprogrammen isolierte morphologische Komponente vorstellte, die das Ergebnis einer jahrzehntelangen Beschäftigung mit der Analyse des Deutschen ist. Sein professionelles computerlinguistisches Wissen ist einmalig empirisch und enthält viele interessante Elemente, wie etwa eine Verbotsliste für unzulässige morphologische Zerlegungen, oder die Unterscheidung von Kurzformnotation und Langformnotation. Die Stärke seines Analysesystems MPRO ist allerdings eher die Robustheit als die linguistische Transparenz. Wortformenerkennung ist unumgänglich mit der Morphologie verbunden und es ist nur positiv zu beurteilen, wenn hierfür die Ergebnisse der deskriptiven Linguistik (traditionell oder modern) aufgegriffen werden. Da die Produktivität von Wortbildungsmustern die Konzeption von MPRO linguistisch motiviert hat, finden in diesem System neben Kompositions- auch Derivationsmechanismen besondere Berücksichtigung. Bei der Derivation werden sowohl Präfigierungsprozesse als auch Suffigierungsprozesse regelgesteuert abgeleitet. Bezogen auf die suffixale Derivation beschreiben 12 Wortbildungsmuster die Ableitung von Adjektiven und Nomina zu Verben und 11 Wortbildungsmuster die Ableitung von Verben, Nomina und Adjektiven zu Substantiven. Zu erwähnen ist auch, daß MPRO als einziges

System Angaben zur Bedeutung von Wortbildungen macht, die sowohl auf semantischen Angaben der einzelnen Einträge im Lexikon als auch auf Angaben zur Bedeutung beruhen, die mit den jeweiligen Wortbildungsregeln verknüpft sind. Da die mit den Wortbildungsregeln verbundenen Bedeutungsangaben im Bereich der Komposition nur die häufigsten Beziehungen zwischen Konstituenten berücksichtigen, erhalten einzelne Komposita bei der Analyse zum Teil jedoch auch keine oder nicht zutreffende semantische Angaben.

4.5 Professionalität

Die konkurrierenden Systeme weisen bezüglich der Professionalität ihrer Urheber erhebliche Unterschiede auf. Sie stammen z. T. von Einzelkämpfern und Amateuren, zum Teil von professionellen Entwicklern und jahrelang arbeitenden Forschungsgruppen. Da die Regularien des 'Wettkampfs' eine Diversifizierung dieser Gruppen nicht vorsahen und auch der' Amateurstatus' nicht vorgeschrieben war, starteten Profis" (wie LINGSOFT) und echte Amateure (wie Wolfgang Lezius) in derselben Gruppe zusammen. Subjektiv haben die Juroren die Leistung der Einzelkämpfer" (insbesondere von Anne Schiller, Stuttgart, Wolfgang Lezius, Paderborn und Gerhard Hanrieder, Erlangen) sehr hoch eingeschätzt, da Professionalität nicht an Gruppenarbeit gebunden ist. Insgesamt mußten sie jedoch den professionellen Kollektiven nachgeordnet werden.

5 Abschließende Bemerkungen

Es wurde schon eingangs bemerkt, dass mit der MORPHOLYMPICS als Veranstaltung, aber auch bezüglich der Art und Weise der wettkampfmäßigen Bewertung linguistischer Beschreibungssysteme, Neuland betreten worden ist. Beinahe zwangsläufig ergibt sich aus dieser Feststellung, daß aus den gemachten Erfahrungen Lehren für künftige ähnliche Veranstaltungen gezogen werden müssen:

> Die Verfahrensweise der Bewertung und die Bewertungskriterien, nach denen die Juroren vorgehen, müssen zu einem sehr viel früheren Zeitpunkt festliegen und nach Möglichkeit den Teilnehmern bekannt gegeben werden.

Den Juroren muß, wenn es, wie an sich für linguistische Gegenstände nicht anders denkbar, bei einer mehrdimensionalen Bewertung (Zeitverhalten, Korrektheit, linguistische Fundiertheit etc.) bleiben soll, mehr Zeit für die Sichtung der Ergebnisse der Testläufe zur Verfügung stehen.

Es ist die Einrichtung mehrerer Startgruppen zu erwägen, in denen dann jeweils ein Preis vergeben werden kann.

Auch wenn in mancherlei Hinsicht bezüglich der Kriterien und der Vorgehensweise bei Evaluierungen direkt gegenübergestellter Systeme noch Wünsche offen sein mögen, so hat sich die Jury der ersten MORPHOLYMPICS doch bemüht, eine ausgewogene Entscheidung zu treffen. Aus der vorgetragenen Begründung dieser Entscheidung sollte Anerkennung für alle vorgestellten Systeme deutlich zu erkennen sein. Es bleibt zu hoffen, dass diese Entscheidung national und international so diskutiert wird, daß auch in Zukunft 'Wettkämpfe' dieser Art für andere Werkzeuge computerlinguistischer Forschung und Anwendung möglich sind.

Results of the 1. Morpholympics

held March 7 and 8 1994
Universität Erlangen – Nürnberg
coordinator: Roland Hausser

The draw of the ballot resulted in the following order of presentations:

1	MORPHY	Wolfgang Lezius, Universität Paderborn
2	PC-KIMMO	Anne Schiller, Universität Stuttgart
3	MORPH	Gerhard Hanrieder, FORWISS, Erlangen
4	MORPHIX	Wolfgang Finkler, Ottmar Lutzy, Universität des Saarlandes
5	PLAIN	Henriette Visser, Heinz-Detlev Koch, Universität Heidelberg
6	LA-MORPH	Gerald Schüller, Oliver Lorenz, Universität Erlangen
7	GERTWOL	Kimmo Koskeniemmi, Mariikka Haapalainen, Lingsoft, Helsinki
8	MPRO	Heinz-Dieter Maas, IAI, Saarbrücken

The judges were:

Istvan Bátori	Universität Koblenz-Landau
Gregor Dogil	Universität Stuttgart
Günther Görz	Universität Erlangen
Winfried Lenders	Universität Bonn
Uta Seewald	Universität Hannover

First prize:	GERTWOL
Second prize:	MORPH
Third prize:	LA-MORPH

Below the results are given in more detail.

1 Number of wordforms per second

As stated in the announcement, the test data were to be analyzed on a workstation, a PC and an Apple Macintosh. However, only two systems ran their programs on all three platforms. Five systems ran only on the workstation and one system ran only on the PC under DOS.

HP9000/735, HP-UX 9.01								
	MORPHIX	LAMORPH	GERTWOL	MORPH	MPRO	PC-K	MORPHY	PLAIN
Text1	4210	3413	722	152	12	11	-	3
Text2	5263	4900	983	204	18	11	-	3
List1	1820	1668	282	106	5	9	-	2
List2	1175	969	473	148	5	7	-	1

Intel 486/33, Linux 99.15
(other systems did not port to this platform)

	LAMORPH	MORPHIX
Text1	357	169
Text2	507	203
List1	167	87
List2	129	67

Intel 486/33, MS-DOS 6.2
(other systems did not port to this platform)

	LAMORPH	MORPHY
Text1	27	7
Text2	37	6
List1	35	6
List2	45	5

MAC IIci, System 7.0
(other systems did not port to this platform)

	MORPHIX	LAMORPH
Text1	78	14
Text2	93	18
List1	38	17
List2	36	25

2 Percentage of word forms recognized

Text1 (roughly 2100 word forms)

MORPH	LA-MORPH	PC-K	PLAIN	GERTWOL	MPRO	MORPHY	MORPHIX
95.5%	93.6%	93.5%	92.9%	92.6%	91.1%	89.2%	86.0%

Text2 (roughly 1620 word forms)

MORPH	LA-MORPH	MORPHY	PC-K	GERTWOL	MORPHIX	MPRO	PLAIN
98.6%	96.2%	95.9%	94.6%	93.2%	92.1%	90.1%	89.6%

List1 (3817 word forms from LIMAS-corpus)

GERTWOL	MPRO	LA-MORPH	PC-K	MORPH	PLAIN	MORPHY	MORPHIX
99.9%	96.8%	95.3%	95.0%	93.8%	93.6%	86.9%	74.3%

List2 (282 words of which about 5% are not wellformed)

PC-K	MORPH	PLAIN	GERTWOL	MPRO	LA-MORPH	MORPHY	MORPHIX
96.8%	89.7%	88.0%	86.9%	84.6%	83.3%	75.8%	73.4%

3 Summary of the results

Speed

	First place	Second place	Third place
HP-UX	4 MORPHIX	6 LAMORPH	7 GERTWOL
Linux	6 LAMORPH	4 MORPHIX	
MS-DOS	6 LAMORPH	1 MORPHY	
MAC S7	4 MORPHIX	6 LAMORPH	

Coverage

	First place	Second place	Third place
Text1	3 MORPH	6 LAMORPH	2 PC-K
Text2	3 MORPH	6 LAMORPH	1 MORPHY
List1	7 GERTWOL	8 MPRO	6 LAMORPH
List2	2 PC-K	3 MORPH	5 PLAIN

GERTWOL

LINGSOFT Oy

QUESTIONNAIRE FOR MORPHOLYMPICS 1994

Name and origin of the participating system 1

The system is called GERWOL (German Two-level system) by Lingsoft, Inc. The linguistic description is done in the two-level framework, and is based on the Collins German Dictionary (CGD) regarding the initial contents of the vocabulary. The two-level description can be run on two distinct finite-state implementations, the Lingsoft TWOL program which has evolved from the original first implementation. A second, more advanced implementation in terms of speed and space, is made using the Xerox Lexical tools. Both the framework for describing German morphology and its implementations are independent of the German language which can be seen from the existence of systems made in the same framework using the same tools (Finnish, Swedish, English, Russian, Swahili, Danish, Estonian, French, all with full scale dictionaries).

The GERTWOL is a result of team work: Mariikka Haapalainen as the linguist, Mikko Silvonen as the computer scientist, Krister Lindin directing the work, and Kimmo Koskenniemi and Fred Karlsson as advisors.

This text is the same as was submitted to the judges, except that a few spelling errors have been corrected, and that the footnotes have been added. We wish to express our most sincere thanks to professor Hausser for the help and assistance he and his staff gave to the Lingsoft team in order to facilitate the participation.

1 Conceptual Criteria

1.1 Declarative specification of the lexical entries and the rules

The two-level description consists of a lexicon and rules. The rule component defines morphophonological (or morphographemic) alternations of the language, and the lexicon identifies the morphemes (or other entries used as units) in the description, and their possible sequences in word forms.

The two-level formalism does not dictate in detail which alternations should be treated with the rule component and which by the lexicon. The designer may choose among different styles of description. GERTWOL is intended to be a full scale system with a wide vocabulary from the beginning. It is not an experimental system for exploring various topics of morphology (such as derivation, historical development of the language, etc.).

It would have been fairly easy to implement e.g.. many more of the alternations in irregular verbs using rules and morphophonemes. After careful consideration, we decided not to follow that path. Instead, most of the irregular or strictly unproductive alternations have been described without resorting to morphophonemes simply in order to keep the description as transparent and verifiable as possible.

The lexical material behind GERTWOL consists of distinct parts:

1. The bulk of word entries converted from the CGD in an intermediary format (called GLEX format) from which the two-level entries are created by a set of Perl conversion programs.
2. A table for irregular verbs which lists all their stems and possible prefixes.
3. A set of entries for closed class words with idiosyncratic properties (such as pronouns) written as two-level entries (and thus not converted from CGD).
4. Proper nouns and other words not in the CGD which have been added to enhance the coverage of GERTWOL.

The full GERTWOL dictionary is generated from this data base automatically using a Unix make program so that only those components will be recomputed where the source has actually changed.

Because of the comprehensive and carefully composed selection of word entries in the Collins German Dictionary, most of the correct derivations of common words are mechanically included in our dictionary. Presence of correct derivations and the absence of incorrect (or overgenerated) derivations is as much true as it is in the CGD (but some corrections have, of course, been made).

In the GLEX file containing cleaned and converted entries from the CGD there is an entry for "maus" (a mouse)

f m>_au<s S7+

This is still technically very much like the entry in the file for typesetting. It marks the gender "f", length "_", and inflectional class (which has been coded as S7), and a mark for umlaut "+" affecting this word. Only the nouns already in the CGD were ever turned to this format. New nouns are added in a format which is used in the TWOL dictionary.

The above GLEX entry is converted into a TWOL lexicon entry:

mau_s S7+/f;

This consists of two fields: "ma_us" is the so called lexical representation of the stem, and "S7+/f" is the continuation class, which determines the name of the lexicon where the next morpheme (or entry) is to be found. The first field may be omitted if the entry is a null entry with no morphophonological material in it. The lexical representation is related to the surface form using two-level rules. All rules operate in parallel and synchronized.

In most entries the first field is like in the above example. The first field might, however, be divided into two parts by a colon (:). Then, the former part is what we like to see in the base form or as features following it (e.g.. "+S+FEM+SG+NOM", "+S+FEM+PL+NOM"). The latter part is, then, the morphophonological representation of a stem or an ending (e.g.. "", "e@U"). The standard case corresponds to the situation when the parts before and after the colon are identical, i.e. the lexical representation to be found as a part of the analysis is the same as is underlying the base form.

There are minilexicons for inflection etc. which were fixed once the inflectional types have been established. In the following only the entry "mau_s S7+/f" is specific to the word "maus". The entries start at the very first lexicon which is called "Root". Each lexicon starts with a keyword and the name of the lexicon, and this is followed by the entries in the lexicon.

```
LEXICON Root
* SUBST;                    ! The "*" marks the next letter as capital
...                         ! Next entry from lexicon SUBST

LEXICON SUBST
mau_s S7+/f;                ! This is the converted entry
                            !  "_" marks the a "long" vowel

LEXICON S7+/f
S7+/f/end;                  !  An entry with only one item is a continuation
SCONT;                      ! Compound formation
s1 SCONT;                   ! Compound formation
CHEN;                       ! Diminutive
LEXICON S7+/f/end
Sg3/f/end;                  ! To singular endings
Pl1+/f/end;                 ! To plural endings
LEXICON Sg3/f/end           ! Here are (null) endings for singular
+S+FEM+SG+NOM:              ! ## as a next lexicon name marks the end
##;
+S+FEM+SG+AKK:
##;
+S+FEM+SG+DAT:
##;
+S+FEM+SG+GEN:
##;
```

```
LEXICON Pl1+/f/end
+S+FEM+PL+-
NOM:e@U ##;
+S+FEM+PL+AKK:e@U    ! @U triggers umlaut
##;
+S+FEM+PL+DAT:en@U
##;
+S+FEM+PL+GEN:e@U
##;
```

Altogether, these entries allow for a sequence of segments as follows:

Base	Lexical	From lexicon	Surface
*	*	Root	0
m	m	SUBST	M
a	a	SUBST	"a
u	u	SUBST	u
-	-	SUBST	0
s	s	SUBST	s
+S	e	Pl1+/f/end	e
+FEM	@U	Pl1+/f/end	0
+PL	0	Pl1+/f/end	
+NOM	0	Pl1+/f/end	

The task of the two-level rules is to relate the second column characters with the surface form which was included in the table as the far right column.

The umlaut trigger is present only in certain noun endings, and the umlaut rule is thus triggered only for those nouns. (Umlaut verbs and adjectives are anyway "irregular" and their stems are listed.) The trigger present in certain endings will force correspondence "a:ä" if and only if there are no instances of word boundaries or "a", "o" or "u" on the lexical level in between. There may, however be exactly one "u" just after the "a". On the other hand, "u" corresponds to "ü" according to the second rule unless it is after an "a".

```
"Uml a~ä"
        a:ä    ⇔    _ \ [%#: | a: | o: | u:]* @U: ;
                    _ u: \ [%#: | a: | o: | u:]* @U: ;
```

The first context does not fit to the above example, because there is a "u" between the trigger and the "a". The second context matches, as there is one "u" and some characters which are not any of the listed.

```
"Uml u~ü"
        u:ü    ⇔    \a: _ [%#: | a: | o: | u:]* @U: ;
```

This rule is for words, where it would be the "u" which needs umlaut. In the example that would not be appropriate, therefore there is a left context "any character but not 'a' ". Thus, the rule accepts the "u~u" correspondence.

The principle in designing this morphological analyzer has been that the CGD has all (or most) necessary distinctions marked, and the distinction is captured from there. The information in the conventional dictionary is related with decisions of the following types:

1. In which lexicon the resulting entry will be placed? This depends e.g.. on the part of speech. In some cases, this is based on extensive global computation, e.g.. in order to determine compounding characteristics which are expressed partly by placing the entries, and partly by continuation classes. Some entries are put to more than one lexicon.
2. What continuation class is assigned to the entry? This depends on the part of speech, and the set of endings listed in CGD.
3. How should the lexical representation in GLEX be constructed? Some extra markings of the CGD are deleted, some characters are decoded from their multi-character representations, some markings are repositioned, etc.

Rules and associated morphophonemes or diacritic markers are used only where they can be used without exceptions, and where they simplify the combinatorics. The combinatory properties of various endings in German are more complex than generally expected, and this seems to result in fairly complicated descriptions, no matter how they are designed (if we wish to design an accurate system).

1.2 Relation between lexical entries and word forms

The GLEX entry for "lächeln" ('smile') is:

el/V la">.<ch i

This has the classification "el/V" which will serve as the continuation class, a multi-character representation "a"" for "ä", ">.<" as stress position indicator, and a marking "i" for inseparability (also used for verbs without any prefixes). Again, this technical format is not really meant for users to enter or study. It is used for generating all necessary entries in the TWOL

lexicon. In that respect, the GLEX entry is the non-redundant expression describing the entry of the lexeme, and the TWOL lexicon is something which has been mechanically compiled (and the runtime binary lexicons results of even more comprehensive compilation and preprocessing).

The following are a few analyses of forms of "anlächeln":

```
anlächle
        anlächeln+V+KONJ+PRÄS+SG3
        anlächeln+V+KONJ+PRÄS+SG1
        anlächeln+V+IND+PRÄS+SG1
anlächeln
        anlächeln+V+KONJ+PRÄS+PL3
        anlächeln+V+KONJ+PRÄS+PL1
        anlächeln+V+IND+PRÄS+PL3
        anlächeln+V+IND+PRÄS+PL1
        anlächeln+V+INF
```

The analyses are based on the following parts of the lexicon. Again, the entry "läch el/V(sep);"is the only one pertaining to this particular lexeme. The lexicon combines "an" which is given at the position where ordinary verbs could start, and "an#läch el/V(sep)" which is where the separable verbs are. The classification of the relevant properties of the verb have been indicated in the GLEX form, and therefore the GLEX entry may result in several entries in the TWOL lexicon (which is a technical matter of compiling the source description into some operational form).

```
LEXICON VERB
an SEPCONT;                ! **
...
LEXICON SEPCONT
# VSEPBASE;                ! **
...
LEXICON VSEPBASE
läch el/V(sep);            ! **
...
LEXICON el/V(sep)
ele2n: V1el/V(sep);        ! ** defines the base
                           form
...                        ! e2 is a morphophone-
                           me

LEXICON V1el/V(sep)
+V+INF:ele2n ##;
+V+IND+PRÄS+SG1:le
##;
+V+KONJ+PRÄS+SG1:le
##;
+V+KONJ+PRÄS-              ! **
+SG3:le ##;
...
```

The setup for the correspondence for "anlächeln+V+KONJ+PRÄS+SG3" — "anlächle" is marked with "!~**" above, and is as follows:

Base	Lexical	Surface	Lexicon of entry
			Root
a	a	a	VERB
n	n	n	VERB
#	#	0	SEPCONT
l	l	l	SEPVBASE
ä	ä	ä	SEPVBASE
c	c	c	SEPVBASE
h	h	h	SEPVBASE
e	0		el/V(sep)
l	0		el/V(sep)
e2	0		el/V(sep)
n	0		el/V(sep)
+V	l	l	V1el/V(sep)
+KONJ	e	e	V1el/V(sep)
+PRÄS	0		V1el/V(sep)
+SG3	0		V1el/V(sep)

Again, the two-level rules relate the second and the third column to each other. The actual base form and analysis shown to the user ("lächeln+V+KONJ+PRÄS+-SG3") is generated with the same rules but in a reverse direction.

The morphophoneme "e2" in the base form has been established and accompanied with two-level rules because the presence or omission of the "e" in infinitives (and other similar endings) can be stated in terms of the phonological context. (The present example word alone does not, of course, justify these decisions.)

L = l r ;

Default correspondence: e2:e

```
"e2:0"
   e2:0 ⇔ [C: | :e :i | [:a | :e | :ä] :u] %_:* e: [L: | r l] _ ;
"e2:0 opt"
   e2:0 ⇒ V %_:* h* _ n ;
"e2:e"
   e2:e ⇐ C i %_:* _ ;
```

The first rule accounts for infinitives like "handel0n", the second the optional omission of "e", and the third one forbids the omission in stems ending in single vowel "i" ("kni-en").

1.3 Transparency and linguistic motivation of the rules

Word forms "Tisch", "Tisches", "Tischen" get the following analyses.

```
Tisch
        Tisch+S+MASK+SG+DAT
        Tisch+S+MASK+SG+AKK
        Tisch+S+MASK+SG+NOM
Tisches
        Tisch+S+MASK+SG+GEN
Tischen
        Tisch+S+MASK+PL+DAT
```

The entry for "Tisch" comes from the

CGD and has the following format in the GLEX file:

m tisch S1(s/es)

This entry is converted trivially into a TWOL entry, and positioned in the SUBST lexicon:

```
LEXICON Root
SUBST;
...
LEXICON SUBST
tisch S1(s/es)/m;
...
LEXICON S1(s/es)/m
S1(s/es)/m/end;
...
LEXICON S1(s/es)/m/end
Sg1s/es/m/end;
Pl1/m/end;
LEXICON Sg1s/es/m/end
+S+MASK+SG+NOM: ##;
+S+MASK+SG+AKK: ##;
+S+MASK+SG+DAT: ##;
+S+MASK+SELTEN+SG+DAT:e ##;
+S+MASK+SG+GEN:s ##;
+S+MASK+SG+GEN:es ##;
LEXICON Pl1/m/end
+S+MASK+PL+NOM:e ##;
+S+MASK+PL+AKK:e ##;
+S+MASK+PL+DAT:en ##;
+S+MASK+PL+GEN:e ##;
```

Note that even rare (SELTEN) forms have been accounted for. These entries can be eliminated if needed automatically. There is a number of distinct series of endings for singular and another set of series for plural. There would be many more distinct series if both were given at the same time.

The next examples "vorbeischwammst", "vorbeischwämme", "vorbeigeschwommenen" exemplify a few points. One is the productive combination of the prefix and the verb, and the other is the need to describe the idiosyncratic behavior of "schwimmen" just once in the description. All irregular (or strong) verbs have been coded once in a tabular form where they are classified (e.g.. U2) and all its stems are given:

U2 schwimm schwimm schwimm schwamm schwömm; schwämm+-SELTEN:V5Ua schwomm ab- an- davon- durch- hinaus- hinüber- ver:* herüber- ab- an- durch:* frei-

Note that the table also includes a still more idiosyncratic form "schwämm" as an alternative to "schwömm". The table also contains a list of prefixes applicable to this verb along with the coding whether the result is inseparable (:), and whether the past participle is formed without prefix "ge" (*).

Here are the analyses of the example words:[2]

```
vorbeischwammst
    vorbeischwimmen+V+IND+PRÄT+SG2
vorbeischwämme
    vorbeischwimmen+SELTEN+V+KONJ+PRÄT+SG3
    vorbeischwimmen+SELTEN+V+KONJ+PRÄT+SG1
vorbeigeschwommen
    vorbeigeschwommen+A(PART)+POS
    vorbeischwimmen+V+PART+PERF
```

```
LEXICON VERB
vorbei SEPCONT;          ! Not automatically
                         from CGD
...
LEXICON SEPCONT
# VSEPBASE;
...
LEXICON VSEPBASE
schw0immen:schwa0mm      ! Generated from
V4U;                     GLEX entry
schw0immen+SELTEN:-      ! Generated from
schwä0mm V5Ua;           GLEX entry
...
LEXICON
VSEPBASE-GE
schw0immen:schwo0mm      ! Generated from
V6U(sep);                GLEX entry
...
LEXICON V4U
+V+IND+PRÄT+SG2:s1t
##;
...
LEXICON V5Ua
+V+KONJ+PRÄT+SG1:e
##;
+V+KONJ+PRÄT+SG3:e
##;
...
LEXICON PARTEND
vorbei#ge#schwomm
V6U(sep)/adj;
...
LEXICON V6U(sep)/adj
A01st-en-part;
LEXICON A01st-en-part
en+A(PART)+POS:en
ADJ;
...
```

[2] Note that the third example word here is a wrong one. The cited entries cover the correct "vorbeigeschwommenen" equally well.

```
LEXICON ADJ
+SG+AKK+MASK:e4n
##;
+SG+GEN+MASK:e4n
##;
+SG+DAT+MASK:e4n
##;
+SG+GEN+NEUTR:e4n
##;
+SG+DAT+NEUTR:e4n
##;
+SG+GEN+FEM:e4n
##;
+SG+DAT+FEM:e4n
##;
+PL+NOM:e4n ##;
+PL+AKK:e4n ##;
+PL+DAT:e4n ##;
+PL+GEN:e4n ##;
...
```

Default correspondence: e4:e

```
"e4:0"
  e4:0   ⇔   e %_:* _ ;
```

The next examples "Hausdächern", "Häusermeers" exemplify compounding and this is the area where significant effort has been spent in designing GERTWOL, because this (rather than derivation) is decisive for the coverage and usability of the system. There is no way of listing enough compounds as ready made entries in a dictionary because compounds are made on the fly when needed.

The following are analyses using the TWOL engine. There is a word boundary (#) between the components which is shown by the program if so desired.

```
Hausdächern
    Haus#dach+S+NEUTR+PL+DAT
Häusermeers
    Häuser#meer+S+NEUTR+SG+GEN
```

The entries for "Haus", "Dach" and "Meer" in the GLEX file have been classified (automatically) as nouns participating the normal compounding, and therefore the TWOL entries were placed in the SUBST lexicon. For a great number of nouns, the glue element after the stem can be determined based on the inflectional classification, or the shape of the stem. For some others, there are more alternatives open. Some of these could be excluded, but the effort needed to carry it out for the whole vocabulary is substantial.

```
LEXICON SUBST
hau_s S4+(es)/nt;
dach S4+(s/es)/nt;
mee_r S1(s/es)/nt;
...
LEXICON SCONT
# SUBST;                    ! recursion to further
                             noun stems
...
LEXICON S4+(es)/nt
SCONT;                       ! Haus-
es SCONT;                    ! Hauses-
er@U SCONT;                  ! Häuser-
...
LEXICON S4+(s/es)/nt
S4+(s/es)/end;
...
LEXICON
S4+(s/es)/nt/end
Pl4+/nt/end;
...
LEXICON S1(s/es)/nt
Sg1s/es/nt/end;
...
LEXICON Sg1s/es/nt/end
+S+NEUTR+SG+GEN:s
##;
...
LEXICON Pl4+/nt/end
+S+NEUTR+PL+DAT:ern@U
##;
...
```

The following example "Unabhängigkeitserklärung" refers to compounding and derivation. The GLEX contains:

f unabhängigkeit S9en

The former comes from CGD as a ready made derivation (probably we would like to make "-keit" into a more productive derivational suffix, but hardly any need has shown up yet). All nouns ending in "-eit" are feminine, and have a fixed and obligatory compounding glue element "s" (no glue and "en" are forbidden).

V01* erkl>_a"<r i

This entry for the verb is converted into both as a verbal entry for "erklären" and as a derived noun "Erklärung". Distinct entries are produced automatically, and separate entries are needed because the compounding characteristics of verbs and nouns are quite different.

Here, again, is the analysis:

```
Unabhängigkeitserklärung
    Unabhängigkeits#erklärung+S+FEM+SG+GEN
    Unabhängigkeits#erklärung+S+FEM+SG+DAT
    Unabhängigkeits#erklärung+S+FEM+SG+AKK
    Unabhängigkeits#erklärung+S+FEM+SG+NOM
```

The lexicon entries are inserted in the expected place, SUBST lexicon. Two entries are generated for "unabhängigkeit", one for the inflectional paradigm and the other for

the selective compound glue element. Because compounding behavior ("-s-") is relatively independent of the inflectional class (S9en/f/end), it is simpler to use separate entries. (There is no penalty for any of the implementations in terms of space or speed.)

```
LEXICON SUBST
unabhängigkeit S9en/f/end;
unabhängigkeit SCONT-s;
erklä_rung S-ung;
...
LEXICON SCONT
# SUBST;
...
LEXICON SCONT-s
s SCONT;
LEXICON S-ung
S9en/f/end;
...
LEXICON S9en/f/end
Sg3/f/end;
...
LEXICON Sg3/f/end
+S+FEM+SG+NOM: ##;
+S+FEM+SG+AKK: ##;
+S+FEM+SG+DAT: ##;
+S+FEM+SG+GEN: ##;
```

The following is the example word "unlesbares". Presently "-bar" derivation is not handled in a productive manner because many of the entries of this kind are already given in CGD. We have doubts whether this derivation is possible for all verbs.

```
LEXICON
ADJ-NONCOMP
un ACONT1;              ! un- permitted only
                        at the beginning of a
                        word, thus here
...
LEXICON ACONT1
| ASTART;               !  | is a boundary
                        (weaker than a word
                        boundary)
...
LEXICON ASTART
le_sbar A01r;           ! From CGD
...
LEXICON A01r
+A+POS: ADJ;
...
LEXICON ADJ
+SG+NOM+NEUTR:e4s
##;
+SG+AKK+NEUTR:e4s
##;
...
```

The following is the example word "durchdachte" where the underlying "denken" is a weak irregular verb, or "dachen" which is a regular verb. Both verbs may follow the initial component "durch". There are two kinds of "durch": one separable and the other inseparable.

```
durchdacht
  durchdacht+A(PART)+POS+SG+AKK+FEM
  ...
  durchdenken+V+IND+PRDT+SG1
  durchdenken+V+IND+PRDT+SG3
  durch#denken+V+IND+PRDT+SG1
  durch#denken+V+IND+PRDT+SG3
  durchdachen+V+IND+PRDT+SG1
  durchdachen+V+IND+PRDT+SG3
  durch#dachen+V+IND+PRDT+SG1
  durch#dachen+V+IND+PRDT+SG3
  durchdachen+V+KONJ+PRDT+SG1
  durchdachen+V+KONJ+PRDT+SG3
  durch#dachen+V+KONJ+PRDT+SG1
  durch#dachen+V+KONJ+PRDT+SG3
  (and 7 incorrect forms with the A(PART)
  with the boundary ...oops!)

LEXICON VERB
durch SEPCONT;          ! ⟶ "durchgedacht"
durch INSEPCONT;        ! ⟶ "durchdacht"
...
LEXICON SEPCONT
# VSEPBASE;
...
LEXICON INSEPCONT
VINSEPBASE;
...
LEXICON VSEPBASE
d000enken:dachte V4R;
dach V01(sep);          ! 'roof' V
...
LEXICON VINSEPBASE
d0000enken:dachte
V4R;
dach V01*;
...
LEXICON V01(sep)
e2n:e1te V4R;
e2n:e1te V5R;
...
LEXICON V01*
e2n:e1te V4R;
e2n:e1te V5R;
...
LEXICON V4R
+V+IND+PRÄT+SG1:
##;
+V+IND+PRÄT+SG3:
##;
...
```

```
LEXICON V5R
+V+KONJ+PRÄT+SG1:
##;
+V+KONJ+PRÄT+SG3:
##;
...
LEXICON PARTEND
durchdacht V6R/adj;
...
LEXICON V6R/adj
e1t A01est/st-part;
LEXICON V6R(sep)/adj
e1t A01est/st-part;
LEXICON A01est/st-part
+A(PART)+POS: ADJ;      ! Adjectivized participle (inflected)
...
LEXICON ADJ
+SG+NOM+FEM:e4
##;
+SG+AKK+FEM:e4 ##;
+SG+NOM+MASK:e4
##;
+SG+NOM+NEUTR:e4
##;
+SG+AKK+NEUTR:e4
##;
+PL+NOM:e4 ##;
+PL+AKK:e4 ##;
...
C1 = b c d f g j k   p q s t v w x s1 t1 ß;
```

Default correspondence: e1:0

```
"e1:e"
  e1:e   ⇔   [C1: | C h] [m | n] - ;
             [t | d] - ;
             C i %-:* - ;
```

Rules for "e2", "e4" as before.[3]

The next examples "gut", "besser", "besten" show how suppletion is typically handled. Instead of one entry, there will be three which all give "gut" as their base form. Note that all remaining characteristics of inflection and compounding can be shared by placing these entries in an appropriate lexicon and assigning a suitable continuation class to them. (The zeroes (0) in the lexical entries are there for technical reasons only, in order to keep the character by character correspondencies simple. The zeroes are mostly added automatically when entries are generated, and the reader may well ignore their presence.)

```
LEXICON ADJ-NONCOMP
gut+ADJ+POS:gut ADJ;
...
LEXICON ASTART
000000gut+A+KOMP:besser ADJ;
000gut+A+SUP:bes00t ADJ;
000gut00+A+SUP2:bes00ten ##;
...
LEXICON ADJ
##;
+SG+AKK+MASK:e4n ##;
+SG+DAT+MASK:e4n ##;
+SG+GEN+MASK:e4n ##;
+SG+DAT+FEM:e4n ##;
+SG+GEN+FEM:e4n ##;
+SG+DAT+NEUTR:e4n ##;
+SG+GEN+NEUTR:e4n ##;
+PL+NOM:e4n ##;
+PL+AKK:e4n ##;
+PL+DAT:e4n ##;
+PL+GEN:e4n ##;
...
```

[3]The "oops" was marked above because we suspected that the adjectivized participle "A(PART)" would have been converted into some incorrect entries in the twol lexicon. It turned out that there was not the kind of error we suspected. Some room for correction was still found.

1.4 Morpho-syntactic analysis (categories)

There is a separate paper which describes the set of morphosyntactic tags or features used in GERTWOL. Where word-forms are homographic, multiple analyses are produced, one for each interpretation. This is intended for the convenience of syntactic processing which might follow the morphological analysis. If only the base form and its part of speech is desired, then the extra features can be deleted, and the user gets only one base form plus part of speech.

Prepositions have valency information (+Dat, +Acc, +Gen), and positional information (+pre, +post). Pronouns have an indication of their possible positions in an NP, i.e.. attributive or independent, (+Det) indicates the possibility of an attributive position which is the normal case for some, and unusual for others.

The syntactic significance of capitalization is preserved, e.g.. "Arbeiten" analyzed as "arbeiten * V IND PRÄS PL3" would be possible only at the beginning of a sentence.

1.5 The handling of generation

The Xerox Lexical Tools includes a program called INFL which is capable of both analyzing word-forms and generating them. This is done by the same program from the same data structure, just by switching the mode. The input for the generation is the same as what the analysis gives as output.

There is also a "prefix" mode where one may drop some of the features at the end, and the INFL program then generates with all possibilities, i.e.. the part of the paradigm, or the full paradigm.

The original TWOL engine is not sensitive to what direction it is being used in. Normally the dictionary is for analysis. With a simple utility program the fields can be switched so that the resulting dictionary will do the generation. Thus the same description is underlying both modes, and so is the same runtime program, but the input file is formatted in a different way.

Bidirectionality has always been one of the fundamental principles of the two-level morphology. Indeed, the normal mode of analysis in TWOL uses the generation mechanism to build the correct base form on the fly (especially when there is no ready-made entry for a derived entry).

1.6 Applications to other languages

The two-level model had been applied to at least 30 languages. The following have a full-scale lexicon and are capable for handling running texts (lexicons with 40-100,000 entries):

> Finnish, Swedish, English, Russian, Swahili, Estonian, Danish, Basque, French, Arabic, Lappish.

There are many others with fairly comprehensive rules and model words for all inflectional types and a few thousand word dictionary (Latin, Mari, Assyrian, Babylonian, Savo dialect of Finnish). Many more less comprehensive descriptions exist, e.g.. for Sanskrit, Nenets, Polish, Turkish, etc.

In principle, most types of languages have been encountered. English can be done without actual two-level rules at all (as the one which is part of ENGCG), or with more extensive rules (the version by Lauri Karttunen at PARC). Finnish is an example of a language where morphophonemes prove to be useful in simplifying the description, because the processes are regular and general. Also some beautiful linguistic generalization may be achieved in the two-level framework which are not easy to express in the rewriting framework. The same comments apply to Sanskrit as well, but the magnitude of overall complexity is far greater there, as the linguists know.

2 Technical design and practical use

2.1 Conceptual goal of the design

GERTWOL was designed to be an industry strength, efficient, wide coverage, general purpose and accurate analyzer/generator which could be used in a wide variety of applications.

Much care was taken to make the basic description as solid as possible so that there would be little need to return to it later for tuning or revision. Thus, the basic inflection is very carefully designed and checked. The lexicon has also been tested on large amounts of corpus text for validation (in cooperation with the University of Stuttgart, some 30 million words of text).

The vocabulary is comprehensive from the very beginning through the use of Collins German Dictionary data as the main source (through an arrangement with HarperCollins). In this way, the need to add more entries is minimized. In spite of this, easy, and even fully automatic inclusion of new entries is provided for.

Efficiency is of primary importance as many application assume the processing of gigabytes (and some terabytes) of text data. The TWOL program can achieve a very satisfactory speed, and the Xerox Lexical Tools allow for speeds of about 250 GB per hour with a highly compressed dictionary.

Hardly any attention has been paid now to the setup where a lexicographer is describing a new language, simply because German morphology is already extremely well accounted for, and there exist very high quality dictionaries. We simply could not imagine of repeating that effort. We would not have the resources to accomplish such a task (of tens or hundreds of man years). Furthermore, German is not a very suitable language for making a theoretically and educationally valuable description, because there is fairly little morphophonol-

ogy which would be regular and productive. Most of the complexity lies in the combinatory part, and that is determined much by convention rather than regular principles.

2.2 Portability of software and data

The TWOL program was originally written in Pascal and later converted into C. The Pascal version was run on Burroughs B6700 and on MSDOS machines. The C version has been ported to a wider set of machines and platforms, at least to: MSDOS machines, Macintosh, Sony NEWS Unix BSD 4.3, IBM RT AIX Unix V, Sun 386i SunOS Unix, Sun 3 SunOS Unix, Sun SPARCstation (various models) SunOS 4.1.X and Solaris 2, HP 9000 Unix V, SCO Unix V, Digital VAX VMS, IBM mainframes MVS, Apollo Domain Unix.

There is also a Prolog version of the TWOL which works on SICSTUS Prolog. It uses a lexicon and a rule component mechanically converted from the C TWOL version into a Prolog clause.

The Xerox Lexical tools can accommodate the GERTWOL lexicon and rules after a mechanical conversion of the format. Those tools and runtimes exist at least on Macintosh and Sun SPARCstations at present.

The lexicons in binary format are portable between various platforms with the limitation of the byte order differences (e.g.. SPARC vs. Intel processors), and the need to handle character code conventions (ISO Latin1, PC code, Macintosh code) with due care. Programs are in no way dependent of any particular character coding.

2.3 Interface to syntax and semantics

GERTWOL is intended to be used in various contexts such as processing of large text corpora, information retrieval, spelling checking, and as the input module for various syntactic or semantic parsers. As a principle, it seems appropriate when designing morphological analyzers for languages with complex morphology to restrict the task of the morphological analyzer to the basic morphology.

The output of the basic morphological analyzer should be used as a key to syntactic and semantic lexicons with richer and theory specific coding. Such postprocessing of the morphological analysis should perhaps occur starting from output where the words and derivational elements are canonized. In this way it would be straight forward to describe the subcategorization features (or case the frame) of derived words and compounds correctly.

For languages with a more trivial morphology, like English, we have chosen to include surface-syntactic features in the lexicon along with the plain morphological tags.

2.4 Aiding the user

Debugging is typically needed for the rule component, inflectional paradigms of model words in each inflectional class, and to some extent, in controlling the continuation patterns in the lexicon.

The rule compiler TWOLC has extensive facilities for verifying the rules, and for tracing the actual behavior of existing rules. One can instantly try what kind of surface forms will be generated:

```
twolc> lex-test
Lexical string ('q' = quit):    lerne1te
                                lernte
l
e
r
n
e1:0
t
e
```

In case the expected surface forms do not show up, or to determine which rules are responsible for blocking certain ungrammatical realizations, one can use "pair-test":

```
twolc> pair-test
Lexical string ('q' = quit):            lerne1te
Surface string ('q' = quit):            lernete
l
e
r
n
e1:e
REJECTED: "e1:e" fails in state 2.
```

The LEXC lexicon compiler has a very useful feature for checking the model words by generating full paradigms of forms for each base form. These are checked manually in order to detect possible missing or extraneous, or incorrectly formed forms.

```
Ihr+PRON+PERS+HÖFLICH+PL2+AKK      Euch
Ihr+PRON+PERS+HÖFLICH+PL2+DAT      Euch
Ihr+PRON+PERS+HÖFLICH+PL2+GEN      Euer
Ihr+PRON+PERS+HÖFLICH+PL2+NOM      Ihr
Ihr+PRON+PERS+HÖFLICH+PL2+NOM+es   Ihr's
```

Once the paradigm has been validated, the verification can be done automatically using the Unix 'diff' program for fast validation after changes to the ending lexicons or rules.

The continuation mechanism of the dictionary is fairly complicated to follow in the actual dictionary files. Therefore, there are specific tools (twol-tree) for viewing the possible sequences of morphemes, e.g..:

```
$ twol-tree -twol -c 3 ger.complete.dic | more
        TwolRoot
        | Root
        | | # [5610]
        | | INTJ [297]
        | | | # [2] ...
        | | iKONJ
        | | uKONJ [53]
        | | SUBST [3]
        | | SUBST-HYPH [2]
        | | SUBST-NCAP [2]
        | | | SCONTH-NCAP
        | | | DIGIT [30]
        | | | | DIGITNUM
        | | | | DIGIT1
        | | | | | DIGITEND
        | | | | | | SCONTH-NCAP [5] ...
        | | S2(-)/nt/end [5]
        | | | Sg1-/nt/end
        | | | Pl2/nt/end
        | | SCONTH--NCAP [445]
        | | | SUBST-NCAP ...
        | | | SUBSTPREF-START
        ...
```

The TWOL engine has some switches for debugging, eg. the following one shows the path along which the analysis proceeds:

```
$ tw-ger -T
ging
"<ging>"
→ #
→ Root
ging → V5U
ging → V4U
ging → #
ging → #
g → #
g ↔ #
    "gehen" V IND PRÄT SG3
    "gehen" V IND PRÄT SG1
```

2.5 Limits to the size of the system

On the whole, there are hardly any limits set by the TWOL program itself. The availability of core memory in the target computer is the actual constraint the size of the lexicon, and this constraints has ever less significance. The runtime TWOL program allocates memory space according to the actual size of the saved lexicon. There is some, high enough limit for input line and input word length (say 1-10 kB) which can be adjusted as needed. Excessively long inputs will be discarded.

The development versions of TWOL programs have parameters according to which one can adjust the upper limits in those areas where longer areas are needed through command line switches.

The versions made using the Xerox lexical tools need (in most cases) less memory than the TWOL.

2.6 Interface to non-ASCII characters

The details which character codes will be used can be adjusted according to the need, and we are prepared to handle three options for German:

```
8 bit ISO Latin1              ä ö ü Ä Ö Ü ß
8 bit ISO Latin1              ä ö ü Ä Ö Ü ss
(as often in Swizerland)
7 bit ASCII                   ae ou ue Ae Oe Ue ss
```

E.g.. to use "ss" instead of "ß" is a matter of excluding the rule which is responsible for this convention.

2.7 User friendliness of the 'turn around'

The vocabulary is extensive and additions are not required often. When large amounts of text are being processed e.g.. for surface syntactic analysis, we have planned to use a separate module which we call "morphological heuristics". This would give each unanalyzed word a set of possible baseforms and morphosyntactic tags. Experience from English suggest that almost all instances can be handled in this way.

In some applications, it is desirable to add new words to the GERTWOL lexicon on the fly. This is common in certain types of information retrieval applications, and it is handled by coupling an entry generator (EGEN) with the Twol program. During the execution, the input may consist either of word-forms to be analyzed, or new word entries which are immediately installed into

the lexicon. New entries may have either the form of a normal GERTWOL entry such as:

*ghotbzadeh NAME-M/F;

In an actual application, the user would add only a brief part of speech classification in front of the base form of the new word, e.g..:

NN:Ghotbzadeh

This entry is converted automatically by the entry generator into the previous format, which is then instantly added to the dictionary. (This facility has been used for about two years in a newspaper archive application of the largest Finnish newspaper publisher for the Finnish texts. By chance, this facility is not in the latest version of TWOL, but it will be reactivated in the near future. In the meantime, one can use the stand-alone entry generator, and cut and paste the resulting entries.)

New entries can be added by editing the master files, if one so wishes. This is typically done by those who maintain the dictionary. The update cycle of rebuilding the whole GERTWOL from its GLEX source and various components might take up to a few hours, but the time depends on how extensive modifications were made. Converting the German TWOL lexicon and rule automata file into a binary file used by the runtime TWOL program takes about 15 minutes.

2.8 The state of the documentation

The morphological two-level model is documented in the Ph.D. dissertation: K. Koskenniemi "Two-level Morphology: A General Computational Model for Word-Form Recognition and Production", Univ. of Helsinki, Dept. of General Linguistics, Publications, No. 11, 1983.

The rule compiler for the two-level rules is documented in L. Karttunen, K. Koskenniemi and R. Kaplan, "A Compiler for Two-level Phonological Rules", in Report CSLI-87-108, CSLI Stanford Univ., 1987. The Xerox lexicon compiler TWOLC is documented in L. Karttunen and K. Beesley, "Two-Level Rule Compiler", Xerox Palo Alto Research Center, ISTL-92-2, 1992.

The current LEXC is documented in L. Karttunen, "Finite-State Lexicon Compiler", Xerox Palo Alto Research Center, ISTL-NLTT-1993-04-02, 1993.

The interface and use of Lingsoft version of the Twol program is documented in a document by Lingsoft, and a few Unix style man pages.

The overall design and description of the GERTWOL and the system of morphosyntactic tags is documented in Mariikka Haapalainen's forthcoming paper, out of which an abbreviated version is attached.

The lexical content from the Collins German Dictionary is, of course, documented in the published dictionary. Other words, such as proper nouns, have been added, but that list is not available.

2.9 Availability and maintenance

The GERTWOL analyzer will be made available for the academic research community at a nominal price which is intended to cover the administration and distribution costs. The academic licenses will permit normal and reasonable research and educational use.

Simultaneously, the product will be available for commercial use at a different type of license where there is a fee or royalty according to the intended use. Queries should be directed to Lingsoft, Inc., Manager, International sales, Mr. Eugene Young, Museokatu 18 A 3, FIN-00100 Helsinki, Finland. There is a separate leaflet describing those aspects.

3 The test data

The GERTWOL has been tested against word forms in a 30 million word corpus. The results were:

99 %	of all correctly spelled word-forms in the corpus
98 %	of all words in the corpus

4 The performance figures on previously known test data

Coverage:[4]

15667	input tokens total (incl. punctuation)
15415	of these tokens got an analysis
252	tokens didn't get an analysis

Speed on a Sun SPARCstation 2 (without preprocessing) with the TWOL engine:

216	words per sec full analysis
994	words per sec if only recognition of correctness is performed

Speed on a HP (without preprocessing) with the TWOL engine[5]:

398	words per sec full analysis
1978	words per sec if only recognition of correctness is performed

Space needed with the TWOL:

41 kB RAM	for the TWOL runtime program
6.2 MB RAM	for the lexicon and rules

5 Portability

On a Intel 80486 based PC the performance figures are similar to those on a SPARCstation 2.[6] On an older 80386 based PC about three times slower than on SPARCstation 2. The TWOL engine takes about the same space on each platform.

Using the Xerox Lexical Tools the space requirement can be reduced into 500 kB to 1 MB. The speeds are nevertheless much higher: at least 2000 words per second for analysis, and some 10,000 words per second if only recognition is needed.[7]

[4]There were problems in porting the preprocessing programs to the HP machines in Erlangen. The faster "flex" version did not work at all because of some unidentified problem (in contrast to the SUN and HP machines in Helsinki). Another, slower Perl version worked partly, but produced somewhat erraneous results. (Again, the error could not be reproduced on a similar machine in Helsinki.) Consequently, in both of the texts used in the benchmark, the first word of each sentence got garbled, and consequently, was rejected.

[5]The HP workstation "sol" was twice as fast.

[6]The MSDOS version was brought to Erlangen, but we were not able to move the files from the Unix machine to the PC in order to install and run it.

[7]A test version of the version using the Xerox Lexical Tools was present on a Macintosh Powerbook, but not ported to the official benchmark compiler, and not officially measured.

MORPH

EIN MODULARES UND ROBUSTES MORPHOLOGIEPROGRAMM FÜR DAS DEUTSCHE IN COMMON LISP

Gerhard Hanrieder
Bayerisches Forschungszentrum für Wissensbasierte Systeme (FORWISS) Am Weichselgarten 7 91058 Erlangen-Tennenlohe E-mail: hanriede@forwiss.uni-erlangen.de

1 Name und Herkunft des Systems:

Das System MORPH wurde 1991 im Rahmen einer Magisterarbeit im Fach Computerlinguistik an der Universität Trier von Gerhard Hanrieder in Common Lisp implementiert. Die Arbeit wurde betreut von Dr. Heinz Josef Weber.

2 Konzeptuelle Kriterien:

2.1 Deklarative Spezifikation lexikalischer Einträge und Regeln

MORPH verwendet zwei lexikalische Wissensquellen: ein Morphemlexikon und ein Flexivlexikon. Das Morphemlexikon ist ein gemischtes Stammform- und Vollformlexikon.

1. Stammformeinträge:
   Die Kodierung flektierender Stämme beruht auf einer strikten Subkategorisierung der Flexionsparadigmen. Die Zuordnung eines Stammes zu seinem Flexionsparadigma geschieht, indem im Lexikoneintrag ein eindeutiges Flexionsklassenkürzel zugewiesen wird, das als Pointer in das Flexivlexikon fungiert. Bsp.:
   (TISCH
   (M3))

   Mehrdeutigkeiten sind als Lesarten innerhalb eines Eintrages kodiert, z.B. repräsentiert der folgende Lexikoneintrag, daß es sich bei *koch* entweder um einen Verbstamm oder um einen Nomenstamm im Singular handelt:
   (KOCH (SWV1) (MSING1 (ALLO KOECH»)

   Innerhalb einer Lesartliste muß das Flexionsklassenkürzel an erster Stelle stehen, danach kann eine beliebige Anzahl weiterer Merkmale folgen.
2. Vollformeinträge:
   Nicht-flektierende Wörter werden als Vollformeinträge kodiert. Hierzu muß an erster Stelle der Merkmalsliste das Merkmal VF angegeben werden, dem wiederum eine - wortartspezifische Attributmenge folgen kann. Der folgende Eintrag zeigt die Kodierung der Präposition *wegen,* die den Genitiv regiert:

```
(WEGEN (VF (P) (KASREK (GEN))
```

   Stammform- und Vollformlesarten werden innerhalb eines Eintrages kombiniert, wie der folgende Eintrag für *laut* mit seinen Adjektiv-, Nomen-, Verb- und Präpositionlesarten zeigt:

```
(LAUT (ADJ2) (M3) (SWV2) (VF
(p) (KASREK (GEN)))
```

Die Flexionsklassen stellen das Bindeglied zwischen Morphem- und Flexivlexikon dar. Das Flexivlexikon enthält Einträge der Form:

```
(FLEXIV {(FLEXIONSKLASSEN) (MERKMALE)})
```

Die Einträge repräsentieren das Wissen, daß ein Stamm `(STAMM (FLEXIONSKLASSE))` mit FLEXIV kombinierbar ist, wenn FLEXIONSKLASSE in FLEXIONSKLASSEN enthalten ist. Die Wortform STAMM+FLEXIV hat die morphosyntaktischen Merkmale MERKMALE. Das folgende Beispiel zeigt den Eintrag für das Flexiv *st*:

```
(ST ((STV4A STV4AA STV5A STV5AA)
     (V) (FLEXION (IND_PRAETERITUM
          (SG 2))))
    ((SWV1 SWV4 SWV5 SWZ1 SWZ4 SWZ5
      STV1A STV3AA STV3AB STV3C STV7C)
     (V) (FLEXION (IND_PRAESENS
                   (SG 2))))
)
```

Auf den oben gegebenen Stammeintrag `(KOCH (SWV1) (MSING1 (ALLO KOECH)))` angewendet, ergibt das folgendes Resultat: die Kombination *koch* und *st* ist gültig, da die erste Lesart von *koch*, `SWV1`, im Flexiveintrag `ST` kodiert ist und festlegt, daß es sich um eine Verbform in der 2. Person Indikativ Präsens handelt:

```
((V) (FLEXION (IND_PRAESENS (SG 2))
```

Die vollständige Definition einer Flexionsklasse ergibt sich aus der Menge aller Flexive, an denen das entsprechende Flexionsklassenkürzel kodiert ist.

Morphem- und Flexivlexikon sind — wie gezeigt — als `Lisp`-Listen repräsentiert. Diese Listen werden in Buchstabenbäume kompiliert mit einem Verfahren, das weitgehend mit dem in [Han89, S. 43ff.] beschriebenen übereinstimmt. Die Lexikonsuche zur Laufzeit von MORPH wird in den Buchstabenbäumen vorgenommen.

Die morphologischen Regeln sind in MORPH als Zustands-Übergangs-Automat implementiert. Der Automat definiert, welche Morphemklassenübergänge möglich sind, und wird bei der Segmentierung einer Wortform verwendet. Das folgende Beispiel zeigt einen Ausschnitt aus den definierten Übergängen, wobei z.B. `(N (N)` bedeutet, daß auf ein Nomen ein Nomen folgen kann. Übergänge können mit *procedural attachments* augmentiert sein, z.B. legt `(N (ADJ (nicht_adjflek 2))` fest, daß ein auf ein Nomen folgendes Adjektiv nicht die Flexionsklasse `ADJFLEK` haben darf. Dies ist die Flexionsklasse nur flektiert vorkommender Adjektive wie *letzt*. Mit diesem Zusatztest wird beispielsweise die Kombination *hausletzt* unterbunden, während *haushoch* zugelassen wird.

```
(setq *wb-rules* '(
 (N (N)
    (V)
    (ADJ (nicht_adjflek 2))
    (FUGE)
    (SUF (agree_kat_wb-subcat 1 2))
    (PRAEF)
    (PARTPRAEF)
    (VPR)
    (TVZ)
    (ORTNUMSUF (=ortsname 1))
    (FL)
 )
 (ADJ (FL)
      (SUF (agree_kat_wb-subcat 1 2))
      (N)
      (ADJ)
      (V (nicht_verbzus 1))
      (FUGE)
      (PARTPRAEF)
      (ALLOV)
      (INFINITIVPARTIKEL)
      (VPR)
 )
 (V (N)
    (ADJ)
    (SUF)
    (FL)
 )
 ...
)
```

Die Zuordnung der Morphemlexikoneinträge zu den im Übergangsautomat verwendeten Morphemklassenkürzeln kann auf zweierlei Art erfolgen:

1. Wenn ein Lexikoneintrag ein Morphemklassenattribut `(MK WERT)` enthält, ist `WERT` die Morphemklasse des Eintrages, z.B. hat das Verbpräfix *be* die Klasse `VPR` aufgrund des Eintrages:

   ```
   (BE (VF (MK VPR) (BOUND)))
   ```

2. Ist keine Morphemklasse kodiert, wird versucht, eine Morphemklasse aufgrund der Flexionsklasse zu ermitteln. Hierzu werden globale Definitionen ausgewertet, die einer Menge

von Flexionsklassen eine Morphemklasse zuweisen, z.B.:

```
(setq *ADJ-FLEXKLASSEN*  ;
      Adjektivstaemme
      '(adj1 adj2 adjpos adjsteig1
        adjsteig2 adjsup adjsyn
        adjflek)
)
```

Diese Definition weist allen im Morphemlexikon verwendeten adjektivischen Flexionsklassenkürzeln die Morphemklasse ADJ zu.

Die Repräsentation der morphologischen Regeln als Zustands-Übergangs-Automat orientiert sich stark an dem im System MARS [BNTW86] verwendeten Verfahren. Es gelten somit auch die in [Thu86, S.89f.] herausgearbeiteten konzeptionellen Nachteile dieses Ansatzes: Der Automat arbeitet nur über Modellen erster Ordnung, d.h. es wird nur das aktuelle Morphem und sein Vorgänger betrachtet. Dadurch können keine Informationen über Wortstrukturen gewonnen werden.

Der Übergangsautomat wird aus diesem Grund nur zur Segmentierung der Wortformen verwendet. Die Generierung der morphosyntaktischen Merkmale erfolgt prozedural im Anschluß an die Segmentierung.

2.2 Bezug zwischen lexikalischen Einträgen und Wortformen

Allomorphe müssen in MORPH als gesonderte Einträge im Lexikon enthalten sein. Allomorpheinträgen ist ebenso wie den Stammeinträgen eine Flexionsklasse zugewiesen, die mögliche Flexivfortsetzungen definiert. Zusätzlich zur Flexionsklassenangabe haben Allomorpheinträge das Merkmal GF_STAMM, das auf den Grundformstamm verweist. Umgekehrt wird vom Grundformstamm mittels ALLO auf Allomorphstämme verwiesen, wie die Einträge für LAECHEL und LAECHL zeigen:

```
(LAECHEL (SWV4 (ALLO LAECHL)))
(LAECHL (SWV4_1SG (GF_STAMM LAECHEL)))
```

Die Analyse einer abgeleiteten Wortform wie *anlächle* erfolgt, indem die Wortform segmentiert und anschließend die Merkmale generiert werden: Der Lexikoneintrag für *an*

```
(AN (VF (P) (KASREK (DAT AKK)))
    (VF (TVZ)))
```

weist u.a. eine Lesart TVZ (trennbarer Verbzusatz) auf. Die Flexionsangabe von LAECHL, SWV4_1SG wird — wie in Abschnitt 2.1 dargestellt — auf die Morphemklasse V abgebildet. Im Morphem-Übergangsautomat ist definiert, daß auf einen trennbaren Verbzusatz (TVZ) ein Verb (V), und auf ein Verb ein Flexiv (FL) folgen kann:

```
(TVZ (PARTPRAEF)
     (V)
     (ALLOV)
     (INFINITIVPARTIKEL)
     (VPR)
     (TVZ)
)
(V (N)
   (ADJ)
   (SUF)
   (FL)
)
```

Die resultierende Segmentierung ("AN" "LAECHL" "E") wird an die Merkmalsgenerierung übergeben: Da Verb und Flexiv kompatibel sind, werden die am Flexiv kodierten Merkmale für die Ausgabe übernommen. Für die Generierung der Grundform wird das Merkmal (GF_STAMM LAECHEL) des Allomorpheintrages LAECHL ausgewertet und die Grundform ANLAECHELN gebildet:

```
(ANLAECHLE (V) (GF ANLAECHELN)
 (FLEXION (IND_PRAESENS (SG1))
 (KONJ_PRAESENS (SG (1 3))))
 (FINIT-STELLUNG)
)
```

MORPH unterstützt die redundanzfreie Lexikalisierung abgeleiteter Stämme: Präfigierte Verbstämme werden mit einem Verweis auf den Simplexstamm kodiert; Allomorphstämme müssen somit nur in der Simplexform ins Lexikon eingetragen werden. Ist ein lexikalisierter komplexer Stamm im Lexikon vorhanden, werden dort kodierte zusätzliche Merkmale von der Merkmalsgenerierungskomponente in die Ausgabeliste übernommen. Im folgenden Eintrag für ANLAECHEL ist exemplarisch ein zusätzliches semantisches Merkmal kodiert:

```
(ANLAECHEL (= LAECHEL (TVZ AN)
              (SEM (TYPE SMILE))))
```

Bei Vorhandensein eines solchen Eintrages würde sich das Analyseergebnis wie folgt ändern:

```
(ANLAECHLE (V) (GF ANLAECHELN)
 (FLEXION (IND_PRAESENS (SG1))
 (KONJ_PRAESENS (SG (1 3))))
 (FINIT-STELLUNG)
 (SEM (TYPE SMILE))
)
```

2.3 Verständlichkeit und linguistische Motivation der Regeln

Flexion, Derivation und Komposition werden in MORPH einheitlich als Konkatenation von Allomorphen behandelt. Wie in Abschnitt 2.1 und 2.2 bereits dargelegt, ist der Analyseprozeß untergliedert in zwei aufeinanderfolgende Schritte:

1. Die Segmentierung in Wortbestandteile unter Zugriff auf den Zustands-Übergangs-Automat.
2. Die Generierung der morphosyntaktischen Merkmale aufgrund des morphologischen Kopfprinzips.

Das Kopfprinzip (s. [SS88, S. 52]) besagt, daß die morphologischen Merkmale komplexer Wortformen vom Kopf ererbt werden, der im Deutschen im unmarkierten Fall rechts ist. Dies gilt gleichermaßen für Flexion, Derivation und Komposition:

- Flexion: Maßgeblich für das Flexionsverhalten einer Wortform ist die Flexionsklasse der am weitesten rechts stehenden Wortkonstituente. In MORPH ist diese Regularität implementiert, indem überprüft wird, ob Kopfkonstituente und Flexiv kompatibel sind.

  Beispiel: Bei der Analyse von *Tisch, Tisches* und *Tischen* gelingt dies, da die im Lexikoneintrag `(TISCH (M3))` kodierte Flexionsklasse `M3` in den entsprechenden Flexiveinträgen definiert ist. Aus den Segmentierungen `("TISCH" "")`, `("TISCH" "ES")` und `("TISCH" "EN")` können somit die entsprechenden Ausgabelisten generiert werden:

```
(TISCH (N) (GF TISCH)
 (FLEXION (MASK (NOM SING)
 (DAT SING) (AKK SING))))
(TISCHES (N) (GF TISCH)
 (FLEXION (MASK (GEN SING))))
(TISCHEN (N) (GF TISCH)
 (FLEXION (MASK (DAT PLUR))))
```

- Derivation: Auch bei der Derivation gilt im unmarkierten Fall das Kopfprinzip, d.h. die morphosyntaktischen Merkmale einer abgeleiteten Wortform werden durch das rechts stehende Suffix bestimmt. Suffixe sind deshalb in MORPH als — gebundene — flektierende Morpheme kodiert, z.B.:

```
(BAR (ADJ1 (MK SUF) (BOUND)))
```

  Präfixe sind dagegen als — gebundene — nicht-flektierende Vollformen kodiert, z.B.:

```
(UN (VF (MK PRAEF) (BOUND)))
```

  Bei der Analyse einer abgeleiteten Form wie *unlesbares* gelingt eine Segmentierung in `("UN" "LES")` `("BAR" "ES")` aufgrund der Morphemübergangsdefinitionen `(PRAEF (V))`, `(V (SUF))` und `(SUF (FL))`.

  Suffix und Flexiv sind kompatibel, da die Flexionklasse von *bar*, `ADJ1` im Flexiveintrag `ES` definiert ist:

```
(ES
 ((ADJ1 ADJ2 ADJPOS ADJSYN ADJFLEK)
  (ADJ) (GRAD POS)
   ----
  (FLEXION (DEKTYPI (NEUT (NOM SING)
   (AKK SING)))
   (DEKTYPIII (NEUT (NOM SING)
(AKK SING)))))
 ((M1 M3 MSING1) (N) (FLEXION
  (MASK
   (GEN SING))))
 ((STN NSING1 N2 N3) (N)
   (FLEXION (NEUT (GEN SING))))
 ((NUM2) (ADJ) (ORDINALZAHL)
   (= FLEXION ES ADJ1)))
```

  Die Merkmale werden somit aus dem Flexivlexikon in die Outputliste übernommen:

```
(UNLESBARES (ADJ)
            (GF UNLESBAR)
            (GRAD POS)
            (FLEXION
             (DEKTYPI
              (NEUT
               (NOM SING)
               (AKK SING)))
             (DEKTYPIII
              (NEUT
               (NOM SING)
               (AKK SING))))
            (KONST (UN (PRAEF))
                   (LES (V))
                   (BAR (SUF))))
```

- Komposition: Die Behandlung der Komposition verläuft analog und sei anhand der Wortformen *Hausdächern* und *Häusermeers* skizziert: Aufgrund der definierten Morphemübergänge

(N (N)), (N (FUGE)), (FUGE (N)) und (N (FL)) gelingen die Segmentierungen ("HAUS" "DAECH" "ERN") und ("HAEUS" "ER" "MEER" "S").

Die Anwendung des Kopfprinzips stellt fest, daß "DAECH" und "ERN" bzw. "MEER" und "S" kompatibel bezüglich der Flexionsklasse sind und generiert die Outputlisten:

```
(HAUSDAECHERN (N)
              (GF HAUSDACH)
              (FLEXION
               (NEUT
                (DAT PLUR)))
              (KONST (HAUS (N))
 (DACH (N))))
(HAEUSERMEERS (N)
              (GF HAEUSERMEER)
              (FLEXION
               (NEUT (GEN SING)))
              (KONST (HAUS (N))
                     (ER (FUGE))
                     (MEER (N))))
```

2.4 Morpho-syntaktische Analyse (Kategorisierung)

Die von MORPH ermittelten morphosyntaktischen Merkmale entsprechen einer traditionellen, wortartspezifischen Charakterisierung:

- Nomina: Die Kategorie wird als N notiert. Genus, Kasus und Numerus sind im Merkmal FLEXION zusammengefaßt, wobei Kasus-Numerus-Listen als disjunktiv verknüpft interpretiert werden, z.B. ist *Tisch* entweder Nominativ Singular oder Dativ Singular oder Akkusativ Singular:

```
(TISCH (N) (GF TISCH)
 (FLEXION (MASK (NOM SING)
                (DAT SING)
                (AKK SING))))
```

- Verben: Die Kategorie wird als V notiert. Unter dem Merkmal FLEXION werden Modus, Tempus, Numerus und Person angegeben, z.B.:

```
(VORBEISCHWAMMST (V)
                 (GF VORBEISCHWIMMEN)
 (FLEXION (IND_PRAETERITUM (SG 2)))
 (FINIT-STELLUNG)
 (KONST (VORBEI (TVZ))
 (SCHWIMM (ALLOV))))
```

Flektierende Partizipformen werden wie Adjektive kategorisiert.

- Adjektive: Die Kategorie wird mit ADJ, bzw. UNFLADJ (unflektierte, prädikativ gebrauchte Adjektive) notiert. Der Komparationsgrad wird unter dem Merkmal GRAD angegeben. Unter FLEXION sind Deklinationstyp, Genus, Kasus und Numerus angegeben. Alternativen sind wiederum als konjunktiv verknüpft zu lesen: So handelt es sich bei der ADJ-Lesart von *besten* entweder um eine stark (DEKTYPI), schwach (DEKTYPII) oder gemischt (DEKTYPIII) flektierende Form. Stark flektierend kann die Form wiederum Maskulinum, Femininum oder Neutrum sein usw.:

```
(BESTEN (L (UNFLADJ) (GF GUT)
           (GRAD SUP))
 (L (ADJ) (GF GUT) (GRAD SUP)
  (FLEXION
   (DEKTYPI (MASK (GEN SING)
                  (AKK SING) (DAT PLUR))
            (FEM (DAT PLUR))
    (NEUT (GEN SING) (DAT PLUR)))
   (DEKTYPII
    (MASK (DAT SING) (AKK SING)
          (NOM PLUR) (GEN PLUR)
          (DAT PLUR) (AKK PLUR))
     (FEM (GEN SING) (DAT SING)
          (NOM PLUR) (GEN PLUR)
          (DAT PLUR) (AKK PLUR))
     (NEUT (GEN SING) (DAT SING)
           (NOM PLUR) (GEN PLUR)
           (DAT PLUR) (AKK PLUR)))
    (DEKTYPIII
     (MASK (DAT SING) (AKK SING)
           (NOM PLUR) (GEN PLUR)
           (DAT PLUR) (AKK PLUR))
     (FEM (GEN SING) (DAT SING)
          (NOM PLUR) (GEN PLUR)
          (DAT PLUR) (AKK PLUR))
     (NEUT (GEN SING) (DAT SING)
           (NOM PLUR) (GEN PLUR)
           (DAT PLUR) (AKK PLUR))))))
```

Morphemgrenzen werden unter dem Merkmal KONST zusammen mit der Kategorie ausgegeben. Allomorphe werden dabei auf den Grundformstamm zurückgeführt, wie am Beispiel *Hausdächern* zu sehen ist, wo das Allomorph *däch* als *dach* in der KONST-Liste steht:

```
(HAUSDAECHERN (N)
              (GF HAUSDACH)
              (FLEXION
               (NEUT (DAT PLUR)))
              (KONST (HAUS (N))
 (DACH (N))))
```

Die Lexikoneinträge von MORPH enthalten keine Valenzinformation, da bei der Konzeption des Systems davon ausgegangen wurde, daß die Aufgabe einer Morphologiekomponente darin besteht, Wortformen morphosyntaktisch zu bestimmen und auf ihre Grundform zurückzuführen. Diese Grundform fungiert als Pointer in ein syntaktisch-semantisches Stammlexikon. Die Lexikonstruktur von MORPH un-

terstützt jedoch auch einstufige Systemarchitekturen: Einträge können um beliebige Zusatzmerkmale erweitert werden, die mit in die Ausgabeliste übernommen werden.

Die Überführung der Analyseergebnisse in das Format spezifischer Formalismen sollte problemlos möglich sein in Analogie zu der bereits implementierten Transduktion des Listenoutputs in eine leichter lesbare Bildschirmdarstellung. Ambiguitäten und Disjunktionen werden dabei explizit aufgelöst, wie das Beispiel *spielen* zeigt:

```
(SPIELEN (L (N)
            (GF SPIEL)
            (FLEXION (NEUT (DAT PLUR))))
         (L (V)
            (GF SPIELEN)
            (FLEXION (IND_PRAESENS
  (PL (1 3)))
                     (KONJ_PRAESENS
  (PL (1 3)))
                     (INFINITIV)))
)
```

```
1. Lesart:
Wortform: SPIELEN
Grundform: SPIEL
Kategorie: NOMEN
Genus: NEUTRUM
Kasus/Numerus: DAT.PL.

2. Lesart:
Wortform: SPIELEN
Grundform: SPIELEN
Kategorie: VERB
Konjugation:
  Modus: INDIKATIV
  Tempus: PRAESENS
  Person/Numerus: 1.PLURAL v 3.PLURAL
 v
  Modus: KONJUNKTIV
  Tempus: PRAESENS
  Person/Numerus: 1.PLURAL v 3.PLURAL
 v
  Infinite Form: INFINITIV
```

2.5 Behandlung der Generierung

Eine Generierungskomponente wurde in MORPH aus Zeitgründen nicht implementiert. Prinzipiell können jedoch dieselben Wissensquellen für die Generierung verwendet werden wie für die Analyse. Dabei fungiert die Flexionsklasse des Stammes, zu dem eine Form generiert werden soll, als Pointer in das Flexivlexikon, dessen Einträge solange durchsucht werden, bis ein Eintrag der gesuchten Flexionsklasse mit den Merkmalen der zu generierenden Form matcht. Die 3. Person Singular Präsens von *lernen* z.B. würde dann wie folgt generiert:

```
(generate '("LERN" (V)
                   (IND_PRAESENS
(SG 1))))
```

Die Lexikonsuche des Stammes `LERN` ergibt `(LERN (SWV1))`. Im Flexivlexikon ist somit ein Eintrag zu suchen, an dem unter `SWV1` die Merkmale `((V) (IND_PRAESENS (SG 1))` enthalten sind. Dies trifft für den (verkürzten) Eintrag `E` zu:

```
(E
 ((SWV1 SWV2 SWV3 SWV4_1SG SWV5 SWZ1
   SWZ2 SWZ3 SWZ4_1SG SWZ5
   STV1A STV1B STV1C STV7A STV7B
   STV7C STV7D)
  (V) (FLEXION (IND_PRAESENS (SG 1))
       (KONJ_PRAESENS (SG (1 3)))
       (IMPERATIV SG)))
 ((STV4AA STV4BB STV4CC STV5AA
 STV5BB STV5CC STVKONJ)
  (V) (FLEXION (KONJ_PRAETERITUM
                (SG (1 3)))))
)
```

Die Zielform *lerne* kann somit durch Anhängen des gefundenen Flexivs gebildet werden.

2.6 Übertragbarkeit auf andere Sprachen

Das System MORPH wurde bislang nur auf das Deutsche angewendet. Eine Übertragung auf eine andere Sprache würde neben dem Austausch der Wissensquellen auch Änderungen am Programmkode erfordern, da die Merkmalsgenerierungskomponente im Hinblick auf die deutsche Morphologie implementiert ist.

3 Technische Konzeption und Einsatzfähigkeit

3.1 Zielsetzung der Konzeption

Bei der Konzeption von MORPH standen zwei Überlegungen im Vordergrund:

- Dem Lexikonentwickler soll es ermöglicht werden, zwischen usuellen und okkasionellen Wörtern (vgl. [Ols86, S. 51]) zu unterscheiden. Usuell gewordenene, voll lexikalisierte Bildungen wie *Junggeselle* oder *Hochzeit* zeichnen sich dadurch aus, daß ihre Semantik nicht mehr aus der Semantik der Wortbestandteile berechenbar ist. In vielen Anwendungsbereichen wird man usuelle Wörter deshalb lexikalisieren und ihre Segmentierung vermeiden wollen.

- Das System soll robust gegenüber nicht analysierbaren Wortformen sein.

Diese Überlegungen haben zu der in Abb. 1 gezeigten, von der Saarbrücker Lemmatisierung [Web76] beeinflußten Architektur geführt.

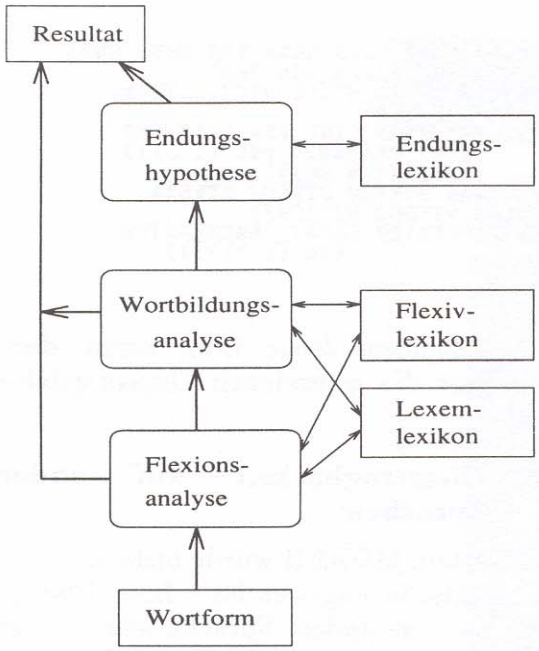


Abbildung 1: Systemarchitektur von MORPH

Die Analyse einer Wortform in MORPH erfolgt defaultmäßig durch maximal drei aufeinanderfolgende Module:

1. Flexionsanalyse:
   Diese Komponente analysiert von rechts nach links und zerlegt die Wortform in mögliche Stamm-Flexiv-Zerlegungen, die auf ihre Flexionsklassenkompatibilität überprüft werden.
2. Wortbildungsanalyse:
   Diese Komponente analysiert von links nach rechts und generiert mögliche Segmentierungen der Wortform, aus denen aufgrund des Kopfprinzips die morphosyntaktischen Merkmale generiert werden.
3. Hypothetische Analyse:
   Diese Komponente analysiert von rechts nach links und prüft, ob die Endung der Wortform eine hypothetische Bestimmung zuläßt.

Sobald eine der Komponenten ein Ergebnis liefern kann, wird die Analyse mit diesem Resultat beendet. Dadurch wird erreicht, daß lexikalisierte Wörter wie *Junggeselle* nicht segmentiert werden, da bereits die Flexionsanalyse ein Ergebnis liefert.

Dieser Defaultprogrammfluß kann durch globale Systemflags geändert werden. Jedes der Module kann ein- und ausgeschaltet werden, womit das System für unterschiedliche Aufgabenstellungen konfigurierbar ist.

3.2 Portabilität der Software und der Daten

MORPH wurde ursprünglich in VAX Lisp 2.2 auf einer DEC VAX 600-410 unter VMS 5.3 implementiert und seither problemlos auf folgende Umgebungen portiert:

- Sun Common Lisp 4.0.1 auf einer SUN Sparc10 unter SunOS 4.3.1
- Allegro Common Lisp 4.1 auf einer DEC 5800 unter Ultrix 4.3
- CLisp auf einem PC 486/33 unter MS-DOS 6.2

3.3 Schnittstellen zur Syntax und zur Semantik

Als allgemeine Schnittstelle zu Programmen, die den Output von MORPH weiterverarbeiten wollen, existiert die Lisp-Funktion **analyze**. Sie nimmt als Parameter den zu analysierenden Wortformstring und liefert das Analyseergebnis in der – mehrfach gezeigten – Listenform:

```
(analyze "wortform")
 (WORTFORM (N) (GF WORTFORM)
           (FLEXION (FEM (NOM SING)
     (GEN SING)
     (DAT SING)
     (AKK SING)))
           (KONST (WORT (N))
                  (FORM (N))))
```

3.4 Hilfestellung bei Benutzerfehlern

Da Lisp-Systeme im allgemeinen sehr gute Debugging-Möglichkeiten aufweisen, wurde bei der Entwicklung von MORPH wenig Zeit in die Entwicklung spezifischer Debugging Utilities investiert. Es wurde lediglich darauf geachtet, daß fehlerhafte Benutzereingaben nicht zu Systemabstürzen

führen, z.B. bei der Eingabe von Dateinamen:

```
Name der ASCII-Datei: test.txt

Die angegebene Inputdatei:
/desklab/hanriede/prog/morph/test.txt
existiert nicht!

Zurueck zum Hauptmenue: <RETURN>
```

3.5 Größenbeschränkung des Systems

Der Quellkode von MORPH umfaßt 127806 KByte, die kompilierte Lexikondatei `lexika.tre` z.Zt. 2572035 KByte.

Bei der Analyse sehr großer Textdateien empfiehlt es sich, diese vor der Analyse in einem separaten Schritt zu segmentieren und die segmentierte Datei an die Analyse zu übergeben, da andernfalls die gesamte Datei in den Arbeitsspeicher eingelesen wird.

3.6 Schnittstelle zu Nicht-ASCII Zeichen

Umlaute in der Eingabe werden durch die Einlesefunktion von MORPH ausgetauscht in die Ersatzdarstellungen `AE`, `OE`, `UE`. In dieser Form sind Umlaute im Lexikon kodiert, z.B.:

```
(PRAESIDENT (SWM1))
```

Ein Textstring wie *Herr Präsident* wird beim zeichenweisen Einlesen überführt in die Liste

```
(("Herr") ("Praesident"))
```

3.7 Benutzerfreundlichkeit des *turn around*

Modifikationen an den Wissensquellen von MORPH betreffen entweder das Lexikon oder die Regeln des Übergangsnetzwerkes. In beiden Fällen werden Veränderungen durch einfaches Edieren der entsprechenden Dateien `lexika.lst`, bzw. `wbrules.lisp` vorgenommen. Bei Lexikonänderungen müssen die Buchstabenbäume anschließend neu generiert werden. Der folgende *trace* zeigt diesen Kompilier-Vorgang für ein Lexikon mit 25600 Einträgen auf einer Sun Sparc10 mit 64 MB RAM. Die benötigte Zeit ist am Ende angegeben:

```
>bfws4e:hanriede 4> lcl
;;; Sun Common Lisp, Development
;;; Environment 4.0.1, 6 July 1990
;;; Sun-4 Version for SunOS 4.0.x
;;; and sunOS 4.1
;;;
;;; Copyright (c) 1985, 1986, 1987,
;;;               1988, 1989, 1990
;;; by Sun Microsystems, Inc.,
;;; All Rights Reserved
;;; Copyright (c) 1985, 1986, 1987,
;;;               1988, 1989, 1990
;;; by Lucid, Inc.,
;;; All Rights Reserved
;;; This software product contains
;;; confidential and trade secret
;;; information belonging to Sun
;;; Microsystems, Inc.  It may not
;;; be copied for any reason other than
;;; for archival and backup purposes.
;;;
;;; Sun, Sun-4, and Sun Common Lisp are
;;; trademarks of Sun Microsystems Inc.

> (setq *gc-silence* t)
T
> (change-memory-management :growth-limit
   30000)
T
> (load "morph")
;;; Loading binary file "morph.sbin"
#P"/desklab/hanriede/prog/morph/morph.sbin"
> (tools)

****************************************
*                                      *
*          MORPH - TOOLS               *
*                                      *
****************************************


  A      : Einzelwortanalyse
  F      : ASCII-Datei analysieren
           (Defaulteinstellungen)
  D      : ASCII-Datei segmentieren
  L      : Lexikoneintrag zeigen
  S      : Lexikondatei (LEXIKA.LST)
           sortieren
  K      : Listenlexika (LEXIKA.LST)
           in Baeume kompilieren
  Z      : Lexikoneintraege zaehlen
  B      : Bildschirmausgabe einer
           Resultatsdatei
  R      : Resultatsstatistik einer
           Resultatsdatei
  E      : Programm beenden

   Ihre Wahl:k

KOMPILIEREN DER LISTEN-LEXIKA IN
BUCHSTABENBAEUME..
Lexemlisten eingelesen..
Lexem-Baum gespeichert!
Flexiv-Baum gespeichert!
Endungs-Hypothesen-Baum gespeichert!

Baum der trennbaren Verbzusaetze
generiert und gespeichert
Baum der Verbpraefixe generiert
und gespeichert


LEXIKON-KOMPILIERUNG BEENDET

LETTER-TREES IM FILE
/desklab/hanriede/prog/morph/lexika.tre
GESPEICHERT
Elapsed Real Time = 399.23 seconds
```

```
(6 minutes, 39.23 seconds)
Total Run Time     = 261.14 seconds
(4 minutes, 21.14 seconds)
User Run Time      = 256.21 seconds
(4 minutes, 16.21 seconds)
System Run Time    = 4.93 seconds
Process Page Faults      =       2,464
Dynamic Bytes Consed     = 26,622,544
Ephemeral Bytes Consed = 30,251,728
There were 10 dynamic GCs
There were 58 ephemeral GCs

Zurueck zum Hauptmenue: <RETURN>
```

Ein Lexikoneditor für MORPH existiert nicht, so daß die Erweiterung des Lexikons eine gewisse Vertrautheit mit der Struktur der Lexikoneinträge voraussetzt.

3.8 Transparenz und Vollständigkeit der Dokumentation

Das System MORPH ist beschrieben in der Magisterarbeit [Han91]. Die Arbeit enthält den vollständigen kommentierten Quellkode und eine Auflistung der definierten Flexionsklassen. Das Vorgehen beim Updaten des Lexikons ist in einem eigenen Kapitel beschrieben.

3.9 Verfügbarkeit und Wartung

Der Lisp-Quellkode des Systems MORPH ist frei verfügbar über anonymous ftp und liegt auf dem ftp-Server des Bayerischen Forschungszentrums für Wissensbasierte Systeme (FORWISS), `ftp.forwiss.uni-erlangen.de` (`131.188.180.1`), im Verzeichnis pub/morph:

1. ftp 131.188.180.1
2. login: anonymous
3. password: your email address
4. cd pub/morph
5. binary
6. prompt
7. mget *
8. quit

Das System wird zur Zeit nicht gepflegt und weiterentwickelt. Support kann in begrenztem Umfang unter der im Titel angegebenen email-Adresse gegeben werden.

Literatur

[BNTW86] Büttel, I.; Niedermair, G.; Thurmair, G.; Wessel, A.: *MARS: Morphologische Analyse für Retrieval-Systeme*, in Schwarz, C.; Thurmair, G. (Hrsg.): *Informationslinguistische Texterschließung*, Hildesheim, 1986, S. 157–216.

[Han89] Handke, J.: *Natürliche Sprache: Theorie und Implementierung in LISP*, Hamburg, 1989.

[Han91] Hanrieder, G.: *Robustes Wortparsing. Lexikonbasierte morphologische Analyse (komplexer) deutscher Wortformen*, Magisterarbeit, Universität Trier, Fachbereich II, Linguistische Datenverarbeitung/Computerlinguistik, Februar 1991.

[Ols86] Olsen, S.: *Wortbildung im Deutschen*, Stuttgart, 1986.

[SS88] Stechow, A. v.; Sternefeld, W.: *Bausteine syntaktischen Wissens. Ein Lehrbuch der Generativen Grammatik*, Opladen, 1988.

[Thu86] Thurmair, G.: *Eine maschinelle morphologische Analyse des Deutschen*, in Schwarz, C.; Thurmair, G. (Hrsg.): *Informationslinguistische Texterschließung*, Hildesheim, 1986, S. 66–107.

[Web76] Weber, H. J.: *Automatische Lemmatisierung — Zielsetzung und Arbeitsweise eines linguistischen Identifikationsverfahrens*, *Linguistische Berichte*, Bd. 44, 1976, S. 30–47.

LA-MoRPH
EIN LINKSASSOZIATIVES MORPHOLOGIESYSTEM
Darstellung für die ersten Morpholympics 1994

LA-MoRPH -kurz für Links Assoziatives Morphologiesystem- wurde an der Abteilung für Computerlinguistik des sprachwissenschaftlichen Instituts der Universität Erlangen-Nürnberg unter der Leitung von Professor Hausser entwickelt.

Folgende Kriterien standen bei der Konzeption und Entwicklung von LA-MoRPH im Vordergrund:

=> Der deklarative und der prozedurale Teil der Grammatik sind streng voneinander getrennt. Der deklarative Teil von LA-MoRPH besteht aus

1. LEXIKONEINTRÄGEN - in der Regel analysierte Grundformen,
2. ALLO-REGELN zur Erzeugung von Allomorphen und 3.

KOMBI-REGELN zur Konkatenation von Allomorphen.

Für diese Komponenten wird eine formale Notation verwendet, die zur Beschreibung unterschiedlicher Sprachen geeignet ist.

Der prozedurale Teil ist in der Programmiersprache C implementiert. Er funktioniert wie ein kombinierter Compiler/ Interpreter für die oben erwähnten Lexikoneinträge und Regeln: diese werden auf syntaktische Korrektheit überprüft und in ein Format übersetzt, das von der LA-MoRPH zugrundeliegenden virtuellen Maschine verarbeitet wird. LAP, der Links Assoziative Parser, benutzt diese Kompilate, um zur Laufzeit vom Benutzer eingegebene (oder von einer Datei eingelesene) Wortformen zu analysieren.

=> In LA-MoRPH werden Allomorphe durch Allo-Regeln aus den Lexikoneinträgen abgeleitet. Außer bei Suppletion müssen daher im Lexikon nur die Grundformen eingetragen werden. I

=> LA-MoRPH ist oberflächenkompositional: Wort formen werden nur durch die Verkettung von lexikalisch analysierten Allomorphen gebildet, wobei die Schreibweise der einzelnen Allomorphe nicht manipuliert werden kann. So wird etwa die Wortform 'bücher' als Verkettung der Allomorphe 'büch' und 'er' analysiert.

1 Z.B. erzeugt eine Allo-Regel aus dem Lexikoneintrag ("buch" (KN ...) buch) automatisch die analysierten Allomorphe ("buch" (KN ...) buch) und ("büch" (KN ...) buch). Daß diese Allomorphe zum gleichen Paradigma gehören, wird durch die gleiche Form an der dritten Stelle der Allomorph-Analysen ausgedrückt. Bei Suppletionsformen wie ("gut" (KA ...) gut) und ("bass"(KA ...) gut) werden die Allomorphe nicht über Regeln abgeleitet, sondern direkt ins Lexikon eingetragen. Auch hier wird der paradigmatische Bezug durch die gemeinsame Form 'gut' an der dritten Stelle der Analysen ausgedrückt.

- ▷ In LA-MORPH wird die Konkatenation der Allomorphe mit den Regeln einer linksassoziativen Grammatik gesteuert. Die Analyse mittels linksassoziativer Regeln ist leicht nachvollziehbar, weil die Wortformen allomorphweise von links nach rechts analysiert oder generiert werden und weil zu jeder Regel ihre Nachfolgeregeln explizit angegeben werden.

- ▷ LA-MORPH ist typentransparent, d.h. die Regeln der Grammatik werden von dem zugeordneten Parser direkt verwendet, wobei Parser und Grammatik die Regeln in derselben Reihenfolge verwenden und bei jeder Regel dieselbe Eingabe nehmen und dieselbe Ausgabe geben. Aufgrund der Typentransparenz besteht eine enge Beziehung zwischen Parser und LA-Grammatik, was die Fehlersuche und die Erweiterung der Grammatik sehr erleichtert.

- ▷ Die in LA-MORPH verwendeten linksassoziativen Grammatiken liegen in der Klasse der C1-LAGs und parsen in linearer Zeit. Sie haben daher in Theorie und Praxis einen niedrigen Speicherbedarf und ein günstiges Laufzeitverhalten.

1 Konzeptuelle Kriterien

1.1 Deklarative Spezifikation

• **Lexikoneinträge** sind in LA-MORPH als geordnetes Tripel definiert, bestehend aus der Oberfläche, der lexikalischen Kategorie und den semantischen Informationen eines Morphems:

```
("bUch"      (KN ES ER_N N - $) buch)
    ^            ^              ^
    |            |              |
Oberfläche  Kategorie   Schlüssel zur Bedeutungsbeschreibung
```

Die mögliche Umlautung wird in der Oberfläche durch das großgeschriebene U spezifiziert. Die Kategorie ist als eine Liste von Kategoriesegmenten definiert, wobei das Kategoriesegment KN die 'Klasse Nomen' bezeichnet, das Kategoriesegment ES beschreibt die Singularformen (*Buch, Buches*), das Kategoriesegment ER_N beschreibt die Pluralformen (*Bücher, Büchern*) und das Kategoriesegment N bestimmt das Genus (Neutrum). Das, was traditionelle Lexika bzgl. Bedeutung und Verwendung beschreiben, wird an der dritten Stelle (unter der Grundform) abgelegt. Da diese Informationen für den Prozeß der Wortformerkennung meist keine Rolle spielen, werden sie in LA-MORPH nur bei Bedarf sichtbar gemacht.

Innerhalb von LA-MORPH ist die Wahl der Kategoriesegmente und die Form der semantischen Repräsentation grundsätzlich frei. Der Benutzer muß nur darauf achten, daß die Form der Lexikoneinträge, insbesondere die Wahl der Kategorien, mit den Mustern der Regeln korrekt interagiert.

• **Allo-Regeln** werden vor der Laufzeit auf die Einträge des Lexikons angewendet und haben folgende abstrakte Form:

Lexikon-Muster ⇒ *1. Allomorph 2. Allomorph ...*

Allo-Regeln spezifizieren ihre Ein- und Ausgaben mit *Mustern.* Diese Muster werden durch reguläre Ausdrücke dargestellt, die um Variablen erweitert wurden. Paßt das *Lexikon-Muster* auf einen eingegebenen Lexikoneintrag, so werden aus ihm die Allomorphe *1. Allomorph, 2. Allomorph ...* gebildet.

Die Ausgaben einer Allo-Regel sind Zeichenketten, in denen nur Variablen auftreten dürfen, die im Lexikon-Muster bereits mit einem Wert belegt wurden. Variablennamen stehen in geschweiften Klammern hinter einem Dollarzeichen. Das folgende Lexikon-Muster paßt auf jeden Lexikon-Eintrag, der ein "U" enthält:

```
"${O1}U${O2}" ==> "${O1}u${O2}" "${O1}ü${O2}"
```

Wird diese Allo-Regel auf eine Oberfläche wie "bUch" angewendet, so wird die Variable *O1* mit dem Wert "b", die Variable *O2* mit dem Wert *ch* belegt. Als Ausgabe erzeugt die Regel die beiden Allomorph-Oberflächen "buch" und "büch".

Neben der Oberfläche können Allo-Regel auch die Kategorie und die Semantik der Eingabe modifizieren. Formal wird dies durch ein tabulares Regelschema ausgedrückt, wobei die Spalten für die Eingabemuster und die ausgegebenen Allomorphe stehen, und in den Zeilen durch die Schlüsselwörter 'surf', 'cats' und 'sem' angezeigt wird, ob Oberfläche, Kategorie oder Semantik beschrieben wird.

Das folgende Beispiel zeigt, wie die Allomorphe für die Ablautreihe von 'singen' durch eine spezielle Markierung in der Oberfläche der Eingabe (**si1ngen**) getriggert werden und für die verschiedenen Allomorph-Oberflächen verschiedene Kategorien zugeordnet werden.

```
RULE verben
...
  # singen
  surf:"${A}i1${B}en"    "${A}i${B}"              "${A}a${B}"      \
                         "${A}ä${B}"              "${A}u${B}"

...
  cats:"KV VS ${vcRest}"  "IV V12 _0 ${vcRest}"  "IV V3  _0 ${vcRest}"\
                          "IV V4  _0 ${vcRest}"  "IV V5  VS  ${vcRest}"

  sem:    "${S}"          "${S}"                  "${S}_i"  \
                          "${S}_k2"               "${S}"

   ...
  rules:
END
```

Zeilenbrüche sind mit '\' gekennzeichnet. In der obigen Regel werden die Oberfläche, Kategorie und Semantik von insgesamt vier Allomorphen (z.B. *sing, sang, säng* und *sung*) charakterisiert. Die Infinitivendung 'en' der Eingabeverben wird von der Allo-Regel entfernt. Die Markierungen **_i** und **_k2** in der semantischen Repräsentation stehen für Imperfekt und Konjunktiv II.

Die regelbasierte Ableitung der Allomorphe aus den Lexikoneinträgen führt zu einer neuen Klassifikation verschiedener Grade von (Ir)regularität in der Morphologie. Dabei basiert der Unterschied zwischen diesen Graden streng strukturell auf der Zahl der Lemmata, der Zahl der Allomorphe und der An- oder Abwesenheit spezieller Markierungen in den Lemmata.

- **Reguläre Paradigmen**

  Das Paradigma wird im Lexikon durch genau ein Lemma dargestellt, dem genau ein Allomorph entspricht, z.B. *schön* ⇒ *schön*. Dabei kann sich das Allomorph vom Lexikoneintrag unterscheiden. So wird bei Verben aus der Grundform der Stamm erzeugt: *lernen* ⇒ *lern*

- **Semi-reguläre Paradigmen**

  Das Paradigma wird im Lexikon durch genau ein Lemma dargestellt, dem mehrere Allomorphe entsprechen. Diese werden ohne spezielle Markierung des Eintrags über Regeln aus der Oberfläche abgeleitet, z.B. *heikel* ⇒ *heikel, heikl* oder *lächeln* ⇒ *lächel, lächl.*

- **Semi-irreguläre Paradigmen**

  Das Paradigma wird im Lexikon durch genau ein Lemma dargestellt, dem mehrere Allomorphe entsprechen. Der lexikalische Eintrag enthält Markierungen, die eine gezielte

Allomorph-Generierung steuern, z.B. *hAus* ⇒ *haus, häus* oder *schla2fen* ⇒ *schlaf, schläf, schlief.*

▷ **Irreguläre Paradigmen**

Das Paradigma wird im Lexikon durch mehr als ein Lemma dargestellt. Dabei entsprechen die Allomorphe meist den verschiedenen Lemmata, z.B. *geh* ⇒ *geh*, *ging* ⇒ *ging*, und *gegangen* ⇒ *gegangen.* Die Allomorphe unregelmäßiger Paradigmen können normal konkateniert werden.

Die strukturellen Kriterien, die die Regularitätsgrade in LA-MORPH unterscheiden, werden im Folgenden noch einmal als Tabelle dargestellt.

	genau ein Lemma pro Paradigma	Lemma ohne Markierung	ein Allomorph pro Lemma
regulär	ja	ja	ja
semi-regulär	ja	ja	nein
semi-irregulär	ja	nein	nein
irregulär	nein	ja	ja

Ein strukturbasiertes Kriterium für die Klassifikation von (Ir)regularitäten ist insofern von allgemeinem Interesse, als die Behandlung der Ausnahmen seit jeher zu den zentralen Aufgaben der Morphologie rechnet.

• **Kombi-Regeln** werden während der Laufzeit verwendet und kombinieren den bereits analysierten (oder generierten) Wortanfang mit einem 'nächsten' Allomorph. Kombi-Regeln haben die abstrakte Form:

Regelname:
Muster_des_Wortanfangs Muster_des_nächsten_Allomorphs
⇒ *Neuer_Wortanfang Regelpaket*

Die Kombi-Regeln, die einen leeren Wortanfang mit dem ersten Allomorph kombinieren, müssen explizit aufgezählt werden (START-Regeln).

Wie die Allo-Regeln beschreiben auch die Kombi-Regeln Eingabe und Ausgabe über Muster, die als reguläre Ausdrücke mit Variablen spezifiziert werden. Auch das tabulare Format und die Verwendung der Schlüsselwörter 'surf', 'cats' und 'sem' haben die Kombi-Regeln mit den Allo-Regeln gemeinsam.

Allo-Regeln und Kombi-Regeln unterscheiden sich jedoch in Bezug auf ihre Ein- und Ausgabe. Eine Allo-Regel nimmt einen Lexikoneintrag als Eingabe und bildet ihn in ein oder mehrere Allomorphe ab.

Eine Kombi-Regel nimmt dagegen einen Wortformanfang und ein Allomorph als Eingabe und erzeugt einen neuen Wortformanfang zusammen mit einem Regelpaket als Ausgabe. In der nächsten Kombination werden dann die Regeln des Regelpakets auf den zuletzt abgeleiteten Wortanfang und ein neues 'nächstes Allomorph' angewendet. Dieser linksassoziative Kompositionsprozess endet, wenn sich in der Eingabe kein 'nächstes Allomorph' mehr findet (erfolgreiche Analyse) oder wenn das 'nächste Allomorph' keine der aktiven Regeln erfüllt (ungrammatische Wortform).

Das folgende Beispiel zeigt die Kombi-Regel für eine bestimmte Pluralbildung bei den deutschen Nomina.

```
#-------------------- pluralS ---------------------------------#
#                                                               #
# Kombiniert einen Nominalstamm mit einer Pluralendung "-s".    #
# mama (F)->mamas                                               #
#---------------------------------------------------------------#

RULE pluralS
```

```
#       START           NEXT        RESULT

surf:   ".*"            "s"         "&"
sem:    "${S}"          ".*"        "${S}"
cats:   ". ${fuge}"     ".*"        "P ${fuge}"

rules: CpVerbpSubnSub^ CSubPrefix^

END # pluralS
```

Die Operationen werden parallel auf die Oberfläche (*surf*), die Semantik (*sem*) und die Kategorie (*cats*) angewandt. Die Start-, Folge- und Ergebniskategorien stehen in den Spalten mit den Überschriften START, NEXT und RESULT. Das Regelpaket mit den Folgeregeln steht hinter dem Schlüsselwort "rules".

1.2 Bezug zwischen lexikalischen Einträgen und Wortformen

Die Analyse der Wortformen *anlächeln/anlächle* basiert zunächst auf dem Eintrag von *lächeln* im Lexikon:

```
("lächeln" (KV VH N GE {zu VH D GE} {an VH A GE} $ <be VH A GE- > ) lächeln)
```

Die Segmente der Kategorie haben folgende Interpretation. KV steht für 'Klasse der Verben' und gibt die Wortart an. VH bedeutet, daß *lächeln* mit dem Auxiliar *haben* kombiniert. N steht für die Nominativ-Valenz, wobei die Abwesenheit weiterer Kasus *lächeln* als intransitives Verb charakterisiert. GE besagt, daß das Partizip Perfekt von *lächeln* mit 'ge' gebildet wird (*hat gelächelt*).

Die dann folgenden Ausdrücke in geschweiften Klammern charakterisieren Formen von *lächeln* mit den trennbaren Präfixen 'zu' und 'an', wobei sich der Valenzrahmen ändert (*lächelte jemandem zu* versus *lächelte jemanden an*). In eckigen Klammern folgt schließlich das nicht-trennbare Präfix 'be', wobei das Segment 'GE-' besagt, daß das Partizip Perfekt ohne 'ge' gebildet wird: *hat belächelt.*

Aus dem Lexikoneintrag erzeugt eine Allo-Regel folgende Allomorphe:

```
("lächel" (RV SEL VH N GE { zu VH -I D GE} { an VH -I A GE} $
                                          < be VH S2i A GE- > ) lächeln)

("lächl" (RV SL _O N GE {zu VH -I D GE} {an VH -I A D GE} $
                                        <be VH S2i A GE- > ) lächeln)
```

Die Allo-Regel modifiziert die Kategorie des Lexikoneintrags, um die Kombination der Allomorphe mit den korrekten Endungen zu steuern. Die Wortform "anlächeln" wird von Kombi-Regeln aus den Allomorphen "an", "lächel" und "n" zusammengesetzt, die Wortform "anlächle" dagegen aus den Allomorphen "an", "lächl" und "e".

Die lexikalische Analyse der Allomorphe, die Abfolge der Kombi-Regeln und die verwendeten Regelpakete können auf dem Bildschirm in verschiedenen Formaten dargestellt werden. Aufgrund der Typentransparenz von LA-MORPH fungieren dabei *traces* des Parsevorgangs als die linguistische Analyse. Das folgende Beispiel zeigt die breadth-first trace der Ableitung von *anlächeln*:

```
lap(29  ) mor> anl"acheln

1  [NIL . a (PX ) a]
1  [NIL . a (PX ) a]
   RP: {CpVerbpSubnSub^ CadjStart^ CverbStart CverbGE_Start^
        CpxStart^ CpartIIStart}
   FIRED: *CpxStart

   [NIL . an (KF D ADP-L ) an]
   RP: NIL

   [NIL . an (PX ) an]
   RP: {CpVerbpSubnSub^ CadjStart^ CverbStart CverbGE_Start^
        CpxStart^ CpartIIStart}
```

```
    FIRED: *CpxStart

    [NIL . an (KF A ADP-D ) an]
    RP: NIL

2     [a (PX ) a . n (SX P13 P _ _ FGN ) pk1]
    RP: {pxGE_ZU^ CPrefixVerb CPrefixPartII^ CPrefixAdj^ CPrefixSub^}
    FIRED: none

      [an (PX ) an . läch (KN DIM ) lache]
    RP: {pxGE_ZU^ CPrefixVerb CPrefixPartII^ CPrefixAdj^ CPrefixSub^}
    FIRED: none

      [an (PX ) an . lächel (RV SEL VH N GE { zu VH -I D GE } { an VH
        -I A GE } $ < be VH S2i A GE- > ) lächeln]
    RP: {pxGE_ZU^ CPrefixVerb CPrefixPartII^ CPrefixAdj^ CPrefixSub^}
    FIRED: *CPrefixVerb

3       [an/lächel (-I A V) pref(an, verb(lächeln)) . n (SX P13 P __ FGN ) pk1]
    RP: {verbSel_n^ CpVerbpSubnSub^ partI^ CVerbStemPx^}
    FIRED: *verbSel_n

4         [an/lächel/n (P13 A V ) pref( an, verb(lächeln) )_pk1 . NIL]
    RP: {PSubConversion}
    FIRED: none

("an/lächel/n" (P13 A V ) pref( an, verb(lächeln) )_pk1)
```

Die Analyse zeigt, daß in der Oberfläche von **anlächeln** neben dem Allomorph 'an' auch das Allomorph 'a' (wie in 'aseptisch', 'asynchron' etc.) gefunden wird. '[x . y]' bezeichnet ein geordnetes Paar aus Wortanfang x und nächstem Allomorph y, wobei Wortanfang x und Allomorph y als geordnete Tripel aus Oberfläche, Kategorie und semantischer Repräsentation erscheinen. '[NIL . x]' charakterisiert den Beginn der Ableitung mit leerem Wortanfang. 'RP' steht für 'Regelpaket' und bezeichnet die Regeln, mit denen versucht wird, die Elemente des geordneten Paars zu kombinieren. Bei bestimmten Kategorien des nächsten Allomorphs werden keine Kombinationen versucht ('RP: NIL').

'FIRED' bezeichnet die Regeln aus dem Regelpaket, die erfolgreich auf ein geordnetes Paar angewendet wurden. Nachdem alle Kombinationen der ersten und zweiten Allomorphe in der Eingabe probiert worden sind und mindestens ein Versuch erfolgreich war, wird der Zähler am linken Rand der Ableitung um eins inkrementiert und der Vorgang wiederholt sich. Dabei erscheinen die Ergebnisse der letzten Kombination auf der linken Seite der neuen geordneten Paare, während auf der rechten Seite neue 'nächste Allomorphe' stehen.

Am Ende der Ableitung steht das Ergebnis. In dem obigen Beispiel wird die Form über die Kategorie (P13 A V) als erste oder dritte Person Plural (P13) mit einer Akkusativ-Valenz (A) und als finites Verb (V) charakterisiert. Die semantische Repräsentation besagt, daß die Form aus dem Präfix 'an' und dem Verb 'lächeln' zusammengesetzt ist. Die Markierung '_pk1' bedeutet, daß die Form semantisch im Präsens und Konjunktiv I steht.

Bei Anwendungen, wo nur das Ergebnis der morphologischen Analyse von Interesse ist, kann die Darstellung der Ableitung ebenso abgeschaltet werden wie die Markierung der Morphemgrenzen. Eine solche Analyse, die nur die analysierte(n) Wortform(en) als Ergebnis liefert, ist in dem folgenden Beispiel mit der Form *anlächle* illustriert (mit Anzeige der Morphemgrenzen):

```
lap(29  ) mor> anl"achle
("an/lächl/e" (S1 A V ) pref( an, verb(lächeln) )_p)
("an/lächl/e" (S13 A V ) pref( an, verb(lächeln) )_k1)
```

Die Analyse ist ambig, weil die Interpretation als Indikativ Präsens nur mit Nominativen der erste Person Singular kongruiert (*sobald ich sie anlächle*), während die Interpretation

als Konjunktiv I auch Nominative der dritte Person Singular nehmen kann (*ob er sie anlächle*).

1.3 Verständlichkeit und linguistische Motivation der Regeln

Als linguistischer Ansatz hat LA-MORPH die charakteristische Eigenschaft, daß Allomorphie und Konkatenation getrennt behandelt werden. Die Vorverarbeitung der Allomorphe erlaubt es, die Zerlegung der Wortformen in ihre Morpheme bzw. Allomorphe sehr einfach und effizient zu gestalten: die vorberechneten Allomorphe werden ohne irgendwelche Transformationen von links nach rechts an die Oberfläche der Eingabe angepaßt.

Zudem hat sich gezeigt, daß die getrennte Behandlung von Allomorphie und Konkatenation die Entwicklung von Morphologiegrammatiken konzeptuell außerordentlich erleichtert.

Bei den deutschen Nomina finden sich zum Beispiel folgende Allomorphietypen:

Umlaut bei (verkürzten) Diminutiva und Pluralformen

LEX	ALLO1	ALLO2	ALLO3
garten	garten	gärt	gärten
boden	boden	böd	böden

Umlaut bei Pluralform

buch	buch	büch
saal	saal	säl
haus	haus	häus
mutter	mutter	mütter
vater	vater	väter

Umlaut bei verkürzten Diminutiva (ohne Plural)

kuchen	kuchen	küch
daumen	daumen	däum
haken	haken	häk
tropfen	tropfen	tröpf
gurke	gurke	gürk
buchstabe	buchstabe	buchstäb
funke	funke	fünk

Umlaut bei nicht-verkürzten Diminutiva (ohne Plural)

staub	staub	stäub
paar	paar	pär
salat	salat	salät
boot	boot	böt
billion	billion	billiön
spur	spur	spür

Verkürzung bei nicht-umlautenden Diminutiva

tante	tante	tant
rübe	rübe	rüb

Plurale bei Fremdwörtern

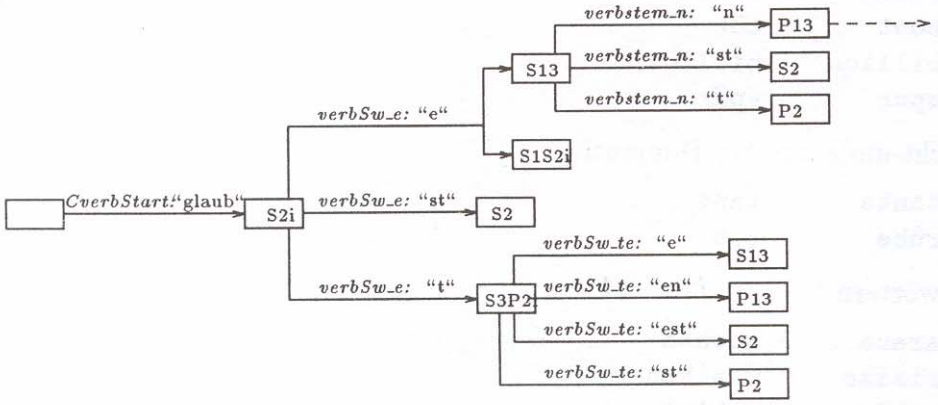
arena	arena	arenen
risiko	risiko	risiken
nukleus	nukleus	nuklei
virus	virus	viren
album	album	alben

Die Ableitung dieser Allomorphe basiert z.T. auf Allo-Regeln, die nur von Oberflächeneigenschaften getriggert werden (semi-reguläre Einträge), z.T. auf Allo-Regeln, die auf besondere Marker im Lexikoneintrag reagieren (semi-irregläre Einträge). Bei der Behandlung der Allomorphie muß der Benutzer nur darauf achten, daß über die Kategorisierung der Allomorphe die korrekten Kombi-Regeln - und damit z.B. die korrekten Endungen - angesteuert werden.

Wie die folgenden Beispiele von Konkatenationsmustern deutscher Nomina zeigen, sind die von den Kombi-Regeln behandelten Phänomene weitgehend orthogonal zur Allomorphie:

```
        es(MG)              es(MG)              es(MG)
       /                   /                   /
schmerz-e(MD)          tag                 leib-e(MD)
 (M-G) \              (M-G)\              (M-G)\
        en(P)               e(MDP-D)-n(PD)      er(P-D)-n(PD)

        s(MG)               s(MG)               s(NG)
       /                   /                   /
 gipfel              stachel             thema
(M-GP-D)\            (M-G)\              (N-G)
        n(PD)               n(P)                themen(P)

        s(MG)               s(NG)
       /                   /
 vAter               auge                uhu-(MGP)
 (M-G)\              (N-G)\              (M-G)
       (P)-n(PD)           n(P)

braten-s(MG)        hAnd-e(P-D)-n(PD)   frau-en(P)
(M-GP)               (F)                 (F)

drangsal-e(P-D)-n(PD)   kenntnis-se(P-D)-n(PD)   mUtter-(P-D)-n(PD)
  (F)                     (F)                      (F)
```

Es folgt ein etwas komplexeres Beispiel, das Konkatenationsmuster des Verbes "glauben":



Die ersten drei Kombinationsmuster unterscheiden sich z.B. durch die Verwendung unterschiedlicher Plural-Allomorphe *en*, *e* und *er*. Bei `Gipfel/n` und `Stachel/n`, andererseits,

sind die Formen zwar analog gebildet, repräsentieren aber verschiedene morphosyntaktische Eigenschaften: `Gipfel/n` ist nur Dativ-Plural (PD) während `Stachel/n` ein Plural ohne Kasusbeschränkung ist (P).

In dieser Weise können auch für die anderen Muster jeweils charakteristische Eigenstrukturen gezeigt werden. Die Behandlung des Diminutivs (*Händchen, Mütterchen*) wurde aus Gründen der Vereinfachung in den Beispielen nicht mit aufgeführt. Die Kategorisierung wird in der folgenden Sektion genauer erklärt.

Die getrennte Behandlung von Allomorphie und Konkatenation erweist sich nicht nur als konzeptuell klar und einfach, sondern wird auch aus empirischer und psychologischer Sicht unterstützt. Empirisch hat sich herausgestellt, daß die regelbasierte Ableitung sämtlicher Allomorphe aus einem gegebenen Grundformlexikon die Zahl der Einträge nur geringfügig erhöht.

So liegt der sogenannte Allomorphie-Quotient z.B. im Deutschen bei unter 30% und im Englischen bei unter 20%. Die Vorberechnung der Allomorphe in LA-MORPH vergrößert das Laufzeitlexikon für das Deutsche also nur um 30%. Dafür wird die Laufzeitanalyse enorm vereinfacht und beschleunigt, denn sie beschränkt sich allein auf die linksassoziative Allomorph-Konkatenation.

Aus psychologischer Sicht stellt sich die Frage, ob die für LA-MORPH charakteristische Behandlung der Allomorphe als eigenständige, vorberechnete Entitäten vertretbar ist. Der unabhängige Status lexikalischer Allomorphe wird von MacWhinney 1978 an Spracherwerbsdaten aus dem Ungarischen, Finnischen, Deutschen, Englischen, Französischen, Lettischen, Russischen, Spanischen, Arabischen und Chinesischen gezeigt. Hooper 1976 untermauert den Begriff der lexikalischen Allomorphe, indem sie die Konkretheit morpho-phonemischer Beschreibungen zeigt und indem theoretische Prinzipien zur Abgrenzung von Morpho-Phonemik und Phonetik aufgestellt werden.

Es folgen die Analysen der in der Ausschreibung genannten Wortformen:

```
lap(29  ) mor> tisch
("tisch" (S2i A D { auf -I A D } V ) verb(tischen))
("tisch" (M-G ) nom(tisch))
```

Die erste der beiden Analysen interpretiert 'tisch' als Imperativ.

```
lap(29  ) mor> tisches
("tisch/es" (MG ) nom(tisch))

lap(29  ) mor> tischen
("tisch/e/n" (P13 A D { auf -I A D } V ) verb(tischen)_pk1)
("tisch/e/n" (PD ) nom(tisch))
```

Die erste der beiden Analysen interpretiert 'tischen' als erste oder dritte Person Plural Präsens bzw. als Infinitiv.

```
lap(29  ) mor> vorbeischwammst
("vorbei/schwamm/st" (S2 V ) pref( vorbei, verb(schwimmen_i) ))

lap(29  ) mor> vorbeischw"ammest
("vorbeischwämm/est" (S2 V ) pref( vorbei, verb(schwimmen_k2) ))

("vorbei/ge/schwomm/en" (S ) vorbeischwimmen)

lap(29  ) mor> hausd"achern
("haus/däch/er/n" (PD ) comp( verb(hausen), nom(dach) ))
("haus/däch/er/n" (PD ) comp( nom(haus), nom(dach) ))

lap(29  ) mor> h"ausermeers
("häus/er/meer/s" (NG ) comp( nom(haus), nom(meer) ))

lap(29  ) mor> unabh"angigkeitserkl"arung
("un/abhängig/keit/s/er/klär/ung" (F ) nom( nom( negabhängig )klären))
```

```
lap(29  ) mor> unlesbares
("un/les/bar/es" (ES ) adj( lesen ))

lap(29  ) mor> durchdachte
("durch/dachte" (S13 A V ) pref( durch, verb(denken_i) ))

lap(29  ) mor> gut
("gut" (ADV ) adj(gut))
("gut" (N-G ) nom(gut))
("gut" (PX ) gut)

lap(29  ) mor> besser
("besser" (S2i RA { aus -I A } { auf -I A } { nach -I A } V ) verb(bessern))
("bess/er" (ADV ) adj(gut)_komp)

lap(29  ) mor> besten
("be/st/en" (EN ) adj(gut)_sup)
("best/e/n" (PD ) nom(best))
```

1.4 Morphosyntaktische Analyse (Kategorisierung)

Der Formalismus der LA-Grammatik im Allgemeinen und von LA-MORPH im Besonderen schreibt keine spezielle Notation für die Kategorien vor. Bei Beschreibungen natürlicher Sprachen hat sich jedoch die Verwendung eines distinktiven Kategoriensystems eingebürgert.

Eine distinktive Kategorisierung unterscheidet sich von der traditionellen exhaustiven Kategorisierung dadurch, daß versucht wird, pro Oberfläche möglichst nur eine einzige Lesart zu haben. Ein exhaustives Kategoriensystem ordnet einer Wortform dagegen alle zugehörigen Paradigmapositionen als Lesarten zu. Z.B. hat das Nomen *Frau* in einer exhaustiven Kategorisierung die Lesarten (FSN), (FSG), (FSD) und (FSA), wobei F für Femininum, S für Singular und N, G, D, A für Nominativ, Genitiv, Dativ und Akkusativ stehen.

In der distinktiven Kategorisierung von LA-MORPH hat *Frau* nur eine einzige Lesart, repräsentiert durch die Kategorie (F), für Femininum. Die Spezifikation des Genus in der Kategorie impliziert, daß die Form im Singular steht. Die Abwesenheit einer Kasusmarkierung impliziert, daß alle Kasus möglich sind. Somit enthält die distinktive Kategorie (F) alle für die Syntax notwendigen Informationen. Entsprechend hat die Pluralform *Frauen* nur eine Lesart, repräsentiert durch die Kategorie (P) für Plural.

Da die Interpretation der Kategorien für die Bewertung der Analysen und für die Frage der Verwendung in anderen Grammatiksystemen wesentlich ist, folgt die Legende der distinktiven Nomensegmente.

M-G = Maskulinum ohne Genitiv (*Schmerz, Tag, Leib*)
MG = Maskulinum im Genitiv (*Schmerzes, Tages, Leibes*)
MD = Maskulinum im Dativ (*Schmerze, Leibe*)
MN = Maskulinum im Nominativ (*Mensch*)
M-NP = Maskulinum ohne Nominativ und Plural ohne Kasusbeschr. (*Menschen*)
M-GP = Maskulinum ohne Genitiv und Plural ohne Kasusbeschränkung (*Braten*)
MGP = Maskulinum im Genitiv und Plural ohne Kasusbeschränkung (*Uhus*)
MDP-D = Maskulinum Dativ und Plural ohne Dativ (*Tage*)
M-GP-D = Maskulinum ohne Genitiv und Plural ohne Dativ (*Gipfel*)

F = Femininum ohne Kasusbeschränkung (*Frau*)

N-G = Neutrum ohne Genitiv (*Kind*)
NG = Neutrum im Genitiv (*Kindes*)
ND = Neutrum im Dativ (*Kinde*)
N-GP = Neutrum ohne Genitiv und Plural ohne Kasusbeschränkung (*Mädchen*)
N-GP-D = Neutrum ohne Genitiv und Plural ohne Dativ *Zimmer*

NDP-D = Neutrum im Dativ und Plural ohne Dativ *Beine*

P = Plural ohne Kasusbeschränkung (*Schmerzen*)
P-D = Plural ohne Dativ (*Leiber*)
PD = Plural im Dativ (*Leibern*)

Auch die Kategorisierung der Adjektive und Adverbien ist im Rahmen einer distinktiven Kategorisierung sehr einfach, wie die folgenden Beispiele zeigen.

```
("gut"       (ADV )      adj(gut))
("bess/er"   (ADV )  adj(gut)_komp)
("be/st/en"  (EN )     adj(gut)_sup)
("gute"      (E )          adj(gut))
```

Die Angabe der Adjektivendung in der Kategorie genügt, um im Deutschen die Kongruenz mit dem Artikel (*das gute Buch* versus *ein gutes Buch*) zu steuern. Die Komparation wird in der Semantik kodiert.

Finite Verbformen sind durch das Kategoriesegment "V" am Ende zu erkennen. Ausserdem enthalten ihre Kategorien die Informationen über die Nominativkongruenz, wobei z.B. "S13" für erste und dritte Person Singular steht, über die nicht-nominativischen Valenzen, wobei z.B. "A" eine Akusativ-Valenz und "D" eine Dativ-Valenz bezeichnet. Endungen wie "_i" und "_k2" in der Semantik stehen für die Modi Imperfekt und Konjunktiv II.

Insgesamt orientiert sich die verwendete Kategorisierung an den traditionellen Konzepten wie Kongruenz und Valenz. Die Knappheit der Notation ist vielleicht anfangs gewöhnungsbedürftig, ist aber gerade bei der Analyse größerer Text von großem praktischen Nutzen, weil die Wortformanalysen in der Regel in eine Zeile passen. Da morphosyntaktische Eigenschaften, die konkret in der Oberfläche markiert sind, in der distinktiven Kategorisierung berücksichtigt sind, steht einer automatischen Transduktion etwa in eine Merkmal-Wertestruktur nichts im Wege.

1.5 Behandlung der Generierung

In LA-MORPH werden für die Generierung der Wortform-Paradigmen die gleichen Regeln wie für die Analyse verwendet. Grundlage hierfür ist die Tatsache, daß sowohl die Äußerung als auch die Interpretation von Sprache zeitlinear, also linear wie die Zeit und in der Richtung der Zeit, abläuft und daß diese Grundstruktur der natürlichen Sprachen in der Linksassoziativen Grammatik formal über die charakteristische Ableitungsordnung erfaßt wird.

So kann zum Beispiel das folgende Fortsetzungsnetzwerk für die verschiedenen Flexionsformen des Verbs *geben* sowohl bei der Analyse als auch bei der Generierung verwendet werden. Der einzige Unterschied zwischen Analyse und Generierung liegt darin, daß bei der Analyse das nächste Allomorph der Eingabeoberfläche geliefert und damit die Regelwahl bestimmt wird, während bei der Paradigmagenerierung die nächsten Allomorphe durch die anwendbaren Regeln bestimmt werden.

Bei der Generierung von Flexionsparadigmen einfacher Formen wird bei den Verben der Infinitiv, bei Adjektiven die Adverbform und bei Nomen der Nominativ Singular eingegeben. Es folgen die Formen, die LA-MORPH bei der Eingabe von "geben" generiert:

```
("gab"          (S13 D A *PX* V)      geben_i)
("gib"          (S2i D A *PX* V)        geben)
("gäb/e"        (S13 D A *PX* V)      geben_k2)
("gäb/en"       (P13 D A *PX* V)      geben_k2)
("gäb/et"       (P2 D A *PX* V)       geben_k2)
("gäb/est"      (S2 D A *PX* V)       geben_k2)
("gäb/st"       (S2 D A *PX* V)       geben_k2)
("gäb/t"        (P2 D A *PX* V)       geben_k2)
("geb/e"        (S1 D A *PX* V)         geben_p)
("geb/e"        (S13 D A *PX* V)      geben_k1)
("geb/t"        (P2i D A *PX* V)        geben_p)
("geb/end"      (ADV )                  geben)
("geb/ung"      (F)                     geben)
("gab/en"       (P13 D A *PX* V)      geben_i)
("gab/st"       (S2 D A *PX* V)       geben_i)
("gab/t"        (P2 D A *PX* V)       geben_i)
("gib/t"        (S3 D A *PX* V)         geben_p)
("gib/st"       (S2 D A *PX* V)         geben_p)
("geb/e/n"      (P13 D A *PX* V)      geben_pk1)
("geb/e/st"     (S2 D A *PX* V)       geben_k1)
("geb/e/t"      (P2 D A *PX* V)       geben_k1)
("geb/end/er"   (ER)                    geben)
("geb/end/er"   (ADV)                 geben_komp)
("geb/end/e"    (E)                     geben)
("geb/end/en"   (EN)                    geben)
("geb/end/em"   (EM)                    geben)
("geb/end/es"   (ES)                    geben)
("geb/ung/en"   (P)                     geben)
("geb/end/er/e" (E)                   geben_komp)
("geb/end/er/en" (EN)                 geben_komp)
("geb/end/er/er" (ER)                 geben_komp)
("geb/end/er/es" (ES)                 geben_komp)
("geb/end/er/em" (EM)                 geben_komp)
("geb/end/st/e" (E)                     geben_sup)
("geb/end/st/en" (EN)                   geben_sup)
("geb/end/st/er" (ER)                   geben_sup)
("geb/end/st/es" (ES)                   geben_sup)
("geb/end/st/em" (EM)                   geben_sup)
```

Die Notation '*PX*' in den Kategorien der finiten Formen besagt, daß die gerade bei *geben* sehr umfangreichen Präfixlisten abgeschaltet worden sind. Wie man sieht, bestehen noch Probleme mit der Generierung der Partizip Perfektformen, also *gegeben, gegebene, gegebenen* etc.

Bei Komposita werden im Prinzip die Grundformen der Teile in der gewünschten Reihenfolge als Eingabe verwendet. Dieser Aspekt der Generierung ist jedoch zur Zeit noch in Arbeit und noch nicht lauffähig. Die Generierung spezifischer Zielformen wird über die Paradigmengenerierung mit einem nachgeschalteten Filter gehandhabt, wobei auch hier Verbesserungsmöglichkeiten bestehen.

1.6 Übertragbarkeit auf andere Sprachen

In seinen verschiedenen Versionen wurde LA-MORPH bisher neben dem Deutschen und Englischen auf das Französische, Spanische, klassische Latein und Chinesische angewendet. Im Rahmen des CHILDES-Projekts an der Carnegie Mellon University wird eine Version mit Merkmalstrukturen für das *tagging* des Englischen, Deutschen und Holländischen verwendet. Anwendungen auf das Koreanische mit seiner sehr ausgeprägten Morphologie haben schnell zu einer umfassenden Beschreibung geführt und werden derzeit in Kooperation mit der Seoul University fortgeführt.

Sprachen mit Vokalharmonie oder Morpheminterkalation wurden bisher nicht im Rahmen von LA-MORPH beschrieben. Die Behandlung dieser Phänomene wurde jedoch theoretisch durchgespielt, wobei sich keine prinzipiellen Hindernisse ergaben.

2 Technische Konzeption und Einsatzfähigkeit

2.1 Zielsetzung der Konzeption

Folgende Aspekte standen bei der Konzeption von LA-MORPH im Vordergrund:

- LA-MORPH sollte auf dem Formalismus der linksassoziativen Grammatik basieren.

▷ Es sollte Wortformen bei geringem Bedarf an Haupt- und Plattenspeicher möglichst schnell analysieren:

Grundformlexika:	ca. 1,3 MB
	ca. 29.000 Grundformeinträge
Laufzeitlexika:	ca. 3 MB Hintergundspeicher
	ca. 70 kB Hauptspeicher
	ca. 37.000 Allomorphe
Allo-Regeln:	ca. 60 kB (ASCII-Datei)
Trie-Structure:	ca. 2 MB Hintergrundspeicher
	ca. 130 kB Hauptspeicher
Kombi-Regeln:	ca. 80 kB (ASCII-Datei)
	ca. 80 kB (Objekt-Datei)
Lexikon/Regel-Compiler:	ca. 440 kB
Parser:	ca. 130 kB
Parser-Speicherbedarf:	ca. 700 kB

▷ LA-MORPH sollte im Rahmen einer kombinierten morphologisch-syntaktischen Analyse arbeiten: "LAP", die Laufzeitkomponente, ermöglicht die morphologische und die syntaktische Analyse mit ein und demselben System.

▷ Das System soll ohne Probleme auf andere Rechner übertragen werden können. Deswegen wurde es in der genormten Programmiersprache "ANSI-C" implementiert.

2.2 Portabilität der Software und der Daten

LA-MORPH ist bislang in ANSI-C auf folgenden Rechnern implementiert:

▷ HP-Rechner der Serie 700 unter dem Betriebssystem HP-UX, Version 9.

▷ Apple Macintosh unter der Betriebssystem-Software Version 7

▷ IBM-PC-Kompatible unter MS-DOS, Version 5 sowie unter Linux, Version 0.99

2.3 Schnittstellen zur Syntax und zur Semantik

Die Kategorien der LA-MORPH-Analysen wurden von vornherein für die Verwendung durch einen Syntaxparser konzipiert. Sie werden bereits zu diesem Zweck verwendet.

Zu jedem Lexikoneintrag können auch semantische Informationen eingetragen werden, die ebenso wie die Kategorien durch die Allo-Regeln und Kombi-Regeln manipuliert werden können. Erfahrungen mit konkreten Semantiksystemen existieren allerdings noch nicht.

2.4 Hilfestellung bei Benutzerfehlern

Syntaktische Fehler in den Lexika, den Allo- oder den Kombiregeln werden bereits zur Übersetzungszeit erkannt und zusammen mit der Zeilennummer, in der der Fehler auftrat, gemeldet. Logische Regelfehler können durch einen *trace* eingegrenzt und und gefunden werden. Eine *trace*-Ausgabe kann den Analysebaum entweder für jede vollständige Analyse getrennt ausgeben (depth first), oder nach der Zahl der Regelschritte geordnet (depth first). Für *EMACS* wurde ein spezieller LAG-Modus entwickelt, der Edierung und Debugging der Regeln erleichtert.

Hier ein *depth first trace* bei der Analyse der Wortform "glaube":

```
lap(29  ) mor> -md
lap(29  ) mor> glaube
 1  [NIL . glaub (RV SV VH N A D GE $ ) glauben]
    RP: {CpVerbpSubnSub^ CadjStart^ CverbStart CverbGE_Start^
         CpxStart^ CpartIIStart}
    FIRED: *CverbStart

     [glaub (S2i A D V ) verb(glauben) . e (SX S1 DS ADJ: E FGE ) pk1]
    RP: {verbSw_e^ CpVerbpSubnSub^ partI^ verb_ung^ PmovER^
         CVerbStemPx^ PverbBAR}
    FIRED: *verbSw_e

     [glaub/e (S1S2i A D V ) verb(glauben)_p . NIL]
    RP: {verbstem_n}
    FIRED: none

 2  [NIL . glaub (RV SV VH N A D GE $ ) glauben]
    RP: {CpVerbpSubnSub^ CadjStart^ CverbStart CverbGE_Start^
         CpxStart^ CpartIIStart}
    FIRED: *CverbStart

     [glaub (S2i A D V ) verb(glauben) . e (SX S1 DS ADJ: E FGE ) pk1]
    RP: {verbSw_e^ CpVerbpSubnSub^ partI^ verb_ung^ PmovER^
         CVerbStemPx^ PverbBAR}
    FIRED: *verbSw_e

     [glaub/e (S13 A D V ) verb(glauben)_k1 . NIL]
    RP: {verbstem_n}
    FIRED: none

 3  [NIL . glaube (KN _2 EN_S M - $ ) glaube]
    RP: {CpVerbpSubnSub^ CadjStart^ CverbStart CverbGE_Start^
         CpxStart^ CpartIIStart}
    FIRED: *CpVerbpSubnSub

     [glaube (MN - ) nom(glaube) . NIL]
    RP: {pluralN^ genitivNS^ CpVerbpSubnSub^ CSubPrefix^ CSubFuge^
         CSubVerbStem^ CungSfugeAdj^ CDiminutiv^}
    FIRED: none

("glaube" (MN ) nom(glaube))
("glaube" (S1S2i A D V ) verb(glauben)_p)
("glaube" (S13 A D V ) verb(glauben)_k1)

lap(29  ) mor>
```

2.5 Größenbeschränkung des Systems

Die Größe des Lexikons und der Umfang der Regeln sind in LA-MORPH unbeschränkt. Ein Lexikoneintrag darf bis zu 8.000 Zeichen lang sein. Im Moment beschränkt LA-MORPH die Zahl der Kategorien auf 200. Ein regulärer Ausdruck darf maximal 256 Zeichen lang sein, eine Variable bis zu 1024 Zeichen.

2.6 Schnittstelle zu Nicht-ASCII-Zeichen

Das System kann zur Zeit alle Zeichen verarbeiten, die im ISO-Latin1-Standard vereinbart wurden. Dazu gehören auch die Umlaute, das "ß" und Zeichen mit Akzenten. Wir planen, auch nicht-lateinische Schriftzeichen wie Hangul über UNICODE zu verarbeiten.

2.7 Benutzerfreundlichkeit des *turn around*

Um neue Grundformen in das Lexikon einzutragen, sind folgende Schritte auszuführen:

1. Lexikon edieren
2. System neu übersetzen (mit dem Kommando "i_m")

3. LAP neu starten.

Zur kompletten Neuübersetzung des Systems benötigt eine HP-Workstation etwa ein bis dreieinhalb Minuten (siehe Anhang A).

Werden die Allo-Regeln geändert, müssen die Schritte 2. und 3. ebenfalls ausgeführt werden.

Hat man nur Kombi-Regeln geändert, so reicht es, wenn man nach dem Edieren der Datei mit den Kombi-Regeln den Befehl "make" eingibt und schließlich LAP neu startet. Dadurch verringert sich die Übersetzungszeit auf zwei bis vier Sekunden.

Auf unbekannte Wortformen reagiert das System mit der Ausgabe des Fehlers "unknown wordform".

2.8 Transparenz und Vollständigkeit des Systems

Zu LA-MORPH existieren folgende Dokumentationen:

- **M1** (*Schüller, Stubert*) Eine formale Spezifikation der Regelsprache
- **MOSAIC - A Tutorial in Computational Morphology** (*Hausser, Schüller, Zierl*) Eine Einführung in die Bedienung von LA-MORPH
- **Complexity in left-associative grammar** (*Hausser*) Eine Analyse der Komplexitätsklassen der LAG

Sie sind in der Abteilung für Computerlinguistik der FAU Erlangen, Bismarckstraße 12, 91054 Erlangen, email: `rrh@linguistik.uni-erlangen.de`, erhältlich.

2.9 Verfügbarkeit und Wartung

Das LA-MORPH-Paket ist für Forschungsinstitute als Quelltext auf Anfrage bei der obigen Adresse erhältlich, unter folgenden Bedingungen:

- Daten und Programme dürfen nicht an Dritte weitergegeben werden
- Das System darf nicht für gewerbliche Zwecke genutzt werden
- Publikationen, zu deren Erstellung LA-MORPH verwendet wurde, müssen einen Zitatvermerk enthalten.

LA-MORPH wird im Moment bei folgenden Projekten verwendet:

- im Institut für Mustererkennung der Informatik der FAU Erlangen-Nürnberg im Rahmen der Automatischen Fahrplanauskunft EVAR
- bei TA zur Entwicklung eines automatischen Übersetzers Englisch/Deutsch und Deutsch/Englisch

Das System wird von uns z.Zt. gepflegt und weiterentwickelt. Technische Anfragen und Fehlerberichte werden von Gerald Schüller (email: `gero@linguistik.uni-erlangen.de`) und Oliver Lorenz (`orlorenz@linguistik.uni-erlangen.de`) bearbeitet.

The Coordinator's Final Report on the First Morpholympics

March 7 and 8, 1994
The University of Erlangen-Nürnberg

Roland Hausser
Gesellschaft für Linguistische Datenverarbeitung *GLDV*
Working Group *Parsing in Morphologie und Syntax*

Most scientific conferences are exciting and unpredictable events for active participants and organizers alike. This proved to be especially true of the 1. Morpholympics, a new type of conference combining the presentation of scientific papers with a competition for testing software on test samples. Being the first event of its kind, participants, jury, and organizers all moved in uncharted waters.

This report analyzes some of the experiences gained during preparation and realization of the 1. Morpholympics. Section 1 discusses different methods of evaluation. Section 2 proposes a standardized procedure for preparing testing devices. Section 3 explains where difficulties with testing devices may arise. Section 4 discusses procedural issues. Section 5 summarizes the experience with a proposal for standardizing procedures in future Morpholympics.

1 Preparations

1.1 Standardized Presentations

To maximize consensus among potential participants at the 1. Morpholympics, a preparatory meeting was held on October 14 and 15, 1993, in Erlangen. It was attended by researchers from 12 universities in Austria, Germany, and Switzerland, all experienced in computational morphology. After the on-line presentation of 10 systems,[1] the group discussed and reached agreement on which aspects of morphological parsers should be evaluated.

The results were published as part 2 of the announcement of the 1. Morpholympics.[2] This *Questionnaire for the standardized presentation of systems* presents detailed questions on the following points:

2.1 *Conceptual criteria*
2.1.1 Declarative specification of lexical entries and rules
2.1.2 Relation between lexical entries and word forms
2.1.3 Transparency and linguistic motivation of the rules
2.1.4 Morpho-syntactic analysis (categories)
2.1.5 The handling of generation
2.1.6 Application to other natural languages
2.2 *Technical design and practical use*
2.2.1 Conceptual goal of the design
2.2.2 Portability of software and data
2.2.3 Interfaces to syntax and semantics
2.2.4 Aiding the user
2.2.5 Limits on the size of the system
2.2.6 Interface to non-ASCII characters
2.2.7 User friendliness of the 'turn around'
2.2.8 The state of the documentation
2.2.9 Availability and maintenance

It was assumed that the information provided on each of these points in combination with the measurements on data coverage and speed would be sufficient to provide a solid, empirical basis for an objective evaluation of the participating systems.

[1] See Seewald 1993, LDV-Forum 10.6, p. 6 - 16.

[2] The German version appeared in LDV-Forum, Vol 10.6 (December 1993), p. 17 - 23, the English version was made public on various electronic lists and is now available via anonymous ftp from sol.linguistik.uni-erlangen.de.

1.2 Approaches to evaluation

Right before the actual contest, members of the jury expressed the feeling that the criteria for evaluating different systems had not been defined clearly enough. Instead of further elaborating, weighting and formalizing the criteria established during the preparatory meeting, the jury decided on a more spontaneous approach, selecting and balancing what they felt to be the most impressive.

This is a possible approach, corresponding to traditional practice. For the longterm success of a competition like the Morpholympics, however, it is important that participants and jury alike operate with a common and widely accepted set of clearly defined assumptions and expectations.

The first and most basic step in this direction is bringing the test data into a canonical form. The second step is to evaluate all competing systems systematically and equally with respect to their performance in each of the tests. The third is to arrive at a decision based on steps one and two.

2 Canonical form of testing devices

2.1 Types of tests

In order to evaluate morphological parsers with respect to their *completeness 01 cover age, speed, quality 01 analysis,* and *quality 01 implementation* they should be tested using the following five types of testing devices:

1. text of written language (w-text)
2. text of transcribed spoken language (s text)
3. list of word forms derived from texts (t-list)
4. questions for evaluating quality of analysis (a-questions) .
5. questions for evaluating quality of implementation (i-questions)

These different types of testing devices make different demands and test for different properties.

Analysis of the w-text requires that the system can handle the structure characteristic of written texts, such as headings, punctuation, line breaks, end-of-the-line hyphenation, and special characters. For the analysis of the s- text, the system must handle the conventions for transcribing spoken language.

The t-list contains each word form only one e. For this reason coverage can be tested for a much larger number of forms than in a text of comparable size. Also, because of the uniqueness of forms, the use of a *cache* for high frequency words or for reusing previously analyzed forms has little or no advantage in the case of t-lists. Therefore, a system's speed will usually be considerably slower on word lists than on a comparable text. t-lists are to be generated automatically from a corpus of texts, possibly using certain structural or statistical principles of selection.

The set of a-questions should probe the quality of linguistic analyses by presenting word forms which relate to the topics presented in the questionnaire under 2.1 *Conceptual Criteria.* Each a-question should be furnished with comments clarifying the purpose of parsing the test forms.

The set of i-questions, finally, should address the topics presented in the questionnaire under 2.2 *Technical design and practical use.* They may be extended into a subtest in which participants develop a grammar for a small, clearly defined set of morphological data from a little known language. By publicly demonstrating the adaptation of each system to the same set of new data, the systems' practical handling, the nature of their rules, and their conceptual approach to morphological analysis may be demonstrated most clearly.

2.2 Preparation of Testing Devices

During the preparatory phase of the 1. Morpholympics, it seemed sufficient to simply use well selected pieces of text or lists of word forms as test samples. After all, the systems should perform in as natural a situation as their normal practical use would require. While this reasoning is generally true, it overlooks that the *comparison* of

different systems is difficult, if not impossible, unless the samples have been carefully prepared for the purpose of testing.

The metamorphosis of a piece of nondescript on-line data into a testing device is brought about by embedding the data into a standard structure. This structure consists of a two part header at the beginning of the test file3 and markers indicating beginning and end of different kinds of data. By stating the header and the begin/end-markers within the formal convention of SGML comments,4 the prepared test file may serve as input to a morphological parser without any need for further editing,5 provided the morphological parser knows how to handle the SGML comment convention.

The data embedded into the structure of a test file are not to be modified in any way. The structure of a test sample depends on its type. We begin with a detailed description of the structure of w-text test samples, to be followed by briefer descriptions of texts and t-lists and a-questions.

2.2.1 Preparation of w-texts

A textual testing sample prepared from a written text consists of part 1 of the header - providing information on the origin of the sample, part 2 of the header - providing essential statistics, part 1 of the data, consisting of the list of ill-formed word forms found in the text and part 2 of the data, consisting of the text itself. The two data parts are clearly set off by begin/end-markers consisting of declarations beginning with upper case letters of the Latin alphabet.

The declarations of the header are filled out by the researcher preparing the sample.

<!- - 1. Type of test sample: w-text - ->
<!- - 2. Name and address of researcher selecting the sample: - - >
<!- - 3. Time, place and occasion of creating the test sample: - - >
<!- - 4. Reason(s) for selecting text as test sample: -->
<!- - 5. Origin of text: (author, date, publisher) ->
<!- - 6. Structure of the text: (e.g. SGML) - ->
<!- - 7. Coding used for special characters: (e.g. ASCII) - - >

<!- - a. Method for counting word forms: - -> <!- - b. Total number of word forms: - ->
<!- - c. Number of well-formed word forms: - ->
<!- - d. Number of ill-formed word forms: - -> <!- - e. Percentage of well-formed word forms: ->
<!- - A. Begin list of ill-formed word forms - - >

form1
form2
form3

<!- - A. End list of ill-formed word forms - ->
<!- - B. Begin w-text - ->
W-TEXT
<!- - B. End w-text - ->

The general information about the origin of the sample begins with a declaration specifying the '1. Type of the sample.

By giving his or her '2. Name and address' a researcher takes charge as the author of the testing device. Stating the '3. Time and place' of creating the test sample and the '4. Reasons for selecting' a given text are useful for developing a taxonomy of different types of test samples.

An exact specification of the '5. Origin of text' is necessary for future comparisons with other test samples. Questions regarding the exact nature of misprints, hyphenations, typographical and dialectal idiosyncrasies and other properties important for the performance evaluation of morphological parsers can only be resolved by being able to go back to the specified origin of the document.

An explicit specification of the conventions used for representing '6. Text structure' and '7. Special characters' tells a user right away whether a given system is able to handle the sample. This information may also be used directly by morphological parsers capable of interpreting SGML declarations, including SGML comments in the position of the header . 6

3 Addition of the header information in the test file itself, rather than a separate Readme file, is also practical for later references, when only certain samples from a test set may be selected for new test runs, discussion or comparison.

4 See Herwijnen *1990*, p. 72.

5 E.g. removal of the header.

6 For example, there are currently at least five different conventions for representing Umlaut in German, according to which, e.g., the preposition

The second part of the header provides numerical information on the number of well-formed and ill-formed word forms contained in the sample. It begins by stating the 'a. Method of counting word forms.,7 If no official method of word count is stated, the numbers for one and the same text arrived at by different systems may vary by as much as 18% (see section 3.1).

Giving the official 'b. Total number of word forms' allows competing morphological parsers to calibrate their word count algorithms. The declaration 'co Number of well-formed word forms' provides the mark morphological parsers should achieve when analyzing the sample. The declaration 'd. Number of ill-formed word forms' is redundant in light of band c, and therefore suited to indicate whether the numbers are consistent.

Stating the 'e. Percentage of well-formed forms' gives a handy guide line for the initial evaluation of coverage by a morphological parser. This is because most modern parsers automatically provide statistical information at the end of an analysis, including the *percentage of analyzed word forms.* The percentage of analyzed word forms provided by the parser should equal the official percentage of well-formed forms (and not 100% of the grand total of word forms).

The two part header is followed by two kinds of data. The first consists of an explicit official list of the ill-formed word forms contained in the text. This list is the basis for the numerical information in declarations d and e. By treating the list as part of the sample data, it is analyzed during parsing.

Finding the candidates for the list of illformed word forms in a larger text is easy enough by using a suitable morphological parser. The final decision on whether a word form is ill- formed or not may sometimes turn out to be a matter of different opinions, however. In this case, the form should be added to the list, followed by a comment consisting of a "?' and the standard variant of the form.

The second kind of data consists of the written text itself. The beginning and the end of the two kinds of data are clearly marked by SGML comment declarations. This helps in the interpretation of the testing device. Marking the end of the text is also useful for checking whether the test sample is complete when transmitted over the net.

2.2.2 Preparation of s-texts

The preparation of s-tests closely resembles that of w-texts. Differences arise only in the lines 1 and 6 of the header and the begin/end-markers of the text data:

```
<!- - 1. Type of test sample: s-text - -> ...
<!- - 6. Method/standard of transcription: (e.g., CHAT) - ->
```

```
<!- - B. Begin s-text - ->
S- TEXT
<!- - B. End s-text - ->
```

The main difference between w-texts and s-texts is that the textual structure of w-texts is provided by the author and/or publisher, whereas the structure of s-texts is imposed during the transcription.

2.2.3 Preparation of t-lists

The purpose of t-lists is to check the coverage of morphological parsers on a large set of word forms automatically derived from a corpus of texts representing a certain domain. The structure of t-lists should be such that each form to be analyzed occurs only once and is written into a separate line.

Like text samples, t-lists consist of a two part header and two kinds of data. The first kind of the data consists of a list of illformed word forms, the second of a list of

für may occur as 'für', 'f\374r', 'für', 'f' 'ur', and 'f}r'. Reading through the header of the text to be analyzed, a suitably extended system may determine and activate the specific convention required by the text. Capabilities like this should be added to the set of i-questions (2.2.5) in future Morpholympics.

1 A widely available method of counting the number of word forms in an on-line text automatically is wc of Unix. It should be noted, however, that wc is in many ways rather empty. For example, wc counts the parts of hyphenated words separately also, punctuation signs surrounded by spaces, e.g. '-' are counted as additional word forms, whereas those following a word without a space, e.g. '... sample. ' are not. Instead of wc a simple, linguistically motivated algorithm should be used, which disregards punctuation signs and treats the parts of end-of-the-line hyphenation as one word form.

well-formed word forms. To facilitate orientation by the user, the two sub-lists should each be structured according to the same ordering principle, such as alphabetical order. t-lists differ from text samples in that both kinds of the data are lists which are moreover disjoint.

```
<!- - 1. Type of test sample: t-list - ->
<!- - 2. Name and address of researcher selecting the sample: - - >
<!- - 3. Time, place and occasion of creating the test sample: - ->
<!- - 4. Reason(s) for selecting list as test sample: -->
<!- - 5. Origin of text or corpus from which list was derived: - - >
<!- - 6. Structure of the list: (e.g. alphabetical order) - - >
<!- - 7. Coding used for special characters: (e.g. ASCII) - - >
<!- - 8. Method by which list was derived from corpus: - - >

<!- - a. Total number of word forms: - ->
<!- - b. Number of well-formed word forms: - ->
<!- - c. N umber of ill-formed word forms: - - >
<!- - d. Percentage of well-formed word forms: ->
<!- - A. Begin list of ill-formed word forms - - >
    Form1
    form2
    form3
<!- - A. End list of ill-formed word forms - - >
<!- - B. Begin list of well-formed word forms - ->
    Form
    1
    form2
    form3
<!- - B. End list of well-formed word forms - ->
```

The first part of at-list header differs from that of w-text and s-text headers in lines 1, 5 and 6, which are straightforward adoptions to the different sample type. Of special interest is the additional declaration in line 8, which specifies the method by means of which the t-list was constructed from a given set of texts.

For example, 'listl', which served as the t-list at the 1. Morpholympics, was constructed automatically as a sub set of word forms of the LIMAS corpus. The subset was formed by selecting only word forms
from the open classes with a frequency of eight or more occurrences in the corpus.

The second part of at-list header does not state the method of counting word forms because this task is trivial, given that lists write each form into a separate line, contain no punctuations signs and do not use end-of-the-line hyphenation. The declarations a, b, c and d correspond to b,c,d, and e in text samples. Because word forms are unique in t-lists, the numbers of (a) word forms, (b) well-formed word forms, and (c) ill-formed word forms have a different status as compared to texts.

2.2.4 Preparation of a-questions

The sample types w-text, s-text and t-list have in common that they contain data or are based on data - that were produced and made public solely for regular purposes of normal communication. Because alterations of these data are not permitted for reasons of scientific method, the researcher has no influence on the specific word forms contained in such a test sample, apart from choosing a particular text or corpus and, in the case of t-lists, a particular method of automatic selection.

In contrast, a-questions are a testing device hand-made by linguists to check a parser's handling of specific phenomena in the morphology and orthography of a natural language. This freedom of chasing whatever forms the author of the list finds interesting should be complemented by comments which clarify the testing purpose of each form.

In order to allow the author of questions to present her/his data in the most transparent and perspicuous manner, the data are loosely organized into a list of 'data items'. Each data item consists of a statement describing its purpose and an open list of word forms. The word forms to be parsed are embedded into a format that can be handled by the systems.

Like the t-list, the a-questions are pre
ceded by a standard header .

```
<!- - 1. Type of testing device: a-questions - -> <!- - 2. Name and address of researcher selecting the sample: - - >
<!- - 3. Time, place and occasion of creating test device: - - >
<!- - 4. Coding used for special characters: (e.g. ASCII) - - >

<!- - a. Total number of word forms: - ->
<!- - b. Number of well-formed word forms: - ->
```

<!- - c. Number of ill-formed word forms: --> <!- - d. Percentage of well-formed word forms: ->

<!- - A. Begin list of data items - ->
<!- - BI. Begin data item - -> <!- - Purpose of data item: - ->
form!
form2 <!- - * - ->
form3
form4 <!- - ? - ->
<!- - BI. End data item - ->

<!- - B2. Begin data item - ->
<!- - Purpose of data item: - ->
form! <!- - * - ->
form2<!- - ? - ->
form3 .
form4
<!- - B2. End data item - ->

<!- - A. End list of data items - ->

Compared to t-lists, part 1 of the header differs in lines 4 and up. This is because a-questions are hand-made. Thus there is neither a text from which they derive, nor any method of derivation.

Each item of a-questions should be interpretable as checking a specific aspect of quality of linguistic analysis. For example,

< - - Purpose of data item: The following examples are intended to show the handling of valency frames in verbs: - - >

or

< - - Purpose of data item: The following examples are intended to show the handling of 'Fugenelemente' in the composition of nouns: - ->

The word lists following the statement of purpose may present well-formed and illformed word forms in any order the author finds suitable to her/his purpose. Because a-questions do not present the ill-formed word forms as a separate list - in contrast to w-texts, s-texts and t-lists -, it is imperative that ill-formed word forms are clearly marked by a comment. Based on the number of such comments, the number of ill-formed word forms can be determined automatically.

In explaining the purpose of an aquestion, the author may have to make explicit which grammaticality judgement or morphosyntactic characterization of a word form is assumed. If such assumptions spark controversies, it is certainly better than leaving the testing purpose of the data unexplained. The a-questions should help guiding and promoting the discussion of central topics in the community of morphological parsing.

2.2.5 Preparation of i-questions

The final test data used at the 1. Morpholympies, called textl, text2, listl, and list 2, corresponded more or less to the concepts of a w-text, an s-text, at-list and a set of a-questions, respectively. The concept of a set of i-questions, on the other hand, had not yet evolved at the time. For this reason, no specific experiences were made at the 1. Morpholympics that would guide in the formulation of standards for i-questions.

In light of the general experience, however, it seems important to test systems for automatic word form recognition with respect to their ability to adapt to new data. This may be done by presenting a small, clearly defined set of morphological data from a little known language in the set of i-questions. The description of the data should be followed by a list of well-formed and ill-formed word forms.

The task of each participating system is to write a grammar for these data and demonstrate its adequacy by parsing them. In this context systems may be evaluated with respect to additional theoretical and practical issues like the following:

- How well does the system separate between the rules used for a specific application and the general parser applying these rules?
- Does the system use a declarative rule format and how readable is its grammar for the set of new data?
- Does the system generate its own error messages or does it rely solely on the debugger of the programming language used?

. How easily does the system adapt to the handling of new special symbols?

. How long does it take to perform all the tasks requested by the i-questions?

The procedure of extending a system to handle the set of i-questions may be organized as a public performance, where the representative(s) of the system explain(s) each step, elaborating the linguistic and technical motives behind it.

3 Testing problems

As final test data for the 1. Morpholympics, the jury provided four files, called textl (2375 word forms8), text2 (1674 word forms), listl (3817 word forms) and list2 (282 word forms). While these files9 corresponded more or less to the concepts of a w-text, an s-text, at-list and a set of aquestions, respectively, they did not support the standards described 2.2.1 - 2.2.4. This created the following problems for evaluating the analyses of the participating systems:

3.1 Official word count totals

The final test samples did not provide numbers of word forms for each of the four test samples and no method for arriving at such numbers was agreed upon. Consequently, the numbers for the total word forms given by the eight participants varied widely. For example, for text! the eight participating systems submitted the following word counts to the judges: 2023, 2082, 2142, 2156, 2375, 2380, 2400.

The difference between the lowest count of 2023 and the highest of 2400 is a 377 word forms and amounts to a whopping 18.5%. This must be seen in light of the fact that the difference between the best percentage (95.5%) and the worst percentage (86.0%) of 'word forms analyzed' is only 9.5%. Thus the range of difference between different systems regarding the word count is twice as large as the difference in their respective percentages of analyzed word forms.

3.2 Official numbers of ill-formed word forms

The final test samples did not specify how many of the word forms in a sample happen to be ill-formed and should therefore not be recognized. This affected the proper evaluation of the percentage of recognized word forms.

For example, a quick examination of textl found the following 23 occurrences of misprints and uncommon abbreviations.

> Pfarr, fur, Schriftum, WV (13)10, syntematischen, Artiekls, Mei-nungeen, Millioen, gefdhreden, staatslichkirchlicher , seitigen

This amounts to 1 % of the actual 2020 word forms of textL Also of interest in this connection is an additional 1 % of unusual names and foreign words.

In list2 with its total of 280 actual word forms the following 15 problem forms can be found:

> molket, geretten, gesagen, geruft, vorbeigerannene, genennt, ankoemmt, voreingenommt, lieferen, schlotzen, schlotzte, geschlotzt, %anrufe, %anrufend, % angerufen 11

This amounts to 5.3% of the word form total of list2.

3.3 Official identification ill-formed ofword forms

ill-formed word forms were neither marked in the test samples nor presented as a separate list as part of the data. Regarding list2, for example, it remains a mystery whether forms like *molket, ankoemmt* or

8These word counts are based on wc. Note, however, that text! had been edited so that the punctuation signs ' . , ; : "_, appear between two spaces, thus being counted as separate words by wc. Therefore, the real number of word forms of text! is only 2020.

9The test data used at the 1. Morpholympics are available via anonymous ftp from so@linguistik.uni-erlangen.de in the directory 'morpholympics. '

l OWV may not be found in the *Wahrig Deutsches Wörterbuch* and is yet to be deciphered.

11 Apparently, three lines were intended to be commented out, which did not stop most parsers.

schlotzte were intended to be well formed or not. Consequently, it is unknown whether their recognition and analysis by a morphological parser should be counted as an achievement or as a mistake.

3.4 Evaluating the parsing of large texts or lists

It turned out that some systems used three pages and more for the analysis of a single word form. Consequently, files containing the analyses of text 1, text2 and list 1 were often huge and their perusal by the jury was frustrating due to a wealth of low level information and a concomitant lack of structure. Obtaining objective results on a system's degree of coverage by browsing through the associated files turned out to be impossible in the short time available.

Nevertheless, the parsing of large test files can serve as a simple, fast, and precise instrument for determining the quality of coverage by any number of competing systems if the following conditions are met:

. The test samples have been properly prepared, providing standardized word counts, explicit lists of ill-formed word forms, and correct official percentage numbers of well-formed word forms.

. The competing systems use the same method of word count and automatically provide the statistics described in 4.3.12

The list of ill-formed word forms provided at the beginning of w-texts, s-texts and tlists will show at a glance whether a system accepts ill-formed input or not.

3.5 Statements of purpose in a questions

As the only sample of the final test data constructed by hand, list2 most dozily resembled a set of a-questions. Unfortunately, however, there were no statements indicating what the word forms in list2 were being tested for. Many possibilities exist: Specification of valency structure? Coding of separable prefixes? Formation of past participle with or without ge-? Classification of certain forms as both indicative and subjunctive?

By. its very nature, the interpretation of a set of a-questions goes far beyond the counting of recognized word forms and should help guide the discussion on issues of linguistic theory and representation. Without a dear statement of purpose, however, it is virtually impossible to evaluate the analyses of word forms in an a-question, as produced by the different systems.

3.6 Balance of phenomena tested in a-questions

List2 contained only inflectional forms of verbs. A thorough evaluation of different morphological parsers with respect to their quality of analysis should be based on a well- balanced and linguistically motivated check list which represents different phenomena from the areas of inflection, derivation and compounding. Furthermore, the inflectional data should take into account nouns, verbs and adjectives/adverbials.

Given the limited time for deliberation and the unwieldy nature of the analysis files for w-text, s-text, and t-lists it is unlikely that a jury can evaluate the quality of word form analyses of different systems simply by browsing through the respective analysis files. Instead, the quality of word form analysis should be based on a carefully prepared set of a-questions.

3.7 Remark

It would have been easy enough to bring the test data used at the 1. Morpholympics into the canonical forms specified in 2.2.1 - 2.2.4, thus avoiding the difficulties of evaluation described above. Also, participants could have easily adapted to a common method of counting word forms. Unfortunately, however, no standards for the format of test samples and their evaluation had yet been written at the time.

12 For reasons of reliability and speed one may want to program a procedure for automatically ranking different systems with respect to their coverage of samples in canonical form.

4 Questions of procedure

4.1 Mode of participation

Participation at the 1. Morpholympics was open to any person or team that followed the rules of registration, installed the system in question via remote login, signed a publication agreement and turned in a standard questionnaire describing various theoretical and practical aspects of the system. This liberal procedure proved to be successful in that it resulted in a lively group of professionals from some very different necks of the woods.

4.2 Advancing distribution of writ ten presentations

All the systems presented were of good quality and raised a wealth of interesting issues. Unfortunately, however, because of a strict time table and because most of the systems were not previously known to the jury as well as the other participants, there was not sufficient opportunity to attain a deeper understanding.

To improve on this state of affairs, the written presentations should be sent to the coordinator four weeks before a Morpholympics (rather than being turned in during the preparatory meeting at the beginning of the conference). The coordinator collects these presentations into a volume and sees to it that duplicates of this volume be sent to each member of the jury.13 Other persons may also obtain copies of this volume, upon request and for a fee to cover copying and postage.

4.3 Supporting standards and automating statistics

The questionnaire requires participating systems to parse a set of preliminary test data and to append the results to the questionnaire. At future Morpholympics, these test runs should be used to calibrate systems to a common method of computing word form totals for arbitrary test samples of canonical form.

Furthermore, it is to be made obligatory for all systems participating at a Morpholympics that they *automatically* provide the following measurements at the end of an analysis file:

Total number of word forms encountered in the sample
N umber of word forms successfully analyzed
Number of word forms not recognized
Percentage of analyzed word forms
N umber of analyzed word forms per second

It is not difficult to add this feature to systems which do not yet have it. Automatic statistics are more reliable and quicker than calculations by hand, especially in the hectic atmosphere of a competition. Also, they will prove to be quite useful for practical work outside the context of the Morpholympics.

4.4 Only one parse per sample on newly loaded system

The measurements of an official parsing test must be based on a newly loaded system, parsing the test files in a given order and using exactly *one* test run per sample. There are at least two scenarios where using data from a second parse of a given sample leads to an improper manipulation of test results.

The first, which happily was not encountered during the 1. Morpholympics, has general technical reasons: Running a system on a sample for the first time requires the reading of data from the disco When parsing the same data a second time without reloading the system, the information read from the disc during the first analysis will still be available in the run time memory and thus result in a considerably faster timing.14 The measurements of the second parse are not relevant because in practical applications the user will not run the system twice just to enjoy a seemingly faster parse.

13 An even simpler method for the coordinator is to install the dvi-files of the presentations in an ftp directory from which the judges can obtain the presentations electronically and print them out at their respective offices.

l4 In the order of 4,000 versus 10,000 word forms/second in the case of the LA-Morph system.

The second scenario is more specific in that it depends on a system-dependent distinction between two separate phases in the parsing of a word form, called 'recognition' and 'analysis'. The first phase consists in a quick check whether the word form at hand is recognizable at all. If it is, a full analysis may be computed in the second phase.

Such a system may be run using only the first phase. In this case, a parse is about 5 times faster than when word forms are really analyzed. If the restriction to exactly one parse per sample is not enforced effectively, the participant running the test might be tempted to compose his results from the best of two different parses and then forget to mention that the timings are not those of the analyses.15

4.5 Variation of test data

In order to give potential participants an idea of what kind of test samples to expect during the competition, a set of preliminary samples was made available two months before the Morpholympics.16 These preliminary samples generated some discussion regarding the conventions of text structure, the coding of special symbols, and the handling of end-of-the-line hyphenation. To get the Morpholympics off to a gentle start, the final test samples were taken from the same domain as the preliminary samples and even edited to regularize and simplify various aspects of coding.

As a consequence, the participating systems varied less than 10% of the total number of word forms in their coverage of text samples. At future competitions, the range of domains should be extended so that a high degree of coverage is more difficult to attain. Furthermore, the systems' handling of different coding conventions should be tested.

In fairness to other participants, care must be taken that the nature of the final test samples is not leaked before the competition. This holds in particular with respect to the t-list, which may be reconstructed from description.

4.6 Testing portability

During the preparatory meeting in October 1993 it was agreed that systems should be tested on three different platforms to demonstrate port ability and to compare measurements. The platforms mentioned were a workstation, a Macintosh, and a PC. Regarding the operating systems nothing specific was said. It was assumed, however, that the operating systems usually associated with these respective types of hardware would be used.

This lack of specificity in the announcement was inconsequential with respect to the choice between DOS and WINDOWS on the PC. But by the time of the l. Morpholympics, the recently introduced LINUX operating system had been discovered to allow easy transfer of programs developed on Unix work stations to the PC, requiring no major adaptations and running at a quite respectable speed.

Even though the running on three different platforms had been announced as an obligatory part of the competition, 5 systems ran exclusively on Unix work stations and 1 system ran exclusively on the PC under DOS. Of the remaining two systems, one tested on the workstation, the Macintosh, and the PC under LINUX, while the other ran an additional fourth test, namely on the PC under DOS.

The practice of routinely adapting a piece of software to run under varying conditions is of great theoretical and practical benefit. It should therefore be encouraged in future Morpholympics, at least in the form of adding or subtracting points in the final evaluation of a system. That the timing of systems on different platforms is meaningful and interesting is demonstrated by the fact that the systems taking first and second place on the workstation under Unix were reversed on the PC under LINUX.17

15 Helping to ensure that the results of test runs need not be questioned is another reason why the various measurements (see 4.3) must be calculated automatically.

16 See 1.1.6 of 'Organization and implementation' in the announcement of the 1. Morpholympics.

17 See Coordinator's announcement of results, available via anonymous ftp from sol.linguistik.uni-erlangen.de.

5 Summary

To ensure that the evaluation of different systems is based on objective criteria, standardized procedures for the quantitative and qualitative measurement of performance should be advanced. The basis for such procedures is the preparation of different testing devices, consisting of three sets of data to be parsed and two sets of questions for checking specific aspects of quality.

Regarding quantitative measurements of software performance, it has been argued that small differences in speed, measured as the number of word forms per second, and coverage, measured as the percentage of word forms recognized, are not really meaningful. The same could be said about sporting events like the 100 meter sprint. From a practical viewpoint of daily life, a difference of a fraction of a second is indeed not really significant, but for winning the competition it is. The important property of such small differences is that it is (1) unquestionably objective and (2) agreed on by all participants. Runners of all shapes and sizes with different views on life and morals can adapt to and focus on this one parameter.

In the case of the Morpholympics, the evaluation is somewhat more complicated and more balanced because there are altogether four parameters relative to which systems are measured. These are (1) coverage, (2) speed, (3) quality of linguistic analysis, and (4) quality of implementation. The first two are measured quantitatively based on parsing the w-text, the s-text and the t-list. The latter two are evaluated qualitatively by using the a-questions and the i-questions, respectively.

Like the test samples, the catalogues of questions should be written by the jury in advance and kept under lock and key till the day of the contest. Just as all competing systems are tested and ranked with respect to coverage and speed on a given set of test samples, all systems should be tested and ranked with respect to explicit sets of questions checking for quality of analysis and implementation.

The preparation of test data add catalogues of questions prior to a Morpholympics requires a certain amount of work from the jury. However, given a jury of 5 judges, each would have to prepare only one of the 5 testing devices.

The benefits resulting would be great. Apart from an evaluation procedure beyond reproach, there would be precise measurements suitable to serve as bench marks for future systems. Even more importantly, the explicit questions for measuring quality of analysis and of implementation will direct attention to concrete problems of linguistic description and set standards for empirical analysis in future research.

Rezensionen

Simon C. Dik:
Functional Grammar in Prolog: An Integrated Implementation for English, French, and Dutch. (Natural Language Processing 2) Berlin: Mouton de Gruyter, 1992. 264 pp.; DM 118,- DM.

Das Buch ist eine Dokumentation zum Programmsystem ProfGlot, einem auf LPA Prolog abgestimmten 'multilingual natural language processor' mit Parsing-, Generierungs-, Übersetzungs- und Inferenzfunktionen (@1991 Amsterdam Linguistic Software). Die Programmoberfläche und die beiliegende Dokumentation präsentieren sich im gleichen Stil wie die ebenfalls von Amsterdam Linguistic Software vertriebene interaktive Einführung in LP A Prolog (@Logic Programming Associates Ltd, London) ProCourse (Dik/KahreI1991).

Das Begleit buch zu ProfGlot gliedert sich in folgende Teile (in Klammern jeweils Kapitel/Seiten): Einführung in ProfGlot (1/13), in Prolog (2/5-17) und in die Funktionale Grammatik [FG] nach Dik 1989 (3/19-30); Beschreibung des Programmsystems ProfGlot im Überblick (4-5/31-44) und nach Einzelmodulen und Funktionen (6-15/45-234); Anmerkungen, Literaturangaben, Verzeichnis der Prolog-Prädikate, Stichwort- und Namensregister (255264). Die Präsentation ist professionell und das Layout ansprechend, es ist eines jener Bücher, die einen gewissen Anreiz zum Blättern und Lesen ausstrahlen, wenn man sie in Händen hält.

Bei der Bestellung für die Institutsbibliothek fragten wir uns, ob denn wohl die Programmdiskette mit geliefert würde. Die Verlagsankündigung schwieg sich darüber aus - Vergessen oder Politik? - eine klare Aussage wäre jedenfalls wünschenswert gewesen. Wir kamen zu dem Schluß, daß, angesichts des stolzen Preises und der Tatsache, daß eine Softwarebeschreibung ohne die beschriebene Software zu wenig nutze sein würde, das Programm enthalten wäre. Erste Enttäuschung: es war es nicht.

Wir bestellten ProfGlot bei Amsterdam Linguistic Software (es kostet ein Vielfaches des Buchpreises). Inzwischen machte ich mich an die Lektüre der gut 250 recht kostbaren Seiten (so etwa DM 0.43 jede) des Druckwerks, von der fidelen Hoffnung getragen, daß das Buch, wenn es schon bona fide - ohne Programmdiskette verkauft wird, es auch einem Nur-Leser - bonae voluntatis - von Wert und Nutzen sein müsse. Es läßt und liest sich tatsächlich gut an: in gewohnt verständlicher Sprache und mit der typischen Dikschen Suggestivität werden einige "Elemente von Prolog" und Grundlagen einer Implementation der FG in Prolog vorgestellt. Alles sieht prima aus: Prolog ist einfach und gut, FG ist einfach und gut, sie sind wie füreinander gemacht, und ihre Verbindung ist selbstverständlich auch einfach und das Beste überhaupt.

Das Grundkonzept von ProfGlot folgt den theoretischen Prämissen in Dik 1989. Das Programm ist der Versuch einer (Teil-) Modellierung der Kompetenz eines (mehrsprachigen) natürlichsprachlichen Sprechers ("M.NLU" - s.auch Dik 1990). Der modulare Aufbau spiegelt den der Theorie mit ihrer Ausrichtung auf 'the construction of linguistic expressions in a quasi-productive, bottom-up way' (Dik 1989:54) wider. Generator- und Parsingkomponente funktionieren nach einheitlichen Regeln (vgl. Dik 1989:51); der Parser ist "eine Art invertierter Generator" mit dem' a sentence is analysed by considering how it could be formed' (33). Die Übersetzung auf der Ebene von 'underlying clause structures' (die, daran sei erinnert, aus einfachen und abgeleiteten Prädikatstrukturen (basic and derived predicates), i.e lexi-

kalischen Einträgen bestehen) funktioniert nach dem theoretisch postulierten Prinzip, daß 'translation will have to be reconstructed in much the same way as is done in a bilingual dictionary' (Dik 1989:86).

Der klare, modulare Aufbau der FG, die konsequente Formalisierung in Verbindung mit einer stets überzeugenden Nähe zur Sprachrealität, die Tatsache, daß auch das durchgehend eingehaltene Bemühen um allgemeine typologische Adäquatheit (fast) nirgends zu unglaubwürdigen konstruktionellen Klimmzügen führt, erweckte in mir, wie vorher schon, die Erwartung, daß eine Computer-Implementierung der FG besonders adäquat, inspirierend und erfolgreich werden würde. Zweite Enttäuschung: sie wurde es nicht.

Die Beschreibung der Lexikonmodule (mit Basisprädikaten Englisch / Französisch / Niederländisch) beginnt mit der Bemerkung, daß es sich nur um ein geschränkte Beispiele handele. Warum, fragt man sich, diese Beschränkung schon beim einfachsten (basic!) Teil? Wäre es nicht interessanter gewesen, sich auf ein oder zwei Sprachen zu beschränken und sich dafür an einer tiefergehenden, auch in ihrer Problematik diskussionswerten Implementation zu versuchen?

Ein Beispiel (aus dem Englischen) für eine einfache Verbprädikatstruktur und ihre faktische Deklaration als Prolog-Prädikat (basic predicate frame verbal):

```
giveV(x1:<anim>(x1»AG(x2)
      GO(x3:<anim>(x3» REC
      (Dik 1989:54)
```

- entspricht:

```
bpredv(eng,[[give],[act],[[[anim],t,[ag]],
      [[inanim] ,t,[pt]],[[anim],t,[rec]]]]) (48).
```

Die dreistellige FG-Prädikatstruktur ist als zweites Argument in ein Prolog-Prädikat integriert, das aus einer Sprachenangabe als erstem Argument und einer Liste mit drei Unterlisten besteht, von denen die letzte sich wiederum aus drei Listen mit je drei Elementen zusammensetzt. Die Liste an zweiter Argumentstelle enthält, jeweils als Liste mit einem Element, die Zitierform ([give)) und den denotierten semantischen Typ (state of affairs (SoA): [act)) und die Liste mit Argumentrollen (semantic functions: [ag], [pt], [rec)) und Selektions beschränkungen ([anim], [inanim]) - abweichend von der Theorie (vgl. Dik 1989:76) nicht als einstellige Prädikate (<anim>(xi)).

Die Listenstruktur wird nur knapp erläutert, warum die FG-Prädikatstruktur so und nicht anders umgesetzt wurde, wird nicht erklärt, ebensowenig Status und Funktion des 't' in den Argument-Unterlisten (es kann in LPA-Prolog jedenfalls keine Variable sein). Nominale und adjektivische Prädikate sind - mutatis mutandis - ebenso deklariert.

Die semantische Beschreibung der Substantivprädikate stützt sich auf wenige marker ([hum], [masc], [fern] ...), deren Status, Auswahl, Erweiterbarkeit nicht angesprochen wird. Sie sind angeblich 'selfexplanatory' (46); einige Redundanzregeln werden durch Implikationsmechanismen realisiert (70).

Bei den Adjektiven werden grammatische ([grad<able>)), FG-relevante ([eval <uative>]) und semantische ([vert<ical>D Kategorien (47) als 'features' angegeben. Es werden vereinzelte 'meaning postulates' ohne weitere Begründung von Auswahl und Zusammenhang angeführt. Daneben werden einige semantische und grammatische Besonderheiten in sog. 'idioms' und 'paradigms ' festgehalten. Die immer wieder beschworene Klarheit, Explizitheit, Gründlichkeit, die, ein Grundstein im Sinnfundament der Computerlinguistik, durch die Implementierung linguistischer Theorien als lauffähige Programme erzwungen würde, bleibt aus. Der Autor meint selbst, daß 'a more extensive and principled system of such restructions must no doubt be set up' (4 7). Wir warten gespannt auf Teil II - Titel: 'Real Functional Grammar in Prolog'?

Im längsten Einzelkapitel (9/69-108) des Buches wird das Kernstück des Systems, der 'universelle Generator' Uni Gen beschrieben, der für eine der auswählbaren

Sprachen (teilweise auf andere Sprachen übertragbare) 'fully specified underlying dause structures' (69) nach Maßgabe von veränderbaren random-Einstellungen generiert. Die allgemeinste Regel, fully -specified_dause(L,X) verzweigt zu 'dause'-, 'proposition'- und 'predication'-Bildungsregeln gemäß dem stratifikationellen Aufbau der FG. Der Aufbau wird, wie in der Theorie, von unten nach oben beschrieben. In den folgenden Kapiteln (10-12/109-167) werden die 'expression rules' zur Morphologie der Konstituenten, Wortstellung und zu sandhi-Erscheinungen (vor allem im Französischen) eingeführt. In den letzten drei Kapiteln (13-16; 169-234) werden das Parser, Übersetzungungs- und Inferenzmodul beschrieben, alle, wie der Generator, leicht hyperbolisch als "universal" bezeichnet: UniPar, UniTra, UniLog. Letzteres baut auf (deklarierten) Hyponymiebeziehungen zwischen Lexemen (rose flower), Bedeutungspostulaten (kiss touch [wiih the lips]) und satzlogischen Theoremen (p (p)) auf.

Die ständige Herausforderung an denjenigen/diejenige, der/die dieses Buch durcharbeitet, ist die, daß er/sie fast alles erraten oder aus der Theorie zu erschließen versuchen muß. Erklärt in irgendeinem interessanten Sinne des Wortes wird so gut wie gar nichts, Beispiele für die Anwendung der Prolog-Regeln mit Instanziierung der Variablen fehlen in Kap. 10 fast und in Kap. 9 völlig. Die in ProfGlot deklarierten Regeln werden, wenn überhaupt, nur äußerst knapp erläutert, so knapp, daß sie weder für einen wenig geübten Prolog-Benutzer wie mich, noch für einen besser trainierten Kenner verständlich sind (s. die Rezension von Patrick Saint-Dizier in Computational Linguistics 19/4 (1993) 695 f.). Kostproben:

(77)
```
Regel:
    predaderpred(L,[[[S,POL],
                          [SE, T, [stand]]] ,

            [state] ,

                          [[SE,t,[zero]]]]) :-
gamble(predform3), member(POL,[pos,neg]),
choose_arb_adj(L,[[S],[gradl_],
                          [[SE,t,[zero]]]]),
term(L,T,SE).
```

Das Prädikat choose_arb_adj. wird sechs Seiten vorher (71), term hingegen 19 Seiten später (96) eingeführt. Darauf fehlt hier ein Hinweis. Es gibt wohl einen Index der Prädikate am Ende, dennoch steht an dieser Stelle der Erklärungszusammenhang für ein später eingeführtes Prädikat aus.

(92)
```
Regel: do_all_terms(L,OP,PRED,PREDi) :-
    PRED = [P2,[Pi,[S,T,A],Si],S2],
    list_perform(L,OP,A,Ai),
    list_perform(L,OP,Si,Sii),
    list_perform(L,OP,S2,S22), PRED = [P2, [Pi,
    [S, T ,11],511],522] .
```

Erläuterung: 'We do a certain operation' "OP" on all terms by "list-performing" OP on the arguments, the satellites-1, and the satellites-2' (ohne Kommentar).

Es fällt mir schwer, den Nutzen dieses Buches zu erkennen, und ein Buch, dessen Nutzen ich nicht erkennen kann, finde ich schlecht. Und wenn ein besonders renommierter Wissenschaftler und Autor ein besonders schlechtes Buch herausbringt, das auch noch besonders teuer ist, so ist das besonders ärgerlich. Sei's drum, tröstete ich mich, vielleicht ist ja die tatsächliche Leistung des Programms überzeugender. Dritte Enttäuschung: sie ist es nicht.

Die erste Profglot- Version 2, die wir bekamen, führte auf einem normalen IBM386-PC mit 640 KB konventionellem und (nicht genutzten) 8 MB erweitertem Hauptspeicher zu vielfältigen Speicherüberlaufproblemen. Nach Reklamation bekamen wir eine modifizierte Version, die durch Ausklammerung von Menüs 23 KB mehr Hauptspeicher für die Ausführung der Regeln zur Verfügung stellt. Trotz vieler Haken und Ösen läßt sie sich soweit bedienen, daß man sich mit der linguistischen statt der technischen Funktionalität befassen kann.

Außer den im Buch referierten Sprachen gibt es in der Version 2 Sprachkomponenten für Dänisch, Italienisch, Spanisch/Galizisch (nur Generator) und Japanisch. Bei der Generierung ist es durchaus amüsant zu verfolgen, wie die erzeugten Sätze peu a peu vom leicht deplaziert Komödiantischen ("ehrlich gesagt, eeh, ich kann klugerweise husten"; «franchement,

eeh, sa enfant qui rigole cause du sac meurt probablemenb) ins stark agrammatisch Unsinnige (" von dat wen wird das buch des sacks das du berührst geschlagen werden?" j «probablement sa enfant qui rigole a. cause du sac du seau qui est donne par nn au professeur moins grand qui embrasse lui-meme meurb) abgleiten. Der Parser kann nicht alle Sätze analysieren, die der Generator vorher erzeugt hat, auch wenn man die Zufallseinstellungen aufeinander abstimmt.

Auch bei der Übersetzung (z.B. Deutsch-Englisch) ist nicht transparent, was wann und warum (nicht) funktioniert: manchmal wird/werden die Übersetzung/en ausgegeben, manchmal nur dause structures, oft auch ein lapidares 'no.' In der Inferenzkomponente erfahren wir, daß (in Auszügen): 'The following sentence: mary kicks john. entails: mary kicks john. john is kicked by mary. that mary does nOt kick john is nOt true. a female person touches a male person. that a person does nOt touch a male person is nOt true. it is nOt true that a person is nOt touched by aperson.' und hunderte ähnlich interessante Dinge.

Am Ende fragen wir uns: Ehrlich gesagt, eeh, kann es sein, daß wir alles in allem etwas veräppelt werden?

Literatur

Dik, Simon C. (1989): The Theory of Functional Grammar. Part1: The Structure of the Clause. Dordrecht / Providence, RI: Foris.

Dik, Simon C. (1990): How to build a natural language user, in: M. Hannay / E. Vester (eds.): Working with Functional Grammar. Descriptive and Computational Applications. Dordrecht / Providence, RI: Foris (203-215).

Dik, Simon C. / Kahrel, Peter (1991): ProCourse: A Prolog Course for Linguists. Amsterdam: Amsterdam Linguistic Software.

Nico Weber, Universität Bonn

Maier-Leibnitz-Preis für Dissertation von Stephan Mehl

Wer sich die Mühe macht, das LDV-Forum Bd. 4, Heft 2 aus dem Jahr 1986 herauszusuchen (Das Heft mit dem Buchstaben "B"), der wird als ersten Beitrag des Themenschwerpunktes "Sprachorientierte KI-Forschung" den von Stephan Mehl finden: *W ord expert parsing und Disambiguierung - Can inquiring experts be helpful?* Das Abstract umreißt knapp den Inhalt:

Unter World Expert Parsing versteht man ein vorwiegend auf lexikalischer Semantik beruhendes Analyseverfahren, bei dem die Auflösung von Mehrdeutigkeiten als zentrale Eigenschaft von Sprachverstehen in den Mittelpunkt gerückt wird. Es wird diskutiert, welche Methoden hierzu verwendet werden und inwieweit sie den an sie gestellten Anspruch erfüllen.

Stephan Mehl hat mittlerweile seine Dissertation mit dem Thema *Dynamische semantische Netze. Zur Kontextabhängigkeit von Wortbedeutungen* an der Universität Konstanz vorgelegt und dafür das Prädikat "mit Auszeichnung" erhalten. Und nicht nur das: Im Juni 1993 wurde seine Dissertation mit dem Heinz-Maier-Leibnitz-Preis (im Fachgebiet Kognitionsforschung) des Bundesministers für Bildung und Wissenschaft ausgezeichnet. Im infix-verlag ist sie im Dezember 1993 erschienen.

Es fällt nicht schwer, die Verbindung des 86-er Beitrags mit der aktuell ausgezeichneten Arbeit zu sehen:

Thema dieser Dissertation ist die Modellierung derjenigen Sprachverstehensprozesse, durch die einzelne Wörter in einem Textzusammenhang eine bestimmte Bedeutung erhalten. Am Beispiel des Verbs "abgeben" wird in der Dissertation gezeigt, daß die in verschiedenen Kontexten realisierten Bedeutungsvarianten eines mehrdeutigen Wortes (Polysems) ein Spektrum bilden. Dieses Spektrum wird in klassischen Lexika zerschnitten und in disjunkte Lesarten aufgeteilt, die durch Bedeutungsparaphrasen beschrieben werden. Dabei entstehen Grenzfälle, die regelmäßig zu Unsicherheiten bei der Zuordnung von Lesarten führen. Ein Vergleich der unterschiedlichen Einteilungen derselben Wortbedeutung in unterschiedlichen Wörterbüchern zeigt die Unsicherheit dieser traditionellen Vorgehensweise. Als Alternative wird in der Dissertation ein Modell vorgeschlagen, bei dem das Verwendungsspektrum eines Wortes durch semantische Relationen zu anderen Wörtern dargestellt wird, mit denen es in Kontexten typischerweise zusammen auftritt. Das gemeinsame Vorkommen einiger dieser Wörter führt dann bei der Verarbeitung eines Textes zu einer unterschiedlich starken, für diesen Text repräsentativen Gewichtung (Aktivierung) eines Teils des Relationenspektrums und weiterer damit verknüpfter Relationen. Auf diese Weise entstehen für jeden Text neue Aktivationsmuster, die als Repräsentation der jeweils neu konstituierten Wortbedeutungen interpretiert werden.

Mittlerweile arbeitet Stephan Mehl an der Universität Duisburg, Institut für Computerlinguistik bei Prof. Dr. W. Hoeppner.

Berufung an die FH Hannover: Annely Rothkegel

Am 1.3.1994 hat Annely Rothkegel (früher Universität Saarbrücken) eine Professur für Textproduktion an der Fachhochschule Hannover im Studiengang Technische Redaktion übernommen.

Dieser Studiengang - der erste und bisher einzige in der BRD - gibt Student/innen die Möglichkeit, -einen Sprachberuf zu erlernen; der technisches Verständnis, Sprachbewußtsein und Sprachfertigkeiten sowie den Umgang mit modernen technischen Hilfsmitteln kombiniert. Die Lehrenden stammen aus den Bereichen Informatik, Technik, verbale und visuelle Kommunikation und schließlich Linguistik, Textlinguistik, Computerlinguistik. Letztere waren bislang die Arbeits- und Forschungsschwerpunkte von A. Rothkegel (u. a. im früheren Sonderforschungsbereich Elektronische Sprachforschung, im MÜ-Projekt EUROTRA und zuletzt mit "Textproduktion" als Hauptthema am Institut für Computerlinguistik an der Universität des Saarlandes.

In dieser Richtung soll es auch weitergehen, sowohl in der Lehre, die nun in Hannover an der FH verstärkt eine Rolle spielt, als auch in der Forschung. Hierzu besteht über ein Esprit-Projekt (" Textrepräsentationen" in DANDELION) noch weiterhin eine enge Verbindung mit der Uni Saarbrücken. Des weiteren durch eine vom DAAD geförderte Zusammenarbeit mit der Universität Straßbourg (Phraseologiedatenbank).

Aber auch in Hannover soll es dazu demnächst mehr Forschung geben. Die neue Anschrift lautet:
Fachhochschule Hannover Bernhard-Caspar-Str.7
30453 Hannover
Tel. 0511/ 2123930
Fax: 0511 / 2103000
email: rotkegel@TR-Server.tr.fh-hannover.de
rothkegel@coli.uni-sb de

DIE ZEIT ALS PROBLEM DER PRAGMATIK

Ringvorlesung an der Universität Berlin

Im Sommersemester 1993 veranstaltete die Arbeitsstelle für Semiotik der Technischen Universität Berlin in Zusammenarbeit mit dem Institut für Linguistik der Zentraleinrichtung Moderne Sprachen eine Ringvorlesung über Die Zeit als Problem der Pragmatik. Sie gab Gelegenheit, Typen der Schlußfolgerung im Rahmen des Kommunikationsprozesses zu untersuchen, die sich auf Eigenschaften der objektiven und subjektiven Zeit stützen.

Als Beispielmaterial dienten unterschiedliche Textgattungen, die von Wissenschaftlern aus verschiedenen Disziplinen thematisiert wurden.

Der Übersetzungstheoretiker Gideon Toury (Universität Tel Aviv) eröffnete die Veranstaltungsreihe mit einer Vorlesung über "The Pragmatics of Translation: In Search of Laws of Translational Behaviour", wobei er die Zeichenprozesse beim Übergang von der Grobübersetzung zur
Feinübersetzung eines Textes thematisiert.

Der Informatiker Peter Bögh Andersen (Universität Aarhus) sprach über "Predictability and Time in Narrative", er untersuchte, inwiefern die Eingriffsmöglichkeiten in die Fortsetzung einer Geschichte, die dem Rezipienten in Computersystemen für Hypertexte zur Verfügung stehen, die zeitlichen Beziehungen zwischen den Einzelereignissen stören, und plädierte für die Entwicklung einer modifizierten Zeitlogik für solche Texte.

Der Volkskundler Vilmos Voigt (Eötvös-Lorand-Universität, Budapest) analysierte "Ancient Problem Solving: The Pragmatics of Folk Riddles" und zeigte dabei unter anderem, welche Rolle die Unumkehrbarkeit der Zeit in den überlieferten Rätseln spielt.

Die Linguistin Deidre Wilson (University College London) behandelte" The Pragmatics of Time and Tense" und erklärte die Tatsache, daß wir bestimmte Textzusammenhänge als zeitliche Reihenfolgebeziehungen konstruieren und andere nicht, mithilfe des Prinzips, daß der Empfänger einer Äußerung ihr mit möglichst geringem Verarbeitungsaufwand den größtmöglichen kognitiven Effekt zuzuschreiben versucht.

Der Slawist Boris Uspenskij (Humanistische Universität Moskau) diskutierte "The Perception of Time as a Semiotic Problem", er faßte die Zeitwahrnehmung sehr breit, indem er auch die Strukturierung der subjektiven Zeit durch den kulturellen Kontext einbezog, und unterschied -zwei gegensätzliche Zeitauffassungen, die er in der Geschichte der Kulturen der Welt nachwies: die historische und die kosmologische Auffassung.

Der Archäologe Klaus Frerichs (Museum Buxtehude) fragte danach, welche Beziehung zur Zeit uns die Gegenstände vermitteln, die uns umgeben; er differenzierte dabei zwischen "Perfektischer und futurischer Bedeutung dinglicher Artefakte am Beispiel des Museums".

Die Abschlußvorlesung hielt der Mediziner Thure von Uexküll (Universität Ulm), der über die Rolle der Zeit bei der Bildung von Selbstkonzepten in Individuen und Gruppen von Individuen sprach, was Anlaß gab zur Diskussion "Pragmatische Prozesse in der Simulation lebender Systeme". Die Reihe wird fortgesetzt.

Quelle: Zeitschrift für Semiotik, Nr. 15/3-4,1993

COMPUTER ALS MEDIUM

Workshop an der Universität Lüneburg

Vom 15. bis 17. Juli 1993 fand im Rechenzentrum- der Universität Lüneburg ein Workshop über "Computer als Medium" statt. Veranstalter waren die Fachgruppe "Computer als Medium" der Gesellschaft für Informatik und das Labor Kunst und Wissenschaft an der Universität Lüneburg. Der Workshop ist der dritte in einer Reihe von Veranstaltungen, über die an dieser Stelle bereits berichtet wurde.

Das Thema des ersten Workshops (1991) waren "Hypersysteme" , also am Computer verfaßte und auch nur mit Computern lesbare Dokumente, die aus assoziativ vernetzten Materialfragmenten in Form von Texten, Bildern, Klängen oder Videos bestehen (siehe auch das Anfang 1994 erscheinende Themenheft "Hypertext und interaktive Systeme" der *Zeitschrift für Semiotik,* in dem Referenten und Veranstalter der Workshopreihe zu Worte kommen).

Der fruchtbare Arbeitszusammenhang, der sich aus der Zusammenkunft von Menschen unterschiedlicher Disziplinen entwickelte, hatte damals Anstoß zur Gründung der Fachgruppe "Computer als Medium" in der Gesellschaft für Informatik gegeben, die es sich zur Aufgabe machte, das Zusammenwachsen von Kultur und Technik an der Stelle des medienintegrierenden Mediums Computer unter einem breiteren Blickwinkel zu beleuchten. Dieses Mal stand der Computer als Klangmedium im Mittelpunkt des Interesses.

Rolf Großmann gab einen einführenden Überblick zum Thema, Heinz W. Burow berichtete aus der Sicht des Musikproduzenten über "Möglichkeiten, Nutzen und Gefahren der digitalen Musikproduktion".

Der' Berliner Komponist und Künstler Arnold Dreyblatt stellte seine Arbeitsform vor: "Zwischen alten und neuen Medien - Musik /Komposition/ Hypertext", die Gruppe "knowbotic research" zeigte ihre mehrfach prämierte Arbeit unter dem Vortragstitel "Echtzeitkompositionen unter Verwendung von" 'öffentlichen Materialien' aus Datenbanken".

Über "Die Grenzen zwischen den Künsten" sprach Heike Staff sie traf damit programmatisch das Kernproblem der Medienintegration am Computer. Der Musikinstrumentenbau ist durch die Einführung der MIDI-Norm unter erheblichen Einfluß der Computertechnik geraten.

Fabio Biasio und Hartmut Kern zeigten "HANDWERK - Projekt virtuelle MIDI-Eingriffe" und Nick Collins von STEIM-Institut in Amsterdam zeichnete die Entwicklungslinien dieses "Exploded View the Musical Instrument in Twilight" nach und gab mit selbst entwickelten hybriden akustisch-elektronischen Instrumenten ein Solokonzert. Dieses fand anläßlich der Finissage der Ausstellung "Die präzisen Vergnügen' - Algorithmus und Kunst" mit den Pionieren der künstlerischen Computergraphik statt, die im Rahmen des internationalen Symposiums INTERFACE II im Februar 1993 in Hamburg zum erstenmal gezeigt wurde und dem Andenken des Informationsästhetikers und Semiotikers Max Bense gewidmet war.

Matthias Lehnhard zeigte eine interaktive Installation und thematisierte damit zusammenhängende Überlegungen unter dem Titel" 'Living Rooms' - Bedeutung, digitaler Kode und Transformation".

Der Mathematiker und Musiktheoretiker Rudolf Wille stellte sein Computersystem "MUTABOR - ein computergesteuertes Musikinstrument zum Experimentieren mit Stimmungslogiken und Mikrotönen " vor, und Bernd Schmeikal zeigte anhand von "BOROTO - morphogenetische Klangbilder" Zusammenhänge zwischen Anthropologie, Chaos-Systemen und Sprachentwicklung auf.

Das Virtuelle Medienzentrum Hamburg, Antje Eske und Nicola Nissen sowie Volker Lettkemann und Werner Justen pflegten dialogische Datenkünste: "Computer als Medium und Meinungsbildung", "Soloparts mit chorischen Anteilen" und "getting love in hyperspaiß".

Neben dem Schwerpunkt Klang gab es noch eine kleine Abteilung zum Thema Bild: Uwe Pirr sprach über "Räumliche Darstellung mit dem Computer", Günther Görz berichtete vom Projekt "Der elektronische Behaim Globus" und Thomas Hölscher startete eine Tour d'horizon zu "Kunst, Natur, Künstlichkeit - Zur Künstlichkeit der Natur in der westlichen Kunst".

Interessenten am Workshop-Material und an der Arbeit der Fachgruppe oder des "Labor Kunst und Wissenschaft" können sich wenden an: Dr. Martin Warnke, Universität Lüneburg, D-21332 Lüneburg.

Martin Warnke, Universität Lüneburg

Quelle: Zeitschrift für Semiotik, Nr. 15/3-4, 1993

PHILOSOPHIE UND DIE KOGNITIVEN WISSENSCHAFTEN

16. Internationales Wittgenstein Symposium 1993

Seit nunmehr fast zwei Jahrzehnten treffen sich alljährlich im August in Kirchberg am Wechsel (Niederösterreich) Philosophen und andere Wissenschaftler, die direkt oder indirekt mit Forschungen zu Ludwig Wittgenstein befaßt sind.

Das diesjährige 16. Internationale Wittgenstein Symposium, das vom 15. bis 22. August 1993 stattfand, hatte das Thema Philosophie und die kognitiven Wissenschaften.

Mit dieser brisanten Thematik war eigentlich schon vorprogrammiert, daß der Kongreß auf großes Interesse der internationalen Wissenschaftlergemeinschaft stoßen wird. Und so war es dann auch keinesfalls überraschend, daß etwa 1000 Wissenschaftler aus allen Erdteilen anreisten.

Aber nicht nur hinsichtlich der großen Teilnehmerzahl unterschied sich dieses Symposium von den vorhergehenden, sondern auch in bezug auf die Vorbereitung der Konferenz. Dazu bemerkte der Präsident der Österreichischen Ludwig-Wittgenstein-Gesellschaft, Rudolf Haller (Graz), in seiner Eröffnungsrede, daß dieses Symposium gewissermaßen einen Generationswechsel vollzog, insofern die Organisation in die Hand eines neuen Organisationsstabes gelegt wurde. In diesem Jahr hatte B. Smith (Schaan) die Betreuung der Konferenz übernommen.

Während des einwöchigen Kongresses wurden 250 Vorträge gehalten. Eine Vielzahl der Beiträge war in einem Pre-Text-Reader gedruckt erhältlich.

Unübersehbar wurde während der gesamten Konferenz deutlich, daß die Forschungen in den kognitiven Wissenschaften zu den aufsehenerregendsten Entwicklungen in der internationalen Wissenschaftslandschaft der letzten drei Jahrzehnte gehören. Anliegen der kognitiven Wissenschaften (wie der kognitiven Psychologie, der kognitiven Linguistik, Computerwissenschaft, Forschungen zur Künstlichen Intelligenz, der kognitiven Semiotik u.a.) ist es, die natürliche Intelligenz biologischer Organismen, insbesondere von Menschen, und die künstliche Intelligenz menschengemachter Maschinen, darunter vor allem von Computern, zu erforschen.

Die in den kognitiven Wissenschaften erzielten Ergebnisse dienen vor allem dem Ziel, die Funktionsweise des menschlichen Geistes besser zu verstehen. Zur Realisierung der Aufgabenstellung der kognitiven Wissenschaften werden unterschiedliche Modelle angeboten. Erwähnt seien hier u. a. Komputationalismus, der Konnektionismus und die Ermergenztheorie.

Verständlicherweise war der Kongreß ein Spiegelbild der miteinander konkurrierenden Theorien. Dies war nun keinesfalls ein Nachteil, sondern belebte die Diskussion. Die Plenarveranstaltungen waren folgenden Themen gewidmet:

1. "Computation and Cognition: Some Important Differences" (J. Searie, Berkeley) ,
2. "Wittgenstein's Conception of Philosophy as Grammar" (N. Garver , Buffalo) ,
3. "Modes of Perceptual Representation"(F. Dretske, Stanford),
4. "Finding the Mind in the Natural World"(F. Jackson, Canberra) und

5. "Thinking with a Word Processor"(J. C. Nyiri, Budapest).

Die Hauptvorträge machen ebenfalls die Vielfalt der in den kognitiven Wissenschaften diskutierten Probleme deutlich.

Zu folgenden Themen wurden Referate gehalten:
"How Grammar Structures Concepts" (L. Talmy, Buffalo), "If God had Looked into Dur Minds he Would not Have Been Able to See There which Logical Operations We are Performing"(J. Hintikka, Boston), "Can a machine Follow a Rule?" (J. Haugeland, Pittsburgh), "Creativity and Representational Redescription" (M. Boden, Sussex), "A New Theory of Content: Solving Frege's Puzzle" (G. Bealer, Boulder), "Semantic Localism: Who Needs a principled Basis?" (M. Devitt, Maryland), "Automated Reasoning and Artificial Intelligence" (N. Tennant, Ohio), "Phenomenology of Perception, Qualitative Physics, and Sheaf Mereology" (J. Petitot, Paris), "Mental Causation and Two Conceptions of Mental Properties" (J. Kim, Brown), "Predication in Natural Language and Formal Logic"(H. Kamp, Stuttgart), "Processing Models for Non-Literal Discourse" (F. Recanati, Paris), "Interpretation as Abduction" (H. Hobbs, SRI-Stanford), "The Role of Contexts in Logic"(J. Sowa, Binghampton), "Cognitive Contents and Cognitive Connections"(H. Hochberg, Austin). Die weiteren Vorträge waren folgenden Themenkreisen zugeordnet:

1. Sprache und Kognition,
2. Methodologien der kognitiven Wissenschaften,
3. Folk Psychology und Naive Physik,
4. Wahrnehmungstheorien,
5. Künstliche Intelligenz,
6. Historische Wurzeln der Kognitionswissenschaft und
7. Wittgenstein und die philosophische Psychologie.

Ergänzt wurden die auf dieser Konferenz gehaltenen Vorträge durch Softwaredemonstrationen von P. M. Simons (Salzburg) und D. W. Smith (Irvine) auf der einen Seite und J. Zeiger (Innsbruck) auf der anderen Seite.

Außerdem war eine eigene Sektion Editionsfragen gewidmet, in der eine Einführung zur "Wiener Ausgabe" der Wittgensteinschriften unter der Leitung von Nedo (Cambridge) gegeben wurde. Diese neue in 37 Bände beim Springer-Verlag, Heidelberg, erscheinende Ausgabe wird von Philosophen und anderen Wissenschaftlern mit Spannung erwartet, wird sie doch weiteren Zündstoff für die Wittgenstein-Interpretation liefern und die nächsten Kirchberger Symposien reichlich mit *Ge*sprächsstoff versorgen.

Evelyn Dölling, Technische Universität Berlin

Quelle: Zeitschrift für Semiotik, Nr. 15/3-4, 1993

VERANSTALTUNGEN

Lexikon und Morphologie

2. GLDV-Herbstschule vom 19. bis 24. September 1994 an der Universität Leipzig, Institut für Informatik

Die 2. GLDV-Herbstschule 94 "Lexikon und Morphologie" wendet sich an Student/inn/en und Doktorand/inn/en, insbesondere der Computerlinguistik, Linguistik, Informatik und verwandter Disziplinen, aber auch an Interessierte aus der beruflichen Praxis.

Das Kursangebot beinhaltet Vorlesungen und praktische Übungen (zum Teil am Rechner) zu folgenden Themenbereichen:

Maschinenlesbare Lexika Automatische Wortformerkennung
und morphologische Algorithmen
MLexD: Ein Standard für die korpusbasierte Lexikographie Produkte der Sprachtechnologie Transfer-Lexika in der Maschinellen Übersetzung
Softwareergonomie und natürliche Sprache

Das Kursangebot

Von den sechs angebotenen Kursen kann jeder Teilnehmer bis zu drei belegen. Da die Veranstaltungen in den Blöcken A, B, C jeweils parallel stattfinden, kann je Block nur höchstens ein Kurs gewählt werden.

Angeboten werden die folgenden Alternativen:

Block A: "Maschinenlesbare Lexika" *oder* "MLexD: Ein Standard für die korpusbasierte Lexikographie"

Block B: "Automatische Wortformerkennung und morphologische Algorithmen" *oder* "Softwareergonomie und natürliche Sprache"

Block C: "Transfer-Lexika in der Maschinellen Übersetzung" *oder* "Produkte der Sprachtechnologie"

Das Rahmenprogramm

Bei der Eröffnungsveranstaltung am Montagvormittag referiert Prof. Lenders (Bonn) über „20 Jahre GLDV-Absichten, Aufgaben, Positionen" und Frau Dr. Walther (Leipzig) über "Zur Stellung der Morphologie in der generativen Grammatik".

Den ersten Abendvortrag am Montagabend hält Prof. Schnelle (Bochum) über "Wortvernetzungen in Computer und Gehirn", den zweiten am Donnerstagabend hält Prof. Bierwisch (Berlin) über "Universalien und Idiosynkrasien im Lexikon".

Am Dienstagabend findet eine Präsentation der Sieger der ersten Morpholympics 1994 statt.

Kurse

Maschinenlesbare Lexika (mit einer Einführung in SGML *[Winfried Lenders / M. Volk]*

Der Kurs bietet in seinem ersten Teil eine Einführung in die Struktur, Verwendungsformen und Konstruktionsmethoden maschinen-lesbarer und maschinenverwendbarer Wörterbücher *(machine rea-*

dable und *machine tractable dictionaries).* Es wird auf folgende Themen auch anhand von Beispielen eingegangen:

. Wörterbücher, Lexika und lexikalisches Wissen (Definitionen)

. Gebrauchswörterbücher Computerterwörterbücher

. Lexikalische Mikrostrukturen und Makrostrukturen

. Wörterbücher in linguistischen Modellen

. Wiederverwertbarkeit lexikalischer Resourcen

. Das Wörterbuch als Netzwerk

. Lexikalische Strukturen und Relationen

. Tools zum Wörterbuchaufbau

. Lexikalische Datenbanken

Im zweiten Teil des Kurses wird, ausgehend vom Beispiel der Standardisierung von Wörterbuchkodierungen, eine Einführung in die heute international allgemein anerkannte Standardisierungssprache SGML (Standard Generalized Markup Language) gegeben. Es wird gezeigt, wie SGML als Grammatik zur Strukturierung von Dokumenten gesehen werden kann.

MLexD: Ein Standard für die korpusbasierte Lexikographie
(Wolf Papprotti)

Der Kurs gibt einen Überblick über die Architektur ein- und zweisprachiger elektronischer Lexika sowie die morphosyntaktische Information, die ein Lexikon abbilden sollte. Die Adäquatheit des Standards hinsichtlich der Sprachen Englisch, Deutsch und Neugriechisch wird diskutiert. Es ist beabsichtigt, Methoden und Werkzeuge für die Extraktion lexikaler Information aus Korpora vorzustellen.

Automatische Wortform-Erkennung und morphologische Algorithmen
(Roland Hausser)

Die linguistische Analyse der Wortformen und Wörter fällt in den Aufgabenbereich der Grammatikkomponente 'Morphologie'. Die traditionell sprachwissenschaftlichen Konzepte der morphologischen Analyse sind die theoretische Grundlage für eine effiziente und linguistisch wohlmotivierte Programmierung der automatischen Wortformanalyse im Rahmen der Computerlinguistik. Die Wortformerkennung wiederum ist die Voraussetzung für praktisch alle weiteren computerlinguistischen Anwendungen wie syntaktisches Parsen, semantische Analyse, Indexierung etc.

Der Kurs gliedert sich in einen theoretischen und einen praktischen Teil. Der theoretische Teil vermittelt die Grundbegriffe der morphologischen Analyse, beschreibt unterschiedliche Methoden der automatischen Wortformerkennung und vergleicht ihre programmiertechnische Eignung. Der praktische Teil vermittelt den Umgang mit dem Softwaresystem LA-MORPH, das bei allgemeinen Texten, z.B. Zeitungsartikeln oder Bundestagsreden, etwa 95% der Wortformen erkennt und auf einem PC 486 / 33 unter Linux eine durchschnittliche Geschwindigkeit von etwa 400 Wortformen pro Sekunde entwickelt.

Softwareergonomie und natürliche Sprache
(Jürgen Krause)

Bei der Mensch-Maschine Interaktion ist die natürliche Sprache nur eine der beiden Hauptmodi, mit dem Computer "natürlich" zu interagieren. Eine zweite - und in den letzten Jahren wesentlich erfolgreichere Alternative sind die grafischen Benutzeroberflächen. Deshalb befaßt sich der Kurs mit beiden Varianten unter dem gemeinsamen Blickwinkel einer softwareergonomischen Gestaltung von Benutzeroberflächen.

Themen sind:

a) Was ist Ergonomie und was Softwareergonomie ?

b) Theoretische Grundlagen "natürlicher" Oberflächen von Computersystemen

c) Grafische Oberflächen auf der Basis von Direktmanipulation und Schreibtischmetapher

d) Natürlichsprachliche Frage- Antwort-Systeme: Grundlagen und kommerzielle Situation heute

e) Computertalk

f) Multimodale Mischformen als Überlebensstrategie für den natürlichsprachlichen Modus bei der Mensch-Maschine-Interaktion (WOB-Modell)

Der Kurs hat nichts mit dem Hauptthema "Lexikon und Morphologie" zu tun. Er ist als fakultive, inhaltliche Erweiterung gedacht, für alle diejenigen, die etwas über den engen Zaun einer an der formalen Linguistik orientierten Computerlinguistik hinausschauen wollen.

Transferlexika in der Maschinellen Übersetzung

[Johann Haller]

. Grundbegriffe der Maschinellen Übersetzung: Analyse, Transfer, Synthese

. Transferkonzepte: direkt, indirekt, Interlingua

. Transfer in älteren Systemen: SYSTRAN, LOGOS, METAL

. Transfer in Forschungssystemen (EU ROTRA, CAT2)

. lexikalischer und struktureller Transfer

. Beispiele von Einträgen in Transferlexika

Die Teilnehmer erarbeiten und erproben selbständig neue Transfereinträge im CAT2-System.

Produkte der Sprachtechnologie

[Gerhard Heyer / Klemens Waldhör]

Im Unterschied zur Computerlinguistik steht in der Sprachprodukttechnologie weniger die Simulation kognitiver Prozesse im Vordergrund, als vielmehr das Ziel, Verfahren und Methoden für eine ingenieursmäßige Entwicklung von Sprachprodukten bereitzustellen. Der Kurs gibt einen Überblick über die wichtigsten Grundlagen, Probleme und Methoden der Sprachprodukttechnologie. Im einzelnen werden behandelt:

1. Das ingenieurswissenschaftliche Paradigma als Grundlage der Sprachprodukttechnologie

2. Überblick und Klassifikation von Sprachprodukten

3. Softwareengineering für Sprachprodukte (insb. elektronische Wörterbücher)

4. Ergonomische Aspekte von Sprachprodukten (insb. elektronische Wörterbücher)

5. Linguistik für Sprachprodukte

Die Herbstschule findet in den Räumen der Universität Leipzig statt.

Hinweise zur Anmeldung:

Die Gebühren für Kursteilnahme und -unterlagen betragen

für Student/inn/en
(mit Immatrikulationsbescheinigung): bis zum 1.7.94. 65,- DM
nach dem 1.7.94 90,- DM für sonstige Teilnehmer/innen:

bis zum 1.7.94 130,- DM
nach dem 1.7.94 180,- DM

Auskunft und Aumeldungsunterlagen über:

2. GLDV-Herbstschule 1994
c/o Prof. Dr. G. Heyer
Universität Leipzig
Institut für Informatik / Abt. ASV
Augustusplatz 10/11
04109 Leipzig
Tel.: 0341 - 7192188 Fax: 0341 - 7192399 e-mail: heyer@informatik.uni-Ieipzig.de

Europäische IR Sommerschule 03.-08
September 1995, Ilmenau/Thüringen

Lokale Organisation: Vorsitz R. Schramm, TH Ilmenau
Programmkommittee: Initiative: Norbert Fuhr, Dortmund
Fachgruppe Information Retrieval der GI
Vorläufiges Programm ab Nov. 94

Hypertext-Information Retrieval-Multimedia '95, HIM '95

05.-07. April, Konstanz

Anforderungen an moderne Informationssysteme lassen aus System- und Nutzersicht eine isolierte Betrachtung von Problemen häufig nicht mehr zu. Die Fachgruppen Hypertext (4.9.1), Information Retrieval (2.5.4 / 4.9.3) und Multimedia (4.9.2) der Gesellschaft für Informatik (GI) führen daher gemeinsam mit der Österreichischen Computer Gesellschaft (ÖCG), der Schweizer Informatik Gesellschaft (SI) und dem Hochschulverband Informationswissenschaft (HI) eine Tagung mit den Schwerpunkten Hypertext, Information Retrieval und Multimedia durch. Die Informationswissenschaft an der Universität Konstanz ist vom 05.04.95 bis 07.04.95 Gastgeber der Tagung. Sowohl fachspezifische als auch fachübergreifende Beiträge in der Form von Vorträgen, Postern und Tutorien sind für alle Teilgebiete und Aspekte von Hypertext, Information Retrieval und Multimedia willkommen. Zu Systemdemonstrationen während der und ergänzend zu den Vorträgen wird ausdrücklich aufgefordert.

Beiträge sind bis spätestens 01.11.94 einzureichen. Tagungssprachen sind Deutsch und Englisch.

Für weitere Informationen oder eventuelle Fragen, wenden sie sich bitte an die Organisationsleitung:

Marc Rittberger
Universität Konstanz
Informationswissenschaft
Postfach 5560
D- 78434 Konstanz
Tel: 49-7531-883595
Fax: 49-7531-882601
EMAIL: ritt@inf-wiss.uni-konstanz.de

TAGUNGSKALENDER

05.07.–30.07.1994 Buffalo, USA
1st International Summer Institute in Cognitive Science
Information: International Institute of Cognitive Science, Center for Cognitive Studies, SUNY, Buffalo, NY 14260, USA.

07.07.–17.07.1994 Urbino, Italien
Semiotisches Sommerinstitut
Information: Centro Internazionale di Semiotica e di Linguistica, Piazza del Rinascimento 7, I–61029 Urbino.

01.08.–06.08.1994 St. Petersburg, GUS
East-West Conference on Human–Computer Interaction
Information: Claus Unger, Praktische Informatik II, Fernuniversität Hagen, Feithstraße 140, D–58084 Hagen.

08.08.–12.08.1994 Amsterdam, Niederlande 11th European Conference on Artificial Intelligence **EOAI 94**
Information: Erasmus Forum, P. O. Box 1738, 3000 DR Rotterdam, Niederlande, Tel.: +31 10 408 2302, FAX: +31 10 453 0784,
e-mail: M.M.deLeeuw@apv.oos.eur.nl

15.08.–20.08.1994 Edmonton, USA
14th Congress of the International Comparative Literature Association
Information: Willie van Peer, Department of Comparative Literature, Stanford University, Stanford, California 94305–2087, USA.

24.08.–27.08.1994 Budapest, Ungarn
4. Internationale Tagung der Gesellschaft für Empirische Literaturwissenschaft
Information: László Halász, Institute for Psychology, P. O. Box 398, H–1394 Budapest.

25.08.–27.08.1994 Turku, Finnland
International Conference on Interpretation
Information: Yves Gambier, Department of Translation Studies, University of Turku, Tykistökatu 4, SF–20520 Turku.

31.08.–01.09.1994 Hamburg, BRD
Fachgespräch über Computer und Künste
Information: Martin Warnke, Rechenzentrum, Universität Lüneburg, Stresemannstraße 6, D–21335 Lüneburg.

07.09.–10.09.1994 Leipzig, BRD
Tagung der Gesellschaft für Analytische Philosophie
Information: Peter Steinacker, Bereich Logik und Wissenschaftstheorie, Universität Leipzig, Postfach 920, D–04009 Leipzig.

12.09.–13.09.1994 Tours, Frankreich
Secondes Rencontres de la Société Francophone de Classification
Information: J. Minet, Laboratoire d'Informatique de l'École d'Ingénieurs en Informatique pour l'Industrie (E31), Université Francois Rabelais, 64, avenue j. Portalis, Technopôle de Tours, Boîte no. 4, F–37913 Tours Cedex 9.

13.09.–15.09.1994 Münster, BRD
Internationale Arbeitstagung zur lexikalischen Semantik
Information: Edda Weigand, Zentrum für Sprachforschung, Bispinghof 3A, D–48149 Münster.

20.09.–24.09.1994 Moskau, GUS
QUALICO – 94. Second International Conference on Quantitative Linguistics
Information: Prof. Reinhard Koehler, Institut für Computerlinguistik, Universität Trier, D–54286 Trier, Tel.: (0651) 201–2270.

21.09.–23.09.1994 Trier, BRD
25. Jahrestagung der Gesellschaft für Angewandte Linguistik **(GAL)**, Thema: Dialogische Formen der Informationsverarbeitung
Information: Prof. Dr. Annely Rothkegel, Fachhochschule Hannover, BID / Studiengang Technische Redaktion, Berhard–Caspar–Str. 7, 30453 Hannover, Tel.: (0511) 212 39 30, FAX: (0511) 210 30 00, e–mail: email@rothkegel coli.uni–sb.de

21.09.–24.09.1994 Trier, BRD
25. Jahrestagung der Gesellschaft für Angewandte Linguistic **(GAL)**, Thema: Sprache — Verstehen und Verständlichkeit
Information: Hartwig Kalverkämper, Auf der Feldkirmes 15, D–57439 Attendorn.

28.09.–30.09.1994 Wien, Österreich
2. Konferenz „Verarbeitung natürlicher Sprache" **KONVENS 94**

Information: Frau Mag. Gerda Helscher, Österreichische Gesellschaft für Artificial Intelligence, Postfach 177, A–1014 Wien, Tel.: +43 1 535 32 81 0, FAX: +43 1 535 06 52, e-mail: sec@ai.univie.ac.at

07.10.–09.10.1994 Blaubeuren, BRD
Symposium über Zeichen Lesen — Lese–zeichen, Thema: Leseweisen in China und Deutschland — Ein kultursemiotischer Vergleich
Information: Jürgen Wertheimer und Susanne Göße, Deutsches Seminar (Komparatistik), Eberhard–Karls–Universität, Wilhelmstraße 50, D–72074 Tübingen.

12.10.–14.10.1994 Dresden, BRD
Hypermedia in der Aus- und Weiterbildung, Dresdner Symposion zum computerunterstützten Lernen
Information: Eric Schoop, TU Dresden, Fakultät Wirtschaftswissenschaften, Lehrstuhl Wirtschaftsinformatik, insbes. Informationsmanagement, Mommsenstraße 13, 01062 Dresden, Tel.: (0351) 463–2845, FAX: (0351) 471–6660, e-mail: schoop@rmhs1.urz.tu-dresden.de

18.11.–20.11.1994 Paderborn, BRD
Paderborner Novembertreffen zu Sprachkybernetik und Interlinguistik
Information: Helmar Frank, Institut für Kybernetik, FB 2, Universität Paderborn , D–33095 Paderborn.

21.11.–22.11.1994 Vienna, Österreich
International MicroISIS Conference
Multilingual Information Management with MicroISIS
Information: Dr. Gerhard Budin, Infoterm, Heinestraße 38, 1020 Vienna, FAX: +43 1 216 3272, Tel.: +43 1 26 75 35 310

01.03.–03.03.1995 Kaiserslautern, BRD
3. Deutsche Expertensystemtagung **XPS–95**
Information: Bernd Bachmann, DFKI, Postfach 2080, D–67608 Kaiserslautern, Tel.: (0631) 205-3482, FAX: (0631) 205-3210, e-mail: bachmann@dfki.uni-kl.de

19.04.–22.04.1995 Maastricht, Niederlande
Maastricht–Lodz–Kolloquium über Übersetzung und Bedeutung
Information: Marcel Thelen, Rijkshogeschool Maastricht, Postbus 964, NL–6200 AZ Maastricht.

06.07.–16.07.1995 Urbino, Italien
Semiotisches Sommerinstitut
Information: Centro Internazionale di Semiotica e di Linguistica, Piazza del Rinascimento 7, I–61029 Urbino.

29.08.–02.09.1995 Vienna, Österreich
LSP 95 10th European Symposium on Languages for Special Purposes, Multilingualism in Specialist Communication
Information: Gerhard Budin, University of Vienna, Infoterm, Tel.: +43 1 26 75 35 310, FAX: +43 1 216 32 72

20.09.–22.09.1995 Goslar, BRD
Herbstakademie der Deutschen Gesellschaft für Semiotik, Thema: Verkehr und Verkehrsformen
Information: Ingrid Lempp, Bahnhofstr. 2, D–25497 Priesdorf.

22.09.–24.09.1995 Lodz, Polen
Maastricht–Lodz–Kolloquium über Übersetzung und Bedeutung
Information: Barbara Lewandowska–Tomaszczyk, Institut für Anglistik, Universität Lodz, Al. Kosciuszki 65, PL–90514 Lodz.

04.08.–09.08.1996 Jyväskylä, Finnland
11th World Congress of the International Association of Applied Linguistics
Information: K. Sajavaara, Department of English, University of Jyväskylä, SF–40100 Jyväskylä.

März 1996 Kobe, Japan 5th Conference of the International Federation of Classification Societies **(IFCS)**
Information: (wird rechtzeitig an GfKI–Mitglieder versandt).

ARBEITSKREISE DER GLDV

Stand der Korpustechnologie in Deutschland und internationale Entwicklungen

Arbeitskreistreffen KORPORA der Gesellschaft für Linguistische Datenverarbeitung (GLDV) im Februar'94 am IDS, Mannheim.

Am 11. Februar 1994 tagte der Arbeitskreis "Korpora" der Gesellschaft für Linguistische Datenverarbeitung, der von Robert Neumann geleitet wird, im Institut für deutsche Sprache, Mannheim, und beschäftigte sich mit der Thematik "Stand der Korpustechnologie in Deutschland und internationale Entwicklungen - Projekte der Europäischen Union".

Rainer Wimmer, Geschäftsführender Direktor des Instituts für deutsche Sprache, begrüßte die Teilnehmer des Treffens und hob die Wichtigkeit der Korpustechnologie für die germanistische Linguistik hervor.

Robert Neumann eröffnete die eintägige Zusammenkunft und umriß das Ziel der aktuellen Sitzung des Arbeitskreises: einen Überblick über die deutschen und internationalen Korpus-Aktivitäten zu gewinnen und die gesammelten Erkenntnisse in Form mehrerer bibliographischer Zusammenstellungen zu Korpus-Arbeiten in Deutsch in einen Listserver zu geben und allen Interessenten so einen Zugang zu gewähren.

In seinem Eröffnungsvortrag stellte R. Neumann das Institut für deutsche Sprache als zentrale außeruniversitäre Forschungseinrichtung vor und sprach über die Aufgaben der Arbeitsstelle "Linguistische Datenverarbeitung" im IDS.

Die Vortragsreihe eröffnete Wolf Paprotté mit der Vorstellung des Projekts 62050 MULTEXT (Multilingual Text Tools and Corpora). Die Diskussion dazu berührte vor allem Fragen der Textauszeichnung (I. Batori), des morphologischen' Taggings (A. Storrer), die Definition von Speech-Dateien und Standardisierungsprobleme (R. Neumann).

Wolfgang Teubert berichtete über das Projekt "Network of European Reference Corpora" (NERC), einen Vorschlag für die Erarbeitung nationaler Korpora.

Angelika Storrer und Helmut Feldweg sprachen über ein System zur Wissensrepräsentation mit der Bezeichnung ELWIS (Korpusunterstützte Entwicklung lexikalischer Wissensbasen).

R. Neumann stellte als drittes Projekt in der Europäischen Union "Multilingual Environment for Corpus-Based Lexicon Building" (MECOLB) vor, das für die Abspeicherung und Verwaltung von großen Textmengen geeignete Werkzeuge entwickelt, die den Aufbau und die Wartung von maschinenlesbaren Lexika unterstützen.

In der anschließenden Diskussion gab R. Neumann bekannt, daß die Arbeitsstelle "Linguistische Datenverarbeitung" des IDS einen Listserver für Korpus-Fragen eingerichtet hat, der ein permanentes Informationsbrett und ein Diskussionsforum der Mitglieder des Arbeitskreises sein soll. Die Bibliographie zum NERC-Projekt soll über den Listserver allen Mitgliedern des Arbeitskreises zur Verfügung gestellt werden, und W. Teubert stimmt diesem Verfahren als Urheber der Bibliographie zu.

Winfried Lenders, Vorsitzender der GLDV, sagte, daß neben dem Informationsaustausch über den Listserver die persönliche Kommunikation der Experten

sehr wichtig sei und daß der Arbeitskreis sich zweimal im Jahr zu Tagungen zusammenfinden sollte. Weiterhin sollen die Mitglieder des Arbeitskreises und des Korpus-Workshops miteinander korrespondieren, und es sollten regelmäßig Informationen über die Projekte und über erzielte Ergebnisse und angewandte Methoden ausgetauscht werden. Materialien, die alle Mitglieder interessieren, sollten über die elektronischen Medien zur Verfügung gestellt werden; die verschiedenartigsten Aktivitäten sollten weitergeleitet werden und darüber regelmäßig in der Zeitschrift "LDV-Forum" berichtet werden.

R. Neumann stellte - die Diskussion zusammenfassend - fest, daß vor allem Clearing-Aufgaben zu den wichtigsten Gegenständen der Arbeit des Arbeitskreises gehören, und schlug eine systematische Aufbereitung von Informationen über die deutschen Textkorpora vor, denn alle deutschsprachigen Korpora sollten zumindest im nationalen Rahmen aufgelistet greifbar sein. Die Mitglieder des Arbeitskreises werden aufgerufen, maschinenlesbare Texte - auch phonetische Korpora verfügbar zu machen und an den Listserver zu melden, so daß eine Liste aller vorhandenen Texte erarbeitet werden kann.

Cyril Belica führte zum Abschluß der Veranstaltung das Computersystem COSMAS (Corpus Storage, Maintenance and Access .System) vor, mit dessen Hilfe Textbelege für einzelne Wortformen und Satzzeichen, für Wortformen in verschiedenen Kombinationen und mit unterschiedlichen Abständen gesucht werden können. COSMAS wurde vom Institut für deutsche Sprache in Zusammenarbeit mit der Firma Makrolog, Wiesbaden, entwickelt, um für sprachwissenschaftliche und andere sprachbezogene Forschungsvorhaben die umfangreichen Textkorpora zur geschriebenen und gesprochenen deutschen Sprache zu erschließen und zu pflegen.

Das nächste Treffen des Arbeitskreises "Korpora" wird am 11. Juni 1994 an der Westfälischen Wilhelms-Universität Münster zum Thema "Aktivitäten zu Korpora der gesprochenen Sprache in der Bundesrepublik Deutschland: VERBMOBIL" stattfinden.

Irmtraud Jüttner, Mannheim

AK - Lexikographie

Der Arbeitskreis Lexikographie möchte in den bevorstehenden Monaten mit folgenden Aktivitäten innerhalb der GLDV in Erscheinung treten:

1. Erstellung einer möglichst vollständigen Bibliographie zur maschinellen Lexikographie *(computationallexicography)* in Deutschland. die Bibliographie soll zentral in Bonn gesammelt werden. Voraussetzung für das Gelingen ist jedoch eine aktive Teilnahme möglichst zahlreicher Beitragender - vornehmlich durch die Bereitschaft, Literaturangaben maschinenlesbar in einem verabredeten Format zur Verfügung zu stellen.

2. Bestandsaufnahme und Bekanntmachung von maschinelesbaren lexikalischen Ressourcen, die für Forschungszwecke verfügbar gemacht werden können. Eine kleine Dokumentation soll Auskunft darüber geben, wo welche Datei wie und unter welchen Bedingungen für Forschungszwecke genutzt werden können.

Dr. Nico Weber (Sprecher) Universität
Bonn
IKP
Poppelsdorfer Allee 47
53115 Bonn
Tel.: 0228/73 56 44
Fax: 0228/73 56 39

G LDV -Mitgliederversammlung

Donnerstag, den 29. September 1994, 16.00 Uhr Universität Wien, Juridicum (Der Sitzungsraum im Juridicum wird bei Tagungsbeginn mitgeteilt werden)

Tagesordnung:

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

Die Beantragung weiterer TOPs durch Mitglieder gern. §17 muß bis zum 21.09.1994 beim Vorstand erfolgt sein.

Zu TOP 9
Die Satzung unserer Gesellschaft muß in der Ihnen zugegangenen fortlaufenden Form als Ganzes verabschiedet werden, nachdem sämtliche bei früheren Jahrestagungen beschlossenen Satzungsänderungen eingearbeitet worden sind. Die als Ganzes zu beschließende Satzung enthält also keinerlei Änderungen gegenüber früheren Satzungen, über die nicht bereits im einzelnen abgestimmt worden wäre. Es ist lediglich aus rechtlichen Gründen erforderlich, die Satzung in ihrer, fortlaufenden Form noch einmal abzustimmen.

Zu TOP 10
Diskussion einer möglichen Satzungsänderung
Der Vorstand möchte der Mitgliederversammlung den Vorschlag unterbreiten, in die Satzung die Möglichkeit einer *Berufung* von Beiratsmitgliedern einzubauen, doch so, daß ein Teil der Beiratsmitglieder weiterhin gewählt wird. Der Hintergrund dieses Vorschlages ist der, daß es sich in der Vergangenheit als äußerst ungünstig, um nicht zu sagen schädlich, erwiesen hat, daß vorn Vorstand aus bestimmten, z.B. wissenschaftspolitischen Gründen, für den Beirat vorgeschlagene Personen nicht gewählt worden sind. Ferner möchte der Vorstand vorschlagen, die Leiter der Arbeitskreise in der Satzung als geborene Mitglieder des Beirats zu verankern.

Im Namen des gesamten Vorstandes bitte ich die Mitglieder der GLDV möglichst zahlreich an unserer jährlichen Mitgliederversammlung teilzunehmen.

W. Lenders

Postrückläufer

Leider haben wir wieder einige Postrückläufer von Mitgliedern, die ihre gültigen Anschriften dem GLDV-Vorstand nicht rechtzeitig mitgeteilt haben. Wer die neue Anschrift eines der nachfolgend aufgeführten GLDV-Mitglieder kennt, möge sie bitte dem Vorstand bekanntgeben:

Bauer, Gabi, M. A., Neuprüll 3 App. 183, 93051 Regensburg

Daiber, Jürgen, Martinskloster, Zimmer-Nr. 214, Martinsufer 1, 54292 Trier

Hasenknopf-Reknes, Adelheid, Johann Seb. Bach-Str. 22, 82140 Olching Kahre,

Anette, Dr. phil., Kattenstrohterweg 137, 33332 Gütersloh Krewer, Monika, Am

Trimmelter Hof 95, 54296 Trier

Kunz, Daniela, Birkenwaldstr. 91, 70191 Stuttgart

Müller, Sonja, Karlsluststr. 4, 69126 Heidelberg

Stallwitz, Gabriele, Triumph-Adler AG, Abt. EF 11, Fürther Str. 212, 90429 Nürnberg

Wenzel, Petra, Dahlbergstr. 2, 93049 Regensburg