EINFÜHRUNG IN DIE COMPUTERLINGUISTIK
Vermittlung computerlinguistischer Grundlagen im Rahmen eines sprachwissenschaftlichen Studiengangs

Uta Seewald
Romanisches Seminar
Universität Hannover

Der Lehrveranstaltungstyp der Einführung kennzeichnet im Rahmen zahlreicher Studiengänge Veranstaltungen, die sich in der Regel an Studienanfänger richten und Grundlagen des jeweiligen Studienfaches vermitteln. Es erstaunt deshalb nicht, daß computerlinguistische Einführungen vor allem in Lehrprogrammen solcher Hochschuleinrichtungen angeboten werden, bei denen das Fach Computerlinguistik als eigenständiges Studienfach etabliert ist oder zumindest im Rahmen eines N ebenfachstudiums betrieben werden kann.

Ohne Zweifel besteht Interesse an computerlinguistischen Fragestellungen jedoch auch über diese Studiengänge hinaus, insbesondere im Rahmen linguistischer Studiengänge sowie der sprachwissenschaftlichen Zweige der Philologien. Verschiedentlich geführte Diskussionen um Ziele und Zukunft der Computerlinguistik führten bei zahlreichen Diskutanten auch zu der Erkenntnis, daß eine linguistische Ausbildung langfristig nur schwer ohne computerlinguistische Anteile vorstellbar ist.

Diese Situation war im Rahmen der Romanistik an der Universität Hannover vor vier Jahren Anlaß, das Angebot an sprachwissenschaftlichen Seminaren um computerlinguistische Veranstaltungen zu erweitern. Die angebotenen Veranstaltungen, die von Anfang an auch Studierenden anderer Philologien offenstanden, waren schließlich auch Ausgangspunkt für die Einrichtung eines sogenannten Anwendungsfaches "Romanistische Linguistik" für Informatiker.

Ein Ausweis für den Mangel an derartigen Veranstaltungen auch im Rahmen anderer Studienfächer ist die Tatsache, daßdie an der Universität Hannover angebotenen computerlinguistischen Einführungsveranstaltungen von einem sehr heterogenen Publikum besucht werden, das sich sowohl aus Romanisten, zum Teil auch Anglisten und Germanisten, als auch aus Informatikern zusammensetzt. Während die Studenten der philologischen Fächer den Computer - wenn überhaupt - nur als Schreibinstrument kennen und in der Regel nicht über die Kenntnis einer Programmiersprache verfügen, operieren die Informatikstudenten, die diese Veranstaltung meist im 3. Studiensemester besuchen, seit Studienbeginn mit dem Computer und haben bereits Kenntnisse in mindestens zwei Programmiersprachen. Dafür verfügen sie zum Zeitpunkt des Besuchs dieser Veranstaltung nur über geringe oder gar keine linguistischen Kenntnisse.

Da der fehlende Umgang mit dem Instrumentarium Computer und Programmiersprache die Formulierung der für alle Teilnehmer gleich gestellten Aufgaben wesentlich erschwert hat, werden die Studenten der philologischen Fächer in einer der eigentlich computerlinguistischen Einführungsveranstaltung vorausgehenden Veranstaltung mit informatischen bzw. programmiertechnischen Grundlagen ver-

traut gemacht. Bei den programmiertechnischen Hilfsmitteln handelt es sich um die Vermittlung der Programmiersprache LISP, die auch im Rahmen des Informatikstudiums von den Informatikstudenten erlernt wird. 1

Im folgenden werden Konzeption und Inhalte der Einführung in die Computerlinguistik skizziert, wie sie derzeit an der Universität Hannover für Romanisten und Informatiker angeboten wird:

1. Aufgaben und Anwendungsbereiche der Computerlinguistik
2. Segmentierung - "vom Text zum Graphem"
3. Lexikon: Modelle und Implementierung
4. Morphologische Analyse und Generierung
5. Parsingverfahren
6. Syntaktische Analyse mit Netzwerkgrammatiken (RTN und ATN)

(1) Da die Teilnehmer der Veranstaltung meistens nur sehr schemenhafte Vorstellungen davon haben, was Computerlinguistik überhaupt ist bzw. mit welchen Aufgabenstellungen man sich in der Computerlinguistik beschäftigt und welche möglichen Anwendungen für Ergebnisse computerlinguistischer Arbeit bestehen, bildet ein Überblick über Forschungsrichtungen und Anwendungsgebiete der Computerlinguistik den Einstieg in die Einführungsveranstaltung. Dieser Überblick wird eingeleitet mit einem Videofilm, der über Forschungszentren und computerlinguistische Projekte informiert.2 Ausgehend von dieser Darstellung werden einzelne Aufgabenstellungen und Anwendungsgebiete, wie *Information Retrieval* und Maschinelle Übersetzung, skizziert. Die Teilnehmer werden auf die Zusammenstellungen in Handke (1988), Hausser (1994), Klavans (1989), Schmitz (1992) und Smith (1991) hingewiesen.

(2) Den ersten Schwerpunkt der Veranstaltung bildet das Thema Segmentierung, daß es sich hierbei um einen der zentralen Prozesse der linguistischen Datenverarbeitung handelt, der den Kern jeder Analyse bildet und Voraussetzung für jede weitergehende Verarbeitung sprachlicher Information ist, sei sie geschrieben oder gesprochen. Hierbei ist von Bedeutung, dass die Studenten ein Problembewußtsein dafür entwickeln, daß jede Information, die sie aus einem Text zur weiteren Verarbeitung extrahieren wollen, mittels Segmentierung isoliert werden kann und daß sich dieser Prozeß auf verschiedenen sprachlichen Ebenen bis hinunter zu einzelnen Phonemen oder Graphemen fortsetzen läßt. Um das Problembewußtsein hierfür zu schärfen, wird zur Übersicht der Artikel "Segmentierung in der Computerlinguistik" , Lenders (1989), behandelt.

Da möglichst viele Überlegungen von den Teilnehmern selbst in gemeinsam erarbeitete oder eigene Lösungen programmiertechnisch umgesetzt werden sollen und das verwendete Programmierwerkzeug hierfür die Programmiersprache LISP ist, wird als begleitende Lektüre und als Arbeitsgrundlage das Buch "Natürliche Sprache: Theorie und Implementierung in LISP" von Handke (1988) herangezogen.

(3) Bei der Verarbeitung natürlicher Sprache ist in der Regel der Zugriff auf ein Lexikon erforderlich, das deshalb einen weiteren Themenschwerpunkt der Einführung bildet. Zunächst wird mit den Teilnehmern die Funktion des Lexikons bei der Verarbeitung natürlicher Sprache erörtert sowie die daraus resultierenden Anforderungen an die Art der Information, die das Lexikon enthalten muß. Da eine der Aufgabenstellungen der Computerlinguistik, die im Grenzbereich der Künstlichen Intelligenz angesiedelt ist, die Simulation menschlicher Sprachverstehensprozesse ist, wird an dieser Stelle zunächst eine Parallele zwischen dem mentalen Lexikon und dem Computerlexikon gezogen. Einzelne Modelle, die die Psycholinguistik für die

1 Als Lehrbuch wird den Teilnehmern das bereits als Standardwerk geltende Buch von Winston/Horn (1987) empfohlen.
Die Entscheidung für LISP oder PROLOG ist arbiträr. Bei einer Wiederholung der Veranstaltung soll anstelle von LISP die Programmiersprache PROLOG eingesetzt werden.

2 „Computer und Sprache". IBM, Videothek der Informationsverarbeitung, 1988.

die Repräsentation des mentalen Lexikons entwickelt hat, werden in Handke (1988) dargestellt. Da dort die Frage der Übertragbarkeit der psycholinguistischen Modelle auf den Computer unmittelbar an die Darstellung der verschiedenen Modelle des mentalen Lexikons anschließt, wird diese Fragestellung anhand des Kapitels "Lexikon", Handke (1988:21-43), mit den Teilnehmern bearbeitet. Weil ein wesentlicher Gesichtspunkt bei der Entscheidung für ein bestimmtes Lexikonmodell in der Computerlinguistik die Effizienz der Darstellung und des Zugriffs ist, wird die Repräsentation des Lexikons als Diskriminationsnetzwerk (*trie structure*) vertieft und auch programmiertechnisch umgesetzt. 3 Ein weiteres Kriterium für die Güte computerlinguistischer Anwendungen ist die Forderung nach einfacher Modifizierbarkeit bestehender Datensammlungen. Um zu einem späteren Zeitpunkt Einträge aus dem Lexikon zu entfernen, Angaben zu verändern oder neue Einträge hinzuzufügen, ist es wichtig, daß das Lexikon bzw. die darin enthaltenen Einträge für den Benutzer (Lexikographen) leicht lesbar sind. Aus diesem Grunde erstellen die Teilnehmer zunächst ein Lexikon, das alle vorgesehenen Einträge listenartig und übersichtlich aufführt. Unter Zuhilfenahme der in Handke (1988) erläuterten LISP-Funktionen wird das Lexikon schließlich in ein Diskriminationsnetzwerk überführt. Um auch von vornher
ein die Modularität als wichtiges Gütekriterium hervorzuheben, wird den Teilnehmern die Aufgabe gestellt, das Lexikon im ersten Schritt als Textdatei anzulegen, diese von LISP in eine Liste einlesen zu lassen und schließlich in eine *trie structure* zu überführen.

(4) Den nächsten thematischen Schwerpunkt des Seminars, bei dem das Lexikon dann erstmalig als Datenbasis eingesetzt wird, bildet die morphologische Analyse sowie die Generierung von Wortformen. Vor der Erstellung eines Analyse und eines Generierungsmoduls, mit dem französische - oder je nach der Zusammensetzung der Teilnehmer auch deutsche, englische oder italienische - Verbformen analysiert bzw. generiert werden sollen, erarbeiten die Teilnehmer in der Diskussion auf der Grundlage vorbereitender Lektüre4 zunächst die Besonderheiten, Vorzüge und Nachteile eines auf einem Vollformenlexikon basierten Verfahrens auf der einen sowie eines auf einem Stammlexikon basierten Verfahrens auf der anderen Seite. Der einfachen programmiertechnischen Umsetzung wegen werden die Teilnehmer zur Implementierung eines paradigmaorientierten Ansatzes5 angeleitet, den jeder Teilnehmer dann eigenständig für eine vorgegebene Liste von Verben erweitert.6 In diesem Zusammenhang wird auch das Phänomen der Allomorphie 7 behandelt und die Alternativen zur Behandlung dieses Phänomens bei der morphologischen Analyse erörtert.

(5) Den letzten Themenbereich des Einführungsseminars bildet die syntaktische Analyse. Obschon das Parsen sprachlicher Einheiten im Sinne von Segmentieren bereits Gegenstand der morphologischen Analyse ist, werden verschiedene Parsingstrategien, wie *bottom-up, topdown, breadth-first, depth-first* , erst an dieser Stelle behandelt.8

(6) Als ein in der linguistischen Datenverarbeitung sehr verbreiteter Formalismus zur syntaktischen Analyse werden schließlich Netzwerkgrammatiken bzw. Netzwerkparser vorgestellt.9 Es wird zunächst

3 Vgl. Smith (1991:111), Sproat (1992:111-113), Handke (1988:43-64) sowie Dei et al. (1990:97108).

4 Guzman et al. (1989), Sproat (1992:1-123), Schaeder/Willee (1989).

5 Bauer (1988:151-163).

6 Als Hilfsmittel zur Systematisierung der Verbformen des Französischen eignet sich *Le nouveau Bescherelle, L 'art de conjuguer* sowie die Flexionstabellen im *Grand Robert,* für die Verbformen des Italienischen *Verbi italiani* von Buratti (1993) sowie die von Cappelletti (1990) in der Collection Bescherelle erschienenen *8000 verbes italiens* und die Übersicht in Dardano/Trifone (1985).

7 Bauer (1988:13-16).

8 Zur Lektüre werden Hellwig (1989a), Klavans (1989), Lenders/Willee (1986), Sabah (1989) sowie Winograd (1983) herangezogen.

9 Auch an dieser Stelle kann zur Einführung in die Thematik und gleichzeitig zur Wiederholung bisher behandelter Themen ein Videofilm eingesetzt werden, der unter dem Titel "Natürlichsprachliche Systeme" neben Mustererkennung auch eine Präsentation von ATN-Grammatiken bietet. ("Einführung in die Künst-

ein einfaches Netzwerk behandelt, das von den Teilnehmern dann in ein LISPProgramm zur Analyse ausgewählter einfacher syntaktischer Phänomene umgesetzt wird. Im nächsten Schritt werden dann rekursive Netzwerkgrammatiken (RTN) vorgestellt. Die Teilnehmer erhalten die Aufgabe, das von ihnen bisher erarbeitete Programm zu modifizieren und daraus eil}.e rekursive Netzwerkgrammatik zu erstellen.

Sofern ausreichend Zeit zur Verfügung steht, wird die Aufgabe gestellt, ein syntaktisches Problem zu bearbeiten, dessen Lösung mit Hilfe eines RTN nicht möglich ist. Im Französischen bieten sich hier u. a. Sätze im *Passe compose* an, bei denen sich das *participe passe* (Partizip Perfekt) in Genus und Numerus nach dem ihm vorausgehenden pronominalisierten oder durch *que* Anschluß aufgenommenen direkten Objekt richtet. Die Einsicht in die Notwendigkeit eines" Gedächtnisses", das die kategoriale Information über das Subjekt aufnimmt, leitet dann die Darstellung erweiterter Übergangsnetzwerke (ATN) ein, deren Behandlung den Abschluß der Veranstaltung bildet. 10

Bei den bisher durchgeführten Veranstaltungen hat sich gezeigt, daß der parallele Besuch einer einstündigen Programmierübung in LISP eine nützliche Ergänzung für die informatisch nicht oder nur wenig vorgebildeten Teilnehmer ist. Diese Programmierübung hat besonderen Effekt, wenn für die Betreuung der Teilnehmer zusätzlich eine studentische Hilfskraft zur Verfügung steht, die den Studenten bei der Programmerstellung, Fehlersuche und Problemlösung Hilfestellung gibt.

Die hier skizzierte Einführungsveranstaltung in die Computerlinguistik hat sich ins- . besondere für die Teilnehmer der sprachwissenschaftlich-philologischen Fächer als von besonderer Bedeutung erwiesen, daß sie durch die programmiertechnische Umsetzung der gestellten Probleme nach entsprechender Programmierarbeit zum einen sichtbare Ergebnisse erzielen und sich zum anderen ein Instrumentarium aneignen, das auch für andere Fragestellungen nutzbar ist. Darüber hinaus ermöglicht eine Einführung in die Computerlinguistik an Hochschulen, an denen trotz ansonsten ausgebauter Philologien Angewandte Sprachwissenschaft nicht vertreten ist, mit der Computerlinguistik jedenfalls in einen Bereich der Angewandten Linguistik Einblicke zu gewinnen.

liehe Intelligenz 3. Natürlichsprachliche Systeme, Teil 1 und 2, von Robert Trappl, Wien. Spektrum Videothek, 1991).

10 Winograd (1983), Winston (1987), Handke (1988), Habert (1983), Charniak/MeDermott (1985), Smith (1991), Hellwig (1989a, 1989b), Sabah (1989).

Literatur

Bei der hier aufgeführten Literatur handelt es sich um Werke, auf die zuvor im Text verwiesen wird, sowie um weitere Literatur, die bei der oben beschriebenen Lehrveranstaltung als Lektüre empfohlen wird.

Art de conjuguer. (Le nouveau BeschereHe) Bearb. v. D. Langendorf. Frankfurt, Berlin, München: Diesterweg.

Barr, Avron/Feigenbaum, Edward A., (eds.) (1981): The Handbook of Artificial Intelligence, Vol. 1-4, Los Altos: Kaufmann.

Blitori, Istvan S. / Lenders, Winfried / Puschke, Wolf gang (1989): Computational Linguistics. Computerlinguistik. Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen. Berlin: Walter de Gruyter.

Bauer, Laurie (1988): Introducing Linguistic Morphology. Edinburgh: Edinburgh University Press.

Buratti, Rosalia (1993): Verbi italiani. Mailand: Garzanti Editore

CappeHetti, Luciano (1990): 8000 verbes italiens. Collection Bescherelle. Paris: Hatier.

Charniak, Eugene/McDermott, Drew (1985): Introduction to Artificial Intelligence. Addison- Wesley.

Dardano, Maurizio /Trifone, Pietro (1985): La lingua italiana. Bologna: Zanichelli.

Deiss, Klaus/Handke, Jürgen/Meyer, Bodo (1990): Professionelles Programmieren mit LISP. Hamburg: McGrawHili.

Görz, Günther (1988): Strukturanalyse natürlicher Sprache. Bonn: AddisonWesley.

Gregor, Bernd/Krifka, Manfred (1986): Computerfibel für Geisteswissenschaften. München: C. H. Beck.

Guzman, V. P. De/O'Grady, W./ Aronoff, M. (1989): Morphology: The Study of Word Structure. In: O'Grady et al. (Hrsg.): Contemporary Linguistics. New York: St. Martin's Press, S. 89-122.

Habert, Benoit (1983): Un regard sur les ATN. Manchester, CCL/UMIST Report No. 83/6.

Handke, Jürgen (1987): Sprachverarbeitung mit LISP und PROLOG auf dem PC. Wiesbaden: Vieweg.

Handke, Jürgen (1989): Natürliche Sprache: Theorie und Implementierung in LISP. Hamburg: McGraw-Hill.

Handke, Jürgen (1994): Zugriffsmechanismen im mentalen und maschinellen Lexikon. In: Börner, Wolfgang/Vogel, Klaus (Hrg.): Kognitive Linguistik und Fremdsprachenerwerb. Das mentale Lexikon. Tübingen: Narr, S. 89106.

Hellwig, Peter (1989a): Parsing natürlicher Sprachen: Grundlagen. In: Batori/Lenders/Puschke (1989), S. 348-377.

Hellwig, Peter (1989b): Parsing natürlicher Sprachen: Realisierungen. In: Batori/Lenders/Puschke (1989), S. 378430.

Hausser, Roland (1993): Aufgaben der Computerlinguistik. In: LDV-Forum 10/2 S. 63-77.

Hausser, Roland (in Vorbereitung): Grundlagen der Computerlinguistik. Erlangen.

Klavans, Judith (1989): Computational Linguistics. In: O'Grady et al. (Hrsg.): Contemporary Linguistics. New York: St. Martin's Press, S. 413-445.

Krifka, Manfred (1988): Die Sprache der Computer und die Sprache der Menschen. In: Forum für interdisziplinäre Forschung 1/1988, S. 22-27.

Lenders, Winfried (1989): Segmentierung in der Computerlinguistik. In: Batori/Lenders/-Puschke (1989), S. 159-166.

Lenders, Winfried /WilIee, Gerd (1986): Linguistische Datenverarbeitung. Ein Lehrbuch. Opladen: Westdeutscher VIg.

Reyle, Uwe/Rohrer, Christian, Hrg. (1987): Natural language parsing and linguistic theories. (Studies in Linguistic and Philosophy). Dordrecht: D. Reidel Publishing Company.

Rolshoven, Jürgen/Seelbach, Dieter, Hrg. (1991): Romanistische Computerlinguistik. Tübingen: Niemeyer.

Sabah, Gerard (1989): L'intelligence artificielle et le langage. Processus de comprehension. Paris: Hermes.

Schaeder, Burkhard/Willee, Gerd (1989): Computergestützte Verfahren morphologischer Beschreibung. In: Batori/Lenders/Puschke (1989), S. 188203.

Schmitz, Ulrich (1992): Computerlinguistik. Eine Einführung. Opladen: Westdeutscher VIg.

Smith, George, W. (1991): Computers and Human Language. New York/Oxford: Oxford Univ. Press.

Sproat, Richard (1992): Morphology and Computation. Cambridge/Landon: MIT Press.

Winograd, Terry (1983): Language as a Cagnitive Pracess, Val 1: Syntax. New York: Addison-Wesley.

Winston, Patrick Henry (1987): Künstliche Intelligenz. Bonn: AddisonWesley.

Winston, Patrick Henry/Horn, Berthold Klaus (1987): LISP. Bann: Addisan- Wesley.