10. Evaluierung komplexer Systeme in der Computerlinguistik: Verfahren und Anwendung

1. Umfeld und Abgrenzung

Im folgenden geht es um eine Untergruppe der Evaluierung von Computersoftware, um die Bewertung realisierter oder geplanter Algorithmen, in denen Sprachdaten bearbeitet werden. Ausgeklammert bleiben Systeme, die dem Sprach- oder Literaturwissenschaftler die Arbeit erleichtern; es geht somit immer um die Verbesserung von maschinellen Informationssystemen durch die Integration von Sprachverarbeitung (vgl. Krause 1984 a).

1.1. Erschwerende Faktoren, Zweckbestimmung und Theorieabhängigkeit

Versucht man Algorithmen der Computerlinguistik (CL) zu bewerten — und sei es nur für die eigene Orientierung —, erweist es sich rasch, daß man relativ ungesichertes Terrain betritt. Ein nach dem Selbstverständnis der Linguistik guter Vorschlag muß keine gute CL-Lösung sein (kann aber) — und umgekehrt. Dies ergibt sich daraus, daß der verfolgte Zweck ein anderer ist und CL-Lösungen in der Regel eingebettet sind in größere Systemzusammenhänge (genauer Krause 1984 a). Das damit notwendig werdende Umdenken erschwert den Zugang, vor allem für Linguisten. Hinzu kommt, daß es keine homogene Evaluierungsmethodologie gibt. Die Evaluierungsvorschläge kommen aus den verschiedensten Forschungsrichtungen. Psychologen, Soziologen, Kognitionsforscher, Mathematiker, Informationswissenschaftler und Informatiker haben Studien mit den ihnen vertrauten methodischen Instrumentarien und Denkansätzen durchgeführt. Einen zumindest partiellen Konsens über ein vernünftiges Vorgehen gibt es nur für einzelne Teilgebiete, nicht für CL-Algorithmen als ganzes. — Die Evaluierungsmethoden unterscheiden sich weiter nach den mit einer Bewertung verfolgten Zielen (die sich gegenseitig nicht ausschließen müssen): (1) Evaluierung kann z. B. in Unternehmen dazu dienen, die Auswahl zwischen konkurrierenden Produkten zu treffen. Liegen Verbesserungen und potentielle Weiterentwicklungen eines Softwarepaketes außerhalb der Möglichkeiten des Käufers (was in der Regel zutrifft), interessiert nur das Urteil „besser“ oder „schlechter“, nicht aber Fehlerursachen. — (2) Evaluierung kann quasi eine verlängerte Systembeschreibung sein; sie soll demonstrieren, was das eigene System leistet, wo die Grenzen liegen (vgl. z. B. Tennant 1979, 2). — (3) Evaluierung kann „Forschungsvehikel“ sein, das Instrumentarium, das zwischen alternativen Hypothesen entscheidet und in einem permanenten iterativen Prozeß die Entwicklung eines Fachgebietes vorantreibt. Letzteres steht hier im Mittelpunkt. Zentrales Bewertungskriterium für eine Evaluierungsmethode ist somit der Nutzen der mit ihr erreichbaren Ergebnisse für die schrittweise Weiterentwicklung einer speziellen Softwarelösung oder eines ganzen Fachgebietes (z. B. Hinweise auf die generelle Architektur von natürlichsprachlichen Mensch-Maschine-Schnittstellen).

Je nach der zugrunde gelegten Theorie über den Zusammenhang zwischen Mensch und Computer bei der Mensch-Maschine-Interaktion (MMI) und zwischen MMI und Sprachverarbeitung unterscheidet sich der Stellenwert von Evaluierungsstudien für die Fortentwicklung: (a) Geht man z. B. davon aus, daß die MMI am erfolgreichsten durch eine Simulation zwischenmenschlichen Verhaltens realisiert wird, hat der empirische Systemtest bestenfalls zweitrangige Bedeutung. Man weiß aus der Beobachtung zwischenmenschlichen Sprachverhaltens, was bei der MMI zu erwarten ist; Evaluierungsstudien kommt der Charakter einer Bestätigung und Funktionsüberprüfung zu. — (b) Zentral werden Evaluierungsstudien bei der theoretischen Grundannahme, die MMI lasse sich nicht generell als Abbild zwischenmenschlicher Kommunikation erklären, sondern laufe nach eigenständigen Gesetzmäßigkeiten ab (z. B. Zoltan/Weeks/Ford 1982;

Krause 1982 a und 1983; Zoeppritz 1985 a und 1985 b). In diesem Fall ist die Evaluierung das entscheidende Instrument, die Gesetzmäßigkeiten der MMI (und den Bereich ihrer Übereinstimmung mit der zwischenmenschlichen Kommunikation) herauszufinden. Erst die empirische Fundierung durch MMI-Evaluierung löst Zustimmungsbereitschaft für Aussagen zu CL-Lösungen aus.

Da der Stellenwert in hohem Maße den tolerierbaren Aufwand an Zeit und Geld für Evaluierungsstudien festlegt und es in bezug auf den notwendigen Aufwand deutliche Unterschiede zwischen den verschiedenen Bewertungsmethoden gibt, beeinflußt die theoretische Grundhaltung einer Forschergruppe — wenn auch indirekt — fast unvermeidlich die Auswahl der Evaluierungsmethode. Allerdings läßt sich in der Forschungsrealität nicht immer der Schluß ziehen, hoher Evaluierungsaufwand bedinge die theoretische Grundhaltung (b). Evaluierungsvorhaben können Ziele (mit)verfolgen, die im sozialen und politischen Umfeld eines Wissenschaftlers begründet sind. Wenn Evaluierungen Politiker oder Firmenleitungen vom Wert eines Systems überzeugen, sind sie dem Wissenschaftler, der Forschungsgelder benötigt, viel Mühe wert. Politiker wiederum scheinen mit Evaluierungen oft andere Ziele als die unter (1)—(3) angeführten zu verfolgen. Die Analysen von Floden/Weiner 1978 („Rationality to Ritual: The Multiple Roles of Evaluation in Governmental Processes„) dürften auch auf die MMI zutreffen. Danach haben Evaluierungsstudien — allein durch ihre Existenz, unabhängig von den erreichten Ergebnissen — ein hohes Konfliktlösungspotential (S. 13 f.), erhöhen die Zufriedenheit der beteiligten Benutzergruppen bei der Einführung von Neuerungen (S. 14 f.) und vermitteln die (notwendige) Suggestion, politische Entscheidungen fielen auf der Grundlage rationeller und wissenschaftlich abgesicherter (Evaluierungs-)erkenntnisse (S. 16 f.).

1.2. Komplexität und CL-Evaluierung

Versteht man das Zusammenwirken mehrerer Komponenten als wesentliches Merkmal von Komplexität, lassen sich die beiden Begriffe unter zwei Gesichtspunkten verknüpfen: (a) Die Komplexität kann systemseitig auftreten. Dies bedeutet je nach der eingenommenen theoretischen Grundhaltung etwas verschiedenes. Geht man davon aus, daß sich linguistische Analysen, ohne nennenswerte Nebeneffekte auszulösen, in maschinelle Informationssysteme integrieren lassen, liegt die CL-Komplexität zwangsweise in einem mehrgliedrigen Aufbau der CL-Komponente; komplexe Systeme sind dann z. B. solche, die nicht nur syntaktische, sondern semantische und/oder pragmatische Aspekte berücksichtigen. Sieht man dagegen, daß CL-Komponenten in der Regel nur ein Teilelement innerhalb größerer Systeme sind und daß es in der Regel nicht fruchtbar ist, CL-Teilelemente ohne ihren Systemkontext zu bewerten, gerade weil von einer hohen gegenseitigen Beeinflussung ausgegangen wird, muß sich CL-Komplexität aus der Komplexität des gesamten maschinellen Informationssystems ableiten. Systemkomplexität für CL-Lösungen läßt sich bei diesem Blickwinkel nicht mehr aus dem Umfang der „Linguistik“ im CL-Teilalgorithmus bestimmen. Entscheidend ist die Beeinflussung des Gesamtsystems. In diesem Sinne sind Dialogsysteme (z. B. auch Referenz-Retrievalsysteme) mit CL-Teilalgorithmen immer als potentiell komplexe Systeme zu betrachten. Da allein diese Sichtweise auf CL-Komplexität bei Evaluierungen sinnvoll erscheint, liegt sie hier zugrunde. — (b) Evaluierung selbst kann mehrschichtig sein. Neuere Entwicklungen zeigen einen klaren Trend zu einer komplexen Evaluierungsmethodologie. Den Schwächen der einzelnen Verfahren, die im Lauf der letzten Jahre immer deutlicher wurden, wird mit einem Verbund verschiedener Methoden begegnet. Der Zusammenhang zwischen komplexer Evaluierung und CL-Komplexität ist sehr stark. Dies scheint nur folgerichtig, wenn man bedenkt, welch verschiedene Aspekte (Psychologie, Arbeitsorganisation, Sprachverhalten, Informatik) bei der MMI zusammenwirken.

1.3. Beispielbereiche

Aus Platzgründen muß die Breite und die Problematik der Bewertung von CL-Lösungen exemplarisch demonstriert werden. Sucht man Bereiche, in denen Evaluierungsstrategien am vielschichtigsten ausgeprägt sind und an denen sich die in 1.1. und 1.2. genannten Probleme deutlich zeigen, fällt die Wahl fast zwangsläufig auf natürlichsprachliche Frage-Antwort-Systeme (FAS), ergänzt um komplexere Evaluierungsansätze bei der Texterschließung für Referenz-Retrievalsysteme. Im ersteren Gebiet hat sich eine große Vielfalt von Methoden herausgebildet, letzteres ergänzt sie durch statistisch-mathematisch geprägte Bewertungen auf der Grund-

lage von Relevanzurteilen. Bei beiden Gebieten ist heute das Dialogverhalten des Benutzers entscheidend, was in der Regel zu komplexen Informationsvorgängen führt.

2. Evaluierung von FAS

Der Zugang wird v. a. für die Zeit bis 1970 über die Diskussion einiger früherer Aufsätze gesucht; danach geben die durchgeführten Evaluierungsstudien eine gute Leitlinie.

2.1. Habitability und analytisches Denken

In dem Aufsatz von Watt 1968 spiegeln sich die grundsätzlichen Entwicklungslinien und -möglichkeiten von FAS bis in die heutige Zeit (vgl. auch Krause 1983), verbunden mit einem ersten Versuch, Meßgrößen für die Erfolgsbewertung festzulegen. Watt sieht die Grundproblematik bei der Entwicklung von FAS in der Divergenz zweier Forderungen: Einerseits soll der Sprachumfang eines FAS möglichst klein sein (Subset), damit es unterhalb der Komplexitätsbarriere bleibt und ökonomisch arbeitet. Andererseits muß der Benutzer in der restringierten Anfragesprache seinen Informationswunsch problemlos ausdrücken können. Letztere Entwicklungslinie, die zu einer Ausweitung der Fähigkeiten von FAS führt, hat sich bis heute ungebrochen behauptet. Auch die Begründungen für einzelne Erweiterungen haben sich seit Watt kaum geändert: Was im zwischenmenschlichen Dialog gebraucht wird, gilt zumindest auf der funktionalen Ebene als notwendig für die MMI. Wie schon bei Watt angedeutet, geht der Weg von der Semantik zur Pragmatik (vgl. auch Kuhlen 1985). Da dieser Standpunkt einerseits eine hohe Plausibilität hat und andererseits 'herrschende Lehre' verkörpert, bleibt der Wunsch, Erweiterungsabsichten durch Evaluierungsstudien zu untermauern, gering. Die Forderung nach Evaluierung stützt sich hier in der Regel auf den Wunsch nachzuweisen, daß das anvisierte Ziel erreicht wurde, und die Erkenntnis, daß immer nur ein Teilbereich der bei der zwischenmenschlichen Kommunikation beobachteten Phänomene in das FAS integriert werden kann. Zu testen bleibt somit, ob der gewählte Ausschnitt sich in Anwendungssituationen bewährt. — Sieht man letzteren Aspekt genereller, fällt er mit der Suche nach Restriktionsregeln in Watt 1968 zusammen. Welche sprachlichen Äußerungen gehören zum Subset und welche nicht? Gibt es Kriterien für diese Auswahl? Watt versucht vor dem Hintergrund der Transformationsgrammatik von Chomsky Reduktionsregeln für eine Subsetdefinition zu finden. Methodisch gesehen arbeitet er qualitativ-analytisch (zum Inhalt vgl. Krause 1982 a, 22—25). In enger Anlehnung an eine Sprachtheorie und unter Rückgriff auf die eigene Sprachkompetenz, die nicht aus der MMI, sondern aus der zwischenmenschlichen Sprachverwendung herrührt, nimmt er eine Bewertung alternativer Design-Entscheidungen vor. Diese Bewertungsmethode ist auch heute noch die verbreitetste. Daß sie isoliert angewendet schnell in Sackgassen führen kann, zeigt sich schon bei Watt 1968 (vgl. Krause 1982 a, 22 f.).

2.2. Übersetzungstests und Simulation

Die hier behandelten Methoden zeichnen sich dadurch aus, daß die Evaluierung ohne (experimentelles) System erfolgt.

2.2.1. Rosenbaum 1968

Rosenbaum 1968 ist die früheste empirische Evaluierungsstudie zum hier behandelten Bereich. Sie basiert auf der Subset-Problematik. Die generell für die MMI entwickelte „English Grammar II" wird danach getestet, ob der Subset erlernbar und relativ fehlerfrei anwendbar ist. In einem ersten Schritt werden Protokolle übersetzt, in einem zweiten formuliert der Proband Beobachtungen über eine im Film gesehene Dialogsequenz, formuliert somit, was sich nach der gedanklichen Durchdringung der Filmszene ergibt. Dies zeigt im Ansatz den heute noch gültigen Ausweg aus dem Zwang, in der Testaufgabe sprachliche Vorgaben kaum umgehen zu können: Wo möglich, ist die verbale durch andere Kommunikationsformen (z. B. Bilder) zu ersetzen; wo dies nicht durchführbar erscheint, sollte die vorgegebene sprachliche Formulierung nur in einem indirekten (deduktiven) Zusammenhang zur Benutzerformulierung stehen.

2.2.2. Simulierte natürlichsprachliche MMI

Sie ist bis heute eine der beliebtesten Evaluierungsmethoden. Das Grundmuster bleibt immer gleich: Das natürlichsprachliche Interface wird durch einen Menschen ersetzt. Technisch gesehen verbindet man zwei Bildschirme mit Tastatur, die in getrennten Räumen stehen. Der Benutzer gibt seine Frage über das Terminal ein (eventuell über zusätzliche Kanäle wie Graphik und gesprochene Sprache); der den Computer simulierende menschliche Dialogpartner empfängt die

Frage über sein Terminal und schickt die Antwort auf den Benutzerbildschirm zurück. Aus dem Dialogverhalten des Benutzers soll geschlossen werden, welche Fähigkeiten ein ideales Interface haben müsse. Der Grundgedanke, natürlichsprachliche MMI auf diese Weise zu simulieren, findet sich bereits als Teil des Turingtests (vgl. Turing 1950). Ein prototypisches Beispiel ist Malhotra 1975. Neuere Beispiele sind Sidner 1980 (vgl. auch Woods 1983 a), Bates/Sidner 1983 und Hofmann 1985. Der Hauptvorteil dieser Methode ist, daß weder Software- noch Hardwarelösungen technisch realisiert sein müssen und Störfaktoren ausgeschaltet bleiben. Der Benutzer wird zudem nicht durch Erfahrungen mit einem stark restringierten Interface dazu verführt, sich während des Tests suboptimal an die Schwächen des zu bewertenden experimentellen Systems anzupassen. Diese Vorteile verbinden sich mit einem relativ niedrigen Kosten- und Zeitwaufwand, was zusammengenommen die Beliebtheit dieser Evaluierungsmethode erklärt.

Allerdings lassen sich auch die Nachteile nicht wegdiskutieren. Im Kern handelt es sich um einen zwischenmenschlichen Dialog via Terminal. Fraglich bleibt schon, ob der den Computer simulierende Mensch bei sich wiederholenden Fragekonstellationen immer gleich antwortet und inwieweit der Benutzer bei diesem Testdesign nicht doch im Mensch-Mensch-Modus verbleibt (vgl. auch Zoeppritz 1986). Geht man davon aus, daß die MMI über eigenständige Gesetzmäßigkeiten verfügt (vgl. Krause 1982 a; 1983; 1984 a; Zoltan/Weeks/Ford 1982; Zoeppritz 1985 a), die zum größten Teil noch unbekannt sind, kann die MMI-Simulation vollends kein adäquates Evaluierungsinstrument sein. — Diese Überlegungen führen dazu, die simulierte MMI nur als erste grobe Annäherung anzusehen. So ist in Kelley 1984 die Simulation ein erster Schritt, dessen Ergebnisse bei der Entwicklung eines Prototypen berücksichtigt werden. Danach findet ein Benutzertest mit dem Prototypen statt; simuliert werden nur noch die fehlenden Teile; die Evaluierung wird bei jedem iterativen Bewertungsschritt realitätsnäher. — Man kann sich weiter fragen, ob die simulierte MMI wesentliche Vorteile gegenüber dem weniger aufwendigen Verfahren hat, Erkenntnisse aus dem zwischenmenschlichen Dialogverhalten auf die MMI zu übertragen. Beispiele aus neuerer Zeit sind Pollack/Hirschberg/Webber 1982 (vgl. auch Webber 1983 b) und Zoltan/Weeks/Ford 1982. Sieht man von der Auswahl der Dialogmitschnitte ab, bleibt als Unterschied der benutzte Kanal; bei der simulierten MMI wird der „richtige“ Kanal (meist Terminaleingabe) benutzt. Inwieweit verschiedene Kanäle Unterschiede im Dialog verursachen, ist selbst Gegenstand von Evaluierungsstudien nach der in Abschnitt 2 angeführten Methodik (Chapanis 1973 und 1975; Zoltan/Weeks/Ford 1982). Da sich deutliche Unterschiede bei der Verwendung verschiedener Kanäle ergeben — und dies Ergebnis auch aus linguistischer Sicht nicht überrascht (z. B. Telefonstil) —, erscheint es sinnvoll, zumindest die Kanalverfälschung aus den Tests herauszuhalten, zumal der Mehraufwand relativ gering ist.

2.3. Empirische Evaluierung mit Prototypen

Die Forderungen nach diesem Evaluierungstyp nehmen seit Mitte der 70er Jahre ständig zu (Petrick 1976, 327 f.; Woods 1977, 523; Waltz 1977, 150; Tennant 1979, 2 ff. und die Arbeiten zur USL-Evaluierung (vgl. 2.4.)). Daß es dennoch bis heute relativ wenige Studien dieser Art gibt, liegt an den schwierigen Voraussetzungen, nicht an der mangelnden Akzeptanz der Methode. — Die erste Systemevaluierung wurde 1971 mit LUNAR durchgeführt (vgl. Woods/Kaplan/Nash-Webber 1972, 5.2 f. und Woods 1977, 557—560). Anläßlich der „Second Annual Lunar Science Conference“ in Houston (Texas) konnten Geologen drei Tage lang Fragen (111 insgesamt) zum Mondgestein der Apollo-Mission an LUNAR richten. Die Fehlerraten und die qualitative Analyse der Fehler bilden die Grundlage für eine noch recht informelle Bewertung. Die Maße ’semantische Adäquatheit’ und ’fully habitability’ von Watt 1968 finden in veränderter Form Verwendung (vgl. Woods/Kaplan/Nash-Webber 1972, 5.6 und Woods 1977, 560—563). Von ähnlich informellen Systemeinsätzen berichten Codd et al. 1978, 82 und Codd 1978, 23—26 für RENDEVOUS (etwa 30 Personen lösten verbal vorgegebene Aufgaben) und Harris 1977 a, 308 für ROBOT.

Die Evaluierungen von Tennant 1979 (zur Planung vgl. Tennant 1978) zu den FAS „Automatic Advisor“ und PLANES gehören dagegen zur Kategorie der kontrollierten Tests. Benutzer waren Universitätsstudenten, die Aufgaben zu lösen hatten. Diese wurden so formuliert, daß die Benutzersprache möglichst zurücktrat (z. B. Tabellen, Diagramme

zum Auffüllen). Tennant 1979 verzichtet auf Fehlerzahlen, Anfragemengen und ähnliches, berichtet jedoch ausführlich über die Arten der Fehler, die vor allem die konzeptuelle Vollständigkeit des Systems betreffen. Die konzeptuelle und linguistische Vollständigkeit sind Meßwerte, die Tennant 1979, 4—8 denen der semantischen Adäquatheit und „fully habitability" von Woods/Kaplan/Nash-Webber 1972, 5.6 entgegensetzt (zur Diskussion vgl. Krause 1982 a, 25 ff.):

"The degree to which the concepts that are expected by a set of users can actually be found in the system's conceptual coverage is the CONCEPTUAL COMPLETENESS of the natural language processor, with respect to the set of users. Similarly, the degree to which the language of a set of users is appropriately analyzed by the system is the LINGUISTIC COMPLETENESS of the natural language processor with respect to that set of users" (Tennant 1979, 6).

Kontrollierte Tests sind v. a. bei der Evaluierung formalsprachlicher Benutzerschnittstellen, die hier ausgespart bleiben, beliebt (vgl. als Überblick Lehmann/Blaser 1979; Reisner 1981). Vergleichstests zwischen formalen und natürlichsprachlichen Abfragemöglichkeiten enthalten Greenblatt/Waxmann 1978; Small/Weldon 1977 und Shneiderman 1978 (vgl. auch Shneiderman/Mayer 1979 und die Kritik in Zoeppritz 1985 b und 1986).

2.4. Empirische Langzeitstudien und komplexe Evaluierungsmethodologie

Die hier dargestellten Studien kennzeichnen den heute erreichten Evaluierungsstand, sowohl im Hinblick auf die zugrundeliegende Methodologie als auch bezüglich der inhaltlichen Aussagen.

Die früheste ausgewertete Langzeitevaluierung eines Systemeinsatzes (12 Monate, 788 Anfragen) fand mit TQA (Transformational Query Answering) statt (vgl. Damerau 1979); Benutzer waren Mitarbeiter eines Planungsbüros. Die Auswertung diente vor allem der funktionalen Überprüfung (vgl. die Diskussion in Krause 1982 a, 83—85). Mehrstufige Konzepte finden sich dann bei der Evaluierung des FAS „User Specialty Languages" (USL); es ist das am extensivsten und am

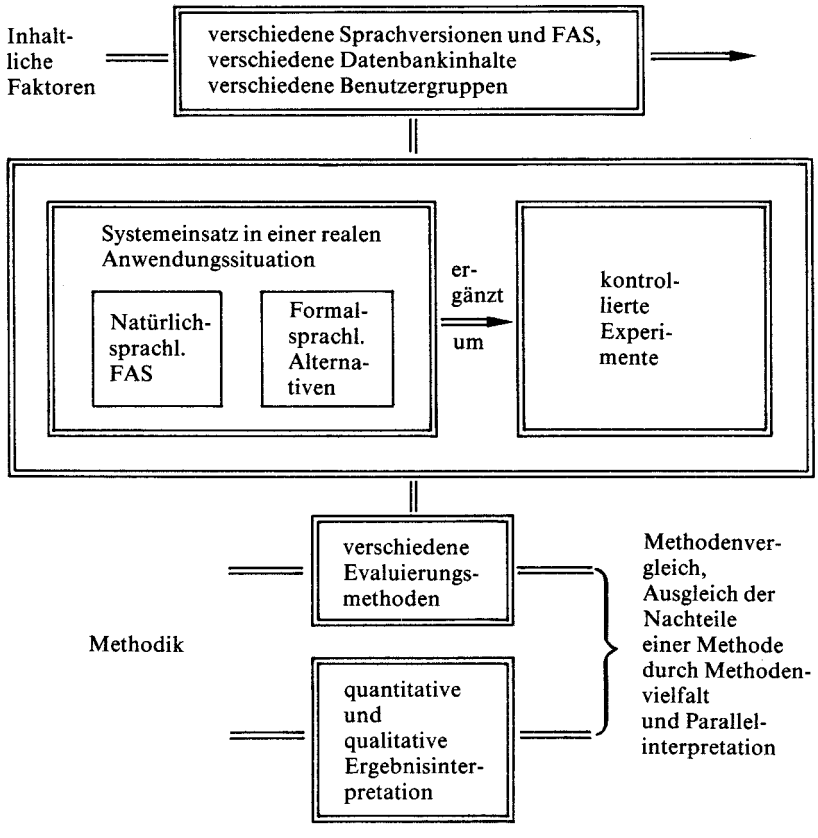

Abb. 10.1: Evaluierungskonzept FAS

häufigsten evaluierte natürlichsprachliche FAS. Die mehrstufigen Evaluierungen zu USL beziehen die Mehrzahl der bisher genannten Evaluierungsmethoden mit ein und verbinden sie zu einem komplexen Methodenverbund; der Schwerpunkt liegt auf empirischen Fallstudien in realen Anwendungssituationen. Insgesamt wurden über 12 000 Anfragen von etwa 100 Benutzern zur deutsch- und englischsprachigen Version von USL ausgewertet. Jarke/Krause/Vassiliou 1986 geben einen ausführlichen Überblick (mit weiterführenden Literaturangaben) zur USL-Evaluierung. Die beiden größten Studien sind die KFG- (Krause 1980, Krause 1982 a) und ALP-Studie (Vassiliou et al. 1983; Turner et al. 1984; Jarke et al. 1985, Jarke/Krause 1985 a; Jarke/Krause 1985 b). Sie können als teilweise Realisierung eines komplexen Evaluierungskonzeptes interpretiert werden, wie es in Abb. 10.1 skizziert ist.

Den Kern der KFG-Evaluierung bildet eine Feldstudie mit der deutschsprachigen USL-Version. Drei Lehrer untersuchten die prognostische Relevanz von Zeugnisnoten auf die Abiturleistung. Sie stellten während 16 Monaten 7 278 Fragen. Schwerpunkt der vor allem qualitativ orientierten Auswertung sind Fehler und Fehlerfolgen. Die Übersetzungstests zum Vergleich natürlichsprachlicher vs. formalsprachlicher Abfragemöglichkeit wurden mit Studenten durchgeführt, in einem späteren Projekt mit einer Sekretärinnengruppe wiederholt (vgl. Krause et al. 1983, Kap. 5). Sie haben im Gesamtrahmen der KFG-Studie eher untergeordnete Bedeutung, im Gegensatz zur ALP-Studie, bei der ein Vergleich mit der formalen Abfragesprache SQL im Vordergrund stand. Grundlage von ALP war eine englischsprachige USL-Version, die mit einer Datenbank der Verwaltung der New York University (Alumni-Spendensystem) eingesetzt wurde. Die ALP-Studie besteht aus zwei kontrollierten Laborexperimenten (8 und 61 Studenten; Aufgabenlösung mit Papier und Bleistift, 1 019 Fragen) und einer Feldstudie, bei der SQL und die natürliche Sprache zur Lösung der gleichen Aufgaben herangezogen wurden (8 Studenten, 1 081 Fragen). Ausgangspunkt waren echte Problemstellungen, die in einem Gespräch zwischen den Bearbeitern der Spendenproblematik und den Probanden festgelegt wurden (request-Ebene). Die Probanden formulierten die Aufgaben für die Terminalsitzung in die eigene Terminologie um (task-Ebene) und versuchten sie durch eine Reihe von Einzelabfragen zu lösen. Die in einem ersten Schritt quantitativ angelegte Auswertung (gelöste Aufgaben, Fehleranzahl, Eingabelänge, Wortwahl u. ä.) wurde in einem zweiten Schritt durch ein qualitativ-interpretatives Vorgehen ergänzt.

3. Fazit

Die Entwicklung bei der Evaluierung natürlichsprachlicher FAS zeigt eine deutliche Hinwendung zu komplexen mehrstufigen Studien, die die verschiedensten Ansätze zusammenfassen. Es wäre falsch, die in 2.1.—2.4. aufgezeigte historische Entwicklung als Ablösung eines Verfahrens durch andere zu interpretieren. Die Methoden ergänzen sich; sie finden Anwendung je nach Entwicklungsstand und verfolgter Absicht. Abgelöst werden dagegen nicht-komplexe durch komplexe mehrstufige Verfahren; nur so kann der zwangsläufigen Komplexität der Systeme adäquat begegnet werden. Eine ähnliche Entwicklung findet bei der Bewertung von Referenz-Retrievalsystemen statt (vgl. hierzu Sparck-Jones 1981; Jochum/Reiner 1983; Fuhr/Knorz 1984; Krause 1984 b; Schneider 1985; Krause 1987). Eine adäquate Bewertung der CL-Komplexität erfordert auch hier zusätzliche Komponenten zur traditionell statistisch-mathematisch ausgerichteten Retrievaltest-Evaluierung auf der Grundlage von Relevanzurteilen.

Bei den FAS dürfte eine der wesentlichsten Fragen für die Zukunft sein, ob sich ein „computer talk“ empirisch belegen läßt, oder anders ausgedrückt, ob sich nachweisen läßt, daß die MMI zumindest in Teilbereichen eigenständigen Gesetzmäßigkeiten gehorcht, die sich nicht aus der zwischenmenschlichen Kommunikation ableiten lassen (vgl. Zoltan et al. 1982; Krause 1982 a; Zoeppritz 1983 und 1985 a, Krause 1989). In diesem Fall käme komplexen empirischen Evaluierungsstudien die entscheidende Rolle bei der Weiterentwicklung der MMI zu.

4. Literatur (in Auswahl)

M. Bates/C. L. Sidner 1983 · A. Chapanis 1973 · A. Chapanis 1975 · E. F. Codd 1978 · E. F. Codd et al. 1978 · F. J. Damerau 1979 · R. Floden/S. Weiner 1978 · N. Fuhr/G. Knorz 1984 · D. Greenblatt/J. Waxmann 1978 · L. R. Harris 1977 a · J. Hofmann 1985 · M. Jarke et al. 1985 · M. Jarke/J. Krause 1985 a · M. Jarke/J. Krause 1985 b · M. Jarke/J. Krause/Y. Vassiliou 1986 · F. Jochum/U. Reiner 1983 · J. F. Kelley 1984 · J. Krause 1980 · J.

Krause 1982 a · J. Krause 1983 · J. Krause 1984 a · J. Krause 1984 b · J. Krause 1987 · J. Krause 1989 · J. Krause et al. 1983 · R. Kuhlen 1985 · H. Lehmann/A. Blaser 1979 · A. Malhotra 1975 · S. R. Petrick 1976 · M. Pollack/J. Hirschberg/B. Webber 1982 · P. Rosenbaum 1968 · P. Reisner 1981 · C. Schneider 1985 · B. Shneiderman 1978 · B. Shneiderman /R. Mayer 1979 · C. L. Sidner 1980 · D. W. Small/L. J. Weldon 1977 · K. Sparck-Jones 1981 · H. Tennant 1978 · H. Tennant 1979 · A. Turing 1950 · J. A. Turner et al. 1984 · Y. Vassiliou et al. 1983 · D. L. Waltz 1977 · W. C. Watt 1968 · B. L. Webber 1983 b · W. A. Woods 1977 · W. A. Woods 1983 a · W. A. Woods/R. J. Kaplan/B. Nash-Webber 1972 · M. Zoeppritz 1983 · M. Zoeppritz 1985 a · M. Zoeppritz 1985 b · M. Zoeppritz 1986 · E. Zoltan/G. D. Weeks/W. R. Ford 1982.

Jürgen Krause, Regensburg
(Bundesrepublik Deutschland)