I. Computational Linguistics I: Development
Computerlinguistik I: Entwicklungslinien

1. Die allgemeine Entwicklung der Computerlinguistik

1. Die Betrachtungsebenen der Computerlinguistik

Die Entwicklung einer wissenschaftlichen Disziplin kann auf drei Ebenen verfolgt werden: 1. auf der epistemologischen, 2. der organisatorisch-förderungspolitischen und 3. der curricularen Ebene.

Die epistemologische Betrachtungsebene ist fundamental; demnach wird unter *Computerlinguistik* (= CL) ein interdisziplinäres Wissensgebiet zwischen Linguistik und Informatik verstanden, das linguistische Erkenntnisinteressen verfolgt und komplettiert, vor allem im Bereich der Sprachanwendungsprozesse und der maschinellen Simulation dieser Prozesse. Sie strebt eine explizite, (deskriptive oder algorithmisch-prozedurale) Sprachbeschreibung an. Eingebettet in einem umfassenden Wissensgebäude sind linguistische und computerlinguistische Erkenntnisse nicht notwendigerweise distinkt, Unterschiede bestehen im Bereich der Methode, also in der Gewinnung der Erkenntnisse.

Im vorliegenden Handbuch wird die epistemologische Betrachtungsweise befolgt.

Es erwies sich jedoch als unumgänglich, die Grundrisse der organisatorischen Dimension der CL auch darzustellen (vor allem Art. 2 und z. T. Art. 3), bzw. gelegentlich darauf hinzuweisen. Während eine Grenzziehung zwischen Disziplinen epistemologisch nicht notwendig, u. U. sogar hinderlich ist, kann es aus praktischen, organisatorischen und förderungspolitischen Gründen trotz potentieller Überlagerungen ihrer Erkenntnisse zweckmäßig sein, Linguistik und CL zu trennen, und die besondere Problematik der CL, (Leistungen, Methoden, Probleme usw.) getrennt darzustellen. Es ist möglich, daß eine und dieselbe Sache in der aktuellen Projektumgebung ganz anders exponiert wird als ihre Manifestation auf der epistemologischen Ebene. Die Forderung nach Interdisziplinarität läßt sich beispielsweise auf der epistemologischen Ebene als Doppelkompetenz begreifen, auf der organisatorischen Ebene erweist es sich mehr praktikabel, sie für Projektgruppen zu fordern und nicht von jedem einzelnen Mitarbeiter individuell zwei Diplome zu verlangen.

In der Entwicklung der CL ist die curriculare Betrachtung, d. h. die CL aus der Sicht der akademischen Lehre, zeitlich nachgeordnet. Während auf der epistemologischen Ebene Wissenstransport durch die übergeordneten Forschungsinteressen begründet sein muß, steht hier neben der Systematisierung der fachspezifischen Grundlagen eben dieser Aspekt im Mittelpunkt. Die Interdisziplinarität der CL manifestiert sich hier vor allem darin, daß die CL in der Lehre auf die Wissensvermittlung in der Linguistik und in der Informatik mit angewiesen ist, die die genuin computerlinguistischen Lehrinhalte komplettieren und verschieben.

Die Interdisziplinarität der CL ist inhärent, sie liegt in der für die CL charakteristischen Verbindung zwischen Sprache (Sprachanwendung) und Computer selbst.

Denn entgegen der populären Annahme, daß Computer und Sprachen (natürliche Sprachen) nichts miteinander zu tun haben, besitzen sie grundlegende Gemeinsamkeiten, sie sind beide Einrichtungen für die Symbolmanipulation (vgl. Newell 1980: 136; Hellwig 1983; Habel 1986: 6—9). Die Wörter der Sprache sind nicht als bloße Folgen von Geräuschsequenzen zu verstehen, sondern sie sind Träger von Inhalten, genau wie auch für den Rechner die dargestellten Inhalte wichtiger sind als die Fluktuation der physikalischen Zustände. Auch wenn in der letzten Zeit (Raskin 1985; Winograd/Flores 1986; Schnelle 1987) die Grenzen dieser Gleichschaltung erkannt werden, bleibt die Symbolverarbeitung als Grundlage für die CL erhalten.

2. Die CL, ihre Benennung und Abgrenzung

Computational Linguistics als Bezeichnung eines neuen Forschungsfeldes wurde am Anfang der 60er Jahre von David G. Hays geprägt und mit Substanz gefüllt (Hays 1967; 1966). Sie fand erstmalig 1963 öffentliche Verwendung in dem Fachverbandsnamen „Association for Machine Translation and Computational Linguistics“. 1968, zwei Jahre nach dem Erscheinen des ALPAC-Berichts, wurde der Vereinsname auf seine heutige Gestalt „Association for Computational Linguistics“ (ACL) gekürzt. Als Folge der Entwicklung in den letzten Jahrzehnten wurde die Bezeichnung unterschiedlich breit angelegt, wobei sich der Umfang der Gebietes änderte und seine Inhalte stets präziser geworden sind.

Rein philologisch läßt sich das Attribut *computational* in der englischen Nominalphase *Computational Linguistics* auf zweierleiweise interpretieren:

— die attributive Bestimmung „computational“ wird als eine Ableitung aus dem Verb *to compute* 'computare' ausgelegt, und dementsprechend versteht man unter Computational Linguistics eine Art Linguistik, die die Sprache oder die Beschreibung der Sprache als ein System von Rechenoperationen ansieht. Oder:

— die attributive Bestimmung wird als eine Ableitung aus dem Substantiv *Computer* angesehen, und demnach versteht man unter Computational Linguistics eine Art Linguistik, die mit Computern verbunden ist.

Die erste Interpretation klingt vor allem in der osteuropäisch-russischen Auslegung der Computerlinguistik als *vyčislitel'naja lingvistika* an, die die Berechnung oder Errechnung der Ergebnisse hervorhebt und daher die besonders intensive Einbeziehung der mathematischen Linguistik in der UdSSR verständlich macht (und die CL in Osteuropa immer noch prägt, siehe Hajičová in 3.1).

Der amerikanische und westeuropäische Gebrauch des Terminus folgt eher der zweiten Interpretationsrichtung, wonach nicht die Berechnung von linguistischen Funktionen, sondern die Rechnerbenutzung bei linguistischen Problemstellungen im Mittelpunkt steht. (Auf die Gefahren in der westlichen Auslegung der CL weist Karlgren in 8.1 hin.)

2.1. CL und die Mathematische Linguistik

Ohne eine vorangehende Definition der *Mathematischen Linguistik* (= ML) bleibt die Zuordnung der CL zu ML leer (vgl. Schnelle 1966; Kiefer 1968; Altmann 1973). Schwierigkeiten ergeben sich in zwei Hinsichten: 1. Wenn unter ML die formale Erfassung der sprachlichen Strukturen verstanden wird, ist CL offensichtlich mehr, da sie nicht nur algebraische Beschreibungen, sondern auch substantielle Aussagen über die Sprache und Algorithmisierung sowie experimentelle Arbeit mit Computern anstrebt. 2. Nicht alle mathematisch möglichen Beschreibungen werden zur CL gezählt, bzw. sind signifikant. Für die CL sind die algorithmisch-prozessualen Beschreibungen durch fachimmanente Erkenntnisinteressen schon legitimiert, während die intervallarithmetischen und sonstige statistischen Untersuchungen eine zusätzliche (externe) Begründung verlangen, und daher für die CL stets am Rande bleiben. Das osteuropäische Konzept der ML ist autonom und es schließt die CL mit ein (vgl. Gladkij/Mel'čuk 1973); in Amerika und in Westeuropa redet man lediglich über die mathematischen Grundlagen der Linguistik (vgl. Hall Partèe 1978) neben einer enger ausgelegten 'Quantitativen Linguistik' (vgl. Art. 9).

2.2. CL und Angewandte Linguistik

Die CL versteht sich als eine wissenschaftliche Disziplin, aber es ist richtig und auf der förderungspolitischen und organisatorischen Ebene hochgradig relevant, daß die von der CL gelieferten Erkenntnisse eng mit der aktuellen Entwicklung der Sprach- und Wissensverarbeitungstechnologie verbunden sind und davon nur schwer getrennt werden kön-

nen. Es wäre jedoch verfehlt, die CL als *Angewandte Linguistik* zu begreifen und sie auf ein Software-Paket oder auf eine Ansammlung von Algorithmen reduzieren zu wollen. Anwendungsrelevanz ist abstufbar, es gibt stärker und weniger stark anwendungsrelevante Bereiche und auch stark und weniger stark verwertbare linguistische Erkenntnisse. Die Anwendungsnähe macht die CL förderungspolitisch interessant (vgl. auch Art. 37).

2.3. CL und die philologische Sprachforschung

Insbesondere durch die Verbreitung der PC-s setzte sich die Computernutzung in der Sprachforschung durch. Es gibt eine Reihe von verdienstvollen Einführungen in die DV und Surveys über geeignete Hardware und Software für Philologen und Textwissenschaftler (Ott/Gabler/Sappler 1982; Krause/Niederehe 1984; Gregor/Krifka 1986). Die Rechnerunterstützung macht jedoch die Sprachforschung nicht automatisch zur CL, auch wenn diese Unterstützung nützlich und förderungswürdig ist. CL beginnt dort, wo die Forschungsergebnisse nicht nur von den Rechnern erbracht werden, oder sich darauf stützen, sondern dort, wo über die Computernutzung und über die *Abhängigkeit der Ergebnisse von den Methoden* systematisch reflektiert wird. Sonst liegen die Erkenntnisinteressen und daher auch die Ergebnisse der Forschung außerhalb der CL.

Die Prämisse führt zu einer breiteren und zu einer engeren Felddefinition der CL:

(1) Nach der breiteren Auslegung genügt die obige Prämisse für die Abgrenzung der CL allein, d. h. keine weiteren Anforderungen sind notwendig. Im Rahmen der CL können also Fragen der Sprachverstehenssysteme, Automatisierung der Lexikologie, Lemmatisierung usw. behandelt werden, aber auch Gesetze des Lautwandels und die der Dialektgeographie, vorausgesetzt daß die Problematik der Verarbeitungsprozesse (mit)untersucht wird. Es ist gleichgültig, ob man sich mit dem Englischen oder beispielsweise mit dem Livischen (einer ostseefinnischen Sprache kurz vor dem Aussterben) befaßt; wesentlich ist die methodische Ausrichtung.

(2) Nach der engeren Auslegung befaßt sich die CL lediglich mit der Sprachanwendungsproblematik und nicht mit allen möglichen linguistischen Fragestellungen. Die Sprachverstehenssysteme werfen selbst die Fragen auf, die in der CL untersucht werden. Die engere Festlegung der CL beansprucht Anwendungsrelevanz für sich und grenzt die sprachlichen Forschungsvorhaben aus, die sich nicht aus dem Informationsverarbeitungsmodell (siehe 3.5.2.1.) ableiten lassen.

Die erste breitere oder integrative Ausrichtung der CL ermöglicht die Einbeziehung der gesamten computergestützten Sprachforschung und wird auch in diesem Handbuch vertreten. Sie ermöglicht eine einheitliche Feldgliederung (siehe 3.5.2.2.). Die Rechtfertigung einer breiteren, nicht auf die sprachliche Simulation beschränkten CL, stützt sich vor allem auf die grundsätzliche Einsicht, 1. daß die Deskription der Sprache eine fundamentale wissenschaftliche Aufgabe ist und prinzipiell nicht aus der CL ausgeschlossen werden darf, 2. daß die qualitativ bessere, präzisere Deskription der Sprache auch eine qualitativ bessere, anspruchsvollere Simulation erlaubt, und 3. daß man nicht immer voraussagen kann, ob oder wie weit die Erforschung eines sprachlichen Problems aus der Perspektive des sprachlichen Informationsverarbeitungsmodells relevant sein wird oder nicht.

2.4. CL und die sprachorientierte KI-Forschung

Die Auseinandersetzung mit der KI verhalf der CL zu einem besseren Verständnis über die eigene Aufgabenstellung in mehreren Hinsichten:

Wie bereits eingangs gesagt, versteht sich die CL als ein *asymmetrisches* interdisziplinäres Forschungsfeld *mit linguistischen* Erkenntnisinteressen. Denkbar sind natürlich Forschungsvorhaben in dem Überschneidungsgebiet zwischen Linguistik und Informatik auch *ohne linguistische* Erkenntnisinteressen, wie dies in der KI-Forschung (und möglicherweise auch in der *sprachorientierten KI*) geschieht. CL und die sprachorientierte KI sind eigentlich keine rivalisierenden Nachbarn, sie sind eher komplementär, sie haben ein gemeinsames Forschungsobjekt, sie verfolgen jedoch unterschiedliche Forschungsziele. In der wissenschaftlichen Forschung ist es nicht selten, daß eine Untersuchung über die engeren Fachgrenzen hinaus Aufmerksamkeit erweckt und gewürdigt wird. Viele Beiträge der sprachorientierten KI und manche der CL gehören zu dieser Kategorie.

Die Objektbereiche der (sprachorientierten) KI und der CL sind jedoch nicht völlig identisch.

Der Objektbereich der KI reicht über den der Linguistik und der CL hinaus und umfaßt über den sprachlichen Bereich hinaus auch die tieferliegende (kognitive) Problemlösungsebene. Die zwei Repräsentationsebenen (die linearisierte, sprachliche und die mehrdimensionale, begrifflich-kognitive) sind selbstverständlich auch für die Linguistik fundamental. Die Aufgabe der Sprache besteht eben darin, daß sie die nicht-linearisierten, mentalen Strukturen linearisiert, bzw. umgekehrt. Aber die linguistischen Erkenntnisinteressen beziehen sich lediglich auf die sprachliche Abbildung, auf die sprachinternen Abbildungsmechanismen und auf die *Oberfläche* der kognitiven Repräsentation, und hören da auf. Für die KI ist hingegen die kognitive Problemlösungsproblematik zentral und eine spezielle Sprachkomponente, die der Problemlösungskomponente vorgeschaltet wird, und eine besondere (nämlich die natürlichsprachliche) Repräsentationsebene beinhaltet, ist nebenseitig und unnötig umständlich.

Durch die Konfrontation mit den Ergebnissen der KI-Forschung wurde es klar, daß eine latente Isomorphie zwischen Komponenten eines Sprachanwendungssystems weder erforderlich, noch zweckmäßig ist. Die CL systematisiert die Sprachverarbeitung und stellt Grundlagen-Domänen, wie *morphologische Analyse, Syntaxparsing, semantische Interpretation usw.* auf, aber sie müssen nicht als isolierbare Komponente in den Sprachverstehenssystemen vorhanden sein (vgl. 3.6.2.). Eben durch die Systematisierung des Vorgehens außerhalb des Verarbeitungsrahmens übersteigt die CL die Grenzen der holistischen KI-Systeme, die eine linguistische Verallgemeinerungsebene nicht kennen. Daher ist es verfehlt wie bei v. Hahn (1987, 57), die CL als eine reine Hilfswissenschaft („only a tool for empirical work“) einzuschätzen.

3. Die Geschichte der Computerlinguistik

3.1. Die Jahre der Vorbereitung

Die Anfänge der CL als eines wissenschaftlichen Unterfangens sind in dem intellektuellen Hochspannungsgebiet an der amerikanischen Ostküste in den Nachkriegsjahren zu suchen. In einer einmaligen offenen Aufbruchstimmung begegneten sich hier Gelehrte und Praktiker, heimkehrende Offiziere und Immigranten, Philosophen und Ingenieure, die die Tragweite der frisch erfundenen *programmierbaren Computer* begriffen hatten. Sie haben die ungeheueren Perspektiven gesehen, die durch die neuen Elektronenrechner eröffnet worden sind. Es entstand ein neues Instrument für die wissenschaftliche und geistige Arbeit, das keine unmittelbare Anwendung finden konnte, da sie den aktuellen Bedarf überstieg. Die Herausforderung der Zeit bestand darin, diesem Potential zu voller Entfaltung zu verhelfen und es für den Fortschritt auszuschöpfen. Man suchte Aufgabenstellungen, die die Leistungsfähigkeit der Rechner demonstrierten, wobei über die inhärente Problematik der Aufgabenstellung selbst zunächst nur wenig nachgedacht wurde. Die Pioniere fühlten sich dem Fortschritt und nicht einer Disziplin verpflichtet. Für Claude Shannon und John von Neumann waren die neu geknüpften interdisziplinären Kontakte und die hierdurch erzielten neuen Einsichten und Erkenntnisse wichtiger als die wissenschaftliche Taxonomie, deren Rubriken eben ihre Arbeiten durcheinanderbrachten. Diese Katharsis erbrachte nicht nur individuelle Spitzenleistungen von einzelnen Wissenschaftlern, wie Claude Shannon, John von Neumann, Rudolf Carnap, Warren McCullogh, Norbert Wiener usw., sondern führte zu einer neuen Partitionierung der Wissenschaft mit *Informatik, Information Science, Psycholinguistik* und auch *CL.* Diese Neugestaltung betraf letztlich die gesamte *moderne Linguistik*.

Insbesondere Shannons Arbeiten über die Informationstheorie (Shannon/Weaver 1949) sind in diesem Zusammenhang relevant. Er wollte mit Hilfe der Wahrscheinlichkeiten die Sprache präziser erfassen, und seine Entropie-Formel befähigt ihn in der Tat, die Fernmeldenachrichten 'billiger' zu übermitteln. Seine informationstheoretisch-nachrichtentechnische Einstellung prägte den Umgang mit natürlichsprachlichen Texten in dem folgenden Jahrzehnt. Unter seinem Einfluß — vermittelt durch Weaver — wurde die Maschinelle Sprachübersetzung (= MÜ) anfangs als eine Dechiffrierungsaufgabe gesehen (Locke/Booth 1957, 24—46). Da die traditionelle, philologisch orientierte Sprachwissenschaft keine brauchbare Theorie zur Verfügung stellen konnte, erhofften die MÜ-Pioniere, durch die Berechnung der sprachimmanenten Übergangswahrscheinlichkeiten und sonstigen sprachstatistischen Indikatoren das Übersetzungsproblem zu bewältigen und die Sprachbarrie-

ren zu brechen. Die Unzulänglichkeit des Shannonschen Modells für die sprachliche Informationsverarbeitung, nämlich, daß es sich auf die Signal-Information bezieht, und nicht auf die semantische Information, wurde allerdings nicht sofort erkannt. Überhaupt war das sprachliche Problembewußtsein in der von der MÜ-Problematik einseitig dominierten Vorphase der CL defizitär. Dies erklärte sich z. T. dadurch, daß die linguistische Schulung der Mitarbeiter niedrig war (für ihre praktische Arbeit hatte die Linguistik damals sowieso nicht viel zu bieten) und, daß die MÜ-Mitarbeiter die Problematik, primär bedingt durch den technischen Stand der Maschinen, in der Rechnerbedienung gesehen haben.

Die Pioniere der ersten Jahre betrachteten sich als Praktiker, sie standen eindeutig außerhalb der philologisch geprägten akademischen Sprachforschung. Viele der aktiven Mitarbeiter des MÜ-Feldes sind in der Tat nach ihrer Ausbildung Ingenieure, Mathematiker, Philosophen usw. gewesen. Insoweit sie doch 'aus der philologischen Ecke' kamen, flüchteten sie sich gerade vor den geisteswissenschaftlichen Unverbindlichkeiten der Zeit.

3.2. Selbstbesinnung: MÜ als der erste Kristallisationskern der CL

Die erste Phase der Entwicklung der CL ist eindeutig von der MÜ geprägt worden. (Zur Geschichte der MÜ im Allgemeinen vgl. Hutchins 1986.) Für die MÜ-Pioniere stand eindeutig der Rechner im Mittelpunkt der Aufmerksamkeit, dessen Bedienung in den 50er Jahren noch recht schwerfällig war, und die die eigentlichen sprachimmanenten Anwendungsprobleme überschattete. Die MÜ-Pioniere neigten dazu, die technischen, maschinenbezogenen Aspekte der Übersetzung überzubewerten. Sie sahen sich gerne gegenüber den traditionellen akademischen Schreibtisch-Linguisten in der Rolle der Praktiker, die mit den Tücken der Maschinen fertig werden. Sie waren nicht theoretisch veranlagt und erkannten das Fehlen der (sprachlichen) Grundlagen nur zögernd. Erstens, weil es für viele die Abstraktion CL in der Zeit noch gar nicht gab; es gab lediglich die einzelnen Anwendungsfelder wie MÜ oder Automatisches Indexing, und man suchte entsprechend MÜ-Grundlagen, Automatische Indexing-Grundlagen usw. jeweils getrennt und zweitens, weil es infolge der komplexen Aufgabenstellungen wie MÜ und Automatisches Indexing selbst für die Pioniere nicht klar war, was sie hier vermißten. Folglich blieb auch die Existenz einer besonderen allgemeinen sprachlich-linguistischen Teilmenge der Grundlagen innerhalb der spezifischen MÜ-, Indexing- und sonstigen Grundlagen lange Zeit unentdeckt. Es gab Illusionisten, die die Hoffnung gehegt haben, daß es sich vielleicht nur um subjektive Wissenslücken handelt, die durch linguistische Lektüre überwunden werden können. Die dominanten Praktiker wollten die fehlenden Grundlagen nebenbei aufbringen, so wie man etwa die Handhabung eines neues Schnelldruckers oder der neuen Version einer Programmiersprache lernen kann, wobei sie die Aufgabe gewaltig unterschätzt haben. Die (anfänglich wohl seltenen) Realisten erkannten richtig, daß die für die MÜ erforderlichen Grundlagen in der traditionellen Linguistik vergebens gesucht werden und projektintern 'nebenbei' nicht aufgebracht werden können. Die Leistung der ersten Entwicklungsphase der CL, die mit dem ALPAC-Bericht zu Ende ging, war die Bereitstellung von paradigmatischen Grundlagen, die allerdings erst in der nachfolgenden Forschung mit Substanz gefüllt worden sind.

3.3. Die Genese der CL

Es wäre jedoch eine unzulängliche Vereinfachung, die CL als Zufallsprodukt eines Gutachtens für die US Academy of Sciences aufzufassen. Richtiger ist der Bericht so einzuschätzen, daß darin eine 'kritische Menge' der Erfahrungen aus der vorangegangenen Forschung zusammengetragen worden ist, und daß hier lediglich die unausgesprochen bereits vorhandenen Erkenntnissen erstmalig explizit artikuliert worden sind. Schließlich haben wichtige Impulse der Linguistik die CL in der Mitte der 60er Jahre erreicht. Demnach verdankt die CL als wissenschaftliche Disziplin ihr Entstehen drei Faktoren: 1. dem ALPAC-Bericht, 2. der vorangegangenen Feldarbeit und 3. den Impulsen der Linguistik (vgl. auch Mey 1971, 43—44).

Von jetzt ab ging es nicht mehr um die Nutzung der Computer in sprachbezogenen Aufgaben, sondern (anders akzentuiert): um die Bewältigung der Probleme, die sich bei der Nutzung der Computer in sprachbezogenen Aufgaben ergeben haben. Es gab einen allgemeinen Konsens darüber, daß

(1) *die Verarbeitung natürlichsprachlich formulierter Informationen ein sprachliches Problem ist* (vgl. auch Montgomery 1969, 2)

und

(2) daß als wissenschaftliche Disziplin *die CL der Linguistik zugeordnet* wird und sich vorrangig auf linguistische Theorien stützt. Die linguistische Orientierung in den 60er Jahren führte fast zwangsläufig zu der Übernahme der *Chomskyschen Sprachbetrachtung* insbesondere der *Generativen Transformationsgrammatik* (= TG) als linguistischer Grundlage für die CL.

Die TG-Orientierung bedingte dann ihrerseits die besondere *Hinwendung zur Syntax*.

Die Genese der CL war in den USA mit einer *linguistischen Rückbesinnung* verbunden. Sie war dadurch bedingt, daß die amerikanische CL — im Unterschied zu der europäischen Entwicklung — überwiegend aus Forschungsprojekten entstand und Bindungen zur akademischen Linguistik entbehrte.

3.3.1. Der ALPAC-Report und seine Folgen

Für die CL spielte der ALPAC-Bericht eine hervorragende Rolle, da in diesem Dokument erstmalig die Notwendigkeit der linguistischen Grundlagenforschung explizit formuliert worden ist. (Der ALPAC-Bericht, erstellt für die National Academy of Sciences, Washington 1965, befaßte sich vor allem mit der Notwendigkeit der Maschinellen Sprachübersetzung. Die Sachverständigen-Kommission des berühmt gewordenen Berichts fand die derzeitigen Ergebnisse der öffentlich geförderten MÜ-Projekte nicht überzeugend und sie erklärte die MÜ als schlechthin überflüssig.)

Die ALPAC-Experten betrachteten die computerorientierte sprachliche Grundlagenforschung als eine neue Entwicklungsphase der Linguistik, die nicht nur eine Anzahl von neuen (sprachlich fundierten) Computeranwendungen ermöglichen, sondern auch qualitativ neue Einsichten in die Struktur und Arbeitsweise der Sprachen mit sich bringen würde.

Es wird in der Literatur mehrfach bedauert (Josselson 1972, 44—49; Henisz-Dostert/MacDonald/Zarechnak 1979, 47—57), daß die großzügige Förderung des MÜ-Vorhabens mit dem ALPAC-Bericht zu Ende ging. Dabei wird übersehen, daß die Verknappung der Mittel sich generell positiv auf das wissenschaftliche Diskussionsniveau ausgewirkt hat. Erst nach der Beendigung der administrativ gelenkten Förderungskampagne und der Lösung der Maschinellen Übersetzung von direkten markt- oder produktorientierten Projektvorhaben begann hier eine normale, problemgesteuerte Entwicklung, die es ermöglichte, die MÜ als wissenschaftliche Forschung zu betrachten (Mey 1971, 38).

3.3.2. Integration des Anwendungsfeldes

Die CL entstand als Ergebnis eines zweistufigen Abstraktionsprozesses, in dem die konstitutiven Impulse von den drei Hauptanwendungstypen 1. MÜ, 2. Anwendungen im Bereich der Automatischen Dokumentation (engl. Information Retrieval = IR) und 3. Frage-Antwort-Systeme (= FA-Systeme) ausgegangen sind. Im ersten Verallgemeinerungsschritt suchte man die Grundlagen für die drei Anwendungstypen jeweils getrennt. Im zweiten Schritt strebte man darüber hinaus nach Schaffung von verallgemeinerten, linguistisch fundierten Grundlagen in einem größeren, abstrakten Raum. Die Integration auf der ersten Abstraktionsstufe erfolgte recht schnell, die Integration auf der zweiten begann erst später, erwies sich als schwieriger und ist vielleicht auch noch heute noch nicht ganz abgeschlossen. Die Schwierigkeiten entstehen dadurch, daß die prototypischen Anwendungen nicht nur auf der Benutzer-Oberfläche, sondern auch auf der abstrakten linguistischen Problemlösungsebene stark divergieren. (Vgl. Art. 37, 5.1.)

In der Anwendungsdomäne spielen drei Anwendungstypen: MÜ, IR und FA-Systeme eine maßgebende Rolle (vgl. Grishman 1986). Es ist natürlich nicht ganz zufällig, daß es eben diese drei Anwendungsfelder gewesen sind, die zu der Profilierung der CL-Grundlagen so intensiv beigetragen haben. Etwas summarisch könnte man sagen, daß die MÜ die fundamentale syntaktische Orientierung mitbrachte, die IR stellte die Sprachinhalte in den Mittelpunkt und FA-Systeme steuerten die zentralen kommunikativen Perspektiven und insbesondere den Partnerbezug bei (vgl. 5.2.).

3.3.3. Chomsky und die CL: Die Axiomatik des Compilermodells

In den 60er Jahren konnte sich keiner in der Linguistik Chomskys Einfluß entziehen, verständlicherweise auch die junge CL nicht. Chomskys Theorie, vor allem seine TG, wurde auch in der CL mehrfach und unkritisch übernommen. Dies führte dann dazu, daß nachfolgend dieselben Thesen ebenso leidenschaftlich bekämpft wurden. Die Auseinandersetzungen um die TG führten dazu, daß Chomskys Bedeutung für die CL einseitig und verfehlt eingeschätzt worden ist.

Es ist ohne Zweifel richtig, daß weder Chomskys *Standard Model* noch sein *Extended Standard Model* sich als Implementierungs-Grundlagen für Frage-Antwort-Systeme besonders geeignet haben, und daß die Transformationsgrammatiken bei der praktischen Parser-Konstruktion unlösbare Probleme aufwerfen. Chomskys prominente Rolle für die CL ist darin zu sehen, daß er exemplarisch für den von ihm favorisierten Syntaxbereich einen Beschreibungsrahmen geschaffen hat, der natürliche Sprachen und Programmiersprachen *commensurabel* macht, ihre Verarbeitung durch dieselben algorithmischen Verfahren vorsieht und hierdurch den direkten Austausch von Knowhow zwischen Informatik und Linguistik ermöglicht. (Vgl. Johnson 1983 a und Klenk, Art. 6.) Die Chomsky-Hierarchie der Formalen Sprachen bildet seither die theoretisch-konzeptuelle Grundlage für die Informatiker im Compilerbau (Hopcroft/Ullman 1969, 51—52; Gries 1971, 46—48; Aho/Ullman[3] 1979, 144).

Chomsky entwickelte ein Sprachbeschreibungsmodell, zentriert um die Syntax, das die Verarbeitung von natürlichsprachlichen Texten analog zu der Kompilierung vorstellt: Es liegt eine natürlichsprachliche Äußerung in der Form einer Zeichenkette (als akustische Signalfolge oder als gleichwertige Schriftzeichenfolge) vor, und die maschinelle Verarbeitung besteht darin, eine andere interne, interpretierbare Repräsentation zu schaffen, eine Zeichenfolge also, die den Informationsgehalt der (ursprünglichen) Eingabekette enthält und dem Empfänger unmittelbar zugänglich (verständlich) ist, wobei das Verstehen natürlichsprachlicher Äußerungen der Ausführung des umgewandelten Codes entspricht (Lenders, Art. 23; Winograd 1983, 15—16). Das 'mapping model' des Verstehens, wonach eine Repräsentation mit Hilfe von Algorithmen in andere Repräsentationen überführt wird, unterliegt auch den noch geltenden Vorstellungen in der KI (vgl. Ramsey, Art. 18 und Habel/Pribbenow, Art. 57).

Chomskys generative Grammatik ging jedoch über das Shannonsche Dechiffrierungsmodell an einer Stelle hinaus und änderte das Forschungsparadigma der CL grundsätzlich. Er führte neue Beschreibungsebenen ein, auf welchen er die Sprache — unabhängig von der Abbildung von Ebene zu Ebene — axiomatisch durch die Grammatiken definierte. So wurde der Compiler durch die Grammatik in seinem Modell im Voraus bestimmt. Die Diskussion über die Unzulänglichkeit der TG für die CL und über Chomskys Kompetenzfetischismus überschattete zeitweilig seine grundsätzliche Leistung und führte zu Mißverständnissen. So überwindet z. B. eine pointierte Zuwendung zu den Performanzaspekten der Sprache — als Alternative zu Chomskys Kompetenzlinguistik — das Chomsky-Paradigma nicht, sondern bleibt darin verhaftet:

(1) Die beanstandete Dichotomie zwischen Performanz und Kompetenz sowie das grundsätzliche Abbildungskonzept als Modell für die Sprachverarbeitung blieben implizit erhalten. Und

(2) jede Theorie muß den Abstraktionsgrad einer Kompetenzbeschreibung anstreben (Hellwig 1983).

Die TG-Orientierung in der CL war keineswegs einheitlich, es gab gleichzeitig auch alternative Modelle, abgesehen davon, daß das Chomsky-Paradigma selbst breiten Spielraum für Variationen erlaubte.

3.4. Die Spaltung der CL

Nach dem initialen Anstoß blieb die CL jedoch weiter gespalten, oder wenn man die Heterogenität der Projekte betrachtet, sogar aufgesplittert. Eine nur durch Generationswechsel überwindbare Zweiteilung, bedingt durch die Interdisziplinarität, durchzog das ganze Feld. Die Mitarbeiter (die neue Selbstbezeichnung 'Computerlinguist' setzte sich nur zögernd durch) hatten entweder einen mathematisch-technischen Hintergrund und arbeiteten sich in die sprachlich-linguistische Problematik hinein, oder aber sie waren Sprachwissenschaftler und hatten eine philologische Ausbildung. Sie lernten durch ihre praktische Arbeit die Rechner kennen. Eine ausgeglichene Doppelkompetenz wurde nur in seltenen Fällen erreicht.

Die zwei Gruppen blieben auch nach der Konstituierung der Genese der CL weiter erhalten, sie wurden unterschiedlich ettiketiert („Theoretiker und Empiriker" bei Pendergraft 1967, „theory oriented und data oriented", „think-hards und work-hards" bei anderen) aber es handelte sich stets um dieselbe Gegensätzlichkeit:

Auf der einen Seite standen diejenigen, die die sprachwissenschaftliche Problematik von der Seite der philologischen Empirie her kannten. Sie erkannten die Möglichkeiten des Computers von ihrem Wissenschaftsverständnis her als Instrumente, sie wollten sie

nutzen. Um ihre Forschungsvorhaben durchführen zu können, lernten sie auch, die Maschinen zu bedienen. Für sie blieben jedoch formale Modelle und grammatische Theorien leere Spekulationen. Konkordanzen, lexikographische Arbeiten, maschinelle Auswertungen von Texten waren für sie durch bereits bestehende übergeordnete Erkenntnisinteressen motiviert, die sie auf der Seite der Sprachverstehensproblematik vermißten.

Die mathematisch-technisch Vorgebildeten (später auch die Informatiker) andererseits kannten die Leistungen der Rechner von der technischen Seite her. Sie haben wiederum nichts gegen abstrakte Grammatikmodelle und formale Beschreibungen einzuwenden gehabt, die einen Teil ihrer Fachkompetenz darstellten. Aber ihnen mangelte es an einer profunden Kenntnis sprachlicher und sprachwissenschaftlicher Probleme.

Es war eigentlich paradox, daß die linguistische Rückbesinnung, die Zuwendung zu Chomsky und zur Sprachverstehensproblematik überhaupt primär von den 'Nicht-Sprachlern', d. h. eben von der mathematisch informatischen Seite her gefördert worden ist, während sich die 'Sprachler' für die neue prozessuale Problematik unempfänglich zeigten.

Unvereinbarkeit der Erkenntnisinteressen, gesteigert durch die potentielle Anwendungsrelevanz der Sprachsimulation, führte schließlich zu einer Absonderung des philologisch motivierten Teils der computergestützten Sprachforschung als eigenes, jedoch nur schwach abgegrenztes Forschungsfeld, der durch die Gründung des Fachvereins: *Association for Linguistic and Literary Computing* (= ALLC 1972) neben Association for Computational Linguistics (= ACL 1968) — auch auf der Ebene der Wissenschaftsorganisation institutionalisiert wurde. Durch die Ausgrenzung der philologischen Forschungsvorhaben aus ihrem Objektbereich entstand die engere anglo-amerikanische Auslegung der CL, die sich vornehmlich der Simulation der Sprachanwendungsprozesse *Natural Language Processing (= NLP)* widmet (vgl. noch 3.6.1.).

Allerdings blieb diese Trennung sowohl terminologisch als auch sachlich unvollkommen. So wollen Evens und Karttunen (1983) „beinahe alle Kombinationsmöglichkeiten zwischen natürlichen Sprachen und Computern" unter CL verstehen und andererseits behandelt auch die ALLC auf ihren Tagungen Themen der MÜ oder der FAS-Systeme.

3.5. Die breitere Auslegung der CL

Während in der angelsächsischen Welt, vor allem in den USA, die praktische Feldarbeit in der CL überwiegend in den computergestützten anwendungsorientierten Projekten geleistet worden ist, kamen die Vertreter der CL in Europa von den Universitäten, und sie verfolgten vornehmlich akademische Forschungsinteressen. In der europäischen Entwicklung der CL gibt es daher keine 'linguistische Wiederbesinnungsphase'. Die niedrigere Konzentration der Forschungsprojekte (aufgesplittert weiter durch die Sprachgrenzen) führte dazu, daß das gesamte Feld der computergestützten Sprachforschung als eine Einheit empfunden wurde. Dies war auch aus förderungspolitischen Überlegungen vorteilhafter. Die engere, amerikanische *Research and Development* (= RD)-orientierte Auslegung der CL warf, wie bereits gesagt (2.3.), auch schwierige Abgrenzungsprobleme auf: Es war nicht immer klar ersichtlicht, ob ein computerorientiertes linguistisches Vorhaben einen Anwendungsbezug besitzt oder nicht und daher zur CL gezählt werden kann oder nicht.

3.5.1. Extensionale Felddefinition der CL

Daher schien es vielen praktikabler zu sein, die CL etwas neutraler als ein Berührungsfeld zwischen Linguistik und elektronischer Datenverarbeitung aufzufassen und mangels einer feldimmanenten Gliederung die thematischen Bereiche der CL einfach aufzulisten.

Die CL wurde danach von zwei Merkmalen bestimmt, nämlich erstens von der sprachlich-linguistischen Natur der behandelten Problematik und zweitens von der Computerbenutzung. Diese extensionale Felddefinition wurde erstmals von Krallmann (1968) präsentiert. Er unterschied zwei Teilbereiche der CL, die er suggestiv als *Linguistik mit Computern* und *Linguistik für Computer* bezeichnete. Linguistik mit Computern umfaßte die linguistischen Forschungsarbeiten, in welchen der Computer als Forschungsinstrument eingesetzt worden ist, und Linguistik für Computer deckte die Vorhaben ab, die im Zusammenhang mit Computeranwendungen vor allem in der sprachorientierten Informationsverarbeitung entstanden sind. Gleichzeitig wurde die Bezeichnung 'Linguistische Datenverarbeitung' eingeführt.

Die polarisierte *mit-und-für*-Feldaufteilung wurde kritisiert, weil sie einen Erkennt-

nisgewinn im Bereich der Linguistik für Computer anzweifeln ließ (Bátori 1977a), prägte jedoch trotz ihrer Unzulänglichkeiten lange die Vorstellung über die CL sowohl in Fachkreisen als auch außerhalb (Lenders 1972 a, 3—4; Schulte-Tigges 1974, 15). Die Aufteilung sagte außerdem nichts über das Wesen der CL aus. Weiterhin sind Linguistik und Computer keine kontrastierbaren Gegensätze, und daher war es verfehlt, Linguistik (= eine Wissenschaft) auf der einen Seite und Computer (= eine Maschine) auf der anderen Seite gegenüberzustellen. Die Bezugnahme auf den technischen Aspekt (auf das Werkzeug 'Computer' und nicht auf die wissenschaftliche Disziplin 'Computer Science/ Informatik') führte zu Mißverständnissen über das Wesen des neuen Faches (v. Hahn 1987) und hemmte die Akzeptanz der gleichzeitig eingeführten neuen deutschen Bezeichnung *Linguistische Datenverarbeitung* (= LDV), denn für die Informatiker suggerierte 'Datenverarbeitung' lediglich eine technische Beschlagenheit und nicht — wie intendiert — eine wissenschaftliche Disziplin.

Als im Laufe der Zeit Computer in der linguistischen Forschung zunehmend eingesetzt wurden, boten sich Klassifizierungsmöglichkeiten an, die sich nach der Art der Computernutzung orientierten; so z. B. unterscheidet Montgomery (1969) vier Tätigkeiten, die in der linguistischen Forschung ausgeführt werden (1. Datensammlung, 2. Datenanalyse, 3. Formulierung der Hypothesen und 4. Testen) und von Computern unterstützt werden könnten.

Martin (1975) sieht vier Nutzungsarten von Computern vor: 1. classifying, 2. calculating, 3. control und 4. simulation. Die Felddefinition erschöpfte sich hier in der Aufzählung der möglichen Einsatzgebiete des Rechners oder der Themenbereiche. Präzisierungsversuche der extensionalen Feldbeschreibung der CL führten jedoch zwangsläufig in die entgegengesetzte Richtung: zu einer mehr inhaltlichen, intensionalen Feldbestimmung.

3.5.2. Intensionale Felddefinitionen der CL

Die extensionale Felddefinition der CL ist eigentlich leer, da es unausgesprochen bleibt, was die durchgehenden Gemeinsamkeiten der aufgelisteten Einzelvorhaben sind. Außerdem haben die aufgezählten Problemdomänen unterschiedliches Gewicht. Selbst wenn die konstitutiven Merkmale der CL 'gefunden' werden, kann man den Sinn der CL nicht aus ihrer bloßen Summe ableiten. Man muß erst die Aufgabenstellung der CL bestimmen und die Feldstruktur daraus ableiten.

3.5.2.1. Das Modell der linguistischen Informationsverarbeitung

Das Modell der linguistischen Informationsverarbeitung der CL ist implizit an den engeren angelsächsischen NLP-Modellen orientiert, inkorporiert jedoch die ganze computergestützte sprachliche Grundlagenforschung. Das Modell existiert in mehreren Varianten, die unterschiedlich explizit formuliert worden sind.

(1) Das *Basismodell von Ungeheuer* (1971) verbindet einen *menschlichen Benutzer*(= M) in dem kommunikativen Rahmen eines abstrakten Frage-Antwort-Systems mit einer Maschine (*Computer* = C). Dem Menschen M liegt ein zu lösendes Problem vor. Er bedient sich dabei der Maschine C und formuliert sein Problem in einer natürlichen Sprache, das auf die Maschine übertragen wird. Die Natürlichsprachlichkeit ist für Ungeheuer eine unabdingbare Voraussetzung eines solchen Systems. Die Maschine C löst (mit Hilfe von internen Analyse- und Interpretationsregeln) das Problem von M und liefert ihm ebenfalls natürlichsprachlich die Lösung.

Das oft zitierte Modell (Dietrich/Klein 1974, 15—17; Lenders 1975, 35—39; Bátori 1977a, 1982 b) ist selbst kein (konkretes) Frage-Antwort-System, sondern hat den Status eines Prototyps, der erst in den einzelnen Anwendungen realisiert wird. Die Komponenten des Modells entsprechen den einzelnen Teilbereichen der CL. Entscheidend ist bei dem M-C-Modell die Natürlichsprachlichkeit. Nach Ungeheuers Auffassung wird das Modell in dem Kommunikationsprozeß mit der Problematik der Natürlichsprachlichkeit direkt konfrontiert. Darüber hinaus kann das M-C-Modell (reflexiv) auf das Problem der Beschreibung der natürlichen Sprache gerichtet sein und so kann es sich auch (sekundär oder reflexiv) in der Problemlösungskomponente mit der natürlichsprachlichen Problematik auseinandersetzen.

Ungeheuers Modell enthält eine Reihe von vage formulierten Stellen, die vielfach durch den schwachen empirischen Bezug bedingt sind. Daher wirkt das Modell eher programatisch. Es war 1971 aber aktuell und richtungsweisend: Es richtete die For-

schungsinteressen auf *das Problem der Problemlösung* in Sprachverstehenssystemen das in seiner vollen Tragweite erst in der nachfolgenden KI-Forschung thematisiert wurde.

(2) Eisenberg (1980) nimmt explizit Bezug auf die anglo-amerikanischen Sprachverstehenssysteme und betrachtet diese als den zentralen Bereich der CL. Er zählt jedoch die linguistischen Forschungsarbeiten vor allem im lexikalischen Bereich und in der maschinellen Syntaxanalyse auch zur CL.

(3) Direkter und nicht weniger engagiert als Ungeheuer argumentiert Hellwig (1983) in diesem Zusammenhang für Natürlichsprachlichkeit. Er setzt auch linguistische Erkenntnisinteressen voraus und meint, daß der eigentliche Durchbruch für die computergestützte Informationsverarbeitung einzig durch die formale Erfassung der natürlichen Sprache erbracht werden kann.

Hellwig betrachtet den Rechner besonders nüchtern und wehrt sich gegen eine unnötige Anthropomorphisierung der Maschine, indirekt also auch gegen Ungeheurs Modell. Er zweifelt jedoch die Mächtigkeit der Maschine keineswegs an, die ihre Leistungen für die Sprachforschung in vier Gebiete erbringt: 1. als Arbeitsmittel, 2. als Test für formale Modelle, 3. als Zwang zu expliziter Formulierung und 4. als eigener Fall von Sprachanwendung.

(4) Thompson (1982) zeichnet ein fein differenziertes Bild über die CL, in dem er oberflächlich ähnlich zu der Aufteilung CL-Grundlagen und CL-Anwendungen (vgl. Art. 37) auch eine Unterscheidung zwischen „theory of linguistic computation“ und „applied computational linguistics“ zieht. Die systematische Ausgrenzung der linguistischen Erkenntnisinteressen führt jedoch zu einer Verarmung der CL und reduziert sie zu einer bloßen „applied linguistic theory“, die nicht mehr ist als „formulae of transition from linguistic theories and models to models and descriptions practically digestible for NLP“ (Raskin 1985: 275), oder zu einer Unterabteilung der KI (Halvorsen 1986).

3.5.2.2. Deskription und Simulation

Aus der Sicht des Basismodells von Ungeheuer erschien die Sprachbeschreibung als eine in dem M-C-Modell zufällige, weglaßbare Aufgabe. Dies entsprach jedoch nicht der Tragweite der computerorientierten Forschung in der deskriptiven Linguistik. Diese Forschung ist nämlich nicht deshalb interessant geworden, weil sie sich auf die Computer stützte, sondern weil sie auf diese Weise neue, früher nicht machbare und daher auch nicht erbrachte Leistungen aufwies. Computerunterstützte Sprachforschung ist allein schon deshalb überlegen, argumentierte Lenders bereits 1974, weil dabei die Kompatibilität der Beschreibungsebenen systematisch beachtet werden muß (Lenders 1974). In einem Survey teilte er (Lenders 1980) die CL (nach seiner damaligen Terminologie: LDV) in zwei große Bereiche, *Deskription* und *Simulation* auf. Die zwei Bereiche sind unterschiedlich ausgerichtet, aber sie sind beide gleich fundamental. Der Bereich der Deskription beinhaltet die auf die Strukturermittlung gerichtete Forschung, hier werden die Texte auf den verschiedenen Ebenen analysiert und beschrieben. Die mit Computerunterstützung vollzogene Beschreibung der Sprache ist explizit und präzis (differenziert und auch quantifizierbar), vollständig, den Erfordernissen der Massendaten gewachsen und nicht zuletzt nachprüfbar (Lenders 1980, 217). Die Deskription stellt bereits in sich neue linguistische Erkenntnisse dar und ihre Qualität setzt überhaupt neue Maßstäbe für die linguistische Forschung. Darüber hinaus sind sie unerläßlich für die Simulation des Sprachverhaltens.

Die Neuigkeit der Erkenntnisse im sprachlichen Simulationsbereich ist noch deutlicher. Die ganze algorithmische Beschreibung der sprachlichen Verstehens- und Formulierungsprozesse entstand erst im Zusammenhang mit computerorientierten Arbeiten. Die maschinelle Simulation der Sprachanwendungsprozesse setzt eine präzise und explizite Deskription der Sprache voraus. Dies gilt auch umgekehrt: die explizite Deskription setzt ebenfalls einen vorangehenden Ermittlungsprozeß voraus.

Lenders breite Auslegung der CL und insbesondere seine Gliederung nach Deskription und Simulation wird in diesem Handbuch verfolgt.

3.5.3. CL in Osteuropa

In Osteuropa, vor allem in der UdSSR und in der Tchechoslowakei, wurde die CL ebenfalls in einer breiteren Interpretation übernommen. Die Entwicklung verlief ruhiger, ohne spektakuläre Höhen und Tiefen. Auffallend war die andauernde starke Bindung an die mathematische Linguistik (vgl. Kiefer 1968), die sich bereits in der MÜ-Phase manifestierte und die linguistische Umorientierung erleichtete (vgl. Art. 3).

Auch die Chomsky-Welle wurde in Osteuropa weniger turbulent erlebt. Vor allem die Prager Schule der CL zeigte sich eigenständiger gegenüber der TG und Chomsky insgesamt. Für die Spitzenvertreter wie Sgall oder Hajičová waren Linguistik und CL niemals Gegensätze, lediglich methodische Aspekte. Die Prager waren die ersten, die die Grenzen der Strukturlinguistik erkannt und eine weiterführende funktionale Alternative angeboten haben (Art. 11).

3.6. Die Konsolidierung der CL

In den 70er Jahren erfolgte die Konsolidierung der CL. Die regelmäßig abgehaltenen Fachveranstaltungen, vor allem die internationalen COLINGS-s (*Conference on Computational Linguistics* seit 1965 zweijährlich), sowie zahlreiche regionale und nationale Tagungen machten auch äußerlich sichtbar, daß hier ein neues Fachgebiet entstanden war.

Nach der *linguistischen Rückbesinnung* in der Mitte der 60er Jahre setzte ein permanenter Informationsfluß von der Linguistik in die Richtung der CL ein. Die Mitarbeiter des CL-Feldes verstanden sich seither als Linguisten und verfolgten die aktuellen Strömungen in der Mutterdisziplin. (Ein Informationsfluß in der entgegengesetzten Richtung, von CL zu Linguistik, gab es jedoch in den 70er Jahre noch nicht, Mey 1971.)

3.6.1. CL als Linguistik der Sprachverstehenssysteme

Um 1970 herum vollzog sich weltweit ein thematischer Wandel in der CL, die Zuwendung zu den Frage-Antwort-Systemen. Insoweit Frage-Antwort-Systeme sich unausweichlich mit Informationen und sprachlichen Inhalten auseinandersetzen mußten, erfolgte eine Felderweiterung, die sich gleichzeitig mit der Zuwendung zur Semantik in der Linguistik abspielte und die auch eine Zuwendung zur Semantik in der CL mit sich brachte. Impulse kamen vor allem von Fillmore, Montague und im Allgemeinen von der Merkmalssemantik.

Die Modularisierung des *sprachlichen Informationsverarbeitungssystems* erfolgte nach einer linguistischen Taxonomie mit Lexikon, Syntax, semantischer Interpretation usw., mit weiterer linguistischen Untergliederung, wobei die Sprachverarbeitungsproblematik als Analyseproblematik verstanden worden ist und Synthese ausgespart blieb. Wichtig und neu war, daß es sich dabei nicht nur um die statische Strukturbeschreibung handelte, sondern um Prozesse (bzw. um Prozeßbeschreibungen), die während des Verstehensvorgang ablaufen, deren Beschreibung und Simulation die zentralen Problembereiche der CL bilden (Bátori 1982 b). Insbesondere wurden zwei Prozeßarten auseinandergehalten und weiter problematisiert: 1. die Syntaxparser und 2. die semantischen Interpreter.

Die CL präsentiert Lösungen für linguistische Probleme, die bis dahin ungelöst oder unbeachtet blieben, erstmalig in dem Bereich des Parsings.

Winograds SHRDLU, Woods LUNAR, Quillians SEMANTISCHE NETZE, das REQUEST System von Plath und Petrick erweckten recht früh Interesse, bereits am Ende der 60er Jahre/am Anfang der 70er Jahre auch außerhalb der engeren Fachkreise, vor allem unter Psychologen und Psycholinguisten, als Modelle für die menschlichen Sprachverstehensprozesse. Auch wenn die Sprachverstehensproblematik weiterhin als unbewältigt angesehen werden muß (vgl. Art. 23), ist hier zu registrieren, daß über die Disziplingrenzen hinweg anregende Impulse der CL der psycholinguistischen Forschung gegeben worden sind. Die interdisziplinäre Zusammenarbeit zur Erforschung des Vorgangs der menschlichen Sprachanwendung zwischen Informatikern, Linguisten, Psycholinguisten und Computerlinguisten wurde inzwischen weiter ausgebaut und erstreckt sich auf mehrere sprachliche Beschreibungsdomänen (für die Syntax vgl. Art. 24, für die Semantik Art. 20). Neben dem Modellieren des sprachlichen Verstehensprozesses wird auch die Problematik der *Spracherzeugung* behandelt (Art. 36). Die CL ist heute untrennbar in den neu entstandenen Wissenschaftsverband *Cognitive Science* eingebunden.

Für die Psycholinguisten handelt es sich vor allem um den Aussagewert der Computersimulation für die psychologische Wirklichkeit des Menschen. Wie weit die Regeln eines Simulationsmodells Schlüsse über die Beschaffenheit von kognitiven Strukturen zu ziehen erlauben, ist eine methodologische Frage, die allerdings in der CL nicht beantwortet werden kann. Nichtsdestoweniger bleibt die Simulation von kognitiven und sprachlichen Prozessen auch für die CL eine Herausforderung.

3.6.2. Prozeßbeschreibung und Streben nach Eigenständigkeit

Die Prozeßproblematik in den Sprachverstehenssystemen besaß eine größere Wichtig-

keit, als dies am Anfang eingeschätzt worden ist. Anfangs neigte man dazu, die eigentliche linguistische Formulierung in der statisch-deskriptiven Deskription zu sehen, die noch für die Computer in eine algorithmisch-programmierte Form umgesetzt werden muß. Später begriff man jedoch, daß die algorithmische Form gleichrangig ist und eigene, wichtige Gesetzmäßigkeiten aufweist, die nicht von der (statischen) Strukturbeschreibung abgeleitet werden können.

Eine prominente Rolle spielten in diesem Zusammenhang die Netzwerke, die die Repräsentation und die (simulative) Realisierung der sprachlichen Prozesse ermöglichten. So sind die vor allem LISP-basierten ATN-Implementierungen in den 70-er Jahren mehr oder weniger zur Standard-Darstellungsform für die Syntax in der CL geworden.

Die Beschäftigung mit den Sprachverstehenssystemen zeigte, daß es eine zwingende direkte Korrespondenz zwischen der linguistisch-theoretischen Beschreibung und der Modularisierung der Sprachverstehenssysteme nicht gibt. Vor allem können in einem Sprachverstehenssystem Moduln vorkommen, z. B. Optimierungskomponenten u. ä., die in der linguistischen Deskription gänzlich fehlen. Es stellte sich heraus, daß die Sprachverstehensproblematik zwar die linguistische Deskription voraussetzt, sich aber daraus mechanisch nicht ableiten läßt. Die Modularisierung eines Diskurs-Modells und die linguistische Repräsentation hierfür sind keineswegs isomorph (von Hahn 1986, 522).

Die Beschäftigung mit den FA-Systemen führte des weiteren zu der Einsicht, daß sich die Sprachsynthese nicht als einfache Inverse der Analysekomponente begreifen läßt, eine eigenständige Problematik besitzt und als ein autonomes Modul aufgenommen werden muß (vgl. Art. 36). Die Sprachverhaltensproblematik warf außerdem auch eine Reihe von Fragen auf, die vorangehend in der Linguistik nicht behandelt worden sind. Ein Teilbereich der Verstehensproblematik, vor allem die Frage des strategischen Vorgehens im Sprachverstehen, berührte die Psycholinguistik (Thompson 1983). Das Problem der satzübergreifenden Zusammenhänge wurde in der jungen Textlinguistik (gleichzeitig mit der CL) in Angriff genommen. Schließlich lagen die unumgänglichen Fragen der Kommunikation und der Dialogsteuerung offen, sie sind in der Linguistik nicht problematisiert worden, man konnte hier am ehesten noch von der Seite der Kommunikationswissenschaft Vorleistungen erwarten.

Die Nachbarschaft zur Psycholinguistik und der besondere Bezug zur Kommunikationswissenschaft stärkten die Eigenständigkeit der CL gegenüber der Linguistik.

3.6.3. Inkrementelle Modellierung

Martin Kays Konzept der *Unifikationsgrammatiken* und die hierdurch inspirierte *Lexikalisch-Funktionale Grammatik* (= LFG) ist in ihrer Tragweite für die CL mit Chomskys Compilermodell vergleichbar (für die LFG vgl. Art. 18 und 32). Während aber Chomsky mit einem axiomatischen System arbeitete (siehe 3.4.), bleibt Kays Modell partiell definiert, d. h. die Grammatiken (Sprachbeschreibungen) werden nicht voll im Voraus bestimmt, sondern sie werden, je nach Bedarf, *inkrementell* erweitert (vgl. auch Kempen/Hoenkamp 1982).

Die Verarbeitung von unvollständigem und unsicherem Wissen führte ebenfalls zur inkrementellen Modellierung. Die inferenzielle Erweiterung der Wissensbasis (mit Rückgriff auf die nicht-monotonen Logiken), so daß dabei trotz interner Inkonsistenzen, Inkompatibilitäten und Widersprüche stets ein lauffähiges System erhalten bleibt, ist nur im Rahmen eines offenen, partiell definierten, erweiterbaren Systems vorstellbar.

Die Entwicklung der *inkrementellen Systeme* hängt entscheidend mit den sog. 'KI Programmiersprachen' wie LISP, PROLOG oder SMALLTALK zusammen, die diese inkrementelle Flexibilität aufbringen. Compiler-Generatoren können programmiert werden, die den formalen Aufbau der Grammatik übernehmen und den Linguisten entlasten.

Das Unifikationsprinzip beruht auf der Erweiterbarkeit: Im Prinzip lassen sich neue Elemente dem System stets zufügen, wodurch Änderungen, Korrekturen oder neue Leistungen des Systems bewirkt werden können, die also nicht im voraus definiert (oder vorhergesehen) werden müssen. Das Denken in der CL (das 'Paradigma' der CL im Sinne von Thomas Kuhn) wird heute durch inkrementelle Systeme geprägt.

3.6.4. Rückbesinnung auf die Informatik: CL und KI

Die sprachorientierte KI-Forschung brachte für die CL — über die Klärung allgemeiner Feldabgrenzungsfragen hinaus (siehe 2.4.) — eine Reihe von neuen Impulsen:

(1) Die KI-Orientierung bedeutet für die CL eine (notwendige) Rückbesinnung auf die *Informatikgrundlagen,* daß nämlich ihre Interdisziplinarität sie zu einer Doppelkompetenz (in Linguistik und Informatik) verpflichtet.

(2) Die KI-Orientierung richtete die Aufmerksamkeit auf die linguistische Problematik der Wissensrepräsentation überhaupt. Wissensverarbeitung liegt zwar außerhalb der Zuständigkeit der Linguistik, sie ist jedoch unerläßlich für die sprachliche Informationsverarbeitung.

(3) Textgenerierung (Antwortgenerierung) wurde als eine eigenständige Komponente des Sprachverstehenssystems entwickelt.

(4) Während die CL die Modularisierung der Systeme anstrebte und mit einer Reihe von Repräsentationsebenen und autonomen Moduln operierte, bringt die KI integrierte Systeme in den alle Elemente der Problemlösung verbunden sind. Dem generellen holistischen Vorgehen entsprechend erfolgt die Modularisierung der Systeme in der KI ohne explizite Festlegung einer linguistischen Ebene und ohne besondere linguistisch motivierte Systemkomponenten (Wahlster 1982 a). Eine denkbare Herausisolierung der Lexikonkomponente oder des Syntaxparsers in einem KI-System wird durch Streben nach höherer Effizienz begründet und nicht, wie in der CL, durch die linguistischen Problemgliederung.

4. CL — Teildisziplin oder Metadisziplin?

Die Diskussionen am Anfang der 80-er Jahre, die in Zusammenhang mit dem Einzug der CL in die akademische Lehre geführt worden sind, brachten neue Anstöße für die Präzisierung der Feldbeschreibung der CL und für ihre Verselbständigung generell (Bátori/Krause/Lutz 1982).

Die systematische Bestandsaufnahmen der CL erwies sich als wichtig nicht nur für die Nachwuchsausbildung, sondern sie ist auch unerläßlich für den wissenschaftlichen Fortschritt überhaupt.

Unter der Bezeichnung *Computerlinguistik* wird ein inhärent interdisziplinäres, jedoch eigenständiges Forschungsfeld zwischen Linguistik und Informatik zusammengefaßt. Die CL verfolgt linguistische Erkenntnisinteressen: Ihr Objektbereich umfaßt die Sprachanwendungsprozesse, insbesondere die Problematik der *Mensch-Maschine-Interaktion* in informationsverarbeitenden Systemen (Winograd/Flores 1986). Die CL strebt über die Erstellung einer expliziten und umfassenden Beschreibung hinaus auch die Simulation dieser Prozesse an. Die Methoden der CL sind formal und algorithmisch und die Erkenntnisse selbst dienen als Grundlage für natürlichsprachlich orientierte Computeranwendungen.

Die CL erreicht eine neue Qualitätsstufe vor allem dadurch, daß sie die für Methodenwissenschaften wie die Linguistik unerläßliche Reflexion über den Zusammenhang zwischen Methoden und Ergebnissen in einem einheitlichen Rahmen durchführt, in dem Forschungsziele, -methoden und -ergebnisse *rekursiv* integriert sind.

Die Behauptung der Eigenständigkeit der CL mit Hervorhebung der erbrachten oder erhofften Leistungen ist auch für die förderungspolitische Ebene wichtig, während auf der epistemologischen Ebene die Ausstrahlung des Feldes auf andere Erkenntnisdomänen und ihr Einfluß auf den allgemeinen Fortschritt entscheidend ist.

Daher ist es in der jüngsten Phase der CL positiv zu registrieren, daß die Isolation der CL durchbrochen wird und die Kommunikation zwischen CL und den anderen wissenschaftlichen Disziplinen, vor allem mit der Linguistik, nicht mehr einseitig verläuft, wie noch in den 70-er Jahren (Mey 1971; Kanngießer 1976, 140), sondern in beiden Richtungen.

(1) Als erste erweckten die Sprachverstehenssysteme mit ihren explizit und algorithmisch formulierten Regeln fachübergreifendes Interesse, vor allem unter den Psychologen und Psycholinguisten (vgl. 3.6.1.).

(2) Die wachsende Komplexität der Grammatikmodelle führte dazu, daß sie mit 'Bleistift und Papier', also mit herkömmlichen philologischen Arbeitsmitteln, nicht mehr effektiv erprobt (verfiziert und falsifiziert) werden konnten. Dies ergab sich bereits bei der *generativen Transformationsgrammatik* (vgl. Art. 22). Noch schwieriger erwies sich die Entwicklungsarbeit bei den nachfolgenden Grammatikgenerationen. Die Unifikationsgrammatiken, insbesondere die *LFG* und die *GPSG (Generalised Phrase Structure Grammar),* die heute zu den anspruchsvollsten linguistischen Vorhaben zählen, existieren nur in computer-implementierter Form, denn sonst wäre die Konstruktion von interessanten, größeren Grammatikmodellen praktisch undurchführbar (vgl. Art. 32).

David G. Hays (so in einem Gespräch auf der COLING 1986 in Bonn) erkennt hier den Beginn einer neuen Phase der Linguistik, symptomatisch für die Entwicklung der Wissenschaft überhaupt, die dadurch charakterisiert wird, daß sich die Forschungsinteressen nicht mehr direkt auf die Objekte richten, sondern um eine Abstraktionsstufe höher, rekursiv auf ihre Beschreibungsmodelle. Die Computerlinguistik sei in unseren Tagen nicht nur ein integrierter Teil der Linguistik geworden, sondern gleichzeitig ihre höchste Manifestation.

5. Literatur (in Auswahl)

ALPAC-Report 1966 a · I. S. Bátori 1977a · I. S. Bátori 1982b · I. S. Bátori/J. Krause/H. D. Lutz 1982 · N. Chomsky 1963 · R. Dietrich/W. Klein 1974 · P. Eisenberg 1978 · R. Grishman 1986 · W. von Hahn 1986 · D. Hays 1966 · D. Hays 1967 · P. Hellwig 1983 · M. Kay 1984 · F. Kiefer 1968 · D. Krallmann · W. Lenders 1972 a · W. Lenders 1980 · R. Mey 1971 · Thompson 1983 b · G. Ungeheuer 1971 · T. Winograd 1983.

István S. Bátori, Koblenz(Bundesrepublik Deutschland)

2. Überblick über die Wissenschaftsorganisation und Forschungseinrichtungen der Computerlinguistik in den westlichen Ländern

1. Betrachtungsbereich und Darstellungsmethode

Bei der Aufgabe, in einem Handbuchartikel eine Übersicht über die Wissenschaftsorganisation und die Forschungseinrichtungen der Computerlinguistik in der westlichen Welt zu geben, erscheint allenfalls auf den ersten Blick die *Vollständigkeit* als Hauptproblem; sie wird sehr bald vom Problem einer sachdienlichen *Informationsreduktion* überlagert. Machbarkeit und Gebrauchswert der Übersicht ergeben sich aus der Art dieser Informationsreduktion.

Darum wird zunächst der Einzugsbereich der Darstellung erläutert. Dann wird die Informationsauswahl beschrieben, die zu dem hier präsentierten Bild der computerlinguistischen Forschungslandschaft geführt hat.

1.1. Einzugsbereich der Darstellung

1.1.1. Definition der Computerlinguistik

Was Computerlinguistik ist, ist umstritten. Einen Eindruck von der *Verschiedenheit der Auffassungen* gewinnt man schon, wenn man beispielsweise Krause 1984 a, Goodwin/Hein 1982, Dietze et al. 1983 und Lenders 1980 heranzieht, die zudem nicht explizit von Computerlinguistik handeln, sondern das zur Diskussion stehende Gebiet Linguistische Datenverarbeitung nennen bzw. zwischen Linguistik und Künstlicher Intelligenz ansiedeln.

Für den Gebrauch in dieser Übersicht wurde die Auffassung von Computerlinguistik, die diesem Handbuch zugrunde liegt, aus der Wissenschaftsorganisation heraus operationalisiert als Aufzählung einschlägiger Kommunikationsmedien (Konferenzen und Zeitschriften), die wichtige Instanzen in der fachlichen Meinungsbildung sind und in den meisten Fällen in Zusammenhang mit Fachverbänden stehen.

Computerlinguistik ist für den Gebrauch in dieser Übersicht, was über diese fachlichen Kommunikationsmedien vermittelt wird. Wo Obermengen der Computerlinguistik behandelt werden (etwa Computeranwendungen in den Geisteswissenschaften oder die gesamte Künstliche Intelligenz), werden nur die sprachbezogenen Beiträge berücksichtigt. Damit vertraut man nicht nur den Medien der Fachkommunikation die Definition der