nition of computational linguistics as linguistics *with* a computer is of course not wrong at all, but not very illuminating as well. However in discussing the use of computers as simulation machines in linguistics, it must have become apparent too that next to linguistics *with* a computer, linguistics *for* a computer is possible also. This is particularly the case when the researcher has to formulate language rules in such a way that a computer can also "understand" them. More simply stated this means that the computer is programmed so that it can read and write natural language as human beings can, i. e. "understanding" what it is doing.

Basically computational linguistics is defined then as *that branch of linguistics that studies linguistic processes by simulating them by computer* or what Bátori (1977a) has called "die Beschreibung der Sprache als Prozeß". As we have seen at the beginning (section 2: theoretical framework) these processes ultimately come down to speaking and hearing so that one also could claim that the basic model underlying computational linguistics is that of an abstract question—answering system in which there is a linguistic interaction between man and machine in such a way that the machine can "understand" the problems it is confronted with and "offer a solution" to them (cf. Ungeheuer 1971). (As computational linguistics simulates this linguistic reality we have put the computational understanding of and solutions to linguistic problems between inverted commas, implying that the latter (the simulation) may differ from the former (reality) but nevertheless is relevant for it).

4. Final Conclusion

One of the goals of this article was to demonstrate that in order to reach the aim of computer simulation of linguistic behaviour not only a new field of problems was "discovered" (process linguistics), but new approaches to solving these problems as well (cf. the discussion of the methodological apparatus implying rules, preferences, procedures, abductions and the like).

On the other hand we also wanted to make clear that an open, dynamic definition of object and method in computational linguistics is the most realistic and fruitful one: computational linguistics is then seen as an autonomous field within linguistics with at its *core* an own object (linguistic processes) and an own methodological apparatus (rules, procedures, preferences, abductions, etc.); however, this core can be extended and move towards a periphery, these *extensions* in their turn having a *core* and an *extension* etc., etc. By defining computational linguistics in this *recursive* way one can go beyond such dichotomies as pure and applied, theory and practice, description and simulation, models and tools.

5. Literature (selected)

I. Bátori 1977a · H. Brandt Corstius 1974 · E. Charniak/D. McDermott 1985 · H. J. Cedergren/D. Sankoff 1974 · M. Evens/J. Vandendorpe/J.-C. Wang 1985 · R. M. Frumkina 1963 · P. Garvin 1962 · M. D. Harris 1985 · W. Lenders/G. Willée 1986 · W. Martin 1973 · W. Martin 1975 · W. Martin 1981 · W. Martin/F. Platteau/R. Heymans 1986 · J. A. Moyne 1977 · G. Ungeheuer 1971 · T. Winograd 1983.

Willy Martin, Amsterdam (The Netherlands)/ Antwerp (Belgium)

5. Problembereiche der Computerlinguistik: Positionierungs- und Abgrenzungsaspekte

1. Einleitung

In den letzten 35 Jahren hat sich ein Forschungs- und Wissenschaftsgebiet entwickelt, in dem als besonderes Forschungsinstrument die elektronische Datenverarbeitung zur Erforschung von Sprache eingesetzt wird: die *Computerlinguistik* (CL). Das vielfältige Spektrum und die verschiedenen For-

schungsrichtungen dieses Fachgebietes spiegeln sich auch in unterschiedlichen Benennungen wider. Namen wie *Automatische Datenverarbeitung natürlicher Sprache (Automatic Language Processing), Sprachdatenverarbeitung (Language Data processing), Philologische Textverarbeitung (Literary and Linguistic Computing)* oder *Elektronische Sprachforschung* bezeichnen dabei eher praktische Forschungsfelder bzw. spezielle Anwendungsgebiete, während Namen wie *Linguistische Datenverarbeitung (Computational Linguistics), Informations-Linguistik* oder *Linguistische Informationswissenschaft* eine Wissenschaftsfundierung anzeigen, die ihre Relevanz in der Praxis von Forschung, Anwendung und Lehre zeigt.

Durch den einheitlichen Untersuchungsgegenstand Sprache hat die CL vielschichtige *Theorie- und Anwendungsbezüge* zu anderen wissenschaftlichen Disziplinen entwickelt: So sind unter anderem insbesondere Bezüge in Theorie und Praxis zur Linguistik, Philosophie, Mathematik, Psychologie und Informatik festzustellen. Überschneidungen mit Fragestellungen der Linguistik waren lange Zeit ausschlaggebend für die Entwicklung dieses Gebietes, gleichzeitig waren und sind CL-Methoden in entscheidendem Maße integriert in die Theoriebildung einer strukturellen, formalen Linguistik; Zusammenarbeit mit einer angewandten Sprachwissenschaft ergibt sich in vielen philologischen Bereichen; Philosophie und CL berühren sich in Fragestellungen der Logik und der Sprachphilosophie; Überschneidungen von Linguistik, Informatik und CL sind in Forschungsansätzen einer sprachorientierten 'Künstliche Intelligenz-Forschung' erkennbar, Problembereiche der CL berühren sich mit Erklärungsansätzen einer Kognitiven Psychologie. Erste Ergebnisse der Zusammenarbeit zwischen CL, Informatik und Informationswissenschaft werden in Form experimenteller Anwendungssysteme als Frage-Antwort-, Übersetzungs- oder Informationssystem sichtbar.

So stellt sich heute die CL als ein *interdisziplinäres Forschungsgebiet* dar, ihre wissenschaftliche Eigenständigkeit begründet sie in einer methodologischen Fundierung der Erforschung von Sprache.

2. Zur Spezifik der Computerlinguistik

Zwei Ereignisse in den vierziger Jahren haben die Entwicklung des Computers von einer einfachen Rechenmaschine zu einer hochentwickelten Datenverarbeitung wesentlich beeinflußt.

McCullough und Pitt stellten die Theorie auf, daß Neuronen im Grunde als logische Verknüpfungen von UND- oder ODER-Schaltungen anzusehen sind; ist diese Grundannahme richtig, so vermuteten sie weiter, lassen sich alle intelligenten Prozesse auf einen einfachen Typ von Verarbeitungsmechanismen zurückführen. Damit war eine gemeinsame Beschreibungsgrundlage für die *Analogiebetrachtung menschlicher und maschineller Verarbeitungsstrukturen und -prozesse* geschaffen. Als zweites in diesem Zusammenhang sind die Arbeiten von Shannon/Weaver zur 'Mathematischen Theorie von Kommunikation' zu nennen. Sie wiesen nach, daß jede Signal- bzw. Symbolform, ob Zahl oder Text, als Spezialform eines allgemeineren Konzeptes, das sie *'Information'* nannten, angesehen werden könne; die Erkenntnis, daß der Informationsgehalt quantifizierbar ist, führte zu einer Reihe interessanter mathematischer und praktischer Anwendungsmöglichkeiten. So können die ersten Ansätze einer 'Elektronischen Sprachforschung', einer ‚Sprachdatenverarbeitung' bis hin zur 'Automatischen Sprachverarbeitung' als direkte methodische Umsetzung dieser Erkenntnisse angesehen werden: die Möglichkeit, mit Hilfe von Computern symbolische Daten durch entsprechende Programme manipulieren zu können, wurde bereitwillig auf geschriebene Texte angewandt, um Wortindizes und Konkordanzen zu erstellen sowie Häufigkeitszählungen durchzuführen. Damit beruhte diese erste 'Oberflächenverarbeitung' (Barr/Feigenbaum/Cohen 1981, 226) von Texten im wesentlichen auf Verfahren des Zählens, des Vergleichens und des Neuordnens bzw. -anordnens symbolischer Daten.

Auch den ersten Versuchen zur *Maschinellen Sprachübersetzung* (MÜ) lag die nachrichtentechnische Überlegung zugrunde, daß Texte oder allgemein Sprache als Information zu behandeln und der Übersetzungsprozeß — als Zuordnungsfunktion eines quellensprachlichen in einen zielsprachlichen Code — als technisches Problem mechanisierbar sei. Diese erste computertechnologische Phase der MÜ (Bátori 1986, 7) vollzog sich unter weitgehender Ausschaltung oder unzulänglicher Beachtung linguistischer Aspekte.

Das offizielle Eingeständnis eines

Fehlschlags dieses Ansatzes zur MÜ in den frühen sechziger Jahren kann als zweite Phase der MÜ und der eigentliche Beginn einer systematisch betriebenen sprachorientierten Datenverarbeitung angesehen werden: Der *ALPAC-Report* von 1966 fordert explizit eine linguistisch-orientierte Grundlagenforschung und initiierte damit letztlich die Computerlinguistik.

Anläßlich des dritten internationalen Kongresses CLIDE '71 (Papp/Szepe 1976) stellt D. G. Hays auf einer Podiumsdiskussion über das *Arbeitsfeld der CL* fest: „... if computational linguistics did not exist it would be necessary to invent it" (Hays 1976, 23); für ihn war damals bereits CL weder ein Teilbereich der Computerwissenschaften, noch war sie einer 'Mutterwissenschaft' Linguistik zu subsumieren, sondern eine Disziplin 'sui generis'. Seine Begründung ist eine methodo-logische.

Es gab und gibt eine Reihe von Disziplinen, in denen Computer eingesetzt und verwendet werden, ohne daß von ihnen deshalb eine eigenständige Disziplin reklamiert würde. Das liegt daran — so Hays —, daß die eingesetzten Computerprogramme auf Methoden und Verfahren beruhen, die nichts mit der internen Struktur des Faches oder fachspezifischen Problemstellungen der Disziplinen zu tun haben.

Innerhalb der CL sieht die Ausgangssituation allerdings anders aus. Die Linguistik hat verschiedene Beschreibungsansätze entwikkelt, die unterschiedliche Lösungsansätze erfordern. John McCarthy hat in diesem Zusammenhang auf einen wichtigen Unterschied aufmerksam gemacht, der für den *methodologischen Ansatz der CL* fundamental ist: die Kenntnis der Lösung eines Problems ist streng zu unterscheiden von der Kenntnis des Lösungsweges eines Problems; das eine läßt sich als Funktion abhängiger (Eigenschafts-)Variablen, das andere als Algorithmus formulieren. Hieraus resultieren für die CL unterschiedliche Vorgehensweisen für die Behandlung und Verarbeitung ihres Untersuchungsgegenstandes. Denn bei aller Verschiedenheit der Auffassungen und Bezeichnungen sowie der Vielfalt der Anwendungen zeichnet sich die CL durch einen *einheitlichen Untersuchungsgegenstand* und ein gemeinsames Forschungsinstrument aus: Sprache und sprachliche Phänomene, bei und für deren Erforschung der Computer eingesetzt wird. Die Analogiebetrachtung maschineller und sprachlicher Strukturen und Prozesse macht die Entwicklung spezifischer, auf das Studium sprachlicher Strukturen, Prozesse und Funktionen ausgerichteter Verarbeitungsprozeduren notwendig. Und so kann Hays mit Recht feststellen, daß die CL nicht einfach Verfahren, z. B. aus den Computerwissenschaften, übernehmen und anwenden kann, sondern im Hinblick auf die besonderen Problemstellungen in diesem Forschungsfeld Darstellungsformate und Verarbeitungsalgorithmen eigener Art entwickeln muß.

Nicht einseitig orientierte Fragestellungen oder Theorieansätze einer strukturellen Linguistik oder sprachorientierte Ausrichtungen anderer Disziplinen sind ausschlaggebend; die Erforschung sprachlicher Strukturen, Prozesse und Regularitäten in unterschiedlichen Bereichen mit dem *Forschungsinstrument 'Computer'* führt zu spezifischen Anforderungen hinsichtlich Beschreibung, Darstellung und Verarbeitungsbedingung. So kann die Untersuchung linguistischen Datenmaterials methodisch 'modelliert' werden als einfache Prozesse des Vergleichens und der Umordnung (z. B. Texte in Indizes, Konkordanzen und Wörterbücher verschiedener Art), neben Fragen einer linguistischen Formalisierung von Grammatiken treten Probleme der computergerechten Methodisierung, und es spielen die in der mathematischen Linguistik ermittelten Isomorphien zwischen der Theorie natürlicher und künstlicher Sprachen eine Rolle.

Hieraus wird deutlich, daß das *Spezifische der CL* nicht primär von Aufgabenfeldern oder Anwendungsbereichen einer DV-orientierten Linguistik bzw. einer linguistisch-orientierten DV (Bátori 1977a; Straßner 1977 b; Lenders 1980) abzuleiten ist, sondern sich aus einem eigenen methodologischen Ansatz selbst begründet.

Ungeheuer hat auf der Grundlage eines 'Basismodells' der linguistischen Datenverarbeitung *zentrale Problemstellungen* formuliert, aus denen dann bestimmte Arbeitsbereiche abgeleitet werden konnten (Ungeheuer 1971, 689). Als zwei Teilprobleme gab er die Prozesse der *Informationserschließung* und die Erforschung des *Problemlösungsverhaltens* in informationserschließenden Systemen an. Theoretische Grundannahmen, Art der Problemstellung und Logik des methodischen Vorgehens bestimmen damit dieses Forschungsfeld, stellen Bedingungsfaktoren für die Erforschung von Sprache mit dem 'Instrument' Computer dar und begründen somit ein eigenes Forschungsparadigma

'Sprachdatenverarbeitung' (vgl. Abb. 5.1). Innerhalb dieses Paradigmas gibt es ein Spektrum möglicher Theoriebildungen über Sprache. Dieses Paradigma erhebt im Unterschied zu klassischen Paradigmen nicht unbedingt einen erklärenden, sondern einen methodologischen Anspruch für die Theoriebildung; konstitutiv für dieses methodologische Paradigma 'Sprachdatenverarbeitung' ist ein Überprüfungs- und Akzeptabilitätskriterium, daß im Rahmen der KI-Forschung zurecht als Kriterium der *Implementabilität* (Hayes 1984, 159) bezeichnet wurde. Das heißt, die Überprüfung der Bedingung der Möglichkeit für Implementierung von Methodentypen stellt ein wesentliches Moment für die Theorieentwicklung dar. So wie Implementabilität kriterial für die Methodologie der Sprachverarbeitung ist, so ist die Wahl von Methodentypen, wie Formalisierung, Operationalisierung und Algorithmisierung symptomatisch für diesen Paradigmenrahmen. Dies bedeutet, daß die auf den Begriff gebrachten Methoden der Implementierung mit dem forschungslogischen Kriterium des Paradigmas korrespondieren. Die Formate der Theoriebildung innerhalb der CL sind danach ebenfalls methodologisch bestimmt und lassen sich nach dem gegenwärtigen Forschungsstand in die Theorietypen *Deskription* und *Simulation* zusammenfassen. Der Theorieanspruch der CL ist so verstanden kein nomologischer Erklärungsanspruch, sondern ein Anspruch auf theoriebildende Rekonstruktion, deren explanative Kraft vom Status der Implementabilität abhängt.

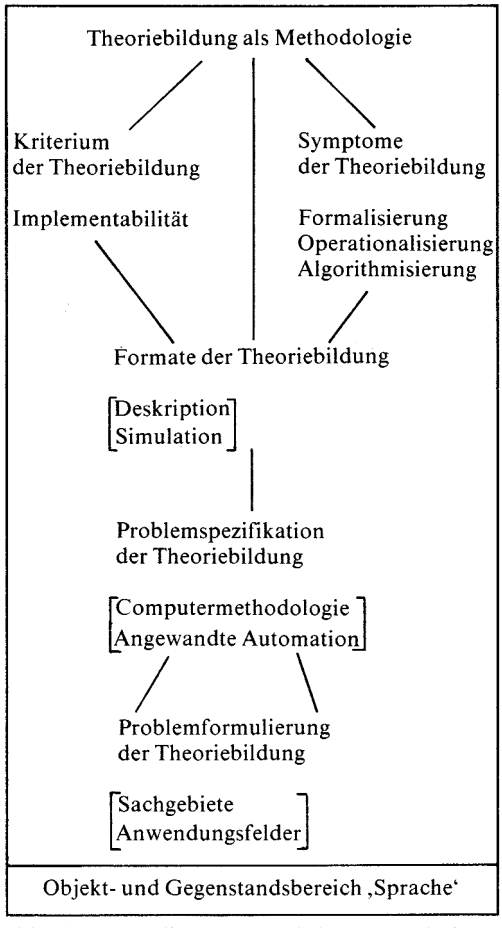

Abb. 5.1: Paradigma Sprachdatenverarbeitung

Im Rahmen einer solchen Methodologie kann es dann eine übliche Formulierung von Problemstellungen nicht geben, sondern nur — in Bezug auf Sach- und Problembereiche des Untersuchungsgegenstandes 'Sprache' — zugeschnittene CL-typische *Problemspezifikationen.* In einem historisch-systematischen Zugriff werden zwei Typen von Problemspezifikationen erkennbar: 1. *Computermethodologie* und 2. *Angewandte Automation,* aus denen dann Problemformulierungen in Sachgebieten und Anwendungsfeldern der CL näher bestimmt und charakterisiert werden können.

3. Problemspezifikationen der Computerlinguistik

Ein Großteil der Schwierigkeiten um die *Eigenständigkeit der Computerlinguistik* resultiert in der unterschiedlichen Beurteilung der theoretischen und/oder der methodischen Fundierung dieses Forschungsfeldes; eine solche Fehleinschätzung schwingt auch teilweise heute noch in der alten Unterscheidung von „Linguistik mit Computern" und „Linguistik für Computer" (Krallmann 1968) mit, in der die Computerlinguistik als Hilfswissenschaft mit Instrumentfunktion einer autonomen Wissenschaft mit eigener Grundlagenforschung gegenübergestellt wird. Gleichwohl blieb lange Zeit ein Desinteresse an einer theoretischen Fundierung dieses Gebietes festzustellen. Für Bátori (1982 b) ist die CL eine angewandte Wissenschaft, die auf eigene Grundlagenforschung angewiesen ist. Als angewandte Wissenschaft übernimmt sie „Lösungsansätze aus anderen Wissenschaften, systematisiert diese und entwickelt eine theoretische Fundierung, die es erlaubt, die übernommenen Lösungsansätze und die in eige-

ner Grundlagenforschung entwickelten ... Lösungsansätze zu integrieren“ (Lutz/ Schmidt 1982, 168). Diese Verbindung ist zu sehen auf der Grundlage der Entwicklung neuer Methoden für die Erforschung von Sprache, die durch die computertechnologische Ausrichtung zu neuen Einsichten der Theorieentwicklung und über eigene Sachgebiete zu neuen Anwendungsfeldern führen.

Während die sprachorientierte KI-Forschung (Wahlster 1982 b) von Anfang an spezifische Problemstellungen, wie zum Beispiel die Modellierung von Verstehen im Rahmen definierter Diskurswelten (Cercone/McCalla 1986), herausgriff und hierfür spezielle Problemlösungen anstrebte, gab es in der CL *keine einheitliche Aufgabenstellung,* sie wechselte vielmehr innerhalb eines weitgesteckten Rahmens je nach definiertem Objektbereich und zugrundegelegtem Theoriebezug.

Ihr Vorgehen ist daher als partikularistisch zu bezeichnen (Lutz/Schmidt 1982, 177). Ihr Untersuchungsgegenstand ‚Sprache’ wird unter verschiedenen Aspekten und Ansätzen mit unterschiedlichen Strukturen und Regularitäten behandelt, so daß als kleinster methodologischer Nenner die *’Modellierung von Informationsverarbeitungsprozessen’* mit Bezug auf den jeweiligen Objektbereich von Sprache und ihre Untersuchungsphänomene angegeben werden kann. Einen derartigen Ansatz teilt sie allerdings auch mit der Informationslinguistik bzw. einer linguistischen Informationswissenschaft.

Einer solchen *Modellierung* werden dabei Hypothesen und Theorieansätze einer teils strukturalen, prozeduralen bis hin zu einer prozessualen Sprachanalyse (Bátori 1982b; Schnelle 1982; Metzing 1982) zugrundegelegt. Problemspezifische Aufgaben- und Problemstellungen der CL ergeben sich aus der methodischen Behandlung linguistischer Bedingungen für die Darstellung sprachlicher Information und diesen zugeordneten Verarbeitungsprozessen; zu spezifizieren sind Eigenschaften der internen Darstellung sprachlicher Strukturen und Regularitäten, Bestandteile von Verarbeitungsprozeduren und -strategien sprachlicher Analysen sowie Modellvorstellungen von Systemkomponenten sprachlicher Analyse- und Syntheseprozesse.

4. Problemstellungen und Aufgabenfelder

Die Entwicklung der CL ist wesentlich durch eine Reihe von wissenschaftlichen Ergebnissen in konkreten Anwendungsfeldern mitbestimmt worden: *Lexikographen* und *Sprachstatistiker* setzten zuerst den Computer systematisch in der linguistischen Forschung ein. Die Möglichkeit, größere Textmengen zu speichern und nach verschiedensten Gesichtspunkten auszuwerten, eröffnete auch für andere Bereiche der Linguistik interessante Perspektiven der Nutzung. Die Zusammenarbeit zwischen einer solchen dateiorientierten CL und den verschiedenen Bereichen der Linguistik war eine forschungspraktische, die auf bestimmten Methoden der CL gründete. Dieser Sachverhalt kommt z. B. in der englischen Bezeichnung ’Literary and Linguistic Computing’ zum Ausdruck.

In hohem Maße in die Theoriebildung integriert sind dagegen CL-Methoden in einer strukturalen, prozeduralen und prozessualen Linguistik, was ihre Beschreibungsmodelle, Testverfahren und Objektbereiche angeht: hier geht es um die *Automatisierung von Grammatiken* und die *Modellierung von Verstehensprozessen* in unterschiedlichem Theorierahmen. Daraus ergibt sich, daß die CL Problemstellungen und Aufgabenfelder in zwei Typen von Problemspezifikationen charakterisiert: *Computermethodologie* und *angewandte Automation* .

4.1. Computermethodologie

Die Kernproblematik der CL, eine natürlich sprachliche Informationsverarbeitung, macht für ein methodologisches Vorgehen bei der Erfassung und Behandlung sprachlicher Prozesse und Strukturen unterschiedliche Problembehandlungsarten notwendig; hier sei zwischen den grundlegenden Arten *Deskription* und *Simulation* unterschieden.

Als Theorietyp *Deskription* sind Beschreibungen Konstruktionen von Strukturen über Sprache, die in Formulierungen von Computerprogrammen eingehen. Sie sind Konstruktionen erster Ordnung und analytischer Natur, was bedeutet, daß sie genau angeben, welche linguistischen Einheiten an sprachlichen Phänomenen beteiligt sind, und welche syntaktischen, semantischen und pragmatischen Beziehungen zwischen ihnen bestehen (Lenders 1980), sei es, daß sie sich auf Sätze, Äußerungen oder Texte, sei es, daß sie sich auf Sprache als System beziehen. Beschreibungen im Sinne des CL-Paradigmas sind Rekonstruktionen, insofern sie wiederum dann als implementierte Regelsysteme Beschreibungsexplikate für die Konstruktionen darstellen. So gesehen sind sie Konstruktio-

nen zweiter Ordnung, analytisch und selbstreferentiell. Sie beziehen sich beispielsweise innerhalb einer theoretischen Linguistik auf Grammatikmodelle, mit denen Gesetzmäßigkeiten der Sprache deskriptiv erfaßt und Eigenschaften sprachlicher Kompetenz charakterisiert werden sollen.

Für diesen Bereich der Sprachbeschreibung sind in der CL eine Reihe von *Verfahren* entwickelt worden, die eine vielfältige und vielschichtige linguistische Analyse ermöglichen; diese Verfahren, von Lenders (1980) als 'komplexe Methoden der Textbeschreibung' bezeichnet, haben das Ziel, explizite Angaben zu den linguistischen Einheiten der einzelnen Sprachebenen zu ermitteln, um damit Sprache als System im strukturalistischen Sinn zu beschreiben. Matching- und search-Verfahren sind Beispiele von Methoden-Typen in diesem Bereich.

Der wichtigste Methodentyp der analytischen Sprachbeschreibung ist sicherlich der des *Parsing* . Ausgehend von den frühen Arbeiten Chomskys zur Formalisierung der Syntax natürlicher Sprachen wurde eine Reihe von Parsing- und Generierungalgorithmen entwickelt, die auf bestimmte Grammatiktypen (kontextfreie Grammatiken oder Transformationsgrammatiken) ausgerichtet waren, um Beschreibungen komplexer Strukturen von Texten zu produzieren.

Die *Klasse der Parser* bildet heute die Grundlage aller nicht-trivialen Aufgaben der CL. Die Parser leisten nach Maßgabe der sie steuernden (grammatischen) Regeln und dem Wörterbuch die Zerlegung eines Satzes oder Textes in seine Konstituenten und deren Klassifizierung. Der Begriff des Parsers hat in den letzten Jahren eine erhebliche Bedeutungserweiterung erfahren; während er zuerst hauptsächlich als syntaktischer Analysealgorithmus begriffen wurde, werden Parser in zunehmendem Maße auch für semantische und pragmatische Textbeschreibungen eingesetzt (Winograd 1983). Eine systematische Zusammenstellung einzelner Parsing-Techniken, wie ATN-, CHART- oder INSEL-Parsing ist in King (1983) und Christaller/Metzing (1979/1980) zu finden.

Die methodologische Forderung nach Robustheit von Parsern hat in *Textanalysesystemen* zu einer Reihe von weiterführenden Lösungsansätzen wie heuristisches und/oder partielles Parsing geführt, um auch außergrammatische Problemstellungen behandeln zu können. Wahlster (1982b) hat die Einsatzmöglichkeiten von Parsern in den einzelnen Komponenten natürlich-sprachlicher Systeme der sprachorientierten KI-Forschung aufgezeigt. Auch in der CL sind eine Reihe von Computer-Systemen zu nennen, in denen Parsing-Verfahren eingesetzt sind. Maschinelle Übersetzungssysteme, Frage-Antwort-Systeme oder Dialogsysteme sind die bekanntesten Typen textanalysierender und -generierender Systeme.

Das *Testen* von deduktiv gewonnenen Sprachbeschreibungsmodellen stellt ein weiteres methodologisches Forschungsmittel der CL dar. Formale Beschreibungsmodelle, verstanden als Regelsysteme, können auf allen sprachlichen Ebenen getestet werden, um die Richtigkeit bzw. Adäquatheit theoretischer Annahmen über die Struktur und Funktion von Sprache zu untersuchen. Ein solcher Adäquatsheitsbegriff beurteilt dann sowohl die Angemessenheit der Beschreibung, als auch der Generierung, also der Theorie gegenüber dem Sprachsystem. Diese theorieorientierten Systeme sind bis heute vornehmlich ausgerichtet auf Testverfahren für Sprachbeschreibungsmodelle im Bereich der Morphologie, der Syntax und Semantik.

Diesen Sprachbeschreibungssystemen sind als besondere Problemspezifikationen die *Sprachsimulationsmodelle* entgegenzusetzen. Die wohl größten Änderungen, was die Theoriebildung und die experimentelle Überprüfung betrifft, haben sich durch die Simulationsmethodologie ergeben. Simulation bezeichnet „den Prozess des Modellierens eines natürlichen dynamischen Systems durch ein künstliches dynamisches System" (Lenders 1980, 221). In der CL wird dieser Problembereich als 'Simulation sprachlichen Verhaltens' bezeichnet, modelliert werden Verstehens- und Verhaltensprozesse, denen sprachlich-fundiertes Wissen zugrunde liegt. Die methodologische Frage zentriert sich auf die Problemstellung: „Wie können die Prozesse des Sprachverstehens und der Sprachproduktion durch Simulation beschrieben und rekonstruiert werden?"

Mit einem solchen simulationsmethodologischen Ansatz nähert sich die CL sehr stark dem Forschungsgebiet der *'cognitive science',* für die die Erforschung von Sprache auf kognitiver Ebene ein zentrales Thema darstellt. Abgrenzungen der CL zur 'cognitive science' und zur Künstlichen Intelligenz ergeben sich aus der modelltheoretischen Behandlung kognitiver Fähigkeiten. Palmer (1984) hat die Künstliche Intelligenz-Forschung als „new style of thinking about cognition" und als

Charakteristikum dieser Richtung die Durchführung von „Gedankenexperimenten“ bezeichnet.

Damit wird die Frage der Theorie und der empirischen Hypothesentestung angesprochen. *Gedankenexperimente* sind zwar Experimente, jedoch keine empirischen; sie befassen sich mit den grundsätzlichen Bedingungen der Möglichkeiten von tatsächlichen Begebenheiten. Die so entwickelten Modellvorstellungen werden in Programme bzw. Programmsysteme implementiert. Dieser Vorgehensweise liegt grundsätzlich die Metapher von der „Psychologie der Informationsverarbeitung“ zugrunde, was bedeutet, daß es hinsichtlich gewisser Eigenschaften interne Struktur- und Funktionsäquivalenzen zwischen Mensch und Maschine gibt, die zu erforschen die Aufgabe der Künstlichen Intelligenz ist. Im methodischen Vorgehen heißt dies, Modellvorstellungen von intelligentem Verhalten werden gefaßt als modulare Konzepte kognitiver Fähigkeiten; die Konzepte führen dann zu Modellentwürfen über kognitive Prozesse. Das Modell wird durch Implementierung in ein Programm bzw. Programmsystem „realisiert“. Der Computer simuliert damit das Verhalten des Modells. Nur so ist es wiederum möglich, das Modell an tatsächlichen Gegebenheiten zu testen. Der Computer generiert Strukturen und Prozesse, die als Erklärungskonstrukte für kognitive Prozesse herangezogen werden.

Zum *Simulationsbegriff* der CL wird auf von Hahn (1978) und Lenders (1980) verwiesen. Simulation sprachlichen Verhaltens beinhaltet als Problemfeld eine Klärung des Bedeutungsbegriffs und des Verstehensprozesses, und zwar durch die Fragestellung, wie die Bedeutung sprachlicher Äußerungen repräsentiert wird und wie Verstehensprozesse sprachlicher Äußerungen rekonstruiert werden. Die modelltheoretische Behandlung sprachlichen Verhaltens hat inzwischen zu einer Fülle von Theorie- und Modellansätzen geführt; so bei der Entwicklung von Darstellungsformaten für Wissen oder der Modellierung von Sprachverstehen. Begriffe wie ’Semantische Netzwerke’, ’Semantische o. Episodische Gedächtnismodelle’, ’frame’-Theorie kennzeichnen grundlegende Ansätze der CL und der sprachorientierten KI-Forschung. Sie stellen wesentliche Komponenten zur Darstellung und Verarbeitung unterschiedlicher Wissensarten, angefangen von lexikalisch-strukturellen über semantisch-kontextuelle bis hin zu thematischen Zusammenhängen, in natürlich-sprachlichen Systemen dar. Dabei erstreckt sich der Gesamtbereich ’Simulierung von Sprachverstehen’ über die Konstruktion einzelner Aspekte textverstehender und -generierender Systeme bis zur Modellierung komplexer Kommunikationsleistungen. Die Akzeptanz der Analogie zwischen Mensch-Maschine und Mensch-Mensch-Kommunikation auch in der CL hat die Forschungsrichtung auf die prozeßhaften Aspekte der Sprache (der sprachlichen Formulierung und des Sprachverstehens) gelenkt. Eine Simulation der gesamten kognitiven Verarbeitung umfaßt dann als Ziel die akustische Signalanalyse, morphologische Prozesse, syntaktische Erkennung, semantische Verarbeitung, sprachliche Problemlösung, kognitive Repräsentation und lexikalisches Wissen. Je nach Zielrichtung einer ablauf- oder ergebnisorientierten Simulation sind unterschiedliche Strategien der Anordnung der einzelnen Verarbeitungsebenen sowie der Verzahnung der formalen Analyse mit der aktuellen Problemstellung notwendig (Bátori 1981 a). Die Verfahren von deklarativer und prozeduraler Wissensdarstellung innerhalb einer sprachorientierten KI-Forschung sind hier als integrativer Ansatz zu nennen, der die Vorteile einer objektorientierten Darstellung mit den Erfordernissen einer vielschichtigen Problembearbeitung verbindet.

Die *Motivation* zur Erforschung der Problembereiche Sprachverstehen und Wissensbehandlung in der CL ist eine zweifache: zum einen erhofft man durch die Simulationsansätze Einsicht in die *Funktionsweise* sprachlich/kognitiver Prozesse zu erhalten, um dadurch wiederum zu neuen computermethodologischen Erkenntnissen in bezug auf Verarbeitungsprozeduren und Systemarchitekturen zu gelangen; zum anderen weist die Konstruktion sprachverstehender Systeme direkte *Anwendungsbezüge* auf, die auf bestimmte Anwendungsfelder zugeschnitten ist: hier sind beispielhaft Frage-Antwort-Systeme, natürlichsprachliche Schnittstellen zu Informations- und Datenbanksystemen und Maschinelle Übersetzungssysteme zu nennen.

4.2. Angewandte Automation

Die CL ist ihrem Eigenverständnis nach und entsprechend der historischen Entwicklung gesehen *anwendungsorientiert* . So sind bestimmte Methoden zur Verarbeitung sprachlicher Daten und ’Instrumente’ zur Textbe-

schreibung das Resultat konkreter Anwendungsprobleme. Das Gesamtgebiet dieser anwendungsorientierten CL ist heute als ein Forschungsfeld zu beschreiben, das Ergebnisse computermethodologischer Grundlagenforschung auf konkrete Anwendungsfelder hin einsetzt. Durch die Problemspezifikation Angewandter Automation sollen dieses Wechselverhältnis einerseits und die Anwendungsorientierung andererseits zum Ausdruck kommen. Dementsprechend werden Darstellungsformate und Verarbeitungskonstruktionen in elementaren Textbeschreibungssystemen (ETS) und komplexen Analyse- und Synthesesystemen (KASS) (Lenders 1980) unterschieden. Mit ETS werden solche Beschreibungsansätze bezeichnet, denen im wesentlichen Operationen des Vergleichens, Zählens und Selektierens zugrunde liegen. Sie beziehen sich auf Sprachdaten, die auf gesprochene oder geschriebene Sprache als Texte zurückgehen. Zahlreiche der hier entwickelten Verfahren gehören heute zur Standard-Software von Textverarbeitungssystemen oder zum Routineeinsatz praktischer Anwendungsfälle. Ihre Anwendungsfelder liegen im Bereich der Sprachwissenschaften, der Philologien, der Informationswissenschaften und zahlreicher wissenschaftsexterner Bereiche. Hierzu gehören statistische bzw. mathematische Verfahren innerhalb des „Information Retrieval“ und der Statistischen Linguistik ebenso wie ausgefeilte Silbentrennprogramme in automatischen Setzverfahren.

Die CL-orientierte *Lexikographie* stellt sich nach ihrem Forschungsstand als vielschichtiges Anwendungsfeld dar. Hier ist die Erstellung von Indizes, Konkordanzen und Wortlisten als Hilfsmittel der Textkritik und -edition zu nennen. Maschinenlesbare Wörterbücher (Hess/Brustkern/Lenders 1983) unterschiedlicher Ausprägungen und Funktionen werden in vielen Systemen mit Sprachverarbeitungskomponenten eingesetzt, zum Beispiel in Datenbanksystemen der Terminologie-Forschung, in Informationssystemen zur Unterstützung der Indexierung und Recherche, in Frage- Antwort-Systemen zur Steuerung natürlich sprachlicher Abfragen zu Datenbanken, in Dialogsystemen zur Strukturierung von Wissen und in MÜ-Systemen zur Unterstützung des Übersetzungsprozesses. Darüber hinaus bieten sich Anwendungsmöglichkeiten in der Bürokommunikation zur automatischen Fehlerkorrektur und im computerunterstützten Unterricht zur Lernüberwachung.

Mit der Bezeichnung „komplexe Analyse- und Synthesesysteme“ werden Problemformen der Angewandten Automation charakterisiert, die unter computermethodologischen Gesichtspunkten als Ansätze und Arbeiten zur Computersimulation natürlicher Sprache in der *Mensch-Maschine-Kommunikation* zu fassen sind. Sie sind konzipiert als automatische Regelsysteme, in denen Verfahren und Methoden der Analyse und Synthese zur Produktion von Textbeschreibung und Modellierung sprachlichen Vermögens inkorporiert sind. Eine solche Orientierung an einer methodologischen Problemspezifikation erlaubt es, CL-typische Fragestellungen und Anwendungsfelder aufzuzeigen, ohne damit gleichzeitig Abgrenzungsaspekte dieses Wissenschaftsfeldes gegenüber der Linguistik, der Informationslinguistik oder Künstlichen-Intelligenz-Forschung thematisieren zu müssen.

Das Spektrum der Forschungsarbeiten auf diesem Gebiet umfaßt alle Systemansätze, die als Realisierung des Ungeheuerschen M-C Modells aufzufassen sind. Ansätze und Verfahren einer solchen simulationsorientierten Behandlung natürlicher Sprache lassen sich aufzeigen im Problemfeld der Automatischen Spracherkennung und -synthese, der Entwicklung von MÜ-Systemen, der Behandlung des Automatischen Indexing und Abstracting in Informationserschließungssystemen und der Konstruktion von Frage-Antwort- und Dialogsystemen.

5. Resumée

Die methodologische Fundierung der CL wird in einem *eigenen Forschungsparadigma* ’Sprachdatenverarbeitung’ begründet, aus dem sich ein Spektrum möglicher Theoriebildungen über Sprache entwickeln läßt. Die Formate der Theoriebildung sind ebenfalls methodologisch bestimmt und lassen sich in die Theorietypen Simulation und Deskription zusammenfassen. Im Rahmen einer solchen Methodologie ergeben sich — in bezug auf Problembereiche des Untersuchungsgegenstandes ‚Sprache’ — Problemspezifikationen besonderer Art, aus denen dann Problemformulierungen in Sach- und Anwendungsfeldern der CL näher bestimmt werden.

Eine solche Wissenschaftsfundierung stellt das Verhältnis der CL zu den Nachbar-

wissenschaften in einen anderen Rahmen: nicht die gegenseitige Abgrenzung, sondern der *wechselseitige Austausch von* Ergebnissen und Erfahrungen stehen im Vordergrund. Er verbindet diejenigen Disziplinbereiche mit der CL, die — wenn auch mit unterschiedlicher Ausrichtung und Zielsetzung — diesem Forschungsparadigma verpflichtet sind.

6. Literatur (in Auswahl)

Barr/Feigenbaum/Cohen 1981 · Bátori 1977a · Bátori 1981 a · Bátori 1982 b · Bátori 1986 · Cercone/McCalla 1986 · Christaller/Metzing 1979/1980 · von Hahn 1978 · Hayes 1984 · Hays 1976 · Hess/Brustkern/Lenders 1983 · King 1983 · Krallmann 1968 · Lenders 1980 · Lutz/Schmidt 1982 · Metzing 1982 · Palmer 1984 · Papp/Szepe 1976 · Schnelle 1982 · Straßner 1977 · Ungeheuer 1971 · Wahlster 1982 b · Winograd 1983.

Dieter Krallmann, Essen
(Bundesrepublik Deutschland)

6. Computerlinguistik und die Theorie der formalen Sprachen

1. Einleitung

Gegenstand der Linguistik ist die Beschreibung der natürlichen Sprachen unter verschiedensten Aspekten (synchronen, diachronen u. a.), Aufgabenbereich der Computerlinguistik sind die maschinelle Analyse natürlich-sprachlicher Texte und die Simulation der Sprachanwendung. Die Computerlinguistik stützt sich dabei auf linguistische Beschreibungen, die unter Zugrundelegung mathematischer Modelle formalisiert sind. Die Frage ist nun, welche Rolle die Theorie der formalen Sprachen, die eine sehr allgemeine mathematische Theorie ist, in diesem Bezugsrahmen spielt.

Die Verbindung von Computerlinguistik, Theorie der formalen Sprachen und Gebieten der Linguistik, besonders der Syntax, erklärt sich zum einen durch die Entwicklung dieser Disziplinen, die entscheidend durch Noam Chomsky geprägt wurde. (Siehe unten 2.) Zum anderen ist die Theorie der formalen Sprachen von der Linguistik und der Computerlinguistik als empirischer bzw. angewandter Wissenschaft a priori unabhängig. Daher ist zu fragen, was die Motive dafür sind, daß sie als eine für diese Gebiete so wichtige Theorie angesehen wird, insbesondere inwiefern sie für Fortschritte in der Computerlinguistik wesentlich ist. Auch hierzu in 2. einige allgemeine Gesichtspunkte. In 3. wird die immer wieder gestellte Frage behandelt, welcher formalsprachliche Typus für die Beschreibung natürlicher Sprachen am besten zugrundezulegen ist. Die Antwort(en) hierauf (in 4.) ist (sind) wichtig für die Konstruktion computerlinguistischer Systeme, da sich daraus Konsequenzen für die Lösbarkeit vieler Probleme und die Komplexität von Verfahren ergeben, d. h. Konsequenzen für die praktische Entwicklung und Ausführbarkeit von Programmen.

2. Linguistik, Theorie der formalen Sprachen und Computerlinguistik

2.1. Entwicklung

Die Theorie der formalen Sprachen entstand in enger Verbindung mit demjenigen Zweig der Linguistik, der unter den Namen 'generative Grammatik' bekannt ist. Der Begründer beider ist Noam Chomsky (Chomsky 1956; 1957; 1959).

In der Linguistik setzt Chomsky an Positionen des amerikanischen Strukturalismus an, die er in neuer Weise interpretiert und weiterentwickelt. Dazu gehören die Konstituentenanalyse (Bloomfield 1933; Harris 1951) und der seit Harris (1952) auftretende Begriff der syntaktischen Transformation. Die Theorie der formalen Sprachen, die nun als mathematische Metatheorie für linguistische Beschreibungen entsteht, orientiert sich an Theorien über Zeichensysteme allgemeinster Art, die aus der Logik und Mathematik bekannt sind, — es sei besonders auf die Arbeiten von Thue, Turing, Kleene und Post hin-