Burghard Rieger

Das Fach Linguistische Datenverarbeitung (LDV) in Lehre und Forschung*

An der Schwelle des Übergangs unserer Industrie- in die Informationsgesellschaft ist das Fach als eine junge wissenschaftliche Disziplin gekennzeichnet von einer durch Denktraditionen noch kaum beschränkten Dynamik während der letzten zwei Jahrzehnte. Mit älteren Nachbardisziplinen wie Philologien und Sprachwissenschaft, Psychologie und Informatik, aber auch mit Mathematik und Philosophie vielfach verwandt, läßt sich die *Linguistische Datenverarbeitung* (LDV) oder *Computerlinguistik* (CL)—wie das Fach in Anlehnung an die im anglo-amerikanischen Bereich geläufige Bezeichnung *Computational Linguistics* auch bei uns genannt wird—durch die besondere Verbindung ihres Erkenntnisinteresses, ihres Forschungsgegenstands und ihrer Untersuchungsmethoden bestimmen:

Ihr *Erkenntnisinteresse* richtet sich auf die Erforschung der mit dem Gebrauch von künstlichen und natürlichen Sprach- und Zeichen- Systemen verbundenen Strukturen, Funktionen und Prozesse.

Ihr *Forschungsgegenstand* ist die natürliche Sprache und ihr kommunikativer Gebrauch, deren funktionale Zusammenhänge sie theoretisch unter systematischen Aspekten (*Kompetenz*) und empirisch unter Aspekten der Realisation (*Performanz*) untersucht.

Ihre *Untersuchungsmethoden* sind die durch die Möglichkeiten des Computers erweiterten Verfahren der theoretischen und empirischen Sprachwissenschaft, der Psychologie, der Mathematik und Statistik sowie der Informatik.

Für die LDV/CL ist Sprache keine bloß zeichen-materielle Gegebenheit, die wie Gegenstände der physikalischen Realität untersucht werden könnte, sondern ein zeichenfunktionaler Prozeß der Kommunikation. Er ist dies aber nur vermöge der an ihm beteiligten Sprecher und Hörer bzw. Schreiber und Leser, die durch den Einsatz ihres sehr komplexen Welt- und Sprach-Wissens die in diesem Prozeß vermittelten sprachlichen Formen überhaupt erst produzieren bzw. erkennen und deren Bedeutungen intendieren bzw. verstehen können. Man spricht deshalb heute allgemein von der *Wissensbasiertheit* solcher *kognitiven* Prozesse, die auf das Erkennen und Verstehen ausgerichtet sind, und von denen die Sprachverarbeitung nur ein—wenn auch sehr prominenter—Teilbereich ist.

Bei der wissenschaftlichen Analyse solcher wissensbasierter Prozesse, ihrer Beschreibung und Rekonstruktion in Modellen hat sich eine neue Form der Untersuchung herausgebildet, die für das *kognitive Paradigma* wissenschaftstheoretisch kennzeichnend

*Erschienen in: *Uni-Journal. Zeitschrift der Universität Trier,* Jg. 19 (1993) Sonderheft Nr. 1, S. 56–59

ist. Danach können die Vielfalt der beobachtbaren Erscheinungen als Strukturen beschrieben, diese Strukturen als Resultate von Prozessen gedeutet und diese Prozesse als Realisierung von Prozeduren verstanden werden. Prozeduren erscheinen dabei als jene Zusammenhänge in Prozessen, welche diese unter Absehung ihres zeitlichen Verlaufs charakterisieren und in Ausdrücken sogenannter Programmiersprachen abstrakt dargestellt werden können. Wenn es daher gelingt, diejenigen Prozeduren zu entwickeln und (in geeigneten formalen Sprachen) als Programme zu formulieren, die—auf relevanten Daten operierend—als Prozesse (in geeigneten Automaten bzw. Computern) ablaufen, so lassen diese im Rechner simulierten Prozesse ihrerseits Strukturen entstehen, welche jenen der beobachtbaren Erscheinungen, die diese modellieren sollen, entsprechen. Ziel solcher *prozeduraler* Modellbildungen in der LDV ist es dabei, Bedingungen und Verlauf der am sprachlichen Zeichengebrauch beteiligten Prozesse zu erkennen und zu analysieren, sowie die Organisation und Struktur des beteiligten Wissens zu rekonstruieren und in seinen Funktionen näher zu bestimmen.

Der Einsatz des Computers ist daher nicht nur *methodisches* Hilfsmittel im Sinne einer bloßen Steigerung von im Prinzip auch ohne ihn möglichen Untersuchungsleistungen, sondern eine *methodologische* Erweiterung der Möglichkeiten zur intersubjektiv überprüfbaren Erkenntnisgewinnung, welche bestimmte—nämlich *prozedural* faßbare—Komponenten unseres Wahrnehmens, unseres Verstehens und unseres kommunikativen Handelns mit und vermöge der natürlichen Sprache quasi-experimentell über Rechner-Simulationen zu untersuchen erlaubt. Dabei wird die Formulierung überprüfbarer Hypothesen und der Entwurf umfassender Theorien zur Erklärung beobachteter Zusammenhänge kognitiver Phänomene die leitende Zielvorstellung bleiben.

Derart als humanwissenschaftliche Disziplin charakterisiert, die sich exaktwissenschaftlicher Ansätze und Methoden bedient, wird das Fach LDV an der Universität Trier in Lehre und Ausbildung wie in Forschung und Entwicklung geprägt von den besonderen Anforderungen seiner Interdisziplinarität. Durch die Verbindung geisteswissenschaftlicher Fragestellungen mit dem formalen Rigorismus natur- und ingenieurwissenschaftlicher Modellbildungen in diesem Fach können die intersubjektive Vermittelbarkeit seiner Problemlösungen erhöht, die praktische Umsetzbarkeit seiner Resultate gesichert und die weitere Erschließung neuer Anwendungsgebiete eröffnet werden.

Lehre und Ausbildung

Nach ersten Anfängen eines Fachs LDV am Fachbereich II Ende der 70-er Jahre und vorläufiger Konzeption eines möglichen Ausbildungsgangs, der eine erprobende Aufnahme des Lehrbetriebs schon erlaubte, konnte in den Folgejahren (1986–1991) das Fach sowohl technisch wie personell ausgebaut und konsolidiert werden. Mit Genehmigung der neuen Studienordnung und der zugehörigen Prüfungsordnungen (1988) durch das Kultusministerium des Landes Rheinland-Pfalz wurde die Universität Trier zur ersten Ausbildunbgsstätte in der Bundesrepublik, die einen Magister-Studiengang mit Hauptfach LDV und Promotionsmöglichkeit anbot, und heute bundesweit die meisten

Studierenden (140 im Wintersemester 1992/93) in diesem Fach ausbildet.

Das Magisterstudium LDV gliedert sich in das *Grundstudium* (4 Semester), das mit der Zwischenprüfung abgeschlossen, und das *Hauptstudium* (4 + 1 Semester), das mit der Magisterprüfung beendet wird.

Im *Grundstudiums* sind 36 Semesterwochenstunden (SWS) mit Veranstaltungen aus der Computerlinguistik (12 SWS) und der Quantitativen Linguistik (4 SWS), den mathematischen und informatischen Grundlagen der LDV (12 SWS) mit Unterweisung in mindestens zwei Programmiersprachen PASCAL und LISP oder PROLOG (6 SWS) zu absolvieren.

Das *Hauptstudium* besteht aus 24 SWS Pflicht- und 12 SWS Wahlpflicht-Anteilen der verschiedenen Ausrichtungen der LDV in Trier; neben den schon bestehenden Fachausrichtungen *Computerlinguistik* (CL), *quantitative Linguistik* (QL) und *Wissenstechnik* (KT) sind im Aufbau oder geplant: *Lehr- und Lernsysteme* (LL) sowie *Fuzzy Linguistik* (FL).

Die im Hauptstudium angebotenen Studieninhalte umfassen dabei in der

Computerlinguistik: Grammatik (Theorien, Formalismen, Modelle); maschinelle Analyse sprachlicher Strukturen (Morphologie, Syntax, Semantik); Sprach-Erkennung und -Erzeugung: Verstehenssimulation; maschinelle Lexikographie/Lexikologie; automatische Übersetzung; Methoden der Softwaretechnik;

Wissenstechnik: Grundlagen der Kognitionswissenschaft; Semantik (Theorien, Formalismen, Modelle); sprachorientierte Künstliche Intelligenz; Wissensrepräsentation; maschinelle Inhalts- und Bedeutungsanalyse (Wort/ Satz/ Text); Datenbanken und Informationssysteme; Datenschutz und Datensicherung;

Quantitativen Linguistik: Methoden und Modelle der QL; Statistik für Linguisten (Probleme, Anwendungen); Techniken der Datenexploration; systemtheoretische Linguistik: Modellierung dynamischer Systeme, synergetische Linguistik.

Neben der Vermittlung theoretischer Kenntnisse ist das Studium der LDV in Trier in gleicher Weise auf den Erwerb praktischer Fertigkeiten im Bereich der maschinellen Verarbeitung natürlicher Sprache gerichtet. So führen während des *Grundstudiums* begleitende *Tutorien* zu den Hauptvorlesungen in die Techniken des selbständigen Wissenserwerbs aus der Lektüre wissenschaftlicher Originaltexte ein. Zahlreiche *Übungsveranstaltungen* sind zu besuchen, die den praktischen Umgang mit dem Erlernten vermitteln und vertiefen.Neben den traditionellen Veranstaltungstypen akademischer Unterweisung haben die Studierenden der LDV in Trier während des *Hauptstudiums* ein mindestens 6-wöchiges *Betriebspraktikum* abzuleisten sowie die erfolgreiche Bearbeitung eines *Studienprojekts* nachzuweisen. Letzteres dient der Einübung arbeitsteiliger Techniken zur Problemlösung im Team, dessen (2 bis max. 5) Mitglieder unter Anleitung (6 SWS) kleinere (Teil-)Projekte (von ihrer Formulierung bis zur Implementation) gemeinsam bearbeiten und dokumentieren. Die Studienprojekte, deren Themenstellungen sich aus laufenden Forschungsprojekten oder Seminarveranstaltungen ergeben, vermitteln den Studierenden dabei einen unmittelbaren Kontakt zur Forschungswirklichkeit. Ab-

solventen des Studiengangs werden so in die Lage versetzt, theoretische wie praktische Probleme der Verarbeitung von Sprache durch Computer aufgrund des erworbenen linguistischen Wissens zu verstehen und zu analysieren und etwa (schon vorhandene) Lösungen—aufgrund ihrer methodischen und programmiertechnischen Fertigkeiten—zu bewerten sowie neue Lösungsansätze für spezielle Problemstellungen zu erarbeiten.

Das Studium qualifiziert—wie die beruflichen Positionen der (bisher 24) Absolventen zeigen—für ein breites Tätigkeitsfeld bei Herstellern und Anwendern der Informationstechnik in Entwicklung und Erprobung, Produktion, Organisation und Dienstleistung, auf dem Informations- und Kommunikationssektor in Industrie und Verwaltung sowie in Lehre und Forschung.

Forschung und Entwicklung

Als Forschungsschwerpunkt der LDV in Trier darf die *Quantitative Linguistik* (QL) gelten, die als Forschungsfeld innerhalb der *Computerlinguistik* derzeit weltweit an Aktualität gewinnt. Zwei Teilbereiche sind dabei in besonderem Maße mit den Forschungen der Trierer LDV verbunden, da deren Fachvertreter diese Bereiche mit ihren Arbeiten und Aktivitäten weitgehend prägten.

Bestimmend war hierfür zunächst die Besetzung der ersten Professur 1987 mit einem Computerlinguisten (Burghard Rieger), der seit Jahren auf dem Gebiet der Semantik und Wissensrepräsentation gearbeitet hatte und schon früh (1973) die Konzepte der Theorie der *unscharfen* Mengen nicht nur zur formal-theoretischen Modellierung *vager* Bedeutungen nutzte, sondern auch zu ihrer empirisch-quantitativen Analyse des Sprachgebrauchs anhand großer Corpora natürlichsprachlicher Texte einsetzte: vor diesem Hintergrund entwickelte sich der Forschungsbereich *Fuzzy Empirical Semantics* mit mehreren Projekten zum Teil sehr unterschiedlicher Akzentuierung. Sodann erlaubte eine gezielte Berufungspolitik die Besetzung der zweiten Professur 1990 mit einem Linguisten (Reinhard Köhler), der in der theoretisch wie empirisch arbeitenden quantitativen Linguistik ausgewiesen war und erstmals (1985) systemtheoretische Konzepte der *Synergetik* zur dynamischen Modellierung sprachliche Phänomene und ihrer Wechselwirkungen eingesetzt hatte: auf dieser Grundlage entstand der Forschungsbereich *Synergetische Linguistik* mit verschiedenen Projekten auch internationaler Beteiligung.

Die folgenden Forschungsprojekte und Aktivitäten stehen in (mehr oder weniger) enger Beziehung zu den beiden genannten Teilbereichen der CL/QL mit wechselseitigen Ergänzungen und vielfältigen Berührungen mit regelbasierten Ansätzen in *Computerlinguistik* und der sprachorientierten Forschung zur *Künstlicher Intelligenz* (KI):

Semantische Analyse von Texten Und Situationen (SATUS):
: Das vom Land RP geförderte Projekt fügt dem formal-theoretischen Apparat der *Situationssemantik* eine empirisch-quantitative Komponente hinzu, welche auf der Basis statistischer Analysen der Gebrauchregularitäten von Wörtern in Texten *pragmatisch-homogener* Corpora entwickelt wurde.

Semiotic Cognitive Information Processing System (SCIPS):

Im Unterschied zu *symbolischen* Darstellungsformaten lassen *verteilte* Repräsentationen der natürlichsprachlichen Semantik sich als Resultate von Prozessen modellieren, die als *Bedeutungs-konstituierend* gelten können. Dabei wurden—mit Förderung durch den *German Marshall Fund of the United States*—verschiedene Prozeduren systematischer Analyse sprachlicher Texte zur Generierung semantischer Repräsentationen entwickelt und erprobt.

Language Learning And Meaning Acquisition (LLAMA):
Das Projekt untersucht die kognitive Leistung sprachlicher Strukturbildung anhand der *syntagmatischen* und *paradigmatischen* Regularitäten von Wortverwendungsweisen in Texten zur Modellierung der System/Umgebung-Unterscheidung für zeichenverarbeitende Systemen mit Lernfähigkeit (DFG-Förderung beantragt).

Datadriven ACquisition of Syntactic knowledge (DACS):
Das Forschungsprojekt zum Daten-orientierten Erwerb syntaktischen Wissens ist der Entwicklung eines Systems zur Generierung von Grammatiken aus Satzkorpora gewidmet. Im Vordergrund stehen dabei regelbasierte und statistische Verfahren zur Konstitution morphologischen und syntaktischen Wissens.

Fuzzy Analysis of very large Linguistic Corpora (FALC):
Mit der Verfügbarkeit *sehr großer linguistischer Corpora* ($> 10^7$ Wörter) stellen sich völlig neuartige Probleme mangelnder Bestimmtheit traditionellerweise akzeptierter sprachlicher Gegebenheiten und linguistischer Kategorien. Im Rahmen des Projekts werden—der Datenintensität in der Hochernergie-Physik vergleichbar—Möglichkeiten der unscharfen (*fuzzy*) Neubestimmung linguistischer Entitäten und Strukturen untersucht.

Synergetische Linguistik (SL):
Bei diesem systemtheoretischen Ansatz geht es um die Beschreibung und Erklärung des Sprachsystems gleichzeitig unter strukturellen, funktionalen und prozessualen Gesichtspunkten: SL betrachtet die Sprache als selbstregulierendes und selbstorganisierendes System, in dem Strukturen, die den äußeren Anforderungen entsprechen, entstehen bzw. sich verän- dern. In der Forschungsgruppe laufen zahlreiche Untersuchungen, z.T. in internationaler Zusammenarbeit (mit Instituten und Wissenschaftlern aus 20 Ländern).

Bibliographie Quantitative Linguistik (BQL):
Das von der DFG geförderte Projekt stellt die weltweit bisher erschienene wissenschaftliche Literatur zur QL zusammen und macht sie bibliographisch zugänglich.

Quantitative Linguistics Series (QLS):
Diese internationale Buchreihe (mit seit 1978 mehr als 50 erschienenen Bänden) wird seit 1989 in Trier von B. Rieger und R. Köhler herausgegeben.

HERbstSchule COmputerLinguistk (HERSCOL):
Das erste einwöchige Kursangebot dieser Art, das die *Gesellschaft für Linguistische Datenverarbeitung* (GLDV) im Oktober 1990 für 70 Teilnehmer organisierte, wurde vom Fach LDV der Universität Trier veranstaltet und ausgerichtet.

International Quantitative Linguistics Conference (QUALICO):
Im September 1991 veranstalteten die Herausgeber der QL-Buchreihe die erste internationale Konferenz zur quantitativen Linguistik, die mehr als 120 Wissenschaftler aus 16 Ländern Europas, Asiens und Nord-Amerikas für 5 Tage in Trier versammelte. Der Erfolg der Konferenz, die als 13. Jahrestagung der deutschen GLDV mit Unterstützung der

DFG vom Fach LDV der Universität Trier organisiert wurde, führte zur Gründung eines ständigen QUALICO-Komitees mit Sitz in Trier.