Henning Lobin

Computerlinguistik und Texttechnologie

W. Fink UTB

UTB **3282**

Eine Arbeitsgemeinschaft der Verlage

Böhlau Verlag · Köln · Weimar · Wien
Verlag Barbara Budrich · Opladen · Farmington Hills
facultas.wuv · Wien
Wilhelm Fink · München
A. Francke Verlag · Tübingen und Basel
Haupt Verlag · Bern · Stuttgart · Wien
Julius Klinkhardt Verlagsbuchhandlung · Bad Heilbrunn
Lucius & Lucius Verlagsgesellschaft · Stuttgart
Mohr Siebeck · Tübingen
C. F. Müller Verlag · Heidelberg
Orell Füssli Verlag · Zürich
Verlag Recht und Wirtschaft · Frankfurt am Main
Ernst Reinhardt Verlag · München · Basel
Ferdinand Schöningh · Paderborn · München · Wien · Zürich
Eugen Ulmer Verlag · Stuttgart
UVK Verlagsgesellschaft · Konstanz
Vandenhoeck & Ruprecht · Göttingen
vdf Hochschulverlag AG an der ETH Zürich

HENNING LOBIN

Computerlinguistik und Texttechnologie

WILHELM FINK

Der Autor:

Henning Lobin, *1964, seit 1999 Professor für Angewandte Sprachwissenschaft und Computerlinguistik an der Justus-Liebig-Universität Gießen, Leiter des Zentrums für Medien und Interaktivität. Promotion 1991 an der Universität Bonn, Habilitation 1996 an der Universität Bielefeld. Letzte Buchveröffentlichungen: „Automatische Textanalyse“ (2004, hrsg. mit A. Mehler), „Texttechnologie“ (2004, hrsg. mit L. Lemnitzer“), „Inszeniertes Reden auf der Medienbühne“ (2009). Forschungsschwerpunkte: Dependenzgrammatik, Grundlagen der Texttechnologie, Text-Parsing, Wissenschaftskommunikation.

Für Aya Philine

Bibliografische Information Der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Gedruckt auf umweltfreundlichem, chlorfrei gebleichtem Papier

(Wilhelm Fink GmbH & Co. Verlags-KG, Jühenplatz 1, D-33098 Paderborn)
ISBN 978-3-7705-4863-7

Internet: www.fink.de

Printed in Germany.
Herstellung: Ferdinand Schöningh, Paderborn
Einbandgestaltung: Atelier Reichert, Stuttgart

UTB-Bestellnummer: ISBN 978-3-8252-3282-5

Inhaltsübersicht

1. Einleitung

Das vorliegende Buch führt in knapper und exemplarischer Form in die Gebiete Computerlinguistik und Texttechnologie ein. Es kann ohne Vorkenntnisse gelesen werden, da auch die verwendeten linguistischen Grundbegriffe jeweils kurz erläutert werden. Aufgrund seiner Kürze ist es nicht dafür gedacht, für ein komplettes Studium – ob auf Bachelor- oder Master-Niveau – als zentrale Referenz zu dienen, vielmehr bietet es einen ersten Einstieg im Umfang eines Moduls von etwa acht bis zehn *Credit Points*, also entsprechend einer Vorlesung oder einem Seminar mit begleitender Übung.

Parallele Darstellung

Mit diesem Buch wird ein neuartiger Ansatz verfolgt: Erstmals werden die Disziplinen der Computerlinguistik und der Texttechnologie konsequent parallel vermittelt – jedem neuen Vermittlungsschritt im Bereich der einen Disziplin entspricht einer im Bereich der anderen. Beide Gebiete speisen sich systematisch und historisch aus denselben Quellen – formale Grammatiken –, haben aber in den letzten Jahren unterschiedliche Wege genommen, die die grundlegenden Zusammenhänge verschleiern. Dabei kann das Verständnis des einen Gebietes das des anderen verstärken und durch Analogieschlüsse befruchten. Die Texttechnologie vermag dem Leser manche Konzepte in anschaulicherer Weise nahe zu bringen, denn für alle beschriebenen Methoden und Techniken können Bezüge zu Anwendungen im Web hergestellt werden. Der Bereich der Grammatik der natürlichen Sprache wird dabei parallelisiert mit dem der Dokumentgrammatik, die Syntaxanalyse mit der Textannotation oder die Übersetzung syntaktischer Konstruktionen in logische Ausdrücke mit der Überführung eines Dokuments in ein bestimmtes Ausgabeformat. Das Prinzip der Parallelisierung ist auf die praktische hochschuldidaktische Erfahrung zurückzuführen, die ich in den letzten Jahren mit einem entsprechenden Anfängermodul machen konnte.

Aufbau

Unterhalb der Ebene der sieben Kapitel des Buches befinden sich insgesamt 15 Abschnitte zu computerlinguistischen und texttechnologischen Themen. Neun davon, bezeichnet mit CL-1 bis CL-9, beziehen sich auf die Computerlinguistik, die übrigen sechs, TT-1 bis TT-6, auf die Texttechnologie. Dabei sind jeweils ein oder zwei CL-Unterkapitel mit einem TT-Unterkapitel zusammengefasst und können in dieser Form gelesen oder im Seminar behandelt werden. Es ist aber auch möglich, sich zunächst zusammenhängend nur mit den computerlinguistischen oder den texttechnologischen Abschnitten zu befassen; dabei sollte jedoch die Reihenfolge der Abschnitte beibehalten werden. Die Gesamtheit aller fünfzehn Abschnitte ist auf die durchschnittlich fünfzehn Wochen eines Semesters zu beziehen, so dass jede Woche ein weiteres Kapitel behandelt werden kann. Falls aufgrund einer Prüfungsphase nur 14 Wochen zur Verfügung stehen,

können die beiden Einleitungsabschnitte (CL-1 und TT-1) zusammengefasst und zum Teil der Eigenlektüre überlassen werden.

Kompromisse

Bei einem so knapp gefassten Einführungswerk mussten natürlich viele Kompromisse eingegangen und Auswahlentscheidungen getroffen werden. Generell wurde in der Darstellung die Lesbarkeit in den Vordergrund gestellt, weshalb die wichtigsten Literaturverweise auch erst in den jeweils die Hauptkapitel abschließenden Abschnitten zur weiteren Vertiefung genannt werden. Viele wichtige Teilbereiche der Computerlinguistik wurden ausgelassen, vor allem Statistik-basierte Verfahren, andere nur kurz angerissen. Auch auf eine selbst nur ansatzweise erfolgende Darstellung einer bestimmten unifikationsbasierten Grammatiktheorie wurde verzichtet. Gleiches gilt für den Bereich der Texttechnologie, wo das Hauptgewicht auf der textuellen Informationsmodellierung in XML liegt, andere wichtige Aspekte werden hingegen nur kurz angesprochen. Das vorliegende Buch kann also auf keinen Fall umfassendere, enzyklopädisch angelegte Einführungswerke in deutscher Sprache ersetzen, etwa Carstensen *et al.* (2004) für die Computerlinguistik oder Lobin/Lemnitzer (2004) für die Texttechnologie. Es will stattdessen eine gut lesbare Kurzeinführung in zwei eng miteinander verwandte Gebiete geben, die auch für ein Studium mit anderen Schwerpunktsetzungen interessante neue Perspektiven eröffnen können.

Die in diese Buch dargestellten Inhalte wurden in den letzten Jahren teilweise in Lehrveranstaltungen an der Justus-Liebig-Universität Gießen in den Master-Studiengängen „Computerlinguistik und Texttechnologie“ sowie „Sprachtechnologie und Fremdsprachendidaktik“ eingesetzt. Ich habe in diesen Lehrveranstaltungen sehr von den Rückmeldungen der Studierenden zur Vorgehensweise profitiert. Das Manuskript wurde von Hans-Jürgen Heringer und Harald Lüngen durchgesehen – für ihre Hinweise, Korrekturen und kritischen Anmerkungen bin ich beiden überaus dankbar.

2. Geschichte und Gebiete

2.0 Ziele und WarmUp

Eine der wichtigsten Erfindung des zwanzigsten Jahrhunderts ist zweifellos der Computer. Mit dem programmierbaren Rechenautomaten ist es dem Menschen das erste Mal in seiner Geschichte gelungen, seine geistigen Fähigkeiten teilweise auf eine Maschine zu verlagern. Ganz am Anfang, als Anwendung der allerersten Computer, wurde das Rechnen mechanisiert – man kann mathematische Prozesse besonders leicht mit der logischen Arbeitsweise des Computers verbinden. Schon sehr bald aber kam auch die Sprache in das Blickfeld der frühen Entwicklungen. Wenn man in einem Computer eine mathematische Formel darstellen und berechnen kann, warum nicht auch einen Satz in einer natürlichen Sprache wie dem Deutschen? Warum sollte man nicht mit dem Computer die Entsprechung dieses Satzes in einer anderen Sprache „berechnen" können? Diese Fragen bildeten den Anfang der Computerlinguistik, und erst nach einigen Jahrzehnten hatte man verstanden, warum sie so schwer zu beantworten sind, so viel schwieriger als für die Mathematik. Und wenn man von Sprache spricht, dann spricht man auch von Texten, denn sprachliche Kommunikation vollzieht sich in gesprochenen oder geschriebenen Texten. Aufgrund dieser zentralen Funktion von Texten in der Kommunikation hat sich parallel zur Computerlinguistik eine Technologie für den praktischen Umgang mit Texten entwickelt, die Texttechnologie.

Ziele

In diesem Kapitel werden wir

- die Computerlinguistik mit der Linguistik und anderen wissenschaftlichen Disziplinen in Beziehung setzen,
- fünf verschiedene Auffassungen von Computerlinguistik als Wissenschaft kennen lernen,
- uns die historischen Entwicklungsphasen der Computerlinguistik ansehen,
- die Entstehung der Texttechnologie nachvollziehen,
- die Teilgebiete der Texttechnologie und ihre Gliederung betrachten und
- die besondere Bedeutung der strukturellen Beschreibung von Inhalt und Ausdruck verstehen.

WarmUp

Wozu verwenden Sie vor allem Ihren Computer? Überlegen Sie, welche Funktion er für Sie erfüllt. Welchen Anteil dieser Verwendung haben Tätigkeiten, die etwas mit Sprache zu tun haben? Stellen Sie sich vor, Sie müssten den Computer der Zukunft beschreiben. Was sollte er können, was heute noch unmöglich ist? Haben diese Dinge etwas mit Sprache im Allgemeinen oder mit Texten zu tun?

2.1 CL-1: Computerlinguistik – Der Computer lernt Sprache

Wir sind umgeben von Sprache, ob gesprochener oder geschriebener. Wir plaudern mit anderen Menschen, führen Telefongespräche, halten Referate, lesen Bücher und Zeitungen, nehmen fast automatisch Beschriftungen und Schilder wahr, schreiben Emails und SMS, erstellen Notizen, Powerpoint-Folien oder Bachelor-Arbeiten. Unser ganzes Leben besteht aus gesprochener oder geschriebener Sprache, oft in Verbindung mit Grafik, Bildern oder Video – Sprache, die uns auf natürliche Weise vermittelt wird oder mit technischer Unterstützung, die von uns produziert oder rezipiert wird.

Computerlinguistik

Die Sprachwissenschaft, oft auch als Linguistik bezeichnet, ist nicht die einzige Wissenschaft, die sich mit diesen Phänomenen befasst. Sprache interessiert auch Psychologen und Philosophen, Kommunikations- und Medienwissenschaftler, ja sogar Biologen, Juristen oder Mediziner. Die Computerlinguistik ist diejenige Wissenschaft, die ganz allgemein die maschinelle Verarbeitung von Sprache mit dem Computer in den Blick nimmt. Im Mittelpunkt stehen dabei Prozesse, die die Erzeugung oder Analyse von gesprochener oder schriftlich fixierter Sprache erlauben. Aber auch die Beschreibung der Sprache selbst in einer Weise, dass der Computer damit umgehen kann, ist Gegenstand der Computerlinguistik. Und schließlich verfolgt man mit der maschinellen Verarbeitung von Sprache meist ein bestimmtes praktisches Ziel, so dass auch die Entwicklung von Software, von sprachverarbeitenden Systemen, ein wichtiges Teilgebiet der Computerlinguistik darstellt.

Ebenen der Linguistik

Gehen wir noch einmal einen Schritt zurück und werfen einen etwas genaueren Blick auf die Linguistik: Diese befasst sich insbesondere mit den Strukturen der Sprache, wie sie sich in Lauten, Wörtern, Wortgruppen, Sätzen, Texten oder Gesprächen ausprägt. Als „Struktur" werden dabei die komplexen Zeichensysteme verstanden, die das „Aussehen", die Form der sprachlichen Einheiten, betreffen, oder deren Inhalt. Hinzu kommen die Regularien für deren Gebrauch. Durch die verschiedenen Betrachtungsebenen der Sprache und die Frage nach Form und Inhalt werden mehrere Teildisziplinen der Linguistik ausgeprägt, die alle die Aufdeckung und Modellierung von Strukturen zum Gegenstand haben:

<table>
<tr><th></th><th colspan="3">Form</th><th>Inhalt</th></tr>
<tr><td>Laut</td><td colspan="3">Phonologie</td><td></td></tr>
<tr><td>Wort</td><td>Morphologie</td><td colspan="2">Lexik</td><td>Wortsemantik</td></tr>
<tr><td>Wortgruppe</td><td colspan="3" rowspan="2">Syntax</td><td rowspan="2">Satzsemantik</td></tr>
<tr><td>Satz</td></tr>
<tr><td>Text</td><td colspan="4" rowspan="2">Textlinguistik</td></tr>
<tr><td>Dialog</td></tr>
</table>

Im lautlichen Bereich der Sprache kann man noch nicht nach Inhalten fragen, dieses ist erst auf der Ebene des Wortes oder der Wortteile möglich. Der Bereich der Lexik verbindet Aspekte der Form und des Inhalts von Wörtern. Im Bereich der Textlinguistik gibt es natürlich ebenfalls Fragestellungen, die stärker auf die formale oder die inhaltliche Seite von Texten abzielen, unterschiedliche Teilgebiete der Linguistik sind daraus allerdings nicht entstanden. Die hier von dem Rahmen umgebenen Gebiete werden gewöhnlich als die Kerngebiete der Linguistik verstanden.

In der Computerlinguistik werden grundsätzlich alle Teilgebiete der allgemeinen Linguistik bearbeitet, wobei auch hier ganz ähnliche Kerngebiete im Mittelpunkt stehen. Dabei geht es immer um Fragen der formalen Modellierung und der algorithmischen Verarbeitung. „Algorithmus" steht für eine detailliert festgelegte Verfahrensvorschrift zur Lösung eines Problems. Typische Fragestellungen der Computerlinguistik im Bereich der Syntax sind deshalb etwa die folgenden:

Grundfragen der Syntax

- Welche formalen Eigenschaften müssen Regeln haben, die einzelne syntaktische Phänomene beschreiben?
- Welchen Aufbau haben und welche übergreifenden formalen Zusammenhänge gibt es bei Regelwerken, die größere Sprachausschnitte beschreiben, bei Grammatiken?
- Wie müssen Grammatiken angewendet werden, um einen gegebenen Satz zu analysieren?
- Welche Gestalt und welche formalen Eigenschaften soll das Ergebnis eines solchen Analyse-Prozesses haben?
- Wie können formale Grammatiken dazu genutzt werden, die Sätze einer Sprache maschinell zu erzeugen?

In der vorliegenden Einführung wird gezeigt, wie diese und ähnliche Fragen für die Computerlinguistik in den Gebieten der Syntax und der Satzsemantik zu beantworten sind, wobei teilweise auch Fragen der Morphologie, Lexik und Wortsemantik angesprochen werden. Viele andere Teilgebiete werden nicht behandelt; auf sie wird lediglich sporadisch verwiesen.

Bei einer wissenschaftlichen Disziplin, deren Name sich aus zwei Teilen zusammensetzt, muss man sich fragen, was dieser Name eigentlich genau bedeutet. Ist die Computerlinguistik der Teil der Linguistik, der auf Computer ausgerichtet ist? Handelt es sich um eine Disziplin, bei der es vor allem um die Entwicklung von Produkten geht? Oder ist womöglich etwas ganz anderes gemeint? Gehen wir dieses Problem an, indem wir uns fünf Auffassungen dazu ansehen, was die Computerlinguistik eigentlich sei.

Auffassungen

Auffassung 1: Die Computerlinguistik ist ein Teilgebiet der Linguistik

Nach dieser Auffassung behandelt die Computerlinguistik die gleichen Probleme wie die Linguistik, nur mit einer anderen, manche würden sagen, präziseren formalen Grundlegung und Methodik. Die Computerlinguistik verfolgt danach das gleiche Erkenntnisinteresse wie die Linguistik, nämlich Struktur und Verwendung der menschlichen Sprache zu untersuchen; dabei setzt sie als Besonderheit das Mittel der Simulation ein, indem sie Sprachverarbeitungsprozesse auf dem Computer durch Programme realisiert und dann das Sprachverhalten des Computers mit dem des Menschen vergleicht. Auf diese Weise können auch linguistische Theorien überprüft werden.

Auffassung 2: Die Computerlinguistik ist eine Ingenieurwissenschaft.

Für informationstechnologische Systeme (IT-Systeme), die mit natürlicher Sprache umgehen können, gibt es einen großen Bedarf. Auch wenn sie sich noch nicht in dem Maße durchgesetzt haben, wie es vor etwa 20 Jahren prognostiziert worden ist, so gibt es doch keinen Zweifel daran, dass wir in der Zukunft die Steuerung technischer Geräte häufig durch sprachliche Interaktion bewerkstelligen werden. Die Entwicklung von technischen Systemen oder Produkten, die einen konkret bestehenden Bedarf decken, ist Gegenstand der Ingenieurswissenschaften; die Computerlinguistik kann demzufolge auch als eine solche angesehen werden.

Auffassung 3: Die Computerlinguistik befasst sich mit den Mechanismen des menschlichen Geistes.

Man kann die Meinung vertreten, dass Sprachverarbeitungsprozesse, die mit dem Computer rekonstruiert werden, umso effizienter und genauer werden, je mehr sie sich an den kognitiven Mechanismen des Menschen orientieren. Wenn man also ein sprachverarbeitendes System entwickelt hat, das in bestimmter Hinsicht der menschlichen Sprachfähigkeit ebenbürtig ist, könnte man behaupten, dass auch der menschliche Geist in ähnlicher Weise wie dieses System arbeiten muss. Das ist eine Ansicht, die seit den siebziger Jahren in der Kognitionswissenschaft und Teilen der Künstliche-Intelligenz-Forschung vertreten wird.

Auffassung 4: Die Computerlinguistik ist eine Hilfswissenschaft für andere wissenschaftliche Disziplinen, die sich mit Sprache befassen.

In neuerer Zeit tritt die Computerlinguistik in immer neuen Zusammenhängen in Erscheinung: bei der Auswertung von Textsammlungen einer bestimmten Sprache oder zu einem bestimmten Gegenstand, bei der Suche von Informationen im Internet oder bei der automatischen Auswertung von digitalen Texten. Computerlinguistische Verfahren spielen dabei häufig die Rolle von Werkzeugen, um an bestimmte Informationen zu gelangen. Damit ähnelt die Compu-