AUFGABEN DER COMPUTERLINGUISTIK

Roland Hausser
Computerlinguistik Universität
Erlangen-Nürnberg

Seit der ersten Entwicklung von elektronischen Rechenanlagen in den vierziger Jahren des 20. Jahrhunderts unterscheidet man zwischen der *numerischen* und der *nicht-numerischen* Informatik. Die numerische Informatik befaßt sich mit der Berechnung von Zahlen und hat in der Physik, der Chemie, der Wirtschaftswissenschaft, der Soziologie etc., zu einer explosionsartigen Wissensexpansion geführt. Aber auch in vielen praktischen Bereichen, wie dem Bankwesen, dem Flugverkehr, der' Lagerhaltung, etc., sind numerische Anwendungen heute nicht mehr wegzudenken. Ohne Computer und ihre Software würde die Funktionsfähigkeit dieser Bereiche zusammenbrechen.

Die nicht-numerische Informatik, andererseits, befaßt sich mit den Phänomenen der Wahrnehmung und der Kognition. Die theoretische und praktische Entwicklung der nicht-numerischen Informatik blieb, trotz hoffnungsvoller Anfänge, weit hinter denen der numerischen Informatik zurück. In neuerer Zeit findet die nichtnumerische Informatik jedoch als KOGNITIONSWISSENSCHAFT und KÜNSTLICHE INTELLIGENZ wieder starkes Interesse. Die *"cognitive science,* " und *"artificial intelligenceJ"* wie sie im Amerikanischen genannt werden, untersuchen die menschliche Informationsverarbeitung unter Einbeziehung der Informatik, der Psychologie, der Linguistik, der Philosophie und der mathematischen Logik.

1 Methoden und Anwendungen

In der Computerlinguistik werden die sehr unterschiedlichen Methoden der theoretischen Linguistik, der Psychologie, der Philosophie und der mathematischen Logik dadurch auf einen Nenner gebracht, daß man theoretische Hypothesen systematisch als Computerprogramme implementiert. Dieses Entwickeln und Testen von theoretischen Hypothesen auf dem Computer bietet die Möglichkeit einer neuen, vereinheitlichten Methodologie und Theoriebildung, die sich deutlich von der der traditionellen Linguistik, Psychologie, Philosophie und mathematischen Logik unterscheidet.

Theoretisch wird die Entwicklung in der Computerlinguistik durch eine neuartige Methodologie (Überprüfung formaler Grammatiksysteme durch den systematischen Einsatz von Computern) und prak-

tisch durch einen ungeheuren Bedarf an einer effizienten automatischen Sprachverarbeitung angetrieben. Im wissenschaftlichen Alltag der modernen Sprachwissenschaft wirkt sich der systematische Einsatz von Computern folgendermaßen aus:

1.1 Methodische Auswirkungen des Programmierens

=> Die Programmierung, z.B. einer Grammatik als Parser, erfordert eine viel detailliertere Analyse des zu behandelnden Phänomens als vorher üblich.

=> Die unterschiedliche Eignung grammatikalischer Formalismen für die Programmierung wird zu einem neuen, wichtigen Faktor im Wettbewerb konkurrierender Theorien.

=> Effizient implementierbare (und implementierte) Formalismen der automatischen Sprachanalyse haben praktische Anwendungen, die eine völlig neue Eigendynamik in die geisteswissenschaftliche Empirie bringen.

In folgenden Bereichen der praktischen Anwendung werden die Methoden der Computerlinguistik in stark zunehmendem Maße eingesetzt:

1.2 Praktische Aufgaben der Computerlinguistik

=> Indexierung von und Abruf aus textuellen Datenbanken

Textuelle Datenbanken speichern Texte in elektronischer Form, zum Beispiel die Jahrgänge einer Tageszeitung, die Publikationen einer Medizinischen Zeitschrift oder sämtliche Gerichtsurteile in den USA seit 1960. Der Benutzer einer solchen Datenbank muß in der Lage sein, all diejenigen Dokumente und Textstellen zu finden, die für seine spezifische Fragestellung relevant sind.

=> Automatisierte Textproduktion

Große Firmen, die ständig neue Produkte wie Motoren, Pumpen, Fernseher etc., herausbringen, müssen hierfür immer wieder neue Produktbeschreibungen und Wartungsmanuale herstellen. Ähnliches gilt für Rechtsanwalt- und Steuerkanzleien, Personalabteilungen, etc., die ein sehr hohes Korrespondenzvolumen haben, wobei sich die Briefe nur an klar definierten Stellen unterscheiden. Die Methoden der automatisierten Textproduktion reichen von einfachen Schablonen zu hochflexiblen und interaktiven Systemen, die auf linguistischem Wissen basieren.

=> Automatische Textüberprüfung

Auch auf diesem Gebiet reichen die Anwendungen von einfachen Orthographie-Checkern auf der Grundlage von Wortformlisten über Morphologiesysteme, die die sprachliche Vielfalt im Bereich der Wortbildung systematisch behandeln, bis zu Syntax-Checkern, die Fehler in Wortstellung, Kongruenz etc. finden können.

=> Automatische Inhaltsanalyse

Es heißt, daß sich die gedruckte Information auf der Erde alle 10 Jahre verdoppelt. Auch auf wissenschaftlichen, rechtlichen, wirtschaftlichen etc., Spezialgebieten ist die Flut der relevanten Literatur so groß, daß die Lebenszeit der Mitarbeiter einfach nicht mehr ausreicht, um ständig auf dem neusten Stand zu sein. Eine zuverlässige automatische Inhaltsanalyse mit kurzen Zusammenfassungen wäre hier von größtem Nutzen. Die automatische Inhaltsanalyse ist auch die Voraussetzung einer *konzeptbasierten Indizierung*, wie sie für den optimalen Abruf aus textuellen Datenbanken notwendig ist, sowie die Vorraussetzung für eine wirklich leistungsfähige maschinelle Übersetzung (s. u.).

=> Maschinelle Übersetzung

Die maschinelle Übersetzung war eine der Hauptanwendungen in den Anfängen der nicht-numerischen Informatik. In der Dekade von 1955 bis 1965 wurde auf diesem Gebiet intensiv geforscht, wobei es in der Öffentlichkeit große Aufmerksamkeit fand. Diese Erwartungen erfüllten sich jedoch nicht, und die Hoffnungen auf kommerzielle Erfolge zerschlugen sich.

Inzwischen ist das Interesse wieder stark gestiegen. HUTCHINS 1986 nennt folgende Gründe für die fortgesetzten Bemühungen um eine maschinelle Übersetzung:

- Wissenschaftler, Techniker, Ingenieure, Manager und viele andere Geschäftsleute müssen täglich viele Briefe und Dokumente in Sprachen lesen und schreiben, die sie nicht beherrschen... Es gibt einfach nicht genug Übersetzer, um mit dieser ständig wachsenden Menge an Material fertig zu werden.

- Viele Forscher betreiben die Entwicklung der maschinellen Übersetzung aus Idealismus. Sie wollen die internationale Kooperation und den Frieden fördern, indem sie Sprachbarrieren überwinden und die Verbreitung technischen, landwirtschaftlichen und medizinischen Wissens in die Entwicklungsländer fördern.

- Die maschinelle Übersetzung wird aber auch von Institutionen gefördert, die Anwendungen im Militärbereich sehen, also z.B. die schnelle Übersetzung gegnerischer Dokumente.

- Als Problem der reinen Forschung stellt maschinelle Übersetzung ein schwieriges Problem dar, dessen Lösung von manchen Forschern als Test ihrer linguistischen Arbeit betrachtet wird.

- Schließlich konstituieren leistungsfähige Systeme der automatischen Übersetzung wertvolle Software-Produkte mit vielfältigen Anwendungen, die entwickelt werden, weil man damit viel Geld verdienen kann.

Gerade in der Europäischen Gemeinschaft, wo derzeit in 12 Mitgliedsländern 9 verschiedene Sprachen gesprochen werden, ist der potentielle Nutzen von automatischen oder auch nur halb automatischen Übersetzungssystemen unübersehbar.

> Automatisierter Unterricht

Es gibt zahlreiche Unterrichtsfächer, in denen viel Zeit für sogenannte Drillübungen verwendet wird, z.B. die mehr oder weniger mechanische Erlernung von regelmäßigen und unregelmäßigen Paradigmen im Sprachunterricht. Diese können mindestens ebenso gut am Computer durchgeführt werden, wodurch dem Lehrer mehr Zeit für andere Aktivitäten, z.B. Konversation, bleibt. In der neueren Forschung beschäftigt man sich intensiv mit weitergehenden Automatisierungen des Unterrichts, zum Beispiel der automatischen Fehleranalyse bei Übersetzungsübungen.

Automatisierte Unterrichtssysteme haben den zusätzlichen Vorteil, daß automatisch über die Interaktionen zwischen Schüler und Computer buchgeführt werden kann. Durch das Wissen darüber, wo die Schüler am meisten Fehler machen und wo die meiste Zeit verbracht wird, erhält man dann eine wertvolle Heuristik für die Verbesserung der Ergonometrie des automatisierten Unterrichts. Dies hat eine Entwicklung vom 'elektronischen Textbuch' alter Prägung zu

neuartigen Lehrprogrammen mit einer eigenständigen, medium-gerechten Pädagogik initiiert.

=> Dialogsysteme und automatische Auskunft

Ein wesentlicher Engpaß bei der Interaktion mit Computern ist die Tatsache, daß die Interaktion entweder aufwendig (z.B. selbstgeschriebene Programme) oder unflexibel (z.B. Menuegesteuerte Interaktion) ist. Deshalb besteht großes Interesse an der Entwicklung robuster, natürlichsprachlicher Systeme, die bei praktisch allen Interaktionen zwischen Mensch und Maschine zum Einsatz gebracht werden können.

Die Zahl der möglichen Anwendungen der Computerlinguistik ist damit keineswegs abgeschlossen. Sie umfaßt vielmehr ganz allgemein sämtliche Bereiche, in denen Menschen (heute und in Zukunft) mit Computern umgehen. Bei all diesen Anwendungen dienen die Erkenntnisse der Sprachwissenschaften, zumindest potentiell, der Optimierung der automatischen Sprachverarbeitung. Umgekehrt spielen die Computer als Hilfsmittel der linguistischen Analyse und Theoriebildung eine immer größere Rolle.

2 Übertragung in das elektronische Medium

Damit natürliche Sprache auf dem Computer automatisch analysiert und verarbeitet werden kann, muß sie als elektronisch repräsentierte Buchstabenfolge gespeichert sein. Texte, die in dieser Form auf dem Computer gespeichert sind, nennt man auch *on-line* Texte.

Texte und Ausdrücke natürlicher Sprachen existieren jedoch meist in verschiedenen nicht-elektronischen Medien: als *Lautzeichen* der gesprochenen Sprache, als *Buchstaben* der geschriebenen oder gedruckten Sprache oder als die *Gesten* einer Taubstummensprache. Während Lautzeichen und Gesten normalerweise nur eine sehr kurze Lebensdauer haben (von Tonoder Videoaufnahmen einmal abgesehen), zeichnet sich die Schrift, konventionell fixiert auf Papier, Pergament oder Stein, durch ihre überdurchschnittlich lange Haltbarkeit aus.

Im Gegensatz zur konventionellen Speicherung von Schriftsprache, werden im elektronischen Medium Magnetband, Diskette oder CD-ROM als Datenträger verwendet. Die Übertragung nicht-elektronisch gespeicherter Sprache in das elektronische Medium ist aufwendig und kann in verschiedener Weise erfolgen.

Eine Möglichkeit ist das Eintippen gesprochener oder geschriebener Sprache (etwa durch eine Sekretärin) in den Computer. Dies ist heute noch eine weit verbreitete Methode, z.B. das Tippen vom Diktiergerät in Büros, das Transskribieren von Tonbandaufnahmen in der Psychologie, oder das Eingeben von Büchern, die bisher nur in gedruckter Form vorlagen.

Hinzu kommen heute Technik-basierte Methoden. Die automatische Überführung von (auf Papier) gedruckter Sprache in das elektronische Medium fällt in den Bereich der *optischen Mustererkennung* und wird mit Hilfe sogenannter *Scanner* vorgenommen. Diese Maschinen machen nicht nur ein Abbild der Seite, wie es auch eine Fotografie tun würde, sondern sie tasten die einzelnen Buchstaben zeilenweise ab und vergleichen sie mit gespeicherten Mustern. Auf diese Weise wird das Druckbild nicht nur in den Computer abgebildet (als sogenannte *bitmap),* sondern buchstabenweise erkannt. Dies ist der Prozess der *optical character recognition* oder OCR.

Nun kann sich das Druckbild von einem Buch zum nächsten sehr stark unterscheiden. Hinzu kommen unterschiedliche Buchstabengrößen und Formatierungen wie bei Überschriften, Fußnoten, Bildunterschriften oder Tabellen. Dies bewältigen moderne Scanner mit Hilfe einer initialen

Lernphase, in der der Benutzer Fehlklassifikationen korrigieren kann, indem er dem Programm eingibt, ob es sich bei einem bestimmten Buchstaben z.B. um ein 'd' oder um ein 'a' handelt.

Zusätzlich verwenden Hochleistungs-Scanner große Lexika, mit deren Hilfe sie in Zweifelsfällen entscheiden, welche von zwei Möglichkeiten eine sinnvolle Wort form darstellt. Auf diese Weise kann man, abhängig vom verwendeten Schrifttyp und der Qualität des Schriftbilds, eine Erkennungsrate von bis zu 99% erreichen, wobei das Gerät für eine Seite zwischen 50 Sekunden und mehreren Minuten benötigt!

Im Vergleich zu dem Abtippen einer Buchseite durch einen Menschen ist die Geschwindigkeit heutiger Scanner bereits durchaus konkurrenzfähig, besonders wenn man bedenkt, daß die Maschine nicht ermüdet und die Bedienung eines Scanners von einer ungelernten Kraft geleistet werden kann. Der wichtigste Faktor ist jedoch die Fehlerfreiheit, und hier erfordern beide Übertragungsformen, daß bei wichtigen Dokumenten nachträglich Korrektur gelesen werden muß.

Die Leistungsfähigkeit von Scannern und ihrer OCR-Software hat sich seit den 80-er Jahren enorm verbessert, bei einem gleichzeitigen Preisverfall, wie er für die Computerbranche charakteristisch ist. Deshalb kann man seit 1991 eine stark steigende Verbreitung von Scannern in Büros beobachten.

Die Übertragung *gesprochener Sprache* in das elektronische Medium gestaltet sich dagegen wesentlich schwieriger. Während das Druckbild klar getrennte Wörter mit verhältnismäßig gleichförmigen Buchstaben aufweist, muß die sogenannte Spracherkennung *(speech recognition)* einen kontinuierlichen Lautstrom analysieren und zudem mit unterschiedlichen Dialekten, Stimmhöhen und Hintergrundsgeräuschen fertigwerden.

Der Qualitätsmaßstab für die automatischen Spracherkennung ist die Spracherkennung des Menschen. Somit ergeben sich folgende Ansprüche an Systeme der automatischen Spracherkennung:

2.1 Desiderata der automatischen Spracherkennung:

=> Sprecher-Unabhängigkeit

Das System soll spontane Sprache verschiedener Sprecher bewältigen, auch wenn deren Aussprache sich in Tonhöhe, Dialekt, Geschwindigkeit etc. unterscheidet.

=> Domänen-Unabhängigkeit

Das System soll in der Lage sein, gesprochene Sprache in geschriebene Sprache zu übertragen, und zwar unabhängig vom Inhalt.

=> Realistischer Wortschatz

Die Zahl der erkennbaren Wortformen soll der eines normalen Sprechers entsprechen.

=> Robustheit

Auch bei Abbrüchen, Kontraktionen und Verschleifungen der gesprochenen Sprache soll das System in der Lage sein, die intendierten Wortformen zu erschließen.

Heutige Systeme der Spracherkennung erreichen eine gewisse Sprecherunabhängigkeit, indem eine Domäne vorgegeben wird (z.B. Zugauskunft), in deren Rahmen nur ganz beschränkte Dialoge sinnvoll sind. Das Wissen über diese inhaltlichen Einschränkungen der verwendeten Domäne wird - in Kombination mit grammatischem Wissen - dazu genutzt, die wahrscheinlichsten Wortfolgen zu erschließen.

Der Wortschatz dieser Spracherkennungssysteme liegt jedoch nach wie vor bei unter 1000 *Wortformen.* Ein normaler Sprecher verwendet dagegen etwa 10000 Wörter, was im Deutschen etwa 100.000

1 Siehe hierzu D. McClelland 1991.

Wort formen entspricht. Das passive Vokabular eines durchschnittlichen Sprechers ist noch einmal drei bis vier mal so groß.

Trotz dieser Schwierigkeiten wird an der automatischen Spracherkennung z. Zt. weltweit intensiv und mit großem finanziellen Aufwand gearbeitet. Der Grund ist, daß das Diktieren wesentlich einfacher (benutzerfreundlicher) ist als das Eintippen. Die praktischen Ziele reichen von der elektronischen Sekretärin über die automatische Zugauskunft per Telephon zum 'Verbmobil' , einem tragbaren Computer, in den man auf deutsch oder japanisch hineinspricht und der dann (über einen kleinen Lautsprecher) eine englische2 Übersetzung ausgibt.

Daß die heutigen Systeme der akustischen Spracherkennung bei der Interpretation der Schallwellen grammatisches Wissen und Domänenwissen sehr stark mit einbeziehen, ist keineswegs als eine Notmaßnahme anzusehen, um mit deren Hilfe überhaupt zu einem Ergebnis zukommen. Vielmehr entspricht diese Strategie der Situation beim Menschen, der ja auch alle ihm zu Verfügung stehenden Informationen bei der Interpretation von gesprochener Sprache mit zum Einsatz bringt.

Dies ändert jedoch nichts an der Tatsache, daß die Aufgabe der Spracherkennung in nicht mehr und nicht weniger als der Übertragung von gesprochener Sprache in das elektronische Medium besteht. Das elektronische Medium ist naturgemäß das eigentliche Medium der computerlinguistischen Analyse von Lexikon, Morphologie, Syntax, Semantik und Pragmatik.

Mit anderen Worten, die computerlinguistische Analyse elektronisch gespeicherter Sprache erfolgt unabhängig von den anderen Sprachmedien. Je höher aber die Qualität und Effizienz dieser allgemeinen, abstrakten Analyse von Sprache auf dem Computer ist, desto leistungsfähiger ist sie als Grundlage der optischen und akustischen Signalerkennung.

3 Technische Vorteile des elektronischen Mediums

Auch ohne den Einsatz sprachwissenschaftlich-basierter Methoden bietet das elektronische Medium den anderen Medien gegenüber ganz wesentliche Vorteile. Die Möglichkeiten der elektronischen Verarbeitung auf dem Computer sind der Grund, warum Texte, die ursprünglich nur im Druckmedium vorhanden waren, systematisch in das elektronische Medium übertragen werden und nun auf CD gekauft werden können. Beispiele sind:

=> sämtliche Texte des klassischen Griechisch
sämtliche Texte des klassischen Latein

die Shakesspeare Gesamtausgabe die
Encyclopedia Britannica der
Brockhaus/Wahrig

Vergleichen wir z.B. die Benutzung der gedruckten Version eines 10 bändigen Lexikon mit der elektronischen Version auf einer CD-ROM. Der Vorteil liegt in der Geschwindigkeit und Bequemlichkeit beim Finden von relevanten Textstellen auf der CD-ROM. Statt mehrere Bände aus dem Regal zu wuchten und nach den richtigen Seiten zu suchen, genügt bei der CD-ROM die Eingabe der Schlüsselwörter.

Mit der geeigneten Software kann man nicht nur nach den Haupteinträgen suchen, sondern sämtliche Vorkommnisse eines Schlüsselwortes in den Einträgen in Sekundenschnelle finden. Und schließlich kann man Kombinationen von Wörtern suchen, etwa alle Stellen, an denen die Wörter 'Maler', 'Venedig' und '16. Jahrhundert' innerhalb einer Länge von 40 Wörtern vorkommen.

2 Als Voraussetzung für die Benutzung wird angenommen, daß die deutschen und japanischen Gesprächspartner eine passive Kenntnis des Englischen besitzen. Auf diese Weise soll nicht nur der Hörer den Sprecher verstehen können, sondern der Sprecher soll auch in der Lage sein, die automatische Übersetzung des Geräts zu überprüfen.

Diese elektronisch-basierten Suchmethoden sind nicht nur bei Benutzung eines Lexikons oder der wissenschaftlichen Arbeit über Shakespeare, die klassischen Texte der Griechen und Römer, etc. von praktischem Nutzen. Auch bei der Vorbereitung auf einen Prozeß mit Hilfe einer juristischen Datenbank, der computergestützten Diagnose einer seltenen Krankheit oder der Wahl eines spezifischen Medikaments sind elektronische Datenbanken traditionellen Schriftstücken und Zettelkästen haushoch überlegen.

Gegenüber der Schreibmaschine bieten Computer zudem die Möglichkeit, Texte elektronisch zu korrigieren, in andere Dateien zu kopieren, zu edieren und formatieren. Aus diesem Grund entstehen die meisten Texte, die heute publiziert werden, schon primär in elektronischer Form und werden erst ganz am Schluß in das sekundäre Medium des Buch- oder Zeitungsdrucks übertragen.

Betrachten wir z.B. eine Tageszeitung. Früher wurden die einzelnen Artikel mit einer mechanischen Setzmaschine im Bleisatz aus einzelnen Buchstaben zusammengesetzt. Der Inhalt der Tageszeitung existierte nur in Form der Druckplatten, die nach dem Druckvorgang wieder zerlegt, bzw. eingeschmolzen wurden, und in Form der Zeitungsexemplare, die auf Papier gedruckt wurden.

Wenn die Redaktion einen Beitrag eines Informationsdienstes übernehmen wollte, mußte der Beitrag Buchstabe für Buchstabe von der Vorlage nachgesetzt werden. Wenn es vor Druckbeginn eine Sensationsmeldung gab, die man unbedingt aufnehmen wollte, mußte man die Druckplatten mit der Hand umbauen, um Platz für den neuen Beitrag zu schaffen.

Heute existiert der Inhalt der Zeitungen in elektronischer Form. Beiträge der Informationsdienste werden nicht mehr auf Papier geliefert, sondern kommen über das Telephon in einer Form, die mit Hilfe eines Modems in das elektronische Medium rekonvertiert werden kann. Eine Zeitungsausgabe in elektronischer Form kann beliebig umformatiert, kopiert und ediert werden. Jede dieser Versionen kann dann automatisch gedruckt werden.

In 3.1 findet sich ein kurzer Artikel aus einer Tageszeitung, wie er in einem Verlagsrechner gespeichert wurde. Dieser Text enthält die charakteristischen Steuerzeichen für die Setzmaschine.

3.1 Zeitungstext mit Steuerzeichen

```
0509636
      / otagD22801P1008501271738otagotag
<01001> <SB15.HOSO.HX2,3.D42.S451.SGS> <ef> po1itik - panorama
vinde1en - vd++ - otag<OO1,OOO3> <01002> <002,0006> <01003>
<sb14> Heinrich Vinde1en, ++ <mp> <S450> <SGS> <sv> <SK> <DZ>
<ef> Bundesminister f}r <01004> Innerdeutsche++ Beziehungen,
sieht 4nzeichen <01005> f}r eine Einigung zwischen Bonn und++
<01006> Ostber1in in der umstrittenen Frage der <01007> DDR-
Staatsbürgerschaft++ . Im Hessischen <0100S> Rundfunk meinte der CDU-
Po1itiker, ohne <01009++> auf Einze1heiten einzugehen, es
verdichteten <01010> sich die Indizien daf}r++ , da' die SED-
F}hrung <01011> offenbar nicht mehr auf einer "vo11en 4ner++<->
<01012> kennung" bestehe, sondern sich mit einer <01013>
"Respektierung++" durch Bonn zufrieden geben <01014>
klnnte.<014,0042> <014,0042>
```

Wie ist nun dieser Text an die Stelle von Beispiel 3.1 gekommen? Linguisten interessieren sich für Tageszeitungen nicht wegen der aktuellsten Informationen, sondern weil sie Zeitdokumente der Sprache sind. Da Tageszeitungen heute primär in elektronischer Form vorliegen, ist es für Computerlinguisten eigentlich nur ein rechtliches Problem (Copyright), beliebige Mengen von elektronisch gespeicherten Zeitungsausgaben zu besorgen.

Sobald die Genehmigung vorliegt, muß man sich nur noch eine Kopie der Druckerbänder besorgen und in den eigenen Computer einspielen. Danach kann man die Information beliebig verarbeiten. So kann man z.B. eine bestimmte Textstelle herauskopieren und in einem anderen Text unterbringen, wie in 3.1. Eine weitere Möglichkeit ist, die Steuerzeichen zu entfernen bzw. zu interpretieren.

3.2 'gereinigter' Zeitungstext

```
05.09.86
politik - panorama vindelen
Heinrich Windelen, Bundesminister fuer Innerdeutsche
Beziehungen, 5ieht Anzeichen fuer eine Einigung zwischen
Bonn und Ostberlin in der umstrittenen Frage der DDR-
Staatsbuergerschaft. Im Hessischen Rundfunk meinte der
CDU-Politiker, ohne auf Einzelheiten einzugehen, es
verdichteten
sich die Indizien dafuer, dass die SEDFuehrung offenbar
nicht mehr auf einer "vollen Anerkennung" bestehe,
sondern sich mit einer "Respektierung" durch Bonn
zufrieden geben koennte.
```

Falls der Text nicht auch als gedruckte Zeitung vorliegt, kann die Interpretation der Steuerzeichen über den textuellen Kontext erfolgen: z.B. soll 'Staatsb}rgerschaft´ in 3.1 offenbar *Staatsbürgerschaft* heißen und' 'kInnte', offenbar *könnte.*

Weltweit existiert heute noch die Schwierigkeit, daß nicht nur jedes Land seine eigenen Konventionen für Steuerzeichen hat, sondern praktisch jede Druckerei. Da es umständlich und zeitraubend ist, ständig wechselnde Steuerkonventionen zu interpretieren, wurde von der INTERNATIONAL STANDARDS ORGANIZATION (ISO) der sogenannte SGML-Standard entwickelt:3

3.3 SGML: *standard generalized markup language.*

> A family of ISO standards for labeling electronic versions of text, enabling both sender and receiver of the text to identify its structure (e.g. title, author, header, paragraph, etc.)
>
> Dictionary of Computing, S. 416
> (Illingworth et al. 1990)

Der SGML-Standard wird auch in Europa, und damit Deutschland, anerkannt und findet mit der Zeit immer weitere Verbreitung.

3 Eine ausführliche Darstellung zum Thema SGML findet sich in Herwijnen 1990.

Denn elektronische Texte, die die Konventionen dieses Standards einhalten, haben den Vorteil, daß ihre Steuerzeichen von allen anderen SGML-Benutzern automatisch interpretiert werden können.4

Ein im Computer gespeicherter Text kann nach den individuellen Absichten und Bedürfnissen des Benutzers elektronisch verändert werden. So kann z.B. der "gereinigte" Zeitungsartikel 3.2 mit Hilfe eines Editors für die Verarbeitung in LATEX wie folgt aufbereitet werden.

3.4 19\TEX-Forma.tierung eines Texts

```
\documentstyle{artikel}
\begin{document}

\noindent
05.09.86\\
{\bf Politik:} - {\it panorama windelen}\\ Heinrich Windelen,
Bundesminister f\"{u}r Innerdeutsche Beziehungen,
sieht Anzeichen f\"{u}r eine Einigung zwischen Bonn
und Ostberlin in der umstrittenen Frage der DDR-
Staatsb\"{u}rgerschaft. Im Hessischen Rundfunk meinte
der CDU-Politiker, ohne auf Einzelheiten einzugehen,
es verdichteten sich die Indizien daf\"{u}r, da{\ss}
die SED-F\"{u}hrung offenbar nicht mehr auf einer
"vollen Anerkennung" bestehe, sondern sich mit einer
"Respektierung" durch Bonn zufrieden geben k\"{o}nnte.
\end{document}
```

LATEX ist eine vereinfachte Version von TEX, welches von D. KNUTH als Programmiersprache für das Schriftsetzen auf dem Computer entwickelt wurde. Nachdem 3.4 durch das LATEX-Programm geschickt worden ist, gibt der Computer folgendes Schrift bild aus:

3.5 GeTEXte Version des Texts

> 05.09.86
> Politik: - *panorama windelen*
> Heinrich Windelen, Bundesminister für Innerdeutsche Beziehungen, sieht Anzeichen für eine Einigung zwischen Bonn und Ostberlin in der umstrittenen Frage

4 Als Vereinfachung der sehr mächtigen SGML wurde inzwischen der TEI Standard vorgeschlagen. TEI ist eine Spezialisierung (Untermenge) der SGML und steht für *text encoding initiative.*

der DDR- Staatsbürgerschaft. Im Hessischen Rundfunk meinte der CDU-Politiker, ohne auf Einzelheiten einzugehen, es verdichteten sich die Indizien dafür, daß die SED- Führung offenbar nicht mehr auf einer "vollen Anerkennung" bestehe, sondern sich mit einer "Respektierung" durch Bonn zufrieden geben könnte.

3.4 und 3.5 illustrieren nur ganz einfache D-TEX Befehle, z.B. für das Fettdrucken ({\bf }, für *bald face)* und das Kursivdrucken ({\it }, für *italic).* Weiterhin finden wir eine Behandlung von Umlauten und scharfem 's', sowie ein rechtsbündiges Druckbild, wobei das Programm Worttrennungen am Zeilenende automatisch vornimmt.

Hinzu kommt die automatische Behandlung der Kapitel- und Sektionsüberschriften, die automatische Erstellung von Inhaltsverzeichnissen und Indices, und vieles mehr. Vor allem bei der Darstellung mathematischer Formeln sind TEX und LATEX außerordentlich leistungsfähig.

Seit ihrer Einführung im Jahre 1984 werden TEX und LATEX immer mehr zur Publikation von wissenschaftlichen Büchern und Zeitschriften verwendet, wobei Wissenschaftler ihre Aufsätze und Bücher nicht nur auf dem Computer schreiben, sondern auch selbst formatieren und in druckfertiger Form beim Verlag abliefern. Die Publikation über das elektronische Medium ist nicht nur wesentlich kostengünstiger als ein konventionell gesetztes Buch, sondern hat auch viele praktische Vorteile, insbesondere die direkte Einflußnahme des Autors auf die Gestaltung und die Tatsache, daß das Korrekturlesen der Setzer arbeit entfällt.

Neben den Möglichkeiten der schnellen, computergestützten Suche von Textstellen und dem *desktop publishing* (DTP) bieten elektronisch gespeicherte Texte auch leistungsfähige Möglichkeiten der linguistischen Analyse. So kann man den Text 3.2 z.B. in wenigen Schritten in eine alphabetische Wortliste verwandeln.

3.6 Alphabetische Wortformenliste des Texts

05.09.86	Windelen	koennte
Anerkennung	auf	mehr
Anzeichen	auf	meinte
Beziehungen	bestehe	mit
Bonn	dafuer	nicht
Bonn	dass	offenbar
Bundesminister	der	ohne
CDU-Politiker	der	panorama
DDR-Staats-		
buergerschaft	der	politik
Einigung	die	sich
Einzelheiten	die	ohne
Frage	durch	sieht
Heinrich	eme	sondern
Hessischen	emer	umstrittenen
hn	einer	und
Indizien	einzugehen	verdichteten
Innerdeutsche	es	vollen
Ostberlin	fuer	windelen
Respektierung	fuer	zufrieden
Rundfunk	geben	zwischen
SED- Fuehrung	m	

Eine Wortliste wie 3.5 zählt jedes einzelne Vorkommnis einer Wortform und bietet somit die Grundlage für statistische Untersuchungen zur Worthäufigkeit in Texten. Man kann aber auch ebenso einfach eine Wortliste erstellen, in der jede Wortform nur einmal vorkommt (und wo kein Unterschied zwischen Groß- und Kleinbuchstaben gemacht wird). Dieser zweite Typ ist dann z.B. für eine lexikalische Kategorisierung das Geeignete.

Die bisher genannten Verfahren der automatischen Suche von Wörtern oder Wortfolgen in großen textuellen Datenbanken, der automatischen Fehlersuche ("spelling checker" über Vergleiche mit Wortlisten), das Formatieren mit Hilfe von Steuerzeichen, die Umformung von Texten in alphabetische Wortformenlisten etc. sind rein technologische Verfahren der Zeichenverarbeitung im elektronischen Medium. Sie basieren in keiner Weise auf linguistischen

Konzepten, Theorien oder Methoden.5

Im Vergleich zu nicht-elektronischen Verfahren (Bleisatz, Zettelkästen, Suche in großen Dokumenten in Form von Durchblättern, bzw. Durchlesen etc.) sind diese elektronischen Verfahren enorm schnell, präzise und bequem zu handhaben. Sie erleichtern nicht nur die praktische Arbeit mit Texten, sondern sie liefern auch wertvolle Daten (alphabetische Listen von Wortformen, statistische Aussagen über die Häufigkeit von Wortformen, Paaren von Wortformen, Tripeln von Wortformen- sogenanten *trigrams* - etc. in großen Texten) für die linguistische Analyse.

Gleichzeitig zeigen sich aber auch deutliche Grenzen. Sie bestehen darin, daß die Technologie-basierten Verfahren rein Buchstaben-orientiert sind. Eine grammatikalische Analyse der Wortformen, der syntaktischen Struktur und, darauf aufbauend, des Inhalts, liegt außerhalb dieser Technologie in der Domäne der Sprachwissenschaft.

4 Komponenten der Grammatik

In welchen Bereichen kann mit den Methoden der Sprachwissenschaft eine substantielle Verbesserung der elektronischen Textverarbeitung erreicht werden? Als erste Grundlage für eine Antwort auf diese Fragen werden im Folgenden die Komponenten der Grammatik und ihre Funktionen beschrieben.

Dabei muß berücksichtigt werden, daß sich innerhalb der Sprachwissenschaft drei unterschiedliche Ansätze der grammatischen Analyse herausgebildet haben, nämlich (a) die TRADITIONELLE GRAMMATIK, (b) die THEORETISCHE LINGUISTIK und (c) die COMPUTERLINGUISTIK. Diese drei Ansätze unterscheiden sich bezüglich ihrer

1. Methoden,
2. Fragestellungen
   (also den deskriptiven bzw.
   explanatorischen Zielen) und
3. Anwendungen.

Bevor wir die Komponenten der Grammatik beschreiben, beginnen wir mit einem schematischen Vergleich der drei verschiedenen Ansätze innerhalb der Sprachwissenschaft.

4.1 Drei unterschiedliche Ansätze der Sprachanalyse

=> Traditionelle Grammatik

Die traditionelle Grammatik ist von der Methode her taxonomisch (deskriptiv-klassifikatorisch) orientiert.

Ihr Ziel ist ein möglichst vollständiges Sammeln und Klassifizieren der sprachlichen Einzelphänomene, insbesondere die Darstellung der sprachlichen Regelmäßigkeiten und der Ausnahmen.

Von der Anwendung her kommt sie aus dem Sprachunterricht.

Für die Computerlinguistik ist die traditionelle Grammatik wegen ihrer empirischen Datenfülle von großem Interesse.

=> Theoretische Linguistik

Die Methode der theoretischen Linguistik ist logisch-mathematisch: es werden formale Regelsysteme formuliert, aus denen alle und nur die wohlgeformten sprachlichen Strukturen ableitbar sein sollen. Dies hat der traditionellen Grammatik gegenüber den methodologischen Vorteil der *expliziten Hypothesenbildung* - allerdings nur theoretisch, denn eine Überprüfung der formalen Regelsysteme an realistischen Datenmengen ist mit Papier und Bleistift praktisch unmöglich.

5 Es zeigt sich aber schon an einem so einfachen Problem wie der automatischen Worttrennung am Zeilenende, etwa im Rahmen des *desk top publishing*, daß Technologie und Linguistik bei der automatischen Textverarbeitung eng zusammenliegen.

0bwohl die theoretische Linguistik nach wie vor in viele verschiedene Schulen zersplittert ist, gilt als gemeinsames Erklärungsziel die formale Charakterisierung des menschlichen Sprachvermögens, und zwar unter Ausgrenzung der Sprachwendung *(Performance)* .

Anwendungsversuche reichen von Erklärungsmodellen in der Psychologie bis zum Sprachunterricht in der Schule.

Für die Computerlinguistik sind vor allem die Untersuchungen zu formalen Sprachklassen und Komplexität relevant.

=> **Computerlinguistik**

Methodisch verbindet die Computerlinguistik das Ziel der traditionellen Grammatik einer möglichst vollständigen Klassifikation natürsprachlicher Phänomene mit dem logischmathematischen Ansatz der theoretischen Linguistik. Hinzu kommt allerdings die wichtige Neuerung, daß die expliziten Hypothesen, repräsentiert durch als Parser implementierte formale Grammatiken, .*automatisch* an großen Datenmengen *überprüft* werden können.

Das deskriptive und explanatorische FERNZIEL der Computerlinguistik ist eine *Modellierung der Informationsübertragung mit Hilfe natürlicher Sprachen.* Auf dem Weg zu diesem Ziel muß eine vollständige morphologische, lexikalische, syntaktische, semantische und pragmatische Erfassung der natürlichen Sprachen in einem funktionalen Rahmen geleistet werden.

Mit dem Erreichen dieses Ziels ergeben sich weitreichende Möglichkeiten in der Anwendung der automatischen Sprachverarbeitung. 6

Trotz ihrer unterschiedlichen Methoden, Zielen und Anwendungen liegt den genannten Varianten der Sprachwissenschaft eine gemeinsame Aufteilung der Grammatik in die Komponenten *Phonologie, Morphologie, Lexikon, Syntax, Semantik* und das zusätzliche Gebiet der *Pragmatik* zugrunde. Allerdings variieren Stellenwert und wissenschaftliche Behandlung dieser Komponenten in den verschiedenen Ansätzen der Sprachwissenschaft:

4.2 Die Komponenten der Grammatik

- **Phonologie**

Wissenschaft von den Sprachlauten.

In der theoretischen Linguistik spielt die Phonologie eine zentrale Rolle als eine Art Grundlagendisziplin, in der universale Prinzipien der Sprachanalyse (distinktive Merkmale, formale Regelapparate) exemplarisch vorgeführt werden. Das Ziel ist eine möglichst allgemeine und elegante Darstellung in Form von Regelsystemen, die (a) historische Veränderungen (Lautwandel) oder (b) synchrone Alternationen in der Aussprache (z.B. die sogenannte "Auslautverhärtung" im Deutschen) beschreiben.

In der Computerlinguistik spielt die Phonologie dagegen, wenn überhaupt, eine untergeordnete Rolle. Der einzige Bereich, wo sie möglicherweise zum Einsatz kommen könnte, ist die automatische Spracherkennung. Allerdings wird dieser Bereich heute mit Hilfe der *Phonetik* (und nicht der Phonologie) bearbeitet. Die Phonetik untersucht die Struktur der (a) artikulatorischen, (b) akustischen und (c) auditiven Abläufe. Im Gegensatz zur Phonologie wird die Phonetik nicht zu den Komponenten der Grammatik gerechnet.

- **Morphologie**

Lehre von den Wortformen einer Sprache. Die Morphologie ist der Hauptbereich der

6 Ein Grammatikformalismus, der von vornherein mit dem Ziel einer effizienten Verarbeitung entwickelt wurde, ist die Linksassoziative Grammatik (LAG). Die Komplexitaetseigenschaften der LAG sind in Hausser 1992 beschrieben.

traditionellen Grammatik, wie sie etwa in Schulgrammatiken (z.B. für Latein) zu finden ist. Sie klassifiziert die Wörter einer Sprache nach ihren Wortarten und beschreibt die Wortformen in bezug auf *Flexion, Derivation* und *Komposition.*

In der Computerlinguistik ist die sogenannte *Computermorphologie* ein zentraler Bereich mit der Aufgabe der automatischen Wortformerkennung. Dies geschieht auf Grundlage eines *on-line* Lexikons und eines morphologischen Analyseprogramms. Die automatische Wortformerkennung ist die praktische Voraussetzung für alle anderen linguistisch basierten Verfahren der automatischen Textanalyse.

. Lexikon

Auflistung der Wörter einer Sprache.

Das möglichst vollständige Sammeln und Einordnen der Wörter einer Sprache fällt in die Lexikographie und Lexikologie. Die Lexikographie beschäftigt sich mit den Prinzipien der lexikographischen Kodierung und der Struktur lexikalischer Einträge und ist ein praktisch orientiertes Randgebiet der Sprachwissenschaften. Die Lexikologie untersucht den Wortschatz einer Sprache in Hinblick auf ihre interne Bedeutungsstruktur und ist in der traditionellen Sprachwissenschaft beheimatet.

Was die theoretische Linguistik betrifft, ist seit Mitte der 60-er ein ständig wachsendes Interesse am Lexikon zu verzeichnen. Die Tendenz dieser Arbeiten ist es, immer mehr syntaktische und semantische Eigenschaften komplexer Ausdrücke aus den lexikalischen Eigenschaften der Teilwörter abzuleiten. Das Ergebnis sind umfangreiche formale Darstellungen einzelner Wörter, die der Erklärung syntaktischer Phänomene dienen sollen.

In der Computerlinguistik fungieren *online* Lexika in Kombination mit Morphologieprogrammen bei der automatischen Wortformerkennung. Das Ziel ist eine größtmögliche Vollständigkeit bei möglichst kompakter Speicherung und schnellem Zugriff. Neben der Erstellung neuer Lexika im Rahmen der automatischen Wortformerkennung besteht großes Interesse dar an, das Wissen traditioneller Lexika wie dem OXFORD ENGLISH DICTIONARY (die inzwischen in elektronischer Form existieren) für die automatische Textanalyse nutzbar zu machen (*"mining of dictionaries").*

. Syntax

Beschreibung der grammatisch legalen Kompositionen von Wortformen.

In der theoretischen Linguistik (generativen Grammatik) ist das Ziel der syntaktischen Analyse die Darstellung der grammatischen Wohlgeformtheit in einer Sprache, und zwar mit Hilfe formaler Regeln, die alle, und nur die wohlgeformten, Ausdrücke einer Sprache generieren (erzeugen) bzw. erkennen. Um aus der Fülle der formalen Möglichkeiten die langfristig richtige Beschreibung zu finden, bemüht man sich dabei primär um eine Charakterisierung des menschlichen Sprachvermögens auf der Grundlage von sogenannten Universalien.

Die Computerlinguistik liefert einerseits die technische Voraussetzung, die generative Kapazität formaler Grammatiken für natürliche Sprachen wirklich an der ganzen Fülle der Daten effektiv zu überprüfen. Andererseits haben sich angesichts des großen praktischen Bedarfs an leistungsfähiger automatischer Syntaxanalyse in den letzten dreißig Jahren immer wieder eigenständige, computer-orientierte Systeme entwickelt, die die Syntaxsysteme der theoretischen Linguistik mehr oder weniger direkt beeinflußt haben.

. Semantik

Analyse der wörtlichen Bedeutung sprachlicher Ausdrücke.

In der theoretischen Linguistik reichen die Aufgaben der Semantik von der Charakterisierung syntaktischer Ambiguität und Paraphrase mit Hilfe von (verschiedenen bzw. gleichen) 'Tiefenstrukturen' zu einer logisch-semantischen Darstellung der Wahrheitsbedingungen mit Hilfe logischer Formeln (z.B. MONTAGUE GRAMMATIK).

Dabei befaßt sich das Teilgebiet der Wortsemantik mit der Bedeutungsanalyse von Wörtern bzw. Wortformen, während die Satzsemantik beschreibt, wie sich die Bedeutung komplexer Ausdrücke aus der Bedeutung ihrer Teile und der Art ihrer Zusammensetzung ableitet (FREGE'SCHES PRINZIP).

Die semantische Analyse sprachlicher Ausdrücke umfaßt u. a. die logische Charakterisierung von Einzahl und Mehrzahl (Quantoren), Konjunktion *(und)* und Disjunktion *(oder),* die Verbergänzung durch Subjekt und Objekt (Kasus und Valenz), die Modifikation von Nomina und Verbalkomplex durch Adjektive und Adverbien, die Neben- und Unterordnung von Teilsätzen und vieles mehr. In der Computerlinguistik reicht die Problematik der Bedeutungsanalyse von der Interpretation von Programmiersprachen über die Konsistenz von Datenbanken zu Verfahren der konzeptbasierten Indexierung und der Desambiguierung in der maschinellen Übersetzung.

. Pragmatik

Theorie von der Verwendung sprachlicher Ausdrücke.

Während sich die bisherigen: Komponenten (Phonologie, Morphologie, Lexikon, Syntax und Semantik) mit den strukturellen Eigenschaften sprachlicher Ausdrücke (Wortformen und Sätzen) beschäftigen, untersucht die Pragmatik, wie sich diese strukturellen Eigenschaften bei der Verwendung der Ausdrücke in einem Äußerungskontext auswirken. Deshalb gehört die Pragmatik streng genommen nicht zu den Komponenten der Grammatik, sondern umfaßt (i) die Strukturanalyse der sprachlichen Ausdrücke (Grammatik), (ii) die Beschreibung des (Äußerungs- und Interpretations-) Kontexts und (iii) die Analyse der Interaktion zwischen Sprache und Kontext.

Wenn es darum geht, die natursprachliche Bedeutungsübertragung zwischen Menschen, bzw. zwischen Mensch und Maschine, theoretisch und praktisch zu modellieren, dann darf die Pragmatik mit ihren drei Teilkomponenten keinesfalls fehlen. Die Verwendung von sprachlichen Ausdrücken umfaßt die Referenz (also den Bezug sprachlicher Ausdrücke auf die vom Sprecher intendierten Objekte), die Interpretation von indexikalischen Ausdrücken (Pronomina, temporale und lokale Adverbien, Verbalflektion), den rhetorisch korrekten Einsatz von Pronomina und die Wortstellung bei der Generierung, sowie die Interpretation nicht-wörtlicher Verwendungen (z.B. Metaphern).

In der theoretischen Linguistik wird die Pragmatik meist als Teil der modelltheoretischen (logischen) Semantik oder im Rahmen der sogenannten Sprechakttheorie behandelt. In der Computerlinguistik findet die Pragmatik wachsendes Interesse aufgrund von praktischen Problemen bei der rhetorisch korrekten Implementierung des *Generierungsaspekts* (etwa bei Dialogsystemen oder Systemen der maschinellen Übersetzung).

Die beschriebene Aufteilung der grammatikalischen Komponenten ist für traditionelle Linguistik, theoretische Linguistik, und Computerlinguistik deshalb gleichermaßen gültig, weil sie sich an unterschiedlichen strukturellen Aspekten natürlicher Sprachen orientiert, nämlich den *Lauten* (Phonologie), den *Wortformen* (Morphologie), den *Wörtern* (Lexikon), den *Sätzen* (Syntax), den *Bedeutungen* (Semantik) und den *Verwendungen* (Pragmatik).

5 Fernziel der Computerlinguistik

Bis heute wurden (und werden) die Grammatiken der theoretischen Linguistik allein nach ihrer Eleganz, Plausibilität, mathematischen Mächtigkeit, Umfang der Datenerfassung oder Kompatibilität mit psychologischen Tests bewertet. Die Methodik der Computerlinguistik erfordert dagegen zusätzlich, daß die formale Struk-

tur der morphologischen, syntaktischen und semantischen Ableitungen eine gute Basis für einen klaren Programmfluß bietet, der eine einfache, schnelle Fehlersuche und Erweiterung erlaubt. Außerdem ist es für die Modellierung des menschlichen Sprachverständnisses auf dem Computer unerläßlich, daß die grammatischen Analysen für eine funktionsfähige pragmatische Interpretation sowohl bei der *Analyse* als auch bei der *Generierung* optimal geeignet sind.

Deshalb liegt aus wissenschaftlicher Sicht das Wesentliche am computerlinguistischen Ansatz nicht so sehr in den möglichen Anwendungen (obwohl die sicherlich wichtig und interessant sind), sondern vielmehr in seiner neuartigen Methodologie. Das Fernziel der Computerlinguistik, nämlich die Modellierung der menschlichen Sprachverwendung auf dem Computer (siehe 4.1), ist von größter methodologischer Bedeutung, weil es eine funktional-holistische Sichtweise erzwingt.

Die neuere Geschichte der theoretischen Linguistik zeigt, daß ihre Vertreter häufig dem Fehler verfallen, vorhandene Komponenten und Formalismen auf Phänomene anzuwenden, für die sie überhaupt nicht entwickelt wurden, und dies nur, weil die Beschreibungsapparate eine solche Ausweitung oberflächlich zuzulassen scheinen. Es gibt hierfür in der Literatur massenhaft Beispiele, etwa die Behandlung semantischer Phänomene in der Syntax, pragmatischer Phänomene in der Semantik, morphologischer Phänomene in der Syntax, usw. Dies hat immer wieder in Sackgassen geführt, bei denen es meist Jahrzehnte dauerte (und dauert), wieder herauszukommen. Diese Gefahr kann nur durch ein einheitliches und funktionstüchtiges Gesamtkonzept wirklich gebannt werden.

Das Fernziel der Computerlinguistik führt zu der Frage, wieweit es prinzipiell überhaupt möglich ist, ein realistisches Modell natursprachlicher Kommunikation
zu entwickeln und zu implementieren. Ich möchte diese Frage anhand einer Analogie beantworten.

Die heutige Situation in der Computerlinguistik entspricht in vielem der Entwicklung des mechanischen Fluges. Viele hundert Jahre lang hat der Mensch die Spatzen und andere Vögel beobachtet, um zu verstehen, wie sie fliegen. Und er hat versucht, sich auf möglichst ähnliche Weise in die Lüfte zu erheben.

Es hat sich dann herausgestellt, daß es mit Flattern nicht geht. Dies wurde gerne zum Anlaß genommen, den menschlichen Flugverkehr für prinzipiell unmöglich zu erklären, häufig mit dem frommen Spruch:

Wenn Gott gewollt hätte, daß die Menschen fliegen könnten, hätte er ihnen Flügel verliehen.

Heute ist das Fliegen für die Menschen selbstverständlich geworden. Außerdem weiß man inzwischen, daß ein Spatz aufgrund desselben theoretischen Prinzips in der Luft bleibt wie ein Jumbojet, nämlich dem Prinzip der "*air foir'*, der Tragflächen.
Es gibt also eine Ebene der Abstraktion, auf der der Flug des Spatzen und der Flug des Jumbojets nach demselben Prinzip ablaufen.

Bei der Modellerierung natursprachlicher Kommunikation in der Computerlinguistik geht es ebenfalls um das korrekte Prinzip auf der korrekten Abstraktionsebene. Dabei besteht naturgemäß die Gefahr, die Ebene entweder zu niedrig oder zu hoch anzusetzen. Zum Beispiel wären geschlossene Signalsysteme, wie man sie etwa bei einem Fahrkartenautomaten findet, sicherlich als Modell ungeeignet.7

Genauso unsinnig wäre es aber auch, die Modellierung natursprachlicher Kommunikation von vornherein mit naiven antropomorphen Vorstellungen *ad absurdum* zu führen. Wer z.B. mit einem Begriff

7 Der entscheidende Punkt der in Frage stehenden Modellierung ist, daß die charakteristische Vielseitigkeit natursprachlicher Kommunikation erhalten bleibt - also die Tatsache, daß dieselben Ausdrücke in den verschiedensten Äußerungskontexten sinnvoll eingesetzt werden können.

von "Verstehen" antritt, wonach sich das System bei der Analyse von FINNEGAN'S WAKE subtil amüsieren muß, liegt ebenso daneben, wie jemand, der Paarungsverhalten und Brutpflege von einem Jumbojet erwartet.

Zum Schluß noch eine zweite Analogie aus der Geschichte des Flugzeugbaus: nachdem man an Doppeldeckern, Propellerflugzeugen und Jets ein immer besseres Verständnis der Flugprinzipien entwickelt hat, analysiert man heute wieder verstärkt natürliche Flugvorgänge, um ihre wunderbare Leistungsfähigkeit zu begreifen und erfolgreich in den Bau leiserer und effizienterer Flugzeuge einzubringen.

Dieses Beispiel zeigt, daß theoretisch-technologische Lösungsversuche in der Computerlinguistik keinesfalls ein fehlendes Interesse an der Analyse menschlicher Sprachfähigkeiten implizieren. Vielmehr ist es so, daß eine Untersuchung des speziell Menschlichen am sprachlichen Kommunikationsprozeß erst dann wirklich sinnvoll wird, nachdem eine prinzipielle Modellierung natursprachlicher Kommunikation geleistet worden ist und sich in massiven Anwendungen bewährt hat.

Bibliographie

Illingworth et al. (Hrsg.) (1990) *Dictionary oi Computing,* Oxford University Press, Oxford.

Hausser, R. (1989) *Computation oi Language,* Springer-Verlag, Symbolic Computation: Artificial Intelligence, Berlin-New York.

Hausser, R. (1992) *"Complexity in Left Associative Grammar, Theoretical Computer Science,* Vol. 103:283-308, Elsevier.

Herwijnen, E. van (1990) *Practical SGML,* Kluwer Academic Publishers.

Hutchins, W.J. (1986) *Machine Translation: Past, Present, Future,* Ellis Horwood Lmt., Chichester .

McClelland, D. (1991) *"OCR: Teaching Your Mac to Read,"* MACWORLD, November 1991:169-175.