1 Computerlinguistik – Was ist das?

Kapitelherausgeber: Kai-Uwe Carstensen, Susanne Jekat und Ralf Klabunde

Die Computerlinguistik ist das Fachgebiet, das sich mit der maschinellen Verarbeitung natürlicher Sprache beschäftigt. Sie ist im Überschneidungsbereich von Informatik und Linguistik angesiedelt, aber die Wurzeln der Computerlinguistik reichen bis in die fünfziger Jahre zurück. In diesem halben Jahrhundert seit ihrem Entstehen hat sie sich mittlerweile national und international erfolgreich etabliert, so dass auf dem Wissen aus der Informatik und der Linguistik aufbauend neue und eigenständige Methoden für die maschinelle Verarbeitung gesprochener und geschriebener Sprache entwickelt wurden.

Unterkapitel 1.1 bringt die in diesem Buch dargestellten Grundlagen, Methoden und Anwendungen in einen umfassenden Bezug zu den verschiedenen Aufgaben der Computerlinguistik.

Anschließend werden in Unterkapitel 1.2 die zwei Verarbeitungsparadigmen der Computerlinguistik, die symbolische und die stochastische Verarbeitung, aus historischer Sicht vorgestellt.

1.1 Aspekte der Computerlinguistik

Jan W. Amtrup

Der Einfluss der Computerlinguistik (CL) auf das tägliche Leben in unserer „Informationsgesellschaft“ wächst. Es ist fast unvermeidlich, dass man mit den aus dieser relativ neuen Wissenschaft entstandenen Produkten in Kontakt kommt, sei es beim Surfen im Internet oder beim normalen Gebrauch des Computers. Ein Achtklässler, der am Computer einen Hausaufsatz schreibt, benutzt morphologische Prozesse (Rechtschreibkorrektur), grammatische Analyse (Grammatiküberprüfung), eventuell auch statistische Informationen über den geschriebenen Text (Häufigkeitsanalysen) oder Lexikographie (Thesaurus). Kommt eine Internet-Recherche dazu, erweitert sich der Kreis der Methoden um Informationserschließung und möglicherweise vollautomatische maschinelle Übersetzung.

Aber selbst wenn man keinen Computer benutzt, wird man mit Anwendungen der Computerlinguistik konfrontiert, etwa beim Lesen der halbautomatisch übersetzten Bedienungsanleitung für den neuen Toaster oder beim Telefonat mit der Bank, an dessen Beginn eine freundliche Maschine nach der Kontonummer fragt.

Diese wenigen Beispiele machen deutlich, welche Bedeutung die Computerlinguistik in den letzten Jahren erfahren hat: Sie erschließt Informationsquellen, erleichtert den Umgang mit Maschinen und hilft, Grenzen zwischen verschiedenen Sprachen zu überwinden.

1.1.1 Computerlinguistik: Die Wissenschaft

Gegenstand der Computerlinguistik ist die Verarbeitung natürlicher Sprache (als Abgrenzung zu z. B. Programmiersprachen) auf dem Computer, was sowohl geschriebene Sprache (Text) als auch gesprochene Sprache (engl: *speech*) umfasst. Computerlinguistik ist im Kern und von ihrer Historie her (siehe Unterkapitel 1.2) eine Synthese informatischer und linguistischer Methoden und Kenntnisse.

Diese Charakterisierung ist bewusst sehr allgemein gehalten, um die verschiedenen Auffassungen von „Computerlinguistik“ zu umfassen, die in diesem Buch vereint werden sollen:

- Computerlinguistik als Teildisziplin der Linguistik (wie Psycholinguistik, Soziolinguistik usw.), die sich, in der Regel theoriegeleitet, mit berechnungsrelevanten Aspekten von Sprache und Sprachverarbeitung beschäftigt (vgl. auch den englischen Terminus für Computerlinguistik, *computational linguistics*), unabhängig von ihrer tatsächlichen Realisierung auf dem Computer. Die Entwicklung von Grammatikformalismen ist ein Beispiel für diese Auffassung von Computerlinguistik.

- Computerlinguistik als Disziplin für die Entwicklung linguistik-relevanter Programme und die Verarbeitung linguistischer Daten („Linguistische Datenverarbeitung“). Diese Auffassung hat ihre Wurzeln in den Anfängen der Informatik und hat insbesondere durch die zunehmende Wichtigkeit empirischer Untersuchungen anhand umfangreicher Sprachdatenkorpora (s. Kapitel 4) eine Renaissance erfahren.

- Computerlinguistik als Realisierung natürlichsprachlicher Phänomene auf dem Computer („maschinelle Sprachverarbeitung“, engl: *natural language processing*). Die Untersuchung vieler dieser Phänomene hat eine lange Tradition innerhalb der Sprachphilosophie bzw. der sprachorientierten formalen Logik. Da Sprache als Teil eines kognitiven Systems aufgefasst werden kann, in dem sprachliche Kenntnis und nicht-sprachliches Wissen, Denkprozesse und Handlungsplanung eng miteinander verknüpft sind, sind insbesondere die Künstliche Intelligenz und die Kognitionswissenschaft an der Untersuchung bzw. der Modellierung dieser Phänomene interessiert. Die Computerlinguistik ist daher untrennbar mit den formalen und/oder kognitiven Disziplinen verknüpft.

- Computerlinguistik als praxisorientierte, ingenieursmäßig konzipierte Entwicklung von Sprachsoftware („Sprachtechnologie“).

Diese Liste verschiedener Auffassungen veranschaulicht *prinzipielle* Unterschiede in der Auffassung von Computerlinguistik. Die Computerlinguistik, die in diesem

Buch vorgestellt werden soll, ist als Summe und Synthese ihrer verschiedenen Ausprägungen zu verstehen.

Hierbei bilden vier Bereiche die Eckpfeiler der Computerlinguistik: Die Entwicklung von Methoden, durch die natürlichsprachliche Phänomene operationalisiert werden; der Aufbau und die Verwaltung großer wiederverwendbarer Korpora sprachlicher Daten, die für empirische, Entwicklungs- und Evaluationszwecke genutzt werden können; die Entwicklung realistischer Anwendungen, die die Relevanz der Computerlinguistik für die moderne Informationstechnologie aufzeigen und die gleichzeitig ihren technologischen Fortschritt widerspiegeln; und die Konzeption effektiver Evaluationsmechanismen, durch die der angesprochene Fortschritt objektiviert wird. Zudem ist die Computerlinguistik in fachlichen Grundlagen verankert, die sie zum Teil aus ihren Mutterdisziplinen erbt und zum Teil von weiteren Nachbardisziplinen übernimmt.

1.1.2 Computerlinguistik und ihre Nachbardisziplinen

Von der *Linguistik* übernimmt die Computerlinguistik den Untersuchungsgegenstand und gleichzeitig das Grundinventar linguistischer Termini und Differenzierungen. Die Strukturierung der Methodenbereiche in der Computerlinguistik orientiert sich daher weitestgehend an den etablierten Teilgebieten der Linguistik: Phonologie, Morphologie, Syntax, Semantik und Pragmatik, welche die Schwerpunktebenen der strukturellen Beschreibung natürlichsprachlicher Äußerungen bilden (vgl. etwa Grewendorf, Hamm und Sternefeld 1987).

Die Computerlinguistik ist aber nicht nur ein Abnehmer linguistischer Theorien und Sachverhalte, sondern sie kann auch ein Stimulus für Erkenntnisgewinn und die Erarbeitung neuer Ansätze innerhalb der Linguistik sein.

Ein erfolgreiches Beispiel für die interdisziplinäre Arbeit zwischen Linguistik und Computerlinguistik stellt die Entwicklung der Optimalitätstheorie dar (vgl. Prince und Smolensky 1993). Ursprünglich hervorgegangen aus der Verbindung von Ansätzen neuronaler Netze und Prinzipien der Universalgrammatik, um eine bessere Beschreibung der Phonologie zu ermöglichen, ist die Optimalitätstheorie neben regelorientierten Ansätzen inzwischen zu einem konkurrenzfähigen Modell für die Beschreibung phonologischer Sachverhalte geworden. Darüber hinaus wird sie zunehmend zur Beschreibung von Phänomenen auf anderen Ebenen, z. B. der Morphologie und der Syntax, benutzt.

Die Anwendung und Evaluation linguistischer Theorien ist eine weitere Aufgabe für die Computerlinguistik. Erst die Applikation von Theorien auf real vorkommende Daten liefert einen Aufschluss über deren Korrektheit und Vollständigkeit und kann teilweise sogar für deren Verwendung außerhalb streng theoretisch orientierter Kreise sorgen. Als ein Vertreter sei hier die Implementierung eines Systems zur Strukturanalyse erwähnt, das auf dem Prinzipien- und Parameter-Ansatz beruht (Fong 1991).

Und schließlich sind einige Zweige der Linguistik stärker als andere auf die Bearbeitung von Material durch Computer angewiesen. Die Korpuslinguistik etwa, die sich mit der Erforschung linguistischer Zusammenhänge durch die Betrachtung von Korpora befasst, ist erst durch die Verwendung von Computern in den

letzten Jahren dazu in die Lage versetzt worden, realistisch große Datenmengen mit einer hohen Abdeckung (oft im Größenbereich von Milliarden von Wörtern) zu untersuchen.

Die *Informatik* steuert zur Computerlinguistik im Wesentlichen das Wissen über Datenstrukturen sowie die Verwendung effizienter Verfahren bei. Neben dem offensichtlichen Zusammenhang zwischen der Untersuchung und Realisierung natürlichsprachlicher Systeme und der Informatik (Systemanalyse, Modellierung, Algorithmik, Implementation) spielen aber auch Aspekte der theoretischen Informatik (Berechenbarkeit, Komplexitätstheorie und der Bereich der formalen Sprachen) eine wichtige Rolle.

Aus der *Philosophie* (insbesondere der Sprachphilosophie und Logik) stammen vor allem Aspekte der Frage, wie sich Sprache, Denken und Handeln zueinander in Verbindung setzen lassen; Sprache an sich kann nicht nur als losgelöstes Phänomen betrachtet werden, sondern steht in enger Relation zu außersprachlichen Gegebenheiten, sowohl der Welt als solches und (in einem engeren Sinn von Welt) der Gemeinschaft der Sprecher einer Sprache (Schmidt 1968). Die formale Logik ist eines der zentralen Mittel in der Computerlinguistik zur präzisen Darstellung natürlichsprachlicher Phänomene.

Eine Reihe wichtiger Verfahren (z. B. Such- und Planungsverfahren) verdankt die Computerlinguistik der *Künstlichen Intelligenz*. Sie werden beispielsweise bei der Spracherkennung (Unterkapitel 5.4), der grammatikalischen Analyse (Unterkapitel 3.5) und der Generierung (Unterkapitel 5.6) eingesetzt. Vor allem für die Semantik (Unterkapitel 3.6) sind die Formalismen zur Darstellung von sprachlichem und nicht-sprachlichem Wissen (Wissensrepräsentation) relevant, die in der Künstlichen Intelligenz entwickelt worden sind (s. auch Unterkapitel 4.6) – ebenso wie Verfahren und Mechanismen, mit denen aus gegebenen Wissensstrukturen weitere Schlüsse (Inferenzen) gezogen werden. Mit der klassischen, symbolischen Künstlichen Intelligenz hat die Computerlinguistik zudem die verbreitete Verwendung zweier höherer Programmiersprachen, LISP und PROLOG, gemeinsam (vgl. auch das Unterkapitel 3.9).

Die Computerlinguistik steht zudem in enger Beziehung zur *Kognitionswissenschaft*. Das lässt sich dadurch erklären, dass die Sprachbeherrschung ein hochspezialisierter Teilbereich der generellen kognitiven Fähigkeiten des Menschen ist und dass sprachliches und nicht-sprachliches Wissen untrennbar miteinander verknüpft sind. Vor diesem Hintergrund erscheint es sinnvoll, bei der Konzeption von Verfahren zur maschinellen Sprachverarbeitung die Eigenschaften menschlicher Sprachverarbeitung und ihrer Beziehung zu allgemeinen Denkprozessen zu betrachten. Bis heute stellt die Fähigkeit zur adäquaten sprachlichen Kommunikation (Turing 1950, siehe auch Unterkapitel 1.2) einen wichtigen Test für die „Intelligenz“ einer Maschine dar, auch wenn der eigentliche Wert solcher Tests umstritten ist (vgl. z. B. Searle 1990).

Zahlreiche theorie- und anwendungsrelevante Facetten der Computerlinguistik fußen stark auf der Grundlage mathematischer bzw. mathematisch-logischer Theorien (Unterkapitel 2.1). Diese werden gegebenenfalls erweitert oder modifiziert, um die Eigenarten natürlicher Sprache adäquat beschreiben zu können. Beispielsweise basiert ein Großteil der semantischen Beschreibung sprachlicher Äußerungen auf der klassischen Prädikatenlogik. Diese zeigt sich jedoch schon

bei der Darstellung des Unterschieds der beiden folgenden einfachen Ausdrücke als unzulänglich.

(1.1) a) *Ein großer Berg*
b) *Eine große Ameise*

Menschen haben keine Schwierigkeit, eine korrekte Skala für diese beiden Instanzen von *groß* zu finden, während das für eine maschinelle Bearbeitung mit einigem Aufwand, etwa mit dem Einsatz von *Fuzzy-Logik* (Zadeh 1965) für die Behandlung der Vagheit des Adjektivs, verbunden ist. Ein weiteres Beispiel zeigt, dass selbst scheinbar widersprüchliche Aussagen manchmal mit Leichtigkeit verstanden werden können:

(1.2) *Vögel können fliegen.*
Pinguine sind Vögel.
Pinguine können nicht fliegen.

Die alltägliche Annahme hier ist die, dass Vögel *normalerweise* fliegen können, Pinguine hingegen nicht. Um diesen Mechanismus in den Griff zu bekommen, werden oft *Default-Mechanismen* der Künstlichen Intelligenz eingesetzt, die es erlauben, Standardannahmen bei Vorliegen von gegensätzlicher Evidenz zurückzunehmen.

Die formale Beschreibung natürlicher Sprachen steht in einem engen Zusammenhang zum Gebiet der Automatentheorie und formalen Sprachen. Hier werden Repräsentationsmechanismen und Berechnungsmodelle für verschiedene Klassen von Sprachen enwickelt. Die Komplexität einer Sprache determiniert hierbei die Ausdrucksmächtigkeit der zu ihrer Beschreibung notwendigen Repräsentationen. Gleichzeitig wird dadurch auch die Klasse von Maschinen festgelegt, die zur Erkennung und Analyse von Ausdrücken in einer Sprache notwendig sind. Unterkapitel 2.2 führt genauer in diesen Problembereich ein.

Ein weiteres prominentes Teilgebiet der Mathematik, das für Computerlinguisten sehr wichtig ist, ist die Graphentheorie. Dieser Zweig der Mathematik beschäftigt sich mit der Beschreibung von Eigenschaften von Graphen, d.h. von Mengen von Knoten, die durch Kanten verbunden sein können. Graphenartige Repräsentationen sind auch im täglichen Leben oft anzutreffen (z. B. stellt das Liniennetz eines öffentlichen Nahverkehrssystems einen Graphen dar, bei dem die Haltestellen durch Knoten repräsentiert werden können, und die Streckenabschnitte zwischen den Haltestellen Kanten sind). Für die Computerlinguistik ist die Graphentheorie auf zwei Ebenen relevant. Zum einen sind die Objekte für eine ganze Reihe von Beschreibungsmechanismen Graphen, etwa die Merkmalsstrukturen in Unterkapitel 2.3, die in Unterkapitel 4.3 beschriebenen semantischen Hierarchien sowie die in Unterkapitel 4.6 vorgestellten Ontologien und semantischen Netze. Zum anderen spielt die Graphentheorie auch bei der Realisierung von anspruchsvollen Anwendungen für geschriebene und gesprochene Sprache eine herausragende Rolle. Die Einsatzgebiete reichen hier von der Darstellung gesprochener Äußerungen in Form von Wort- oder Phonemgraphen über die Modellierung syntaktischer Analyse als ein Suchproblem in Graphen

bis hin zur Architektur großer Systeme als gerichtete Graphen, die Komponenten und Datenströme beschreiben. Unterkapitel 2.3 befasst sich u.a. mit dieser Problematik.

Neben Logik, Automatentheorie und Graphentheorie spielt die Statistik eine immer größer werdende Rolle für die Computerlinguistik. Diese ist imminent für das Gebiet der automatischen Erkennung gesprochener Sprache, die heutzutage fast ausschließlich mittels stochastischer Automaten betrieben wird (Unterkapitel 5.4). Zusätzlich ist in den letzten Jahren die korpusorientierte Computerlinguistik stark gewachsen, die statistische Aussagen über die tatsächliche Verwendung von Sprache anhand großer Datenmengen extrahiert und Verarbeitungsverfahren zugänglich zu machen versucht (Unterkapitel 4.1, 4.2, 4.5, 5.3). Unterkapitel 2.4 führt genauer in dieses Gebiet ein.

1.1.3 Teilbereiche der Computerlinguistik

Wie viele Disziplinen, hat auch die Computerlinguistik eine theoretisch und eine praktisch ausgerichtete Seite. Die praktische Computerlinguistik ist der im Wesentlichen nach außen sichtbare Anteil: Hier werden neue Anwendungen erforscht und entwickelt, die sich möglicherweise auf dem lokalen Computer anfinden. Die theoretische Computerlinguistik hingegen untersucht die einer maschinellen Verarbeitung zugrundeliegenden Strukturen im Hinblick auf prinzipielle Fragestellungen wie deren Berechenbarkeit, Adäquatheit und Erlernbarkeit. Die Relevanz beider Aspekte wird in den folgenden Abschnitten erläutert.

Praktische Computerlinguistik

Entscheidende Fragen im Bereich der praktischen Computerlinguistik sind die folgenden:

1. Wie konstruiert man ein Softwaresystem zur Verarbeitung natürlicher Sprache?
2. Welche Formalismen scheinen relevant?
3. Welcher Gegenstandsbereich wird modelliert?
4. Welche interessanten einzelsprachlichen oder anwendungsbezogenen Eigenheiten sollen modelliert werden?
5. Was ist das globale Ziel der Entwicklung?

Das Hauptziel besteht somit darin, (sprachliches) Wissen erfolgreich auf einer Maschine zu modellieren und relevante praktische Probleme zu lösen, z. B. die Übersetzung eines Satzes vom Koreanischen ins Englische oder die Erkennung und Analyse einer telefonischen Pizza-Bestellung. Auf dem Weg zu diesem Ziel sind zahlreiche Aufgaben zu erfüllen, von denen einige den Kern der praktischen Computerlinguistik bilden:

- Die Entwicklung von Formalismen, die dazu genutzt werden können, bestimmte Aspekte natürlicher Sprache zu modellieren. Derartige Formalismen finden sich auf allen Ebenen der Beschreibung natürlicher Sprache, mit unterschiedlicher Ausdrucksmächtigkeit und Zielsetzung. Der Einsatz eines Formalismus, der unabhängig von einer bestimmten Sprache deklarativ die Modellierung sprachlicher Gegebenheiten erlaubt, ist von unschätzbarem Vorteil und hat konsequenterweise die direkte Implementierung von sprachverarbeitenden Algorithmen für die Behandlung bestimmter Phänomene in einer bestimmten Sprache weitgehend verdrängt.

- Die Bereitstellung von Wissen über individuelle Sprachen bzw. bestimmte Aspekte einer Sprache. Dazu gehört neben der Lexikographie (Unterkapitel 5.2) vor allem die grammatische Beschreibung einzelner Sprachen (normalerweise noch weiter eingeschränkt auf bestimmte Anwendungszusammenhänge oder Verwendungsformen). Ein wichtiges Teilgebiet ist die Beschäftigung mit realen Sprachdaten (d.h. die Sammlung, Aufbereitung und Verwaltung von Texten und Sprachaufnahmen, Unterkapitel 4.1–4.5). Die Menge und Verfügbarkeit solcher computerlinguistischer Ressourcen nimmt ständig zu, insbesondere deswegen, da sich die statistischen Eigenschaften bestimmter Phänomene anhand großer Datenmengen besser untersuchen lassen.

- Die Entwicklung von Algorithmen und Methoden zur Bearbeitung natürlichsprachlicher Äußerungen. Die Aufgabenfelder reichen hier von der Erkennung gesprochener Sprache über den Parserbau bis hin zum Design von Dialogsystemen für spezielle Anwendungen (vgl. die Unterkapitel 3.5, 5.4, und 5.5).

- Die Evaluation natürlichsprachlicher Systeme. Um die Performanz und Bandbreite eines Algorithmus oder Systems zu bewerten, reicht es normalerweise nicht aus, einige wenige Beispiele zu verarbeiten. Vielmehr ist es das Ziel, real vorkommende Daten in hinreichender Menge zu untersuchen. Dies gilt uneingeschränkt für Systeme, die auf einer statistischen Modellierung beruhen; aber auch für rein symbolische Verfahren werden Evaluierungen immer wichtiger. Kapitel 6 führt genauer in die Verfahrensweisen ein.

Ein Beispiel für ein Anwendungssystem, das hier prototypisch für den Einsatz praktischer Computerlinguistik genannt werden soll, ist **SmartWeb** (Reithinger, Herzog und Blocher 2007). Dies ist ein multimodales Zugangssystem zum *semantic web*, einem Ausschnitt des Internets, dessen Inhalte durch Metainformationen so angereichert sind, dass Korrelationen einfach hergestellt werden können. Für den Benutzer stellt sich SmartWeb schlicht als eine Applikation auf dem Mobiltelefon dar, die bei einigen täglichen Verrichtungen helfen kann, etwa bei der Auswahl eines Restaurants für den Abend und der Planung einer Autoroute dorthin mit einem kurzen Zwischenstopp an einer Tankstelle. Die

zugrundeliegenden Informationen sind sämtlich im Internet vorhanden; das Auffinden und Verknüpfen der Daten zu einem kohärenten Plan jedoch ist manuell mit einiger Mühe verbunden.

SmartWeb benutzt bereits vorhandene semantisch annotierte Informationsquellen direkt. Um den Zugang zu konventionellen Web-Seiten zu ermöglichen, wurden Verfahren entwickelt, deren Inhalt zumindest in Grenzen automatisch zu verstehen und maschinell zu annotieren.

Zur Realisierung eines solch umfangreichen Projekts sind nicht nur theoretische Einsichten der Computerlinguistik erforderlich; daneben müssen nahezu alle Teilgebiete der praktischen Computerlinguistik herangezogen werden.

Zunächst gilt es, gesprochene Sprache zu erkennen; für die hier angesprochene Anwendung wird das noch kompliziert durch die Vielzahl an Namen (Straßen, Orte, Restaurants usw.), für die das Spracherkennungssystem nicht vorher explizit vorbereitet werden kann. Außerdem kann die sprachliche Eingabe durch andere Modalitäten unterstützt werden, etwa durch Gesten oder über die Tastatur. Diese multimodalen Eingabeäußerungen müssen auf multiplen Ebenen analysiert werden: Syntaktisch, semantisch, und im Hinblick auf ihre Funktion innerhalb des Dialogkontextes. Das Ziel des Benutzers muss erschlossen werden, um die adäquaten Daten aus dem Semantic Web abzurufen. Und schließlich ist es erforderlich, die Resultate multimodal passend aufzubereiten, sei es als Text, in Form einer Landkarte, als Bild, Video oder Ausgabe über einen Lautsprecher.

Über die Entwicklung der Formalismen und Verarbeitungsmechanismen für einzelne Teilbereiche einer Gesamtanalyse hinaus muss allerdings auch dafür gesorgt werden, dass alle Einzelbausteine korrekt und effizient zusammenarbeiten können. Hier werden dann Fragen der Architektur von großen natürlichsprachlichen Systemen und softwaretechnische Aspekte der Integration von Komponenten sowie deren Kommunikation untereinander relevant.

1.1.4 Theoretische Computerlinguistik

Innerhalb der theoretischen Computerlinguistik geht es um die Frage, wie natürliche Sprache formalisiert und maschinell verarbeitet werden kann, ohne dass der Blickwinkel durch die Notwendigkeit, ein tatsächlich funktionierendes System bauen zu müssen, eingeschränkt wird. Abhängig vom tatsächlichen Fachgebiet sind Logik, formale Linguistik und Compilerbau wichtige Grundlagen für erfolgreiche Forschung, während Detailwissen um anwendungsrelevante Aspekte nicht zentral erscheint.

Formalismen spielen auch hier eine große Rolle, allerdings weniger unter dem Blickwinkel, Grammatiken mit einer hohen Abdeckung für eine konkrete Sprache anzufertigen. Vielmehr stehen prinzipielle Fragen wie die Eignung eines Formalismus zur Beschreibung verschiedener Phänomene oder die Komplexität einer Berechnung mittels eines Formalismus im Mittelpunkt. Wichtige Fragestellungen sind etwa:

- Welche Komplexität weist natürliche Sprache an sich auf, und inwieweit kann diese Komplexität durch heutzutage verfügbare Maschinen effektiv bewältigt werden? (vgl. Unterkapitel 2.2)

- Welche Eigenschaften muss ein Formalismus aufweisen, um relevante Aspekte natürlicher Sprache angemessen repräsentieren zu können? Diese Frage stellt sich ebenenübergreifend, so dass zum Teil unterschiedliche Formalismen zur Darstellung von Phonetik, Phonologie, Morphologie, Syntax, Semantik und Pragmatik entwickelt werden. Dies wirft wiederum die Frage auf, bis zu welchem Grade die Repräsentation ebenenübergreifend stattfinden kann, und welche Vor- und ggfs. Nachteile dies mit sich bringt.

Als ein Beispiel für die Forschung in der theoretischen Computerlinguistik sei hier die adäquate Modellierung syntaktischer Strukturen für natürlichsprachliche Äußerungen genannt. Beginnend mit Chomsky (1959) werden verschiedene Komplexitätsklassen formaler Sprachen unterschieden (siehe Unterkapitel 2.2). Diese Klassen entsprechen unterschiedlich komplexen Methoden zur Erkennung und Strukturanalyse. Gemeinhin wird angenommen, natürliche Sprachen seien zwischen den kontextfreien und kontextsensitiven Sprachen angesiedelt; sie sind „schwach kontextsensitiv“. Allerdings sind die Phänomene, die es notwendig machen, über den kontextfreien Rahmen hinauszugehen, eher selten (vgl. Sampson 1983, Shieber 1985). Ein wesentliches Motiv für die Entwicklung komplexer, merkmalsbasierter Formalismen ist denn auch weniger deren prinzipielle theoretische Notwendigkeit, sondern vielmehr ein stärkeres Bestreben nach der adäquaten Beschreibung natürlichsprachlicher Phänomene. Wichtige linguistische Merkmale (wie Kongruenz, Koreferenz oder Spuren) lassen sich kontextfrei analysieren, allerdings verliert die Modellierung an Allgemeingültigkeit dadurch, dass nicht über die Werte bestimmter Merkmale (Kasus etc.) abstrahiert werden kann. Auf der anderen Seite besteht die Gefahr, durch einen zu mächtigen Formalismus Effizienz (und manchmal sogar Berechenbarkeit) einzubüßen. Daher wird innerhalb der theoretischen Computerlinguistik nach Wegen gesucht, komplexe Beschreibungsformalismen zu entwickeln, die gleichzeitig handhabbar und eingängig sind. Im Laufe der Zeit sind zahlreiche Vertreter solcher Modelle entstanden, die in der Folge auch innerhalb der praktischen Computerlinguistik (und zuweilen in kommerziellen Anwendungen) populär geworden sind (Lexical Functional Grammar (Bresnan 1982), Head Driven Phrase Structure Grammar (Pollard und Sag 1987), und Tree Adjoining Grammar (Joshi 1985), um nur einige Beispiele zu nennen).

Ein immer wichtiger werdender Anteil der theoretischen CL beschäftigt sich mit der Frage, ob und wie eine signifikante Untermenge sprachlicher Konstrukte und Konzepte automatisch erlernt werden kann[1]. Dies hängt neben der Verfügbarkeit hochgradig leistungsfähiger Computer vor allem mit der ständig wachsenden Menge an Text zusammen, die leicht zugänglich ist.

Das initiale Problem ist das der Umwandlung von natürlichsprachlichen Eingaben in interne Repräsentationen oder direkt in andere natürlichsprachliche Ausgaben. Dies kann sich auf mehreren Ebenen abspielen: z. B. kann eine morphologische Analyse oder die Zuweisung von Wortarten (*Part-of-Speech Tagging*) als ein Klassifikationsproblem verstanden werden, bei dem jedes Wort der Ein-

[1] Die Erlernbarkeit durch Maschinen steht hier im Vordergrund, nicht die Untersuchung der Mechanismen, die es Menschen erlauben, eine Sprache zu lernen (Spracherwerb).

gabe zu einer von mehreren Dutzend unterschiedlichen Kategorien zugewiesen wird. Im Rahmen der syntaktischen Analyse kann es als eine Transformation von einer linearen Struktur (der Eingabeäußerung) in eine Baum- oder Graphenförmige Struktur (der Analyse) behandelt werden. Und schließlich kann man es in der Maschinellen Übersetzung als eine Transformation und Umdeutung von einer linearen Eingabe in eine (anderssprachige) lineare Ausgabe ansehen.
Gängige Methoden zum Erlernen solcher Umwandlungen sind normalerweise sehr stark an statistische Prozesse gebunden (z. B. an stochastische Automaten für Morphologie, Markov-Modelle für Wortartenzuweisung, stochastische Grammatiken für Syntaxanalyse, oder *noisy channel models* für Übersetzung). Diese beruhen darauf, eine Menge von manuell mit dem gewünschten Resultat annotierten prototypischen Eingaben als Trainingsmaterial zu benutzen. Statistische Lernalgorithmen konsumieren das Material und produzieren Modelle, die von den einzelnen Eingaben abstrahieren und Generalisierungen über die vorkommenden Phänomene darstellen. Laufzeitkomponenten benutzen diese Modelle dann, um bisher ungesehene Eingaben zu analysieren und die gewünschten Resultate herzustellen. Kritische Fragestellungen im Umgang mit Lernalgorithmen sind u.a.:

- Wie gut ist der Lernmechanismus? Im Vordergrund steht hierbei natürlich, welchen Erfolg ein System bei der Analyse von unbekannten Eingaben hat: Wieviele Eingaben können überhaupt verarbeitet werden, wieviele Antworten werden erzeugt, und wieviele davon sind richtig (vgl. Kapitel 6)?

- Wie schnell ist der Mechanismus? Für diese Frage sind zunächst Aspekte der Komplexitätstheorie relevant, um festzustellen, ob ein Lernalgorithmus oder die Anwendung der generierten Modelle prinzipiell möglich scheint. Darüber hinaus ist es interessant abzuschätzen, welche Menge an Trainingseingaben notwendig ist, um ein akzeptables Modell zu erstellen (z. B., wenn man sich Gedanken über sog. *low density languages* macht, Sprachen, für die nur ein kleines Korpus verfügbar ist). Dies ist die Frage nach der Generalisierungsfähigkeit des Algorithmus, nach der Balance zwischen sturem Auswendiglernen von Trainingseingaben und der Extraktion von abstrakten Eigenschaften aller Trainingseingaben. Und schließlich ist wichtig zu untersuchen, wie schnell potentielle neue Eingaben in das Wissen des Mechanismus integriert werden können. Kann z. B. eine gerade analysierte und verifizierte Äußerung dazu benutzt werden, die Qualität des benutzten Modells inkrementell zu verbessern?

- Wie adäquat ist der Mechanismus? Hier sind (normalerweise zu einem kleineren Anteil) philosophische Aspekte zu betrachten, etwa der Art, ob der automatische Lernalgorithmus ein ähnliches Fehlerprofil wie Menschen aufweist. Wichtiger erscheint eine Abschätzung darüber, ob die untersuchte Methode relativ einfach auf eine neue Domäne, eine andere Sprache, oder ein anderes Teilgebiet linguistischer Phänomene angewendet werden kann.

Die angedeuteten Fragestellungen deuten darauf hin, dass das (theoretische) Feld der Lernalgorithmen eng mit dem Vorhandensein von Trainings- und Testkorpo-

ra zusammenhängt. So ist es kein Zufall, dass in den letzten Jahren zahlreiche regierungsfinanzierte Projekte zur Sammlung und Annotierung von Sprachdaten initiiert wurden. Diese umfassen zahlreiche Sprachen, Anwendungsdomänen und Modalitäten. Der diesen Anstrengungen innewohnende Aufwand hat zudem zu einem stärkeren Fokus auf unüberwachte Lernalgorithmen geführt, Algorithmen, die kein annotiertes Trainingskorpus benötigen, sondern Regularitäten ausschließlich basierend auf Eingabeäußerungen ableiten. Manchmal ist dies schon ausreichend, etwa im Bereich der Disambiguierung von Wortbedeutungen; meist werden die gefundenen Regularitäten allerdings einem weiteren, manuellen Analyseschritt unterworfen, um deren Korrektheit sicherzustellen und ihnen eine symbolische Bedeutung zuzuordnen.

Ein relativ neuer Bereich der Forschung ist der der hybriden Systeme. In der vorangegangenen Diskussion war davon ausgegangen, dass ausschließlich extensional gearbeitet wird: Paare von Eingabeäußerungen und den mit ihnen assoziierten korrekten Antworten wurden dazu benutzt, Regularitäten zu finden. Im Gegensatz dazu sind konventionelle Grammatiken stark intensional orientiert, in dem man direkt Abstraktionen formuliert, basierend auf der Intuition der Grammatikschreiber oder einer subjektiven Analyse eines Beispielkorpus. Die Proponenten beider Ansätze haben gewichtige Argumente für die Überlegenheit der eigenen Sichtweise. Intensionale Grammatikschreiber argumentieren, dass mit einer Regel eine ganze Klasse von Äußerungen abgedeckt werden kann, und dass sich feine Unterschiede in Strukturen einfach handhaben lassen, während extensionale Statistiker hervorheben, dass stochastische Methoden stärker an der realen Benutzung von Sprache orientiert sind, und dass die Verfügbarkeit von Sprachmaterial die Anwendung auf unterschiedliche Domänen und Sprachen enorm erleichtert. In den letzten Jahren haben sich diese beiden Schulen aneinander angenähert, insbesondere im Bereich der Maschinellen Übersetzung (s. z. B. Charniak, Knight und Yamada 2003). Statistische Methoden werden benutzt, um Übersetzungsmuster im Trainingstext zu finden, während linguistisch orientierte Strukturregeln die Validität von bestimmten Satzmustern hervorheben.

1.1.5 Wissensbereiche

Die Wissensbereiche innerhalb der Computerlinguistik sind weitgehend an den von der Linguistik angenommenen Beschreibungsebenen natürlicher Sprache orientiert. Dies erscheint aus methodischer Sicht zunächst unvermeidlich und sinnvoll, auch wenn aus theoretischen oder praktischen Erwägungen heraus diese Einteilung häufig aufgehoben wird.[2]

Generelles Paradigma der Computerlinguistik sollte das Streben nach Erkenntnissen über bedeutungsdefinierende und bedeutungsunterscheidende Merkmale sein. Insofern sind die Resultate der theoretischen Linguistik von weit stärkerer

[2] Etwa bei der Entwicklung von Übersetzungssystemen, die ausschließlich statistische Information nutzen (Brown et al. 1990). Hier wird versucht, ein zusammenhängendes Modell für alle relevanten Verarbeitungsschritte zu berechnen, so dass auf den Einfluss einzelner Ebenen nicht mehr geachtet werden muss.

Bedeutung für die Computerlinguistik als der Bereich der rein deskriptiven Linguistik, von dem überwiegend nur die Bereitstellung von initialen Daten über Sprachen von Interesse ist.

Eine vertikale Einteilung der Computerlinguistik umfasst zumindest die folgenden fünf Bereiche:

- **Phonetik** und **Phonologie** (Unterkapitel 3.1): Sie untersuchen die artikulatorischen Merkmale sowie die Lautstruktur natürlicher Sprachen und kommen in der Computerlinguistik vor allem im Bereich der Erkennung und Produktion *gesprochener* Sprache vor. Ziel ist u.a. zu modellieren, welche Segmente ein Wort enthält und wie sich deren Struktur auf die Aussprache auswirkt, z. B. wenn ein im Prinzip stimmhafter Konsonant am Wortende stimmlos wird (Auslautverhärtung):

  (1.3) *Dieb* vs. *Diebe*
  /Diep/ /Diebe/

- Die **Morphologie** (Unterkapitel 3.3) beschreibt die Bildung und Struktur von Wörtern. Untersucht wird hier, welche lexikalische Wurzel einzelne Wörter haben, welche Prozesse für die unterschiedlichen Erscheinungsformen an der Oberfläche verantwortlich sind, und wie diese Oberflächenmodifikationen die Verwendung und Bedeutung des Wortes verändern. Die Morphologie ist durch eine vorwiegend anglozentrische Forschung innerhalb der Computerlinguistik lange Zeit unterrepräsentiert gewesen; erst mit der Untersuchung stärker flektierender Sprachen gewann sie an Gewicht. Eine morphologische Analyse des Deutschen muss etwa erkennen können, dass das Suffix *-e* im folgenden Beispiel eine Pluralmarkierung darstellt:

  (1.4) *Dieb-e*
  Dieb-pl
  „Mehr als ein Dieb“

- In den Bereich der **Syntax** (Unterkapitel 3.5) fällt alles, was mit der Strukturbildung von Sätzen zu tun hat. Sie ist die traditionell am stärksten vertretene Teildisziplin der Computerlinguistik. Eine strukturelle Analyse von Äußerungen ist unverzichtbar für die erfolgreiche Erkennung von Grammatikalität und eine darauf folgende Bedeutungserschließung. So muss im folgenden Gegensatz nicht nur erkannt werden, dass (1.5b) ungrammatisch ist, auch der Zusammenhang zwischen den einzelnen Wörtern und die daraus gebildete Struktur sind relevant (ungrammatische Sequenzen werden mit einem Stern „*“ eingeleitet):

  (1.5) a. *Der gewitzte Dieb stahl das Geld.*
  b. **Der Dieb gewitzte stahl das Geld.*

- Die **Semantik** (Unterkapitel 3.6) befasst sich mit der Bedeutung sprachlicher Einheiten. Dabei wird sowohl versucht, die Aspekte der Bedeutung

von lexikalischen Einheiten zu beschreiben (in der lexikalischen Semantik), als auch die Bedeutungszusammenhänge von größeren strukturellen Einheiten zu repräsentieren. Z. B. kann beiden Sätzen in Beispiel (1.6) dieselbe prinzipielle Bedeutungsstruktur zugewiesen werden, obwohl die Wortstellung unterschiedlich ist:

(1.6) a. *Die Polizei beschlagnahmte das Diebesgut.*
b. *Das Diebesgut beschlagnahmte die Polizei.*

- Die **Pragmatik** (Unterkapitel 3.7) untersucht sprachliche Ereignisse daraufhin, welchen Zweck eine Äußerung in der Welt hat. Die Frage

  (1.7) *Ist das Fenster auf?*

  mag schlicht eine einfache Informationsfrage sein. Weitaus wahrscheinlicher ist jedoch, dass der fragenden Person kalt ist, oder dass es zieht. In diesem Zusammenhang muss die Frage dann als Aufforderung verstanden werden, das betreffende Fenster doch bitte zu schließen. Die Abschnitte in Unterkapitel 3.7 befassen sich unter anderem mit der automatischen Bestimmung des Antezedens einer Anapher wie in *Die Katze*$_1$ *schnurrt. Sie*$_1$ *hat Hunger.* (Abschnitt 3.7.2), die Äußerungen innewohnenden impliziten Annahmen (Präsuppositionen, Abschnitt 3.7.3) und der Frage, welche Annahmen eine Maschine über einen Benutzer machen kann und sollte (Benutzermodellierung, Abschnitt 3.7.4). Auch der Bereich der Konstruktion sprachlicher Oberflächenrepräsentationen durch eine Maschine (Generierung, Unterkapitel 5.6) ist pragmatisch motiviert.

Zusätzlich lassen sich einige Bereiche erfassen, die ebenenübergreifend von Relevanz sind: Ein Beispiel hierfür ist die Prosodie, deren Einfluss auf praktisch alle oben genannten Gebiete nachgewiesen werden kann.

Neben dieser vertikalen Einteilung der hier aufgeführten Wissensbereiche lassen sich zwei weitere, mehr horizontale Unterscheidungskriterien herausarbeiten:

- Es muss zwischen der Repräsentation von Wissen und der Modellierung der Prozesse, die dieses Wissen benutzen, um ein bestimmtes Phänomen zu untersuchen, unterschieden werden. Beide sind gleichermaßen notwendig und wichtig, um erfolgreich natürliche Sprache zu erforschen und funktionierende Systeme zu deren Verarbeitung zu konstruieren.

- Alle hier genannten Wissensebenen spielen sowohl bei der Analyse als auch der Produktion natürlicher Sprache eine Rolle. So ist beispielsweise die *Analyse* der syntaktischen Struktur einer Äußerung der Kernbereich des Parsing (vgl. Unterkapitel 3.5), während die *Erzeugung* einer Oberflächenstruktur ausgehend von einer syntaktischen Beschreibung als Generierung im engeren Sinne bezeichnet wird (vgl. Unterkapitel 5.6).

1.1.6 Industrielle Anwendungen

Ergebnisse aus der Computerlinguistik-Forschung haben bereits Einzug gehalten in einen weiten Bereich industrieller Anwendungen. Das Paradebeispiel hier ist *Google*: Die Suchanfragen nach Webseiten werden z. B. normalerweise einer morphologischen Analyse unterzogen, um die Menge an potentiell relevanten Seiten zu erhöhen. Findet man eine Seite in einer Sprache, die man nicht versteht, kann Google diese übersetzen. Eine andere Anwendung, *Google News*, benutzt unüberwachte Clustering-Methoden und Textzusammenfassung, um einen Überblick über die augenblickliche Nachrichtenlage zu ermöglichen.

Das Internet enthält eine sehr große Menge an Information (vgl. Unterkapitel 4.7). Das bedeutet aber nicht, dass diese Information immer leicht zugänglich ist. Im Gegenteil, sie ist hochgradig unstrukturiert, so dass ein direkter Zugang zu relevanten Daten unwahrscheinlich ist. Um einen Zugriff auf Information für einen weiten Kreis von Benutzern verfügbar zu machen, oder bestimmten Aufgaben in einer einfacheren, natürlicheren Art und Weise gerecht zu werden, scheinen natürlichsprachliche Schnittstellen sinnvoll. Eine Anfrage wie *„Wie kann ich am billigsten nach Amerika telefonieren“* ist in vielen Fällen einfacher zu stellen als die ungefähr äquivalente Form *„+telefon +amerika +preis +vergleich“*. Folglich arbeitet eine beachtliche Anzahl von Firmen an der Frage, wie natürlichsprachliche Anfragen dazu benutzt werden können, Information aus einer Menge von Dokumenten zu extrahieren. Ein solches Verfahren ist insbesondere dann extrem anspruchsvoll, wenn die Eingabe nicht mehr oder weniger direkt auf eine syntaktisch äquivalente Datenbankanfrage abgebildet werden kann, sondern versucht werden muss, Teile der Bedeutung von Dokumenten zu modellieren, so dass auch eine Frage, die nicht aus relevanten Kennwörtern besteht, Aussicht auf erfolgreiche Beantwortung haben kann (vgl. Unterkapitel 5.3).
Als zweites Beispiel für den immer wichtiger werdenden Einfluss der natürlichsprachlichen Verarbeitung sei die Einführung von Dialoganwendungen genannt (vgl. Unterkapitel 5.5). Diese können einen relativ einfachen Zugang zu komplexen Systemen realisieren, bei denen eine Reihe von Informationen vom Benutzer zum System geleitet werden müssen. Als Paradebeispiel hierfür gilt normalerweise die Bestellung eines Bahn- oder Flugtickets, aber auch die Interaktion mit der eigenen Bank. Während hier Telefonsysteme, die auf dem Eingeben numerischer oder alphabetischer Daten mit Hilfe der Tastatur des Telefons beruhen, inzwischen weite Verbreitung gefunden haben, sind natürlichsprachliche Anwendungen, innerhalb derer der Benutzer verbal mit einer Maschine kommuniziert, noch selten. Allerdings existieren bereits seit einigen Jahren beachtenswerte prototypische Systeme hierzu (vgl. Unterkapitel 5.5).

Übersetzungssysteme erlangen stärkere Marktdurchdringung. Dies ist nicht nur motiviert durch den Wunsch von Endbenutzern, Web-Seiten in anderen Sprachen lesen zu können. Der Trend zur Globalisierung zwingt Anbieter von Produkten und Maschinen, Information in mehreren Sprachen anzubieten (z. B. in der Form von Gebrauchsanweisungen) oder dazu in der Lage zu sein, solche zu konsumieren (in der Form von Anfragen, Serviceanforderungen usw.). Geopolitische Realitäten zwingen insbesondere Regierungen dazu, in Übersetzungs-

systeme zu investieren, um Personal dazu in die Lage zu versetzen, erfolgreich mit Personen und Gruppen in anderen Ländern zu kommunizieren. Dies hat in den letzten Jahren zur verstärkten Forschung und Produktentwicklung von Übersetzungssystemen vor allem für nichteuropäische Sprachen geführt.

Schließlich sei auch angemerkt, dass eine Reihe von Geschäftsprozessen bereits durch die CL unterstützt sind. Z. B. ist es wahrscheinlich, dass ein Bewerbungsbrief und Lebenslauf, der an eine sehr große Firma geschickt wird, zunächst von einer Maschine untersucht wird, um relevante Qualifikationen zu extrahieren und möglicherweise die am besten passende Stelle zu ermitteln. Auch werden die in einem Konzern eingehenden Briefe vielfach gemäß ihres Inhaltes klassifizert, um die richtige Abteilung in einer großen Organisation zu identifizieren.

Die hier zitierten Schwerpunkte der Anwendung computerlinguistischen Wissens in der Industrie bedeuten, dass vor allem drei Bereiche stark nachgefragt sind:

- Die Verbindung von Sprachkenntnissen mit Computerlinguistik-Wissen, insbesondere im Bereich der Lexikographie und Korpusbearbeitung. Die Erweiterung einer Anwendung auf eine neue Sprache verlangt zunächst nach einem Muttersprachler für diese Sprache. Aus praktischen Erwägungen heraus ist es von unschätzbarem Vorteil, wenn dieser darüberhinaus über die notwendigen Grundlagen zur effektiven Modellierung sprachlichen Wissens verfügt. Dazu gehören neben dem prinzipiellen Aufbau eines Lexikons und den Eigenschaften von Einträgen (Argumentstrukturen, lexikalische Semantik) auch Fertigkeiten im Bereich des Grammatikentwurfs (Linguistik und Formalismen) und die Fähigkeit, Korpora aufzubauen oder zusammenzustellen und daraus relevante linguistische Fakten abzuleiten.

- Dialogsystembau. Zum gegenwärtigen Zeitpunkt sind kommerzielle Dialogsysteme noch meist einfach strukturiert. Der Ablauf eines Dialogs ist weitgehend vorher festgelegt, ohne dass der Benutzer die Möglichkeit hat, großen Einfluss auf dessen Inhalte und Strukturen zu nehmen. Es ist folglich umso wichtiger, dass das Design eines Dialogs umfassend und korrekt ist, und auf ungewöhnliche Phänomene vorbereitet ist. Zur Modellierung von Anwendungen werden eine Reihe von Designtools benutzt, deren prinzipielle Möglichkeiten und Begrenzungen bekannt sein müssen. Ein Computerlinguist bringt hier sein Wissen um Dialogstrukturierung und die genannten linguistischen Teilgebiete Syntax, Semantik und Pragmatik ein.

- Erfahrung in der Entwicklung natürlichsprachlicher Systeme. Die genaue Ausrichtung hängt selbstverständlich von dem jeweiligen Anwendungszweck ab, doch läßt sich feststellen, dass ein umfassendes Querschnittswissen für die Entwicklung der meisten Systeme unumgänglich ist. Um nur ein Beispiel zu nennen: Für die erfolgreiche Entwicklung eines Systems zur Informationsrecherche im Internet sind zumindest die Teilbereiche Morphologie und Syntax (um Anfragen zu analysieren), Semantik (vornehmlich zur Modellierung des Wissens in Dokumenten), und statistische

Computerlinguistik (erneut zur Inhaltsmodellierung und Abschätzung von Relevanzfragen) wichtig.

In der Zukunft wird sich die Interaktion von Konsumenten mit Produkten und die Handhabung von Information weiterhin stark verändern. Es ist abzusehen, dass immer mehr Funktionen unter Zuhilfenahme persönlicher Assistenten erledigt werden. Insbesondere die Möglichkeit zur Eingabe natürlich gesprochener Sprache sowie die immer besser werdenden Systeme zur Informationsextraktion, Plansynthese und dynamischer Textzusammenfassung bedeuten, dass das Internet immer weniger als eine passive Informationsquelle angesehen werden muss, sondern dass man quasi mit ihm kooperiert. Während man heute relativ einfach nach günstigen Flugpreisen nach Miami suchen kann, könnte die Reiseplanung in Zukunft beinhalten, dass der persönliche Assistent Alternativen vorschlägt (*„Du bist letztes Jahr schon nach Miami geflogen. Wie wäre es mit Jamaica? Ähnliches Klima, aber wesentlich exotischer.“*), Nachrichten zusammenfasst (*„Das Hotel ist in einer Gegend mit hoher Kriminalität. Ich weiss, es ist billig, aber vielleicht solltest Du doch besser dieses hier nehmen.“*), und komplexe Prozesse übernimmt (*„Ok, soll ich das jetzt buchen?“*).

Auch Haushaltsgeräte könnten mit Sprachtechnologie ausgerüstet werden (dann kann der Kühlschrank mitteilen, was er enthält, und einen Einkaufszettel vorschlagen). Das Hauptproblem hier könnte das Überangebot an sprachlicher Kommunikation sein, und folglich könnte die Aggregation und Priorisierung von Information im Vordergrund stehen. Natürlichsprachliche Zugangssysteme zu Fahrzeugen existieren bereits rudimentär, hauptsächlich in Form von Kommandosystemen und in niedriger Zahl als sogenannte Sprachdialogsysteme. Auch in diesem Bereich kann erwartet werden, dass die Bandbreite an relevanter Information, die mit Hilfe natürlicher Sprache abgefragt und kontrolliert werden kann, stetig wächst. Eine kluge Anwendung von Computerlinguistik kann hier dazu führen, dass die Ergonomie solch komplexer Systeme stark verbessert wird.

Auch in der Geschäftswelt wird sich der Einfluss der CL erhöhen. Während ein Teil der Kommunikation zwischen Unternehmen stark formalisiert ist (Rechnungen usw.) und mit relativ einfachen Mechanismen gehandhabt werden kann, so ist ein weiterer großer Teil natürlichsprachlich (Anfragen, Beschwerden, Notizen, Memos usw.) und erfordert computerlinguistische Methoden, um wenigstens partiell automatisch behandelt werden zu können.

1.1.7 Berufsfelder für Computerlinguisten

Die Computerlinguistik/Sprachtechnologie eröffnet vielfältige Anwendungsbereiche innerhalb einer modernen Informationsgesellschaft – das Kapitel 5 stellt die wichtigsten Anwendungen vor. Es ist abzusehen, dass die Verarbeitung gesprochener Sprache für die Interaktion mit Computern und für die Steuerung intelligenter Geräte an Bedeutung gewinnen wird, und dass die Verarbeitung von Texten als allgegenwärtigen Trägern von Information ohne texttechnologische Anteile (z. B. Klassifikation, Retrieval, Übersetzung, Zusammenfassung) kaum denkbar sein wird. Schon jetzt verfügen weltweit operierende Softwareanbieter

in der Regel über eigene Sprachtechnologie-Forschungslabore, während die Zahl eigenständiger Computerlinguistik-Firmen stetig zunimmt (allein für den Bereich der maschinellen und computergestützten Übersetzung listet Hutchins und Hartmann (2002) mehr als 160 Firmen auf).

Neben diesem Bereich der Computerlinguistiksoftware-*Entwicklung* finden Computerlinguisten und Computerlinguistinnen ihre Berufsfelder vor allem im Rahmen des *Einsatzes* bzw. der *Verwendung* sprachtechnologischer Software und Ressourcen (in Verlagen, Übersetzungsbüros, Verwaltungen etc.) und, insbesondere langfristig gesehen, auch in deren *Wartung/Support* und *Vertrieb* (zu detaillierteren Informationen siehe auch `http://berufenet.arbeitsamt.de` mit dem Suchwort „Computerlinguistik“).

1.1.8 Literaturhinweise

Es existieren mittlerweile eine Reihe von Einführungen und Handbüchern zur Computerlinguistik und Sprachtechnologie. Der „Klassiker“ ist in dieser Hinsicht Allen (1995), das 1987 zuerst erschienen ist. Neuere englischsprachige Alternativen hierzu sind insbesondere Jurafsky und Martin (2009) sowie Mitkov (2003). Die erste umfassende und gute Einführung in die statistische Computerlinguistik stellt Manning und Schütze (2003) dar. Weiterhin sind Cole et al. (1997), Dale et al. (2000) sowie Hausser (2001) (das auch in deutscher Sprache als Hausser 2000 vorliegt) zu nennen.

Eine sehr grundlegende deutschsprachige Einführung ist Schmitz (1992). Die für die (Computer)linguistik notwendigen Statistik-Kenntnisse vermittelt anschaulich und fundiert Gries (2008). Der Sammelband Batori und Lenders (1989) dokumentiert den Kenntnisstand in der Computerlinguistik aus den 80er Jahren, ist aber immer noch teilweise lesenswert. Heyer et al. (2006) führen in praxisorientierte Aspekte der Textverarbeitung ein, während Lobin und Lemnitzer (2004b) eine Mischung aus Grundlagen, Methoden und Anwendungen in der Texttechnologie präsentiert. Carstensen (2009b) bietet einen Überblick über die komplexen Anwendungen in der Computerlinguistik.

Görz et al. (2003) ist eine allgemeine Einführung in die Künstliche Intelligenz, die auch einen Teil über Sprachverarbeitung enthält. Für Darstellungen von aktuellen Entwicklungen sei auf die Zeitschrift *Computational Linguistics* verwiesen, das Organ der ACL (*Association for Computational Linguistics*). Es ist online unter `http://www.aclweb.org/anthology-new` verfügbar, zusammen mit elektronischen Versionen von Beitragsbänden zahlreicher CL-Konferenzen.

Die Referenzadresse zur Sprachtechnologie im (deutschsprachigen) Web ist `http://www.lt-world.org`. Hier finden sich Neuigkeiten und nach Sparten geordnete Informationen zur praxisorientierten Sprachverarbeitung.

1.2 Zur Geschichte der Computerlinguistik

Wolfgang Menzel

1.2.1 Die Ursprünge

Die frühen Entwicklungen zur Computertechnologie in den dreißiger und vierziger Jahren des 20. Jahrhunderts waren sehr stark durch die Hinwendung zu numerischen Problemstellungen geprägt. Dieser Umstand spiegelt sich recht deutlich in den ursprünglichen Namensgebungen wider: computational machinery, machine à calculer, ordinateur, ная машина, Elektronenrechner usw. Allerdings wurde auch damals schon das enorme Potential der neuen Technologie für die Behandlung rein symbolischer Verarbeitungsaufgaben erkannt. Ausschlaggebend hierfür war wohl nicht zuletzt der erfolgreiche Einsatz zur Dechiffrierung verschlüsselter Nachrichtentexte, der letztendlich auch die maschinelle Übersetzung der natürlichen Sprache als Spezialfall einer Dekodierungsaufgabe realisierbar erscheinen ließ (Weaver 1949). Zugleich wurden erste Überlegungen zu den prinzipiellen Möglichkeiten der maschinellen Informationsverarbeitung angestellt (Turing 1950). Auch wenn es sich dabei anfangs noch um reine Gedankenexperimente handelte, so bezogen sie sich doch ebenfalls auf ein Szenario, das dem Bereich der maschinellen Sprachverarbeitung zuzuordnen ist, und setzten damit die prinzipielle Realisierbarkeit eines natürlichsprachlichen Dialogs zwischen Mensch und Maschine indirekt schon einmal voraus.

In diesen frühen Überlegungen weisen die sich abzeichnenden Lösungsansätze zur maschinellen Sprachverarbeitung durchaus noch eine gemeinsame Wurzel auf, die stochastische Informationstheorie (Shannon und Weaver 1949). Aus deren Perspektive erscheint ein fremdsprachlicher Text als das Ergebnis der Übertragung einer Nachricht über einen gestörten Kanal. Die Aufgabe etwa der maschinellen Übersetzung besteht dann darin, den ursprünglichen Nachrichtentext unter Verwendung der sprachspezifischen Symbolwahrscheinlichkeiten und der Kanalcharakteristika beim Empfänger zu rekonstruieren.

War zu diesem Zeitpunkt die Einheit des methodischen Inventariums noch weitgehend gewahrt, so konnte man schon bald darauf eine stärkere Aufspaltung in stochastische Verfahren einerseits und symbolische Ansätze andererseits beobachten. Während erstere vor allem im Bereich der Informationswissenschaft, aber auch zur Verifizierung der Autorenschaft eines Textes zum Einsatz kamen, wurden letztere geradezu zum Synonym der späteren Computerlinguistik und dominierten die Entwicklung des Gebiets über einen erstaunlich langen Zeitraum.

Für diese recht einseitige Entwicklung lassen sich sicherlich mehrere Gründe identifizieren. Zum einen war da Chomsky's Diktum (Chomsky 1957), dass prinzipiell kein statistischer Ansatz in der Lage sein kann, den fundamentalen Unterschied zwischen den beiden Sätzen

(1.8) *Colorless green ideas sleep furiously.*

(1.9) *Furiously sleep ideas green colorless.*

zu erfassen, da man mit einiger Sicherheit davon ausgehen darf, dass keiner von beiden jemals in einem englischen Diskurs auftreten würde, und somit einer stochastischen Beobachtung *per se* nicht zugänglich ist. Es hat letztendlich mehr als vier Jahrzehnte intensiver Forschung benötigt, um erkennen zu können, dass diese Annahme grundfalsch war, und dass sich unter Zuhilfenahme versteckter Variablen durchaus stochastische Modelle auf ganz gewöhnlichen englischen Korpusdaten trainieren lassen, die tatsächlich einen Unterschied von mehr als fünf Größenordnungen zwischen den Wahrscheinlichkeiten für diese beiden Sätze vorhersagen (Pereira 2000).

Auf der anderen Seite hatte die einseitige Bevorzugung symbolischer Verfahren aber sicherlich auch ganz praktische Gründe, die vor allem in der mangelnden Leistungsfähigkeit der damals verfügbaren Hardware zu suchen sind. Derartige Beschränkungen bevorzugen in der Tat symbolische Ansätze in ganz entscheidender Weise: So lässt sich etwa die prinzipielle Idee eines symbolischen Verfahrens immer auch anhand eines extrem stark vereinfachten Modells (wenige Regeln, geringer Abdeckungsgrad usw.) demonstrieren, wobei sich die eigentlichen Schwierigkeiten dann natürlich bei der Verallgemeinerung auf größere Sprachausschnitte einstellen. Dagegen muss bei einem vergleichbaren stochastischen Ansatz bereits für das allererste Experiment ein ganz erheblicher Aufwand im Bereich der Datensammlung und der sehr ressourcenintensiven Schätzverfahren (Training) geleistet werden.

1.2.2 Symbolische Sprachverarbeitung

Die frühen Arbeiten zur symbolischen Sprachverarbeitung orientierten sich einerseits sehr stark an den vorhandenen linguistischen Beschreibungsebenen (Morphologie, Syntax, Semantik), zum anderen aber auch an den unmittelbaren Bedürfnissen praktischer Anwendungen, wie Maschinelle Übersetzung und Informationsrecherche. Im Mittelpunkt standen daher Untersuchungen zur lexikalischen Repräsentation und morphosyntaktischen Analyse von Wortformen, sowie zur syntaktischen Struktur von Sätzen.

Auf der Ebene der Morphotaktik lässt sich ein starker Trend hin zu elementaren Techniken aus dem Bereich der Endlichen Automaten bereits seit den frühesten Ansätzen nachweisen. Hinsichtlich der lexikalischen Beschreibungen konzentrierten sich die Bemühungen stark auf die syntaktischen Auswirkungen von Wortbildungs- und Flexionsprozessen, während die semantischen Aspekte lange Zeit eher ausgeklammert wurden. Seit den achtziger Jahren wurden verstärkt Anstrengungen unternommen, die Redundanz im Lexikon zu reduzieren. Einen ersten Schritt hierzu stellte die systematische Nutzung von Transducern zur Modellierung der phonologischen Variation (Koskenniemi 1983) dar. Durch geeignete Vererbungsmechanismen konnte auch auf der Seite der Lexikoninformation eine kompaktere Beschreibung erreicht werden. Um dabei dem Spannungsverhältnis zwischen Regel und Ausnahme angemessen Rechnung zu tragen, kamen dabei zumehmend auch Techniken der nichtmonotonen Vererbung zum Einsatz (Evans und Gazdar 1989).

Wichtigster Motor für die Aktivitäten zur syntaktischen Analyse waren sicherlich die Bedürfnisse der Maschinellen Übersetzung, wo man sich von dem Rückgriff auf syntaktische Repräsentationen einen deutlichen Forschritt gegenüber den rein wortformbasierten Ansätzen versprach. Zum anderen lag hier ein enger Berührungspunkt mit parallelen Entwicklungen im Bereich der Programmiersprachen vor, wo beim Compilerbau durchaus vergleichbare Techniken zum Einsatz kamen. Dadurch gab es insbesondere in den sechziger und siebziger Jahren eine starke wechselseitige Befruchtung.

Kontrovers wurde vor allem die Frage nach dem jeweils geeignetsten Grammatiktyp diskutiert, wobei im wesentlichen Ansätze zur Modellierung der Phrasenstruktur (Chomsky 1957) bzw. der Abhängigkeitsbeziehungen (Tesnière 1959), aber auch Kategorialgrammatiken (Bar-Hillel 1954) verwendet wurden. Besonders einflussreich war hierbei die Schule der Transformationsgrammatik (Chomsky 1957; Chomsky 1965), obwohl diese wegen der zugrundeliegenden generativen Sicht letztendlich keinerlei praktikable Sprachanalysesysteme hervorgebracht hat. Breiten Raum nahmen Untersuchungen zur effizienten Realisierung der syntaktischen Analyse (Parsing) ein. Wichtige Meilensteine stellen der Nachweis eines polynomialen Algorithmus für beliebige kontextfreie Grammatiken (Earley 1970), sowie die Idee der Wiederverwendung partieller Analyseergebnisse beim Chart-Parsing (Kaplan 1973; Kay 1973) dar.

Waren die frühen Systeme zur Sprachverarbeitung im wesentlichen *ad hoc*-Implementierungen bestimmter algorithmischer Ideen, so ist seit den siebziger Jahren eine zunehmende Tendenz hin zu generischen Formalismen zu verzeichnen, die dank ihres hohen Abstraktionsgrades dann auch für ganz unterschiedliche Verarbeitungsaufgaben eingesetzt werden können. Diese Entwicklung vollzog sich über spezielle Programmiersprachen mit teilweise noch stark prozedural orientierter Semantik (z. B. der durch gezielte Erweiterung aus den Endlichen Automaten entstandene Formalismus der Augmented Transition Networks, ATN; Woods 1970), über stärker deklarativ angelegte Formalismen zur Darstellung linguistischen Wissens (z. B. die Baum- und Graphtransformationssprachen ROBRA; Boitet, Pierre und Quèzel-Ambrunaz (1978) bzw. Systèmes-Q; Colmerauer 1970), bis hin zu den rein deklarativen Formalismen auf der Basis der Unifikation (z. B. die unifikationsbasierten Grammatikformalismen mit kontextfreiem Grundgerüst, wie PATR-II; Shieber 1986). Mit den constraint-basierten Unifikationsformalismen (Shieber 1992) liegt nunmehr auch ein rein deklaratives und dennoch berechnungsuniverselles Modell vor, das einerseits hohen Ansprüchen im Hinblick auf eine prinzipienorientierte und damit erklärungsadäquate Modellierung der Grammatik gerecht wird (Chomsky 1981; Pollard und Sag 1994), andererseits aber auch die Brücke zum Paradigma der Logikprogrammierung in der Informatik schlägt.

Generell sind durch die verstärkte Hinwendung zu universell verwendbaren Formalismen auch deren formale Eigenschaften verstärkt ins Blickfeld geraten. Ziel dieser Untersuchungen ist es vor allem, diejenigen Modellklassen zu identifizieren, die es gestatten, eine gegebene Problemstellung mit minimaler Mächtigkeit und größtmöglicher Effizienz zu lösen.

Universell verwendbare Formalismen eröffnen darüber hinaus auch die Möglichkeit zur Realisierung ebenenübergreifender Modelle, die sehr unterschiedliche Aspekte des sprachlichen Wissens integrieren können. Ein Beispiel hierfür ist die Konstruktion einer semantischen Repräsentation auf der Grundlage der Montague-Grammatik (Montague 1974), die dann mit den Mitteln der Unifikation in einem constraint-basierten Formalismus emuliert werden kann (Bouma et al. 1988). Vergleichbare Erweiterungen sind auch zur Einbeziehung satzübergreifender Phänomene auf der Grundlage der Diskursrepräsentationstheorie (DRT; Kamp und Reyle 1993) möglich.

1.2.3 Korpusstatistische Verfahren

Das Wiedererwachen des Interesses an stochastischen Verfahren steht in engem Zusammenhang mit den deutlichen Fortschritten bei der Erkennung gesprochener Sprache seit Anfang der achtziger Jahre. Gerade in diesem Gebiet hat sich gezeigt, dass die automatische Ermittlung von Modellparametern aus einem speziell aufbereiteten Korpus von Sprachdaten (oftmals als Training bezeichnet), einen entscheidenden Schritt zur Lösung des Wissensakquisitionsproblems darstellt. Letztendlich wurde erst durch den konsequenten Einsatz solcher Trainingsverfahren die Erkennung mit großen Wortschätzen und mehreren Sprechern überhaupt ermöglicht (Jelinek 1976).

Für die erfolgreiche Anwendung stochastischer Techniken müssen mehrere, teils widersprüchliche Forderungen erfüllt sein:

- Zum einen muss die Struktur des Modells so gewählt werden, dass die Zahl der zu schätzenden Modellparameter und die verfügbaren Trainingsdaten in einem ausgewogenen Verhältnis stehen.
- Zum anderen sollte das Modell über genügend Freiheitsgrade verfügen, um die Struktur der Daten angemessen widerspiegeln zu können, gleichzeitig aber beschränkt genug sein, um eine Generalisierung über den Trainingsdaten zu erzwingen und ein „Auswendiglernen“ der Einzelbeispiele zu vermeiden.

Ausgangspunkt des Modellentwurfs ist hierbei also nicht ein extern vorgegebener Adäquatheitsanspruch, wie dies für die symbolischen Verfahren charakteristisch ist, sondern vor allem die Frage der wirksamen Trainierbarkeit eines Modells auf einem vorgegebenen Datensatz.

Diese grundlegende Besonderheit teilen die generativ orientierten, stochastischen Verfahren mit anderen Klassen von trainierbaren Modellen, zu denen mit den konnektionistischen Ansätzen, den Support-Vektor-Maschinen, und den Entscheidungsbaum- bzw. Regelinduktionsverfahren auch Systeme zum diskriminativen, sowie zum rein symbolischen Lernen gehören. Wesentliches Charakteristikum ist also nicht so sehr die wahrscheinlichkeitstheoretische Fundierung des Ansatzes, sondern vielmehr die Tatsache, dass in der Trainingsphase die für die jeweilige Aufgabe relevanten statistischen Eigenschaften der Daten zur Modelladaption ausgenutzt werden.

Die wohl erste computerlinguistische Aufgabe, die Ende der achtziger Jahre mit korpusstatistischen Methoden erfolgreich bearbeitet wurde, war die Wortartendisambiguierung (Tagging; DeRose 1988). Angespornt von diesen Anfangserfolgen wurden dann zunehmend anspruchsvollere Zielstellungen verfolgt und Erfahrungen mit komplexeren Modellstrukturen gesammelt. Zu diesen Aufgaben gehören

- die syntaktische Analyse (Parsing) unter Verwendung unterschiedlich stark strukturierter Repräsentationen, z. B. (Briscoe und Waegner 1992),
- die strukturelle syntaktische Disambiguierung, z. B. PP-Attachment (Hindle und Rooth 1993),
- die semantische Lesartendisambiguierung,
- die automatische Ermittlung lexikalischer Information und
- die bilinguale Übersetzung (Brown et al. 1990).

Auch wenn bei den vielfältigen Experimenten zur Entwicklung korpusstatistischer Verfahren oftmals die klassischen Modellvorstellungen der strukturellen Linguistik Pate gestanden haben, so hat sich jedoch bald gezeigt, dass die elementaren Modellstrukturen der traditionellen Ansätze (z. B. kontextfreie Regeln) für eine direkte Übernahme in das neue Paradigma nur bedingt geeignet sind. Dies hat zu einer Reihe von Akzentverschiebungen geführt:

- In vielen Fällen kann eine stochastische bzw. konnektionistische Modellierung besser über die elementaren Operationen des zugrundeliegenden Entscheidungsprozesses (z. B. Transformation von Symbolsequenzen, Parseraktionen, ...) erfolgen, als auf der Ebene der Modellstrukturen selbst (Magerman 1995, Nivre et al. 2006). Somit rückt die Perspektive der Performanz wieder stärker in den Mittelpunkt.
- Das klassische Ideal einer redundanzarmen Beschreibung bringt gleichzeitig eine massive Verletzung der stochastischen Unabhängigkeitsannahme mit sich, so dass sich für eine erfolgreiche Modellierung vielfach sehr komplexe und hochgradig redundante Modellstrukturen besser eignen (Bod 1995).
- Es hat sich herausgestellt, dass sich die verschiedenen Arten von Strukturbeschreibungen unterschiedlich gut mit bestimmten Lernparadigmen (generativ vs. diskriminativ, struktur- vs. operationsbasiert) behandeln lassen. Dies hat u.a. zu einem so völlig unerwarteten Wiedererwachen des Interesses an Dependenzmodellen geführt (McDonald et al. 2005).

Zunehmende Aufmerksamkeit wird nunmehr auch der Frage nach möglichen Synergieeffekten durch die Integration symbolischer, stochastischer und konnektionistischer Verfahren in hybriden Systemlösungen gewidmet. Dies betrifft sowohl die Kopplung von Modellen auf der Basis unterschiedlicher Lernparadigmen (z. B. Nivre und McDonald 2008), als auch die Kombination trainierbarer Verfahren mit klassischen Ansätzen zur manuellen Grammatikentwicklung (z. B.

Foth und Menzel 2006). Eine besondere Herausforderung stellt dabei die optimale Zusammenführung von tiefen und flachen Analyseverfahren dar. Hierdurch kann erreicht werden, dass Verarbeitungskomponenten, die auf den im vorangegangenen Abschnitt behandelten ausdrucksmächtigen Repräsentationsformalismen beruhen, von der Effizienz und breiten sprachlichen Abdeckung flacher Analysetechniken (vgl. Unterkapitel 3.4) profitieren können, auch wenn diese Informationsbeiträge nicht immer sehr zuverlässig sind.

1.2.4 Anwendungen der Computerlinguistik

Obwohl das anwendungsbezogene Problem der Maschinellen Übersetzung bereits am Anfang der Arbeiten zur Computerlinguistik stand, zieht es auch ein halbes Jahrhundert später noch ein unvermindert starkes Forschungsinteresse auf sich, das nur gegen Ende der sechziger Jahre durch die recht pessimistischen Prognosen des ALPAC-Reports (siehe Hutchins 1986) für kurze Zeit abgeschwächt worden war.

Dass trotz einer jahrzehntelangen und intensiven Forschungsarbeit auf diesem Gebiet noch immer wesentliche Fragen der Übersetzungsqualität, sowie der Portierbarkeit auf neue Anwendungsbereiche und Sprachpaare offen sind, zeigt zum einen, dass es sich bei der Maschinellen Übersetzung um ein überaus schwieriges Sprachverarbeitungsproblem handelt. Zum anderen wird aber auch deutlich, dass wir es hier mit einer typischen technologischen Fragestellung zu tun haben, die immer durch einen Kompromiss zwischen Anspruch und Wirklichkeit gekennzeichnet ist, und dass damit so etwas wie eine *endgültige* Lösung des gegebenen Problems auch gar nicht erwartet werden darf. In diesem Sinne steht die Maschinelle Übersetzung gleichberechtigt in einer Reihe mit anderen technologischen Aufgabenbereichen, die sich in einer ganz ähnlichen Situation befinden: Zwar existieren nach nunmehr schon mehreren Jahrhunderten intensiver Entwicklungsarbeiten zahlreiche brauchbare Lösungsansätze für das Problem des Transports von Personen und Gütern, dennoch sind auch hier keinerlei Aussichten auf eine abschließende Behandlung dieser Aufgabenstellung zu erkennen.

Analog hierzu haben seit den achtziger Jahren einige Übersetzungssysteme durchaus auch die Reife zum Einsatz in speziellen Anwendungsszenarien erlangt. Ein Weg hierzu führte über die Beschränkung auf sehr spezielle Textsorten (z. B. Wetterberichte; Thouin 1982). Alternative Ansätze setzen stärker auf eine manuelle Nachbereitung der Übersetzungsresultate. Andere Entwicklungen wiederum zielen vor allem auf eine optimale Unterstützung des Humanübersetzers, dem eine Reihe von Werkzeugen zur Sicherung der terminologischen Konsistenz, zur Wiederverwendung bisheriger Übersetzungsresultate, sowie zur partiellen (Roh-) Übersetzung bei Routineaufgaben an die Hand gegeben werden soll.

Parallel zu den Arbeiten an der Maschinellen Übersetzung ist in den letzten drei Jahrzehnten eine erstaunliche Vielfalt von Anwendungssystemen auf der Grundlage computerlinguistischer Verfahren entwickelt und teilweise auch schon zur Einsatzreife gebracht worden. In vielen Fällen sind diese Arbeiten erst durch die bedeutenden Fortschritte auf anderen Gebieten der Informationstechnologie initiiert bzw. vorangetrieben worden. So wurde die wohl erste erfolgreiche An-

wendung morphologischer Analysetechniken zur automatischen Silbentrennung ganz wesentlich durch den umfassenden Übergang zum Photosatz im Druckereigewerbe Anfang der sechziger Jahre forciert. Erst mit der flächendeckenden Verbreitung der Mikrorechner seit den achtziger Jahren steht diese Technologie als standardmäßiger Bestandteil aller Textverarbeitungssysteme auch einem Massenpublikum zur Verfügung. Vergleichbare Entwicklungen waren auch im Bereich der Hilfsmittel zur Rechtschreibprüfung und -korrektur zu verzeichnen (Peterson 1980).

Recht deutlich lässt sich der Einfluss externer Faktoren auch auf dem Gebiet der Informationssuche nachvollziehen, wo durch die zunehmende Verbreitung des WWW eine deutliche Belebung der diesbezüglichen Forschungsaktivitäten zu verzeichnen ist (Baeza-Yates und Ribeiro-Neto 1999). Durch die explosionsartig anwachsende Menge der digital verfügbaren Information sind in diesem Zusammenhang eine Reihe von Anwendungsszenarien mit zum Teil ganz neuartigen Anforderungen entstanden:

- die Online-Recherche, die sich insbesondere durch extreme Effizienzerwartungen auszeichnet und durch das kontinuierliche Wachstum der online verfügbaren Textinformation mit ständig steigenden Qualitätsanforderungen konfrontiert ist,
- die Informationsfilterung und -klassifikation zur Zuordnung relevanter Dokumente z. B. bei der E-Mail-Sortierung bzw. als Grundlage hochgradig individualisierter Informationsangebote (vgl. das Unterkapitel 5.3),
- die Informationsextraktion zur inhaltlichen Erschließung von Textdokumenten im Hinblick auf stark spezialisierte Informationsbedürfnisse (vgl. ebenfalls das Unterkapitel 5.3) oder aber
- die Beantwortung von beliebigen Fragen aufgrund der in großen Textkorpora enhaltenen Information.

Ein Bereich, der vor allem von der gewaltigen Steigerung der Hardwareleistungsfähigkeit seit Beginn der neunziger Jahre profitiert hat, ist die automatische Spracherkennung, die insbesondere in Form von Diktieranwendungen zunehmende Verbreitung findet. Ein wesentlicher Berührungspunkt mit computerlinguistischen Forschungen ergibt sich hierbei durch die Notwendigkeit, Prädiktionen über Wortformsequenzen (Sprachmodellierung) in die Ermittlung des Erkennungsergebnisses einfließen zu lassen. Benötigt werden hierzu vor allem Verfahren zur leichteren Modelladaption an neue Nutzer und unbekannte Textsorten, sowie Techniken zur besseren Einbeziehung nichtlokaler Abhängigkeiten auf den verschiedenen sprachlichen Ebenen.

Dass sich die fundamentalen Trends der Informationstechnologie durchaus nicht immer förderlich auf die Entwicklung computerlinguistischer Anwendungen auswirken müssen, lässt sich etwa am Beispiel des natürlichsprachlichen Zugriffs zu Datenbanken beobachten, an den Mitte der achtziger Jahre erhebliche kommerzielle Hoffnungen geknüpft waren. Hier wurde die Entwicklung jedoch durch

Abbildung 1.1: Zeittafel

das Aufkommen graphischer Nutzerschnittstellen vollständig überholt. Für spezielle, aber typische Anwendungskontexte, wie Fahrplan- und Produktauskünfte, konnte alternativ zur geschriebenen Sprache ein Kommunikationskanal bereitgestellt werden, der eine bequemere und zugleich robustere Mensch-Maschine-Interaktion ermöglicht. Wichtige Aspekte dieser Technologie erfahren allerdings bereits heute eine Neuauflage in Dialogsystemen zur automatischen Telefonauskunft bzw. durch aktuelle Entwicklungsarbeiten zur automatischen Beantwortung von E-Mail im Servicebereich.