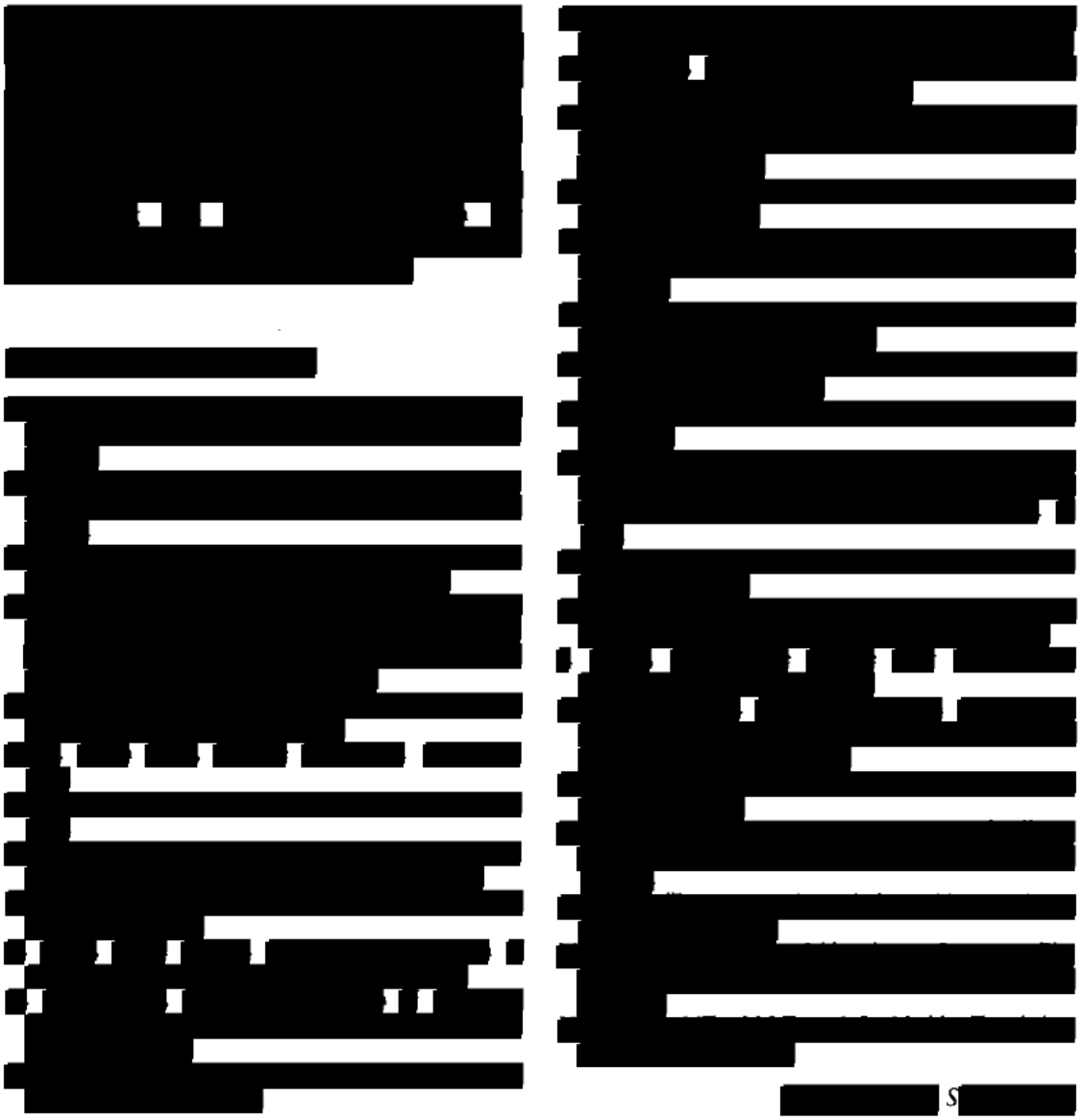

99. Computerlinguistik

1. Linguistik und Datenverarbeitung
2. Maschinelle Sprachanalyse
3. Bearbeitung großer Datenmengen
4. Mensch – Maschine – Kommunikation
5. Bibliographie (in Auswahl)

1. Linguistik und Datenverarbeitung

Die Computerlinguistik (linguistische Datenverarbeitung, automatische Sprachbearbeitung) (im folgenden CL) wird je nach Standpunkt und Abgrenzung des Objektbereiches als Zweig der angewandten Informatik (Hays 1967), als Zweig der angewandten Linguistik (Bott 1970, ALPAC-Report, Systematik des vorliegenden Lexikons), als Abkömmling der kognitiven Psychologie (Wilks 1972), als Teildisziplin der Kommunikationswissenschaft (Ungeheuer 1971) oder aber pragmatisch als Menge der Aktivitäten verstanden, die sich mit Linguistik und Datenverarbeitung oder Sprache und Datenverarbeitung befassen und damit eher durch Äußerlichkeiten als durch systematische Gegebenheiten zu einer Einheit werden (Dietrich/Klein 1974, vgl. die Diskussion in Bátori 1977).

Versucht man, das Feld über das Verhältnis von Linguistik und Computerwissenschaft (Informatik) zu erfassen, so ergibt sich folgendes Bild. Insbesondere in den 50er und frühen 60er Jahren haben sich Linguistik und Informatik teilweise gemeinsam entwickelt. Die Entwicklung von z. T.

sehr komplexen formalen Sprachen zur Erfassung von Problemen aus den verschiedenen Wissenschaften (Programmiersprachen) und von Algorithmen zu ihrer Analyse und Synthese sowie zur Übersetzung zwischen formalen Sprachen ging Hand in Hand mit der Entfaltung der algebraischen Linguistik, als deren Abkömmlinge die verbreitetsten Grammatikmodelle der neueren Sprachwissenschaft zu gelten haben (Floyd 1964).

Vielfältig ist der Einfluß, den die linguistische Datenverarbeitung selbst auf die Entwicklung der Sprachwissenschaft nehmen kann. Diskutiert wurde z. B. eine Mechanisierung oder Teilmechanisierung des linguistischen Forschungsprozesses (Karlgren/Brodda 1976). Das Testen und Vergleichen von umfangreichen und damit unübersichtlichen Grammatiken auf Adäquatheit und Konsistenz, aber auch auf Einfachheit und Effektivität der Verarbeitbarkeit hin ist erklärtes Ziel vieler Analyseprogramme seit den 60er Jahren (Londe/Schoene 1968, Friedman 1971, vgl. auch Pause 1976). Ein etwas anderer Aspekt steht im Vordergrund, wenn die Datenverarbeitung benutzt wird, um große Materialmengen zu sortieren und auszuwerten, wie sie etwa in der Lexikographie anfallen (Josselson 1967). In allen diesen Fällen wird die Datenverarbeitung im Grundsatz für die Linguistik nicht anders eingesetzt, als sie für andere Wissenschaften auch eingesetzt werden kann. Schließlich hat die CL der Sprachwissenschaft zu theoretischen Impulsen verholfen, insbesondere aufgrund von Erfahrungen in den Bereichen maschinelle Syntaxanalyse (2) und künstliche Intelligenz (4). Gliedert man das Feld in Aufgaben- oder Anwendungsbereiche, die die Bearbeitung von Sprache mit Hilfe von Rechenmaschinen betreffen, so stellt sich heraus, daß die Linguistik nur zum Teil eine Rolle für die Computerlinguistik spielt. Während in manchen Bereichen, namentlich der maschinellen Übersetzung (vgl. Art. 98), Fortschritte augenblicklich nur erzielbar erscheinen, wenn linguistische Erkenntnisse berücksichtigt werden (Josselson 1971, Stachowitz 1973), ist der Nutzen der Linguistik in anderen Bereichen, etwa bei den Informationswissenschaften, unklar (Sparck Jones/Kay 1976) oder, wie bei den Sprachverstehensmodellen der künstlichen Intelligenz, nach wie vor umstritten (Computational Semantics, Eisenberg 1976).

Die folgenden Abschnitte beschäftigen sich mit der ausdrücklich linguistisch fundierten Bearbeitung von Sprachmaterial mit Hilfe von Computern (2), mit der Bearbeitung von Sprachmaterial, bei dem der quantitative Aspekt im Vordergrund steht (3) und mit der teilweise über linguistische Konzeptionen hinausgehenden Sprachbearbeitung im Rahmen der Mensch-Maschine-Kommunikation (4).

Eine gut lesbare Übersicht zur Computerlinguistik ist Dietrich/Klein 1974, einen Literaturbericht gibt Straßner 1977. Der Stand wichtiger Aktivitäten in der Bundesrepublik und der Schweiz ist in ‚Kolloquium zur Lage [...]' dokumentiert.

2. *Maschinelle Sprachanalyse*

Unter maschineller Sprachanalyse wird meist der an der Wissenschaftssystematik der Linguistik orientierte Zweig der CL verstanden („Linguistik mit dem Computer"). Die maschinelle Sprachanalyse bemüht sich also darum, die Ergebnisse linguistischer Arbeit „auf den Computer zu bringen". Deshalb ist es auch nicht möglich, eine allgemeine Zielsetzung für die maschinelle Sprachanalyse anzugeben.

Im phonetisch-phonologischen Bereich sind vor allem Arbeiten zu nennen, die sich mit der Überführung von akustisch gegebenen Sprachsignalen in phonetische, phonologische (oder auch graphematische) Repräsentationen beschäftigen, die also ein kontinuierliches Signal in eine Folge von diskreten Signalen umsetzen (automatische Spracherkennung, IKP 1971, Speech Communication) und damit die Voraussetzung für die eigentliche Sprachverarbeitung auf Digitalrechnern schaffen. Obwohl das Ziel der automatischen Spracherkennung eine Analyse auf der Ebene der Lautsegmente ist, ist sie praktisch nicht ohne Information von höheren Beschreibungsebenen (lexikalisch-syntaktisch und semantisch) durchzuführen.

Arbeiten zur Computermorphologie haben meist nicht die Realisierung isolierter morphologischer Analyse- und Synthesealgorithmen zum Ziel, sondern sind morphosyntaktisch motiviert. Entweder in dem Sinne, daß die Flexionsmorphologie als Voraussetzung zur syntaktischen Analyse (Kongruenz und Rektion) betrieben wird (IdS 1974) oder in dem, daß im Bereich der Derivationsmorphologie mit tiefenstrukturellen Analysen gearbeitet wird (vgl. Höppner 1978). Als ein besonderer Zweig der maschinellen morphologischen Analyse kann die automatische Lemmatisierung (Zuordnung von Wortformen aus Texten zu ihren Lemmata) angesehen werden (vgl. 3.2.).

In der Computerlinguistik hat es relativ wenige Entwürfe von Systemen zur linguistischen Semantik gegeben. Beispiele sind Friedman 1973 und Dietrich 1973 für Analysen im Sinne der Kasusgrammatik, Stachowitz 1973 für die Verwendung einer Katz/Fodor-Semantik und neuerdings (nachdem es bereits in den 60er Jahren eine von der Linguistik weitgehend unabhängige Logikphase in der automatischen Sprachbearbeitung gegeben hatte) etwa Habel/Schmidt 1978 und Schmidt/Schneider 1978 für die Verwendung einer modallogischen Semantik. Der Hauptgrund für

die geringe Wirksamkeit der linguistischen Semantik ist, daß Semantiktheorien fast stets als Teil integrierter Grammatikmodelle entwickelt worden sind, das Hauptinteresse der Computerlinguistik sich jedoch auf die jeweilige syntaktische Komponente richtete (vgl. unten). Semantische Analysen wurden vor allem im Rahmen von Projekten der künstlichen Intelligenz durchgeführt, hier jedoch weitgehend unabhängig von der linguistischen Semantik (vgl. 4 sowie Eisenberg 1976).

Verfahren zur automatischen syntaktischen Analyse (Parsing) sind für viele der einschlägigen neueren Grammatiktypen sowie zahlreiche Varianten davon realisiert oder zumindest entworfen worden. Ersetzungsgrammatiken sind intensiv z. B. von Kuno und Öttinger (1963) sowie von Greibach (1964, 1965) bearbeitet worden. Dependenzgrammatische Ansätze gehen meist auf die Arbeiten von Hays (1964), kategorialgrammatische auf die von Bar-Hillel (1964) zurück. Verfahren für die schwerer zugänglichen Transformationsgrammatiken wurden z. B. von Friedman (1969, 1971), Petrick (1965) und Kuno (1969) vorangetrieben.

Neben der Syntaxanalyse im Sinne eines bestimmten Grammatiktyps bestand immer auch ein großes Interesse an der Entwicklung allgemeiner Analyseverfahren, die für verschiedene Grammatiktypen verwendet werden konnten. Besondere Verbreitung erlangten der Cocke-Algorithmus, bei dem zu jeweils zwei Symbolen (nur binäre Verzweigungen zugelassen) sämtliche nach den Regeln der Grammatik möglichen Reduktionen aufgesucht und geeignet gespeichert werden, sowie die prädiktive Analyse, bei der aus der Kategorie eines Elementes auf die möglicherweise mit ihm zu einer Konstituente verketteten Elemente geschlossen wird (vgl. die Übersicht in Kuno 1969). Ein neueres, sehr allgemeines Verfahren zur syntaktischen Analyse und Generierung ist in Kaplan 1973 beschrieben. Die Gründe für Umfang und Intensität, die die Arbeiten zur maschinellen Syntaxanalyse insbesondere in den späten 50er und den 60er Jahren auszeichnen, sind vielfältig. Zunächst war die Syntax für die Linguistik selbst zentral und daher der nächstliegende Gegenstand für die Computerlinguistik. Die vorherrschende Auffassung, daß eine Syntax bzw. Grammatik als Menge von Regeln anzusehen sei, die mengentheoretisch rekonstruierbar sind und insgesamt (als Regelwerk) zwar nicht einen eindeutig festliegenden Algorithmus darstellen, aber die ‚Realisierung' durch einen Algorithmus nahelegen, forderte zur Programmierung vorliegender Syntaxen geradezu heraus. Die zentrale Stellung der Syntax in der Linguistik entsprach auch der Auffassung, daß wichtige Gebiete der angewandten Linguistik im wesentlichen oder vollständig mit syntaktischen Mitteln zu bewältigen seien. Aus dem Bereich der Computerlinguistik sind hier vor allem die bereits erwähnte Beschreibung und maschinelle Übersetzung von Programmiersprachen (Compilerbau), die Mensch-Maschine-Kommunikation (Dialogsysteme), die automatische Übersetzung natürlicher Sprachen und die maschinelle Dokumentation zu nennen.

Im Bereich der linguistischen Theoriebildung selbst spielt der automatische Grammatiktester zwar eine wichtige, letztlich aber auf die eines technischen Hilfsmittels beschränkte Rolle. Dagegen hat z. B. die Tatsache, daß man zur automatischen syntaktischen Analyse neben der Grammatik selbst immer auch ein Verfahren für die Handhabung der Grammatik (Algorithmus) zur Verfügung haben muß, insofern Folgen für die Theorie gehabt, als dadurch die Reflexion über das Verhältnis des statischen und des dynamischen Teils eines linguistischen Verarbeitungsmechanismus mit ihren Folgen für das Verhältnis von Kompetenz und Performanz in Gang gekommen ist (vgl. z. B. Winograd 1972, Bátori 1976, 1978). Weitere wichtige Einzelpunkte betreffen die von der CL ausgegangene Diskussion über die Umkehrbarkeit von Transformationsgrammatiken (Petrick 1965, Pause 1976) sowie die Arbeiten über die Äquivalenz von Grammatiken verschiedenen Typs (Bar-Hillel 1960, Greibach 1964, Gaifman 1965).

3. Bearbeitung großer Datenmengen

3.1. Dokumentationssysteme

Größere Mengen gleich oder ähnlich strukturierter Daten werden maschinell mit Hilfe sog. Datenbanksysteme verwaltet. Ein Datenbanksystem enthält i. a. eine Komponente zur Speicherung von Daten, einen Datenspeicher (Datenbank) und eine Komponente zur Rückgewinnung der Daten aus der Datenbank. Datenbanksysteme werden seit langem in sehr verschiedenen Gebieten eingesetzt, etwa in Bibliotheken, zur Verwaltung von großen Warenlagern, zur Erfassung von Personaldaten usw. Nicht alle Datenbanksysteme gehören in das Arbeitsgebiet der CL. Der für uns wichtigste Typ von Datenbanksystemen sind die Dokumentationssysteme (DS) (Salton 1968, Niedermeyr 1976). DS dienen zur Verwaltung von Dokumenten, die in natürlicher Sprache formuliert sind, z. B. wissenschaftliche Literatur, Gesetzestexte, Gerichtsurteile, historische Quellentexte. Ein vollständig automatisches DS kann mit der Sprachwissenschaft an drei Stellen Berührungspunkte haben (Sparck-Jones/Kay 1976): (1) die Analyse der Dokumente zur Gewinnung der Information, die gespeichert werden soll (Indexierung oder Indexing) ist eine Analyse natürlicher Sprache (2) die extrahierte Information wird häufig in einer

elaborierten formalen Sprache formuliert (Indexierungssprache) (3) der Zugriff zur gespeicherten Information (Retrieval) kann über natürliche Sprache erfolgen.

3.1.1. Indexierung

Obwohl die Indexierung immer eine Analyse natürlicher Sprache ist, bedient man sich dabei keineswegs überwiegend linguistisch fundierter Verfahren. Die Art der Analyse richtet sich ganz wesentlich nach der Indexierungssprache. Die Indexierungssprache soll es erlauben, einen Extrakt des Inhalts der bearbeiteten Dokumente darzustellen. Bei vielen DS besteht das Lexikon der Indexierungssprache (Thesaurus) lediglich aus einer wenig verbundenen Menge von Wortformen oder Wörtern (Schlüsselwörter, Deskriptoren). Bei der Indexierung werden die Dokumente nach Deskriptoren abgesucht. Das Ergebnis der Suche wird statistisch ausgewertet (Maron/Kuhns 1960, Salton 1968, 1970). Die Relevanz eines Deskriptors für den Inhalt eines Dokumentes kann etwa danach gewichtet werden, wie häufig er dort im Vergleich zu anderen Dokumenten auftaucht und wie häufig er gemeinsam mit anderen Deskriptoren auftaucht.

Eine Anreicherung der Indexierungssprache soll die Möglichkeiten verbessern, den Inhalt eines Dokumentes genauer wiederzugeben, ohne daß dazu mehr Speicherplatz verbraucht wird als bei einfachen Stichwortverfahren. Einige der verwendeten Techniken sind Speicherung von Schlüsselwörtern mit einem bestimmten Kontext (KWIC = keyword in context), sorgfältige Auswahl der Deskriptoren für eine Klasse von Dokumenten (etwa wissenschaftliche Texte aus einem Fachgebiet), Kennzeichnung morphosyntaktischer Klassen (Lemmatisierung von Wortformen, Zerlegung von Komposita), Kennzeichnung syntaktischer Relationen zwischen Deskriptoren, Kennzeichnung semantischer Klassen und insbesondere Kennzeichnung semantischer Relationen (Antonymie, Hyponymie, Bedeutungsüberschneidung verschiedener Art). (Coyaud 1972, Soergel 1974).

Eine vergleichende Bewertung konkurrierender Analyseverfahren ist bisher nicht gelungen (Sparck Jones/Kay 1976), insbesondere ist nicht gezeigt worden, daß linguistisch fundierte Systeme besser arbeiten als statistisch orientierte (speziell zur Syntax z. B. Salton/Lesk 1968).

3.1.2. Retrieval

Beim document retrieval kommt es darauf an, daß ein möglichst hoher Anteil der Dokumente, die ein System auf eine Frage hin nennt, im Sinne der gestellten Frage von Bedeutung ist (precision), und daß andererseits möglichst viele von den Dokumenten gefunden werden, die für die gestellte Frage relevant sind (recall). Recall und precision sind einander in der Regel umgekehrt proportional und liegen in der Praxis bei 0.5 oder etwas darüber (Optimum für beide 1.0) (Salton 1968, Robertson/Sparck Jones 1976). Die Effektivität der Suche nach relevanten Dokumenten wird häufig dadurch gesteigert, daß die Struktur der Datenbank an die Struktur der Indexierungssprache angepaßt wird. Beispielsweise können Dokumente entsprechend den in der IS gekennzeichneten semantischen Relationen zu Blöcken (clustern) zusammengefaßt werden (Niedermeyr 1976). Datenbankstrukturen dieser Art machen nicht nur die Suche effektiver, sie erleichtern auch die Interaktion zwischen System und Benutzer, etwa indem das System anbietet, die gestellte Frage zu präzisieren, zu erweitern oder zu paraphrasieren (feedback) (Salton 1968, Attar/Fraenkel 1977).

Ein weites Feld für die CL eröffnet sich, wenn die Interaktion zwischen Benutzer und System in natürlicher Sprache ablaufen soll. Ob ein Datenzugriff über natürliche Sprache möglich und wie restringiert diese Sprache sein soll, richtet sich wesentlich nach dem Benutzerkreis (vgl. dazu die in Kuhlen 1978 zusammengestellte Literatur über künftige Entwicklungen im Dokumentationswesen). Entscheidend ist weiter das Verhältnis von Benutzersprache und Indexierungssprache. Die Benutzersprache muß mindestens so reich sein, daß Information unter allen den Gesichtspunkten erfragt werden kann, die die Struktur der IS bestimmen. Entsprechend hat sich die Analyse der Benutzersprache an dem zu orientieren, was in der IS ausgedrückt werden kann.

Dokumentationssysteme, die einen Datenzugriff in natürlicher Sprache erlauben, sind den Frage-Antwort-Systemen zuzurechnen (Simmons 1970). Der wesentliche Unterschied zu den in der künstlichen Intelligenz entwickelten Frage-Antwort-Systemen (vgl. 4) besteht darin, daß bei den ersteren die Effektivität leitender Gesichtspunkt ist, während letztere Sprachverstehen simulieren wollen. Erstere suchen nach Dokumenten, die einen bestimmten Inhalt haben (Referenz-Retrieval), letztere direkt nach Inhalten (Fakt-Retrieval). Dennoch wird eine zukünftige Synthese beider Typen von Dialogsystemen nicht ausgeschlossen (Smith 1976, Kuhlen 1978).

3.2. Wörterbucharbeiten

Maschinell gespeicherte Wörterbücher haben gegenüber konventionellen Wörterbüchern (vgl. Art. 94) den Vorzug, daß sie schnell und systematisch gelesen, bearbeitet und ausgewertet werden können. Sie werden in der Regel für bestimmte, relativ wohldefinierte Zwecke erstellt und erfas-

sen die Wörter oder Wortformen einer vorgegebenen Menge von Texten (Corpuswörterbücher). Nach der zu einem Stichwort bereitgestellten Information unterscheidet man Indizes und Konkordanzen. Konkordanzen enthalten zu jeder Belegstelle einen genau festgelegten Teil des Kontextes, in dem das Stichwort an der jeweiligen Stelle steht. Damit wird es möglich, anhand des Wörterbuches allein Stichwortvorkommen nach formalen und inhaltlichen Gesichtspunkten zu vergleichen. Konkordanzen werden vor allem für literarische Corpora erstellt (Wisbey 1967, Spevack 1968/70). Indizes gibt es zu allen Textgattungen. Sie können neben dem Belegstellennachweis Angaben über absolute und relative Häufigkeit, Rang und Wortklasse des Stichwortes enthalten (Lutz 1971, Klein/Zimmermann 1971). Eine strenge Unterscheidung von Indizes und Konkordanzen ist nicht immer möglich oder sinnvoll. Beide Typen von Wörterbüchern können häufig für die gleichen Zwecke verwendet werden, auch wenn die Information, die ein Index explizit enthält, aus einer Konkordanz erst errechnet werden muß und umgekehrt. Corpuswörterbücher werden in mehreren Reihen systematisch publiziert (Indices zur deutschen Literatur, Deutsche Wortindices) und in maschinenlesbarer Form zur Verfügung gehalten (z. B. beim IKP in Bonn und IdS in Mannheim). Typische textwissenschaftliche Analysen, die auf Corpuswörterbüchern aufbauen, fragen nach Worthäufigkeiten, dem Verhältnis von Wortklassen zueinander, Wortlängen, Satzlängen, Satzkomplexitäten, Häufigkeiten bestimmter Konstruktionen, Übergangswahrscheinlichkeiten zwischen unterschiedlichen Einheiten usw. (vgl. Literatur und Datenverarbeitung).

Bei der maschinellen oder maschinenunterstützten Herstellung von Wörterbüchern (automatische Lemmatisierung) treten, ähnlich wie beim automatischen Indexing, sehr unterschiedliche Probleme auf. Je nach Typ des bearbeiteten Corpus und gewünschten Wörterbuches müssen Codierungsfragen gelöst (Vereinheitlichung der Texte, Homographenauflösung, Codierungskonventionen, Bereitstellung eines ausreichenden Zeichenvorrats) sowie morphosyntaktische und semantische Analysen durchgeführt werden (Wortklassenbestimmung, Flexionsklassenbestimmung, Desambiguierung, Bestimmung der Paradigmenzugehörigkeit). Die Entwicklung auf diesem Gebiet ist noch voll im Fluß (Rath 1971, Dietrich 1973, Droop u. a. 1976).

4. Mensch – Maschine – Kommunikation

Aufbau und Arbeitsweise von Systemen für die Mensch-Maschine-Kommunikation hängen stark von der jeweils vorgesehenen Verwendung bzw. dem zugrundeliegenden Erkenntnisinteresse ab. Trotzdem haben die meisten Dialogsysteme gewisse strukturelle Ähnlichkeiten miteinander, z. B. enthalten sie bestimmte Bausteine (Komponenten) mit vergleichbaren Funktionen (Wilks 1977). Ein dialogischer Zyklus umfaßt in der Regel die syntaktisch-semantische Analyse eines Eingabesatzes, seine meist über eine deduktive Komponente vermittelte Einordnung in ein Wissenssystem und seinen Bezug auf die Redesituation sowie eine dem entsprechende Generierung einer Antwort der Maschine. Der Umfang und die Differenziertheit einer syntaktischen Analyse sind von zwei Hauptfaktoren abhängig: (1) von der Komplexität, die auf semantischer Ebene erreicht werden soll, d. h. zumindest bis zu einem gewissen Grade ist die Komplexität des Gegenstandes, über den geredet werden soll, ausschlaggebend für die Komplexität der Syntax; (2) von der gewünschten Flexibilität des Systems. Die Erfordernisse bezüglich Flexibilität und Robustheit des Systems hängen wesentlich vom Benutzerkreis ab. Wir können also hinsichtlich der Gründe für die Komplexität der Syntax im Extrem zwei Typen von Systemen unterscheiden. Systeme, die eine komplizierte Syntax haben, weil sie eine komplizierte Semantik haben, streben nach eindeutigen Abbildungen zwischen Syntax und Semantik. Hierher gehören z. B. die frühen in erster Linie semantisch motivierten Systeme mit oft bewußt restringierter Ein- und Ausgabesprache (Raphael 1968) sowie tendentiell die ersten Mikrowelt-Systeme (Winograd 1972). Auf der anderen Seite stehen Systeme, deren syntaktische Fähigkeiten unabhängig vom Umfang der Semantik ausgedehnt werden. Wichtigste semantische Fähigkeit ist das Paraphrasieren (Schank 1977). Das extremste Beispiel für fast ausschließlich syntaktische Analyse ist Weizenbaum 1966.

Unter der semantischen Analyse eines natursprachlichen Satzes wird am allgemeinsten seine Übersetzung in einen oder mehrere Ausdrücke einer internen Repräsentationssprache verstanden. Damit kann im Sinne einer linguistischen Semantik je nach Aufbau des Systems und Typs der verwendeten Repräsentationssprache folgendes erreicht werden: (a) Begriffe werden in eine Struktur integriert, die Begriffe direkt (Quillian 1968, Simmons 1973) oder unter Verwendung lexikalischer Zerlegungen (HAM-RPM: v. Hahn u. a. 1976 und Wahlster/v. Hahn 1976, Schank 1977) vernetzt und insgesamt das lexikalisch-semantische System einer Sprache repräsentiert. (b) Sätze werden in Makrostrukturen integriert, die den Inhalt von Texten enthalten. Damit wird ein fortlaufender, kohärenter Dialog für die Maschine erst möglich (Schank 1975, Norman/Rumelhardt 1975, Hays 1977). (c) Sätze oder Teile von Sätzen lösen Such- und Deduktionsvorgänge über der gespeicherten Information aus. Damit können etwa bestimmte

Begriffe oder Sätze gesucht werden (z. B. als Antwort auf eine Frage) oder Präsuppositionen und Folgerungen einer Menge von Sätzen bestimmt werden (Schwarcz u. a. 1970, Charniak 1976).

Die Hauptschwierigkeiten beim Entwurf intelligenter Dialogsysteme liegen auf der semantischen Ebene einerseits in der mangelhaften Effektivität der Analyse und der Weiterverarbeitung ihrer Ergebnisse (Kombinatorische Explosion). Darüber hinaus sind aber auch Grundfragen strittig, die insbesondere das Verhältnis der semantischen zu einer weitergehenden Analyse betreffen. Zu nennen sind vor allem das Repräsentationsproblem und das Wissensproblem. Beide sind nicht unabhängig voneinander.

Das Repräsentationsproblem beinhaltet die Frage, wie komplexe Sachverhalte und ihre Veränderung maschinenintern darzustellen sind. Klassisch geworden ist in der AI die Unterscheidung von propositionalen und prozeduralen Ansätzen. Erstere erfassen Sachverhalte statisch-beschreibend, etwa in einer ausgearbeiteten formalen Sprache (logischer Ansatz, vgl. Simmons 1970, Habel/Schmidt 1978 und Schmidt/Schneider 1978) oder mit Hilfe einer – nicht notwendig logisch fundierten – Netzwerkdarstellung (Simmons 1973, Woods 1975, Wittig 1978). Letztere erfassen Sachverhalte dynamisch-beschreibend, indem z. B. der Ablauf von Ereignissen direkt in Form von Elementarhandlungen, die im Verlauf eines Ereignisses zu vollziehen sind, gespeichert wird (Winograd 1972, 1977). Häufig wird jetzt die Meinung vertreten, beide Ansätze unterschieden sich nicht prinzipiell, weil, in der Sprache der Informatik, die Unterscheidung von Daten und Programmen letztlich willkürlich sei (Hays 1977). Oder man macht geltend, daß beide Darstellungsweisen gleichzeitig benötigt werden (Winograd 1975). Unabhängig von der Unterscheidung propositional/prozedural wurde in den letzten Jahren mehr und mehr mit dem von Minsky (1975) explizierten Begriff des frame operiert. Ein frame ist eine Struktur, in der begriffliches Wissen, Wissen über Situationen und Zustände, Wissen über Ereignisse und Vorgänge auf verschiedenen Stufen der Allgemeinheit festgehalten ist. Frames enthalten Leerstellen, deren Füllung z. B. den Übergang von einer allgemeinen (typisierten) zu einer singulären Situation bedeuten kann. Frames können miteinander verbunden werden, einander aktivieren und ineinander überführt werden. Das frame-Konzept ist ein Versuch, aktuelles und allgemeines Wissen im richtigen Verhältnis zu halten (Kuipers 1975, Wilks 1977).

Das Wissensproblem beinhaltet vor allem die Frage, welches und wieviel Weltwissen eine Maschine braucht, um intelligent über einen Gegenstand sprechen zu können. Damit ist gleichzeitig die Frage nach der Abgrenzung von Semantik und Weltwissen aufgeworfen. Man unterscheidet sinnvoll zwei Modi des Gebrauchs von Weltwissen in Dialogsystemen: (a) Der Einsatz von Wissen zur Simulation sprachlicher Fähigkeiten im engeren Sinne, also insbesondere zur lexikalischen und grammatischen Desambiguierung, zum Feststellen von Sprechakttypen (Fraser 1977) und zum Erkennen und Gebrauch formaler und semantischer Mittel zur Herstellung von Text- bzw. Dialogkohärenz, insbesondere des Gebrauchs kataphorischer und anaphorischer Wörter (Charniak 1977, Schank 1975, Norman/Rumelhardt 1975). Das Weltwissen hat hier die Funktion, einen zumindest teilweise extraverbalen Kontext zu bilden oder zu ersetzen, der als Bezugsgröße für die jeweilige Fixierung des Sprachlichen benötigt wird. (b) Die Wissensbasis hat unmittelbar als Abbild der Welt zu gelten, über die die Maschine redet und in der sie – zumindest im Idealfall – als Roboter handelt. Das Wissen begrenzt damit nicht nur den Gegenstand, über den die Maschine sprechen kann, sondern es bestimmt auch, was über den Gegenstand und damit überhaupt gesagt werden kann. Die Tendenz von der sprechenden Rechenmaschine zum sprechenden Roboter (ansatzweise realisiert bei Winograd 1972, vgl. z. B. auch HAM-RPM) beruht auf der Erkenntnis, daß „Sprechen" unter zwei Aspekten in umfangreichere Systeme zu integrieren ist: einmal als ein spezieller Informationskanal neben anderen, insbesondere dem visuellen und dem taktilen (Minsky 1975, Hays 1977); das heißt: Information muß zwischen den verschiedenen Kanälen übersetzbar, jedenfalls aber aufeinander beziehbar sein (vgl. Art. 22). Zweitens ist Sprechen nur eine spezielle Form des Handelns. Sprachliches Handeln ist unabhängig von außersprachlichem nur begrenzt zu realisieren. Aus diesen Gründen besteht das Ziel des Entwurfes einer „sprechenden Maschine", die nichts kann als Sprechen, zumindest perspektivisch und als Aufgabe der Grundlagenforschung nicht mehr.

5. *Bibliographie (in Auswahl)*

R. Attar u. A. S. Fraenkel, Local Feedback in Full-text Retrieval Systems. In: JACM 24. 1977, 397–417.

ALPAC-Report, Language and Machines – Computers in Translation and Linguistics. Washington 1966.

Y. Bar-Hillel, On Categorial and Phrase Structure Grammars. 1960. In: Bar-Hillel 1964, 99–115.

Y. Bar-Hillel, Language and Information. Reading. Mass. 1964.

S. Bátori, Teleologie der Grammatiktypen: Generative Grammatik und Analysegrammatik. In: Maschinelle Sprachanalyse, 49–75.

S. Bátori, Linguistische Datenverarbeitung. Computerunterstützte Sprachforschung und EDV für Philologen. In: SDV 1. 1977, 2–11.

S. Bátori, On the Procedural Motivations of Natural Lan-

guage Structures. IBM Wiss. Zentrum Heidelberg, TR 78.08.006.
M. F. Bott, Computational Linguistics. In: Lyons, J. (Hrsg.): New Horizons in Linguistics. Harmondsworth 1970, 215–288. [Deutsch: Neue Perspektiven der Linguistik. Reinbek 1975.]
E. Charniak, Inference and Knowledge I und II. In: Computational Semantics, 1–21 und 129–154.
E. Charniak, Referenz und Fragebeantwortung in einfachen Erzählungen. In: Semantik und künstliche Intelligenz, 21–38.
Computational Semantics. Hrsg. von E. Charniak und Y. Wilks. Amsterdam 1976.
M. Coyaud, Les langues documentaires. Paris 1972.
Deutsche Wortindices. Hrsg. von L. E. Schmidt, Berlin 1970ff.
R. Dietrich, Automatische Textwörterbücher. Tübingen 1973.
R. Dietrich u. W. Klein, Computerlinguistik. Eine Einführung. Stuttgart 1974.
H. Droop, W. Lenders und M. Zeller, Untersuchungen zur grammatischen Klassifizierung und maschinellen Bearbeitung spätmittelhochdeutscher Texte. Hamburg 1976.
P. Eisenberg, Die Bedeutung semantischer Theorien für die künstliche Intelligenz. In: StL. 2. 1976, 28–43.
R. W. Floyd, The Syntax of Programming Languages A Survey. In: IEEE Transactions on Electronic Computers, vol. EC–13 (1964), 346–353. Auch in: Programming Systems and Languages. Hrsg. v. S. Rosen. New York 1967, 342–358.
B. Fraser, Pessimistische Ausblicke auf die Möglichkeit zur Verbesserung der Mensch-Maschine-Kommunikation. In: Semantik und künstliche Intelligenz, 39–58.
J. Friedman, A Computer System for Transformational Grammar. In: CACM 12. 1969, 341–348. Deutsch in: Maschinelle Sprachanalyse, 29–48.
J. Friedman, A Computer Model of Transformational Grammar. New York 1971.
J. Friedman, A Computational Treatment of Case Grammar. In: Approaches to Natural Language. Hrsg. v. J. Hintikka, J. M. E. Moravcsik u. P. Suppes. Dordrecht 1973, 134–152.
H. Gaifman, Dependency Systems and Phrase Structure Systems. In: IC 8. 1965, 304–337.
S. Greibach, Formal Parsing Systems. In: CACM 7. 1964, 346–353.
S. Greibach, A New Normal-Form Theorem for Context-Free Phase Structure Grammars. In: JACM 12. 1965, 42–52.
C. Habel u. A. Schmidt, Eine modallogische Repräsentationssprache zur Darstellung von Wissen. 13. Ling. Koll. Gent 1978.
W. v. Hahn, D. T. Henskes, W. Höppner u. W. Wahlster, HAM-RPM: ein Redepartnermodell als Simulationsprogramm. In: Grammatik. Akten des 10. Ling. Koll., Bd. 2. Tübingen 1976, 337–357.
D. G. Hays, Dependency Theory: a Formalism and Some Observations. Language 40. 1964, 511–525.
D. G. Hays, Introduction to Computational Linguistics. New York 1967.
D. G. Hays, Kognitive Netzwerke: Formen und Prozesse. In: Semantik und Künstliche Intelligenz, 86–112.
W. Höppner, Derivative Wortbildung in automatischen Analysesystemen. Diss. Hamburg 1978.
IDS, Zur maschinellen Syntaxanalyse I und II. Mannheim 1974.
IKP, Beiträge zur automatischen Spracherkennung. Hamburg 1971.
Indices zur deutschen Literatur. Hrsg. v. H. Schwerte u. H. Schanze. Frankfurt 1970ff.
H. H. Josselson, Automatic Translation of Languages Since 1960, A Linguistis View. In: Advances in Computers 11. 1971, 1–58.
H. H. Josselson, Lexicography and the Computer. In: To Honor Roman Jacobson II. Paris 1967.
R. M. Kaplan, A General Syntactic Processor. In: Natural Language Processing, 193–241.
H. Karlgren u. B. Brodda, Computer als Hilfsmittel bei der Lösung nichtformalisierbarer Probleme. In: Maschinelle Sprachanalyse, 9–28.
W. Klein u. H. Zimmermann, Index zu Georg Trakls Dichtungen. Frankfurt 1971.
Kolloquium zur Lage der linguistischen Datenverarbeitung. Hrsg. v. D. Krallmann. Essen 1978.
B. J. Kuipers, A Frame for Frames: Representing Knowledge for Recognition. In: Representation and Understanding, 151–184.
R. Kuhlen, Information Retrieval. Vortrag auf dem Kolloquium Methodological Problems in Automatic Text Processing. Bielefeld 1978.
S. Kuno, Computer Analysis of Natural Language. In: Proc. of Symposia in Applied Mathematics, AMS, Bd. 19, 1969, 52–111. Deutsch in: Maschinelle Sprachanalyse, 167–203.
S. Kuno u. A. G. Öttinger, Multiple Path Syntactic Analyzer. In: Information Processing 1962. Amsterdam 1963, 306–311.
Literatur und Datenverarbeitung. Hrsg. v. H. Schanze. Tübingen 1972.
D. L. Londe u. W. J. Schoene, TGT: Transformational Grammar Tester. In: Proc. AFIPS 1968. Spring Joint Comp. Conf., Part 1, 1968.
H. D. Lutz, Übersicht zur maschinellen Analyse altdeutscher Texte. In: ZDP 90. 1971, 336–355.
M. E. Maron u. J. L. Kuhns, On Relevance, Probabilistic Indexing and Information Retrieval. In: JACM 30. 1960, 216–243.
Maschinelle Sprachanalyse. Hrsg. v. P. Eisenberg. Berlin 1976.
M. Minsky, A Framework for Representing Knowledge. In: The Psychology of Computer Vision. Hrsg. v. P. H. Winston. New York 1975, 211–277.
Natural Language Processing. Hrsg. v. R. Rustin, New York 1973.
W. Niedermeyr, Dokumentation und Information Retrieval. In: Zur Theorie und Praxis des modernen Bibliothekswesens. Hrsg. v. W. Kehr, K. W. Neubauer u. J. Stoltzenburg. München 1976, 268–312.
D. A. Norman u. D. E. Rumelhardt, Explorations in Cognition. San Francisco 1975.
E. Pause, Adäquatheitstests und Syntaxanalyse. In: Maschinelle Sprachanalyse, 76–97.
S. R. Petrick, A Recognition Procedure for Transformational Grammar. Diss. MIT 1965.
M. R. Quillian, Semantic Memory. In: Semantic Information Processing, 227–270.
B. Raphael, SIR: a Computer Program for Semantic Information Retrieval. In: Semantic Information Processing, 33–145.
R. Rath, Probleme der automatischen Lemmatisierung (AL). In: ZPSK 24. 1971, 409–425.
S. E. Robertson u. K. Sparck Jones, Relevance Weighting of Search Terms. In: JASI 27. 1976, 129–146.

Representation and Understanding. Hrsg. v. D. G. Bobrow u. A. Collins. New York 1975.
G. Salton, Automatic Information Organization and Retrieval. New York 1968.
G. Salton, Automatic Text Analysis. In: Science 168. 1970, 335–343 (1970a).
G. Salton, Evaluation Problems in Interactive Information Retrieval. In: ISR 6. 1970, 29–44 (1970b).
G. Salton, The SMART Retrieval System. Englewood Cliffs 1971.
G. Salton u. M. Lesk, Computer Evaluation of Indexing and Text Processing. In: JACM 15. 1968, 8–36.
R. Schank, The Structure of Episodes in Memory. In: Representation and Understanding, 237–272.
R. Schank, Computer, primitive Aktionen und linguistische Theorien. In: Semantik und künstliche Intelligenz, 113–141.
A. Schmidt u. H. J. Schneider, Natürlich-sprachliche Frage-Antwort-Systeme – Bedeutung, Realisierung und Ausblick. In: Kolloquium zur Lage [. . .], 216–229.
D. Soergel, Indexing Languages and Thesauri. Los Angeles 1974.
Speech Communication with Computers. Hrsg. von L. Bolc. London 1979.
R. M. Schwarcz, J. F. Burger u. R. F. Simmons, A Deductive Question-Answerer for Natural Language Inference. In: CACM 13. 1970, 167–183.
Semantic Information Processing. Hrsg. v. M. Minsky. Cambridge (Mass.) 1968.
Semantik und künstliche Intelligenz. Hrsg v. P. Eisenberg. Berlin 1977.
R. F. Simmons, Natural Language Question Answering Systems: 1969. In: CACM 13. 1970, 15–30. Deutsch in: Maschinelle Sprachanalyse, 204–242.
R. F. Simmons, Semantic Networks. In: Computer Models of Thought and Language. Hrsg. v. R. Schank u. K. M. Colby. San Francisco 1973, 63–113.
L. C. Smith, Artificial Intelligence in Information Retrieval Systems. In: IPM 12. 1976, 189–222.
K. Sparck Jones u. M. Kay, Linguistik und Informationswissenschaft. München 1976.
M. Spevack, A Complete and Systematic Concordance to the Works of Shakespeare. Hildesheim 1968–1970.
R. Stachowitz, Voraussetzungen für maschinelle Übersetzung. Frankfurt 1973.
E. Straßner, Linguistische Datenverarbeitung (LDV). Anwendungsbereiche und Forschungsstand. In: Sprachwissenschaft 2. 1977, 433–470.
G. Ungeheuer, Linguistische Datenverarbeitung – die Realität und eine Konzeption. In: IBM Nachrichten 21. 1971, 688–694.
W. Wahlster u. W. v. Hahn, Einige Erweiterungen des natürlich sprachlichen AI-Systems HAM-RPM. In: Dialoge in natürlicher Sprache und Darstellung von Wissen. Workshop Freudenstadt 1976, 204–225.
J. Weizenbaum, ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine. In: CACM, vol. 9, Nr. 1, Jan. 66, 36–45.
J. Weizenbaum, Die Macht der Computer und die Ohnmacht der Vernunft. Frankfurt 1977.
T. Winograd, Understanding Natural Language. Edinburgh 1972.
T. Winograd, Frame Representations and the Declarative – Procedural Controversy. In: Representation and Understanding 185–210.
T. Winograd, Ein prozedurales Modell des Sprachverstehens. In: Semantik und künstliche Intelligenz, 142–179.
Y. Wilks, Grammar, Meaning, and the Machine Analysis of Language. London 1972.
Y. Wilks, Sprachverstehende Systeme in der künstlichen Intelligenz. Überblick und Vergleich. In: Semantik und künstliche Intelligenz, 180–230.
R. A. Wisbey, Vollständige Verskonkordanz zur Wiener Genesis. Berlin 1967.
T. Wittig, Zur Abbildung von Sachverhalten in einfache semantische Netze. In: SDV 1. 1978, 68–74.
W. A. Woods, What's in a Link: Foundations for Semantic Networks. In: Representation and Understanding, 35–83.

Peter Eisenberg, Hannover