



## Die maschinelle Simulierbarkeit des Humanübersetzens

Evaluation von Mensch-Maschine-Interaktion  
und der Translatqualität der Technik

Markus Ramlow

**F** Frank & Timme

Markus Ramlow

Die maschinelle Simulierbarkeit des Humanübersetzens

Hartwig Kalverkämper/Larisa Schippel (Hg.)

TRANSÜD.

Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens

Band 27

Markus Ramlow

# Die maschinelle Simulierbarkeit des Humanübersetzens

Evaluation von Mensch-Maschine-Interaktion  
und der Translatqualität der Technik

**F** Frank & Timme  
Verlag für wissenschaftliche Literatur

Umschlagabbildungen: Federzeichnung © Irmgard Bornemann (2009),  
wohnhaft im Landkreis Lüchow-Dannenberg

ISBN 978-3-86596-260-7

ISSN 1438-2636

© Frank & Timme GmbH Verlag für wissenschaftliche Literatur  
Berlin 2009. Alle Rechte vorbehalten.

Das Werk einschließlich aller Teile ist urheberrechtlich geschützt.  
Jede Verwertung außerhalb der engen Grenzen des Urheberrechts-  
gesetzes ist ohne Zustimmung des Verlags unzulässig und strafbar.  
Das gilt insbesondere für Vervielfältigungen, Übersetzungen,  
Mikroverfilmungen und die Einspeicherung und Verarbeitung in  
elektronischen Systemen.

Herstellung durch das atelier eilenberger, Leipzig.

Printed in Germany.

Gedruckt auf säurefreiem, alterungsbeständigem Papier.

[www.frank-timme.de](http://www.frank-timme.de)

# Inhaltsverzeichnis

Abkürzungsverzeichnis	10
<b>Vorwort</b>	<b>11</b>
<b>THEORETISCHER TEIL</b>	<b>15</b>
<b>1. Kulturhistorische Annäherung</b>	<b>15</b>
<b>2. Maschinisierung menschlicher Sprachverwendung</b>	<b>17</b>
2.1. Teilbereiche der Computerlinguistik	17
2.2. Anwendungen der Computerlinguistik	18
2.2.1. Textkorrektur	18
2.2.2. Volltextsuche	21
2.2.3. Textzusammenfassung	23
2.2.4. Sprachsynthese	26
2.2.5. Mensch-Maschine-Dialog	29
2.2.6. Weitere Anwendungen	33
2.3. Stellung der Computerlinguistik in der Geschichte der Maschinisierung	34
<b>3. Konsequenzen der Maschinisierung</b>	<b>43</b>
3.1. Funktionen der Arbeit und Konsequenzen des Nichtarbeitens	43
3.2. Konsequenzen des Einsatzes von Maschinen	49
<b>4. Geschichte der Maschinellen Übersetzung</b>	<b>54</b>
4.1. Pioniere der Maschinellen Übersetzung vor der Zeit des Computers	54
4.2. Die Anfänge der Maschinellen Übersetzung	56
4.3. Der ALPAC-Bericht	61
4.4. Die 70er und 80er Jahre	63
4.5. Neuere Entwicklungen und gegenwärtiger Stand der Forschung	67

<b>5.</b>	<b>Klassifizierung der Übersetzungssysteme nach der Transferstrategie</b>	<b>73</b>
5.1.	Regelbasierte Übersetzungsstrategien	73
5.1.1.	Direkte Systeme	73
5.1.2.	Transferbasierte Systeme	75
5.1.3.	Interlingua-basierte Systeme	77
5.2.	Neuere Ansätze	79
5.2.1.	Korpusbasierter Ansatz	79
5.2.2.	Wissensbasierter Ansatz	82
<b>6.</b>	<b>Einsatz von Übersetzungssystemen in der Praxis</b>	<b>84</b>
6.1.	Bedeutung der Maschinellen Übersetzung	84
6.1.1.	Europäische Union	84
6.1.2.	USA	90
6.1.3.	Asien	92
6.2.	Maschinelle und maschinengestützte Übersetzung in der DGT	95
<b>7.</b>	<b>Menschliche und maschinelle Übersetzung im Vergleich</b>	<b>99</b>
7.1.	Menschliche Kommunikation	99
7.2.	Definitionen der Humanübersetzung	102
7.3.	Definitionen der Maschinellen Übersetzung	107
7.4.	Fokus: Maschinengestützte Übersetzung	113
7.5.	Fokus: Anwendungsverfahren	116
7.6.	Menschliche und maschinelle Informationsverarbeitung	120
7.7.	Menschliche und maschinelle Textanalyse und -produktion	123
<b>8.</b>	<b>Evaluierung menschlicher und maschineller Übersetzungen</b>	<b>129</b>
8.1.	Ansätze zur Evaluierung im Bereich der Humanübersetzung	129
8.1.1.	Übersetzungsorientierte Texttypologie	129
8.1.2.	Übersetzungsrelevante Textanalyse	131
8.1.3.	Der didaktische Übersetzungsauftrag	133
8.1.4.	Translatkritischer Ansatz	135
8.1.5.	Funktional-pragmatisches Modell	136

8.2.	Ansätze zur Evaluierung im Bereich der Maschinellen Übersetzung	138
8.2.1.	van Slype (1979)	138
8.2.2.	Lehrberger / Bourbeau (1988)	141
8.2.3.	Hutchins / Somers (1992)	143
8.2.4.	Arnold (1994)	146
8.2.5.	Hutchins (1997)	148
8.2.6.	Falkedal (1998)	149
8.2.7.	White (2003)	151
8.3.	Anwendbarkeit der Evaluierungskriterien der Humanübersetzung auf die Evaluierung maschinell übersetzter Texte	155
8.4.	Bewertung der Kriterien im Bereich der Maschinellen Übersetzung	158
8.4.1.	Verständlichkeit	158
8.4.2.	Inhaltstreue	159
8.4.3.	Kohärenz	160
8.4.4.	Brauchbarkeit / Akzeptabilität	161
8.4.5.	Stil	162
8.4.6.	Posteditationsaufwand	163
8.4.7.	Fehleranalyse	164
8.4.8.	Zusammenfassung	165
<b>9.</b>	<b>Forschungskontext zur Translatevaluierung im Bereich der Maschinellen Übersetzung</b>	<b>168</b>
9.1.	Überblick über für MÜ-Evaluierungen gewählte Methoden	168
9.2.	Eigenortung	172
	<b>EMPIRISCHER TEIL</b>	<b>181</b>
<b>1.</b>	<b>Methodik der in dieser Arbeit durchgeführten Evaluierung</b>	<b>181</b>
1.1.	Evaluierungsmethodik für den Bereich der privaten Nutzung	181
1.2.	Evaluierungsmethodik für den Bereich der professionellen Nutzung	184



<b>2.</b>	<b>Evaluierung auf der Grundlage der Satzbedeutung</b>	<b>190</b>
2.1.	Übersetzungsrichtung Französisch-Deutsch	190
2.1.1.	Satzbedeutung korrekt	191
2.1.2.	Satzbedeutung teilweise korrekt	193
2.1.3.	Satzbedeutung inkorrekt	196
2.1.4.	Satzbedeutung unklar	200
2.1.5.	Bewertung der Titel	204
2.1.6.	Zusammenfassung	205
2.2.	Übersetzungsrichtung Deutsch-Französisch	206
2.2.1.	Satzbedeutung korrekt	207
2.2.2.	Satzbedeutung teilweise korrekt	209
2.2.3.	Satzbedeutung inkorrekt	212
2.2.4.	Satzbedeutung unklar	216
2.2.5.	Bewertung der Titel	219
2.2.6.	Zusammenfassung	220
2.3.	Übersetzungsrichtungsübergreifende Zusammenfassung	221
<b>3.</b>	<b>Evaluierung auf der Grundlage der Fehlertypologie</b>	<b>227</b>
3.1.	Übersetzungsrichtung Französisch-Deutsch	227
3.1.1.	Lexik	227
3.1.2.	Morphologie	240
3.1.3.	Syntax	242
3.1.4.	Pronominalisierung	251
3.1.5.	Formale Fehler	255
3.1.6.	Zusammenfassung	258
3.2.	Übersetzungsrichtung Deutsch-Französisch	261
3.2.1.	Lexik	261
3.2.2.	Morphologie	272
3.2.3.	Syntax	276
3.2.4.	Pronominalisierung	288
3.2.5.	Formale Fehler	291
3.2.6.	Zusammenfassung	292

3.3.	Übersetzungsrichtungsübergreifende Zusammenfassung	295
3.3.1.	Bewertung der Fehler nach Fehlertypen	295
3.3.2.	Bewertung der Fehler nach dem Posteditationsaufwand	297
3.3.3.	Bewertung der Qualität der Übersetzungssysteme	300
<b>4.</b>	<b>Ganzheitliche Auswertung</b>	<b>303</b>
4.1.	Ansätze zur Reduzierung der Fehlerzahl in maschinell übersetzten Texten	303
4.1.1.	Simulierung des Vorgehens beim Humanübersetzen	303
4.1.1.1.	Lexik	304
4.1.1.2.	Fokus: Kulturspezifika als Übersetzungsproblem	307
4.1.1.3.	Morphologie	324
4.1.1.4.	Syntax	328
4.1.1.5.	Pronominalisierung	331
4.1.1.6.	Zusammenfassung	333
4.1.2.	Maschinelle Kompensationsstrategien	335
4.2.	Chancen der interdisziplinären Methodik	337
4.3.	Pragmatische Perspektiven	340
4.3.1.	Gegenwärtiger Stand und Perspektiven der Maschinisierbarkeit des Humanübersetzens	340
4.3.2.	Perspektiven im Bereich der privaten Nutzung	342
4.3.3.	Perspektiven im Bereich der professionellen Nutzung	344
	<b>Bibliographie</b>	<b>350</b>

# Abkürzungsverzeichnis

<b>A-Satz</b>	Ausgangssatz
<b>AT</b>	Ausgangstext
<b>FAMT</b>	Fully Automatic Machine Translation
<b>FI</b>	Fehlerindex
<b>MAT</b>	Machine Aided Translation
<b>MT</b>	Machine Translation
<b>MÜ</b>	Maschinelle Übersetzung
<b>PT</b>	Personal Translator
<b>Z-Satz</b>	Zielsatz
<b>ZT</b>	Zieltext

## Vorwort

Schon seit dem 17. Jahrhundert träumt der Mensch davon, eine Maschine zur Verfügung zu haben, die es ermöglicht, Texte von einer natürlichen Sprache in eine andere natürliche Sprache zu übersetzen, davon also, den Übersetzungsprozess zu automatisieren. Die ersten Versuche, diesen Menschheitstraum zu verwirklichen, wurden allerdings erst vor etwas mehr als 60 Jahren unternommen.<sup>1</sup> Anfangs herrschte großer Optimismus. Man ging davon aus, dass es ohne größere Schwierigkeiten möglich sei, eine Übersetzungsmaschine zu konstruieren, die ebenso gut wie der Mensch übersetzen kann.<sup>2</sup> Doch schon bald stellte sich heraus, dass die maschinelle Simulierung des Übersetzungsprozesses von zahlreichen Schwierigkeiten begleitet ist und dass es einer Maschine nicht möglich ist, ebenso wie ein Mensch zu übersetzen. Dennoch sind inzwischen Fortschritte erzielt worden, und zwar derartige Fortschritte, dass mithilfe von Übersetzungssystemen heute zumindest Rohübersetzungen generiert werden können.

Zudem ist die Maschinelle Übersetzung trotz der bestehenden Mängel aus mehreren Gründen von Bedeutung.

Sie ist gesellschaftlich und politisch bedeutend, da Übersetzungen in allen Gemeinschaften, in denen Sprecher unterschiedlicher Muttersprachen zusammenkommen, unerlässlich sind. Dies betrifft zum einen Staaten, in denen mehr als eine Sprache gesprochen wird, wie Italien und Kanada, zum anderen supranationale und internationale Organisationen wie die Europäische Union oder die Vereinten Nationen. Übersetzungen sind dabei in praktisch allen Lebensbereichen notwendig, z. B. in den Bereichen Wissenschaft, Wirtschaft, Industrie, Technik, Medizin und Recht.

Die Alternative zur Übersetzung besteht in der Verwendung einer Lingua franca. Allerdings ist die Entscheidung für eine Lingua franca gleichbedeutend damit, dass eine ausgewählte Sprache und Kultur eine Vormachtstellung gegen-

---

<sup>1</sup> Vgl. Hutchins 1995: 431, 433

<sup>2</sup> Vgl. z. B. Delavenay 1960: 114 - *The translation machine (...) is now on our doorstep. In order to set it to work, it remains to complete the exploration of linguistic data by means of comparative lexical and structural analyses...*

über anderen Sprachen und Kulturen gewinnt. Die Sprecher dieser anderen Sprachen werden in einem solchen Fall also ihres Rechts beraubt, Informationen in ihrer Muttersprache zu erhalten und zu senden.

Möchte man den Menschen das Recht zugestehen, in ihrer eigenen Muttersprache zu kommunizieren, kann dies nur geschehen, wenn Texte in verschiedenen Sprachen zur Verfügung gestellt werden. Der Bedarf an Übersetzung ist dann sehr groß, und zwar so groß, dass die Humanübersetzer diesen Bedarf nicht vollständig decken können. Die Produktivität eines Übersetzers lässt sich nämlich nur bis zu einem gewissen Grad steigern. Soll eine darüber hinausgehende Produktivitätssteigerung erfolgen, so muss zusätzlich auf maschinelle und maschinengestützte Übersetzungsverfahren zurückgegriffen werden.<sup>3</sup> So kann ein erfahrener Übersetzer bis zu 5 000 Wörter pro Tag übersetzen, doch sind für die Übersetzung ergänzende Recherchen durchzuführen und soll sie von höchster Qualität sein und abschließend noch einmal überarbeitet werden, wird die zu erreichende Wortzahl eher bei etwa 1 500 Wörtern oder noch darunter liegen.<sup>4</sup>

Sie ist wirtschaftlich aus zwei Gründen bedeutend. Einerseits ist es für ein Unternehmen, das seine Produkte oder Dienstleistungen auch im Ausland verkaufen möchte, von entscheidender Bedeutung, dass sämtliche Texte, die mit dem Verkauf des Produkts oder der Dienstleistung in Zusammenhang stehen, in der Sprache der potentiellen Kunden verfasst werden. Dies betrifft beispielsweise Gebrauchsanweisungen, Produktbeschreibungen und Werbeprospekte. Stehen diese Informationen nicht in der Muttersprache der Zielgruppe zur Verfügung, so werden viele potentielle Kunden nicht auf das Produkt oder die Dienstleistung aufmerksam werden bzw. nicht in der Lage sein zu verstehen, wie das Produkt zu bedienen ist. Selbst wenn ein potentieller Kunde es vermag, ein in einer Fremdsprache verfassten Text wie eine Gebrauchsanweisung zu verstehen, ist die Bereitschaft, derartige Dokumente in einer anderen als der eigenen Muttersprache zu lesen, wohl bei den meisten Menschen nicht vorhanden. Werden also Informationen über Produkte und Dienstleistungen nicht in der Sprache der Zielgruppe zur Verfügung gestellt, verringert sich die Zahl der Kunden und damit der Umsatz der Unternehmen erheblich.

---

<sup>3</sup> Vgl. Arnold 1994: 4

<sup>4</sup> Vgl. Goshawke 1987: 40

Andererseits stellen Übersetzungen an sich einen Kostenfaktor dar. So entstehen nicht nur Kosten durch die Bezahlung der Übersetzer, sondern auch durch Verzögerungen bei der Übersetzung von Texten, die dringend benötigt werden. Wenn eine Übersetzung nicht pünktlich fertiggestellt wird, kann sich dadurch beispielsweise die beabsichtigte Einführung eines Produkts am Markt verzögern. Dies führt dazu, dass das betreffende Unternehmen einen Wettbewerbsnachteil gegenüber konkurrierenden Unternehmen hinnehmen muss und in der Folge mit Umsatzeinbußen konfrontiert ist.<sup>5</sup>

Sie ist wissenschaftlich bedeutend, weil die Forschung und Entwicklung auf dem Gebiet der Maschinellen Übersetzung einen interdisziplinären Charakter besitzt und damit der Forschung in anderen Bereichen wie Informatik, Künstliche Intelligenz, Linguistik und Computerlinguistik Impulse geben kann.<sup>6</sup>

Sie ist philosophisch bedeutend, weil mit dem Versuch, den Übersetzungsprozess zu automatisieren, ein Prozess durch eine Maschine simuliert werden soll, der beim Menschen nur gelingt, wenn er umfangreiches Wissen anwendet, über Problemstellungen nachdenkt und zu Problemlösungen gelangt. Insofern zeigt die Maschinelle Übersetzung die Möglichkeiten und Grenzen der Automatisierung des Denkens auf.<sup>7</sup>

Um weitere Fortschritte bei der Qualität maschinell übersetzter Texte herbeiführen zu können, ist es erforderlich, den Status quo der Maschinellen Übersetzung zu begutachten, bestehende Schwächen aufzuzeigen und aus den erkannten Schwächen Schlussfolgerungen zu ziehen.

Die vorliegende Arbeit – Dissertation zur Translationswissenschaft an der Philosophischen Fakultät II der Humboldt-Universität zu Berlin (2009) – möchte einen Beitrag dazu leisten. Zu diesem Zwecke werden verschiedene Ausbaustufen der Übersetzungssysteme von zwei Herstellern vergleichend evaluiert. Dabei handelt es sich zum einen um zwei für den professionellen Anwender konzipierte Systeme, nämlich @prompt Professional 2006 und Personal Translator 2008 Professional, zum anderen um zwei für den privaten

---

<sup>5</sup> Vgl. Arnold 1994: 4

<sup>6</sup> Vgl. Arnold 1994: 5

<sup>7</sup> Vgl. Arnold 1994: 5

Anwender konzipierte Systeme, nämlich @prompt Personal 7.5 quattro und Personal Translator 2008 Home.

Dementsprechend besteht die in dieser Arbeit vorgenommene Evaluierung aus zwei Teilevaluierungen, und zwar einer vergleichenden Evaluierung zweier für den professionellen und zweier für den privaten Anwender konzipierten Systeme. Die beiden Teilevaluierungen, die für das Sprachenpaar Französisch-Deutsch vorgenommen werden, sollen sich abschließend zu einer Gesamtevaluierung zusammenfügen. Diese wird die Stärken und Schwächen der jeweiligen Übersetzungssysteme aufzeigen und verdeutlichen, inwieweit derzeit verfügbare Übersetzungssysteme den Anforderungen verschiedener Nutzerkreise – im vorliegenden Fall: denen des professionellen und denen des privaten Anwenders – gerecht werden können. In ihrer Gesamtheit werden die von den verschiedenen Übersetzungssystemen generierten Übersetzungen, vor allem die der leistungsfähigsten, d. h. die der beiden für den professionellen Übersetzer konzipierten Systeme, somit auch aufzeigen, inwieweit der Prozess des Humanübersetzens maschinell simulierbar ist.

Auf den ersten Blick mag es unzureichend erscheinen, verallgemeinernde Aussagen zur maschinellen Simulierbarkeit der menschlichen Übersetzung auf der Grundlage einer Evaluierung von nur zwei Übersetzungssystemen treffen zu wollen. Doch die Auswahl der beiden Übersetzungssysteme beruht auf einer zuvor durchgeführten vergleichenden Evaluierung von sechs Übersetzungssystemen, im Rahmen derer sich die beiden dieser Arbeit zugrunde liegenden Systeme als vergleichsweise leistungsstark erwiesen haben.<sup>8</sup> Somit ist die Möglichkeit gegeben, aus der Evaluierung dieser beiden Systeme verallgemeinerbare Aussagen zum gegenwärtigen Stand der Maschinellen Übersetzung abzuleiten.

---

<sup>8</sup> Vgl. Ramlow 2008: 130-140

# THEORETISCHER TEIL

## 1. Kulturhistorische Annäherung

Schon immer kommunizieren die Menschen. Im Laufe der Jahrtausende und Jahrhunderte haben sich die Möglichkeiten zur Kommunikation erheblich erweitert: zum einen durch den Bau von Verkehrswegen (Straßen, Brücken usw.) und Transportmitteln (Kutschen, Schiffe, Autos usw.), zum anderen durch eine immer weiter fortschreitende Entwicklung von Techniken und Technologien, die eine von Raum und Zeit unabhängige Kommunikation ermöglichen.

Anfangs erfolgte die Übermittlung von Informationen ausschließlich in mündlicher Form. Mit der Entwicklung der Schriftsprache – angefangen mit der Keilschrift der Sumerer und den Hieroglyphen der Ägypter – wurden Informationen unabhängig von Ort und Zeit der Textproduktion verfügbar gemacht.

Mit dem Wagen, einem zumeist zweiachsigen Räderfahrzeug, stand bereits im 4. Jahrtausend v. Chr. ein einfaches Transportmittel und damit ein Instrument zur Verbreitung von Informationen über größere Distanzen zur Verfügung. Im Laufe der Jahrhunderte kamen zahlreiche weitere Transportmittel, und damit Möglichkeiten der immer schnelleren Informationsübermittlung, auf, wie beispielsweise das Segelschiff (ab etwa 4 000 v. Chr.), die Kutsche (ab dem 15. Jahrhundert), der Heißluftballon (1783), die Dampflokomotive (1803), das Dampfschiff (1807), das Automobil (1885/86) und das Flugzeug (1903)<sup>9</sup>.

Ein erster Schritt hin zur Massenkommunikation auf der Grundlage der Schriftsprache gelang Mitte des 15. Jahrhunderts mit der Erfindung des Buchdrucks mit beweglichen Metalllettern. Weitere Erfindungen und Entwicklungen im Dienste einer verbesserten Informationsübermittlung durch die Technik waren beispielsweise der *Semaphor*, ein optischer Flügeltelegraf (1791/1792), elektrische (ab 1804) und elektromagnetische Telegrafen (ab 1833), das Telefon (1876), das Mikrofon (1861), der Rundfunk (Radio und Fernsehen) (ab 1894/95) und der Computer (1941)<sup>10</sup>.

---

<sup>9</sup> Vgl. Rehm 2000a; 2000b; 2000c; 2000d; 2000f; 2000g

<sup>10</sup> Vgl. Rehm 2000b; 2000c; 2000e; 2000f; 2000h



Durch den stetigen Ausbau von Verkehrswegen, Transportmitteln und Kommunikationstechnologien entstanden vielfältige Möglichkeiten zur Kommunikation sowohl innerhalb von Kultur- und Sprachräumen als auch darüber hinaus. So ist die heutige Welt in großem Umfang durch mono- und plurikulturelle Beziehungen in Politik, Wirtschaft, Wissenschaft usw. geprägt. Die schnelle und kostengünstige Übermittlung von Informationen ist zu einem entscheidenden Faktor nicht nur in monokulturellen, sondern auch und vor allem in interkulturellen Kommunikationsprozessen geworden.

## 2. Maschinisierung menschlicher Sprachverwendung

Um Mitte des 20. Jahrhunderts sind die ersten Großrechenmaschinen, die Computer, entwickelt worden. Dies veranlasste die Menschen dazu, mithilfe dieser Rechenanlagen nicht nur mathematisch-logische Operationen durchzuführen, sondern auch die menschliche Sprachverwendung zu simulieren. Der Versuch, Letzteres zu tun, ist Gegenstand der Computerlinguistik.

### 2.1. Teilbereiche der Computerlinguistik

Die Computerlinguistik befasst sich mit der "Verarbeitung natürlicher Sprache (...) auf dem Computer, was sowohl geschriebene Sprache (...) als auch gesprochene Sprache (...) umfasst."<sup>11</sup> Diese allgemeine Definition umfasst verschiedene Auffassungen von Computerlinguistik, nämlich:

- Computerlinguistik als Teilgebiet der Linguistik: Sie befasst sich mit berechnungsrelevanten Gesichtspunkten der Sprache und der Sprachverarbeitung, wobei dies vorrangig auf theoretischer Ebene und somit unabhängig von der tatsächlichen Umsetzung auf dem Computer geschieht.
- Computerlinguistik als Forschungsvorhaben, das auf die Entwicklung von für die Linguistik bedeutsamen Computerprogrammen und die Verarbeitung linguistischer Daten abzielt. Dieser Bereich beruht vor allem auf der empirischen Auswertung umfangreicher Sprachdatenkorpora.
- Computerlinguistik als Simulation natürlichsprachlicher Phänomene auf dem Computer.
- Computerlinguistik als praxisorientierte Aufgabe, die auf die Entwicklung von Sprachsoftware abzielt.<sup>12</sup>

Die Computerlinguistik gliedert sich in einen theoretisch ausgerichteten und einen praktisch ausgerichteten Teil.

---

<sup>11</sup> Amtrup 2001: 10

<sup>12</sup> Vgl. Amtrup 2001: 10-11

Gegenstand der theoretischen Computerlinguistik ist die Untersuchung der Strukturen, die der maschinellen Verarbeitung natürlicher Sprachen zugrunde liegen. Hierbei stehen grundsätzliche Fragen wie die Berechenbarkeit, Adäquatheit oder Erlernbarkeit dieser Strukturen im Vordergrund. So setzt sich die theoretische Computerlinguistik beispielsweise mit den folgenden Fragen auseinander: Wie komplex ist natürliche Sprache und inwieweit kann man diese Komplexität mit derzeitig zur Verfügung stehenden Maschinen simulieren? Wie muss ein Formalismus beschaffen sein, damit er es ermöglicht, natürlichsprachliche Phänomene in geeigneter Form maschinell nachzuahmen? Ist es einer Maschine möglich, sprachliche Konzepte automatisch zu kategorisieren und zu lernen?<sup>13</sup>

Gegenstand der praktischen Computerlinguistik ist die Erforschung und Entwicklung von Anwendungen, die es ermöglichen, sprachliche Phänomene, wie beispielsweise das Übersetzen von einer natürlichen Sprache in die andere, auf dem Computer zu simulieren. Um dies erfolgreich tun zu können, muss Verschiedenes unternommen werden, vor allem:

- Entwicklung von Formalismen, die es ermöglichen, natürlichsprachliche Phänomene zu modellieren,
- Beschreibung von Einzelsprachen bzw. von bestimmten Aspekten der Einzelsprachen, vor allem im Hinblick auf Lexikon und Grammatik,
- Entwicklung von Algorithmen, die es möglich machen, natürlichsprachliche Äußerungen zu bearbeiten,
- Evaluierung natürlichsprachlicher Systeme.<sup>14</sup>

## **2.2. Anwendungen der Computerlinguistik**

### **2.2.1. Textkorrektur**

Um Fehler in Texten zu korrigieren, kommen drei Arten von Programmen zum Einsatz, nämlich Programme zur Korrektur von Nichtwörtern, Programme zur kontextabhängigen Korrektur und Programme zur Grammatikkorrektur. Ein

---

<sup>13</sup> Vgl. Amtrup 2001: 15

<sup>14</sup> Vgl. Amtrup 2001: 14-15

Programm, das alle drei Aufgaben in sich vereint, steht derzeit nicht zur Verfügung.<sup>15</sup>

Im Rahmen der Korrektur von Nichtwörtern wird ein Text nach Wörtern durchsucht, die dem Korrekturprogramm unbekannt sind. Findet das Programm ein solches Wort, wird es als unbekannt gekennzeichnet. Durch Anlegen eigener Wörterbücher bzw. Ergänzung des Systemwörterbuchs kann der Systemwortschatz erweitert werden. Für die Worterkennung wird nicht nur ein Lexikon, sondern meist auch eine Liste verwendet, aus der die Veränderungen hervorgehen, die Stammwörter durch morphologische Prozesse (Flexion, Derivation, Komposition) erfahren können. Auf diese Weise können Rechtschreibfehler identifiziert werden.

Um diese auch korrigieren zu können, wird in der Regel eine Minimal Edit Distance genannte Funktion verwendet. Diese Funktion ermöglicht es, zu einem im Lexikon nicht vorgefundenen Wort die ähnlichsten vorhandenen Wörter zu finden und dem Benutzer als Korrekturvorschläge anzuzeigen. Bei der Bestimmung der Korrekturvorschläge geht das Programm davon aus, dass ein Großteil der Rechtschreibfehler auf Auslassung, Einfügung oder Änderung eines Buchstabens oder auf Vertauschung zweier aufeinanderfolgender Buchstaben (Transposition) zurückzuführen ist. Werden vom Programm mehrere Vorschläge unterbreitet, so obliegt es dem Benutzer zu entscheiden, welches im gegebenen Kontext der richtige Begriff ist.<sup>16</sup>

Ein Eingabefehler führt allerdings nicht zwangsläufig zu einem Nichtwort. Es ist auch möglich, dass dadurch andere korrekt geschriebene Wörter zustande kommen. Derartige Fehler können nur im Rahmen einer kontextabhängigen Korrektur gefunden werden. Hierzu werden zwei Verfahren angewendet.

Das erste Verfahren besteht darin, ähnlich geschriebene Wörter, die vom Benutzer häufig verwechselt werden (z. B. engl. *their*, *there*, *they're*) zu Verwechslungsmengen zu gruppieren. Wird in einem Text ein der Verwechslungsmenge zugehöriges Wort verwendet, werden auf das konkrete Wort ausgerichtete Analyseverfahren gestartet, um zu ergründen, ob das verwendete Wort tatsächlich

---

<sup>15</sup> Vgl. Flidner 2001: 411

<sup>16</sup> Vgl. Flidner 2001: 412-413

das im gegebenen Kontext erforderliche ist. Allerdings erlaubt dieses Verfahren nur die Korrektur von Wörtern, die der Verwechslungsmenge angehören. Andere Fehler werden nicht korrigiert.

Das zweite Verfahren beruht auf einer statistischen Auswertung des Kontexts, bei dem durch die Analyse umfangreicher Korpora die Wahrscheinlichkeit bestimmt wird, dass jeweils drei aufeinanderfolgende Wörter zusammen auftreten. Zu einem gegebenen Satz werden mithilfe der bereits erwähnten Verfahren Auslassung, Einfügung, Änderung und Transposition ähnliche Sätze generiert, wobei für jeden Satz ermittelt wird, wie hoch die Wahrscheinlichkeit ist, dass dieser korrekt ist. Wenn ein vom Programm generierter Satz eine höhere Wahrscheinlichkeit erzielt als der vom Benutzer eingegebene Satz, wird er als Korrekturvorschlag angezeigt.<sup>17</sup>

Teilweise ist für die Erkennung und Korrektur eines Fehlers die Analyse eines ganzen Satzes erforderlich, vor allem dann, wenn es sich um Grammatikfehler wie fehlende Kongruenz zwischen verschiedenen Wörtern (z. B. zwischen Subjekt und Prädikat) oder fehlende Wörter (z. B. Artikel) handelt. Im Normalfall werden syntaktisch fehlerhafte Sätze im Rahmen einer maschinellen Syntaxanalyse gänzlich zurückgewiesen. Um nun aber einem Benutzer bei grammatisch fehlerhaften Sätzen Korrekturvorschläge unterbreiten zu können, muss die Syntaxanalyse dahingehend verändert werden, dass ungrammatische Sätze nicht abgelehnt werden und dass stattdessen Korrekturmöglichkeiten aufgezeigt werden. Hierzu werden zwei Verfahren eingesetzt, nämlich Constraint Relaxation und Fehlerantizipation.

Die Constraint Relaxation beruht darauf, dass gewisse Grammatikalitätsbedingungen (Constraints) nicht unbedingt erfüllt sein müssen. Kann bei einem eingegebenen Satz keine Analyse vorgenommen werden, bei der alle Grammatikalitätsbedingungen erfüllt sind, so werden diese systematisch gelockert. Die Satzanalyse wird mit jeweils unterschiedlichen gelockerten Bedingungen durchgeführt, bis eine Satzanalyse mit einer der durchgeführten gelockerten Bedingungen gelingt. So erschließt das Programm, welche Grammatikalitätsbedin-

---

<sup>17</sup> Vgl. Flidner 2001: 413-414

gung im gegebenen Satz nicht erfüllt ist, und schlägt eine entsprechende Korrektur vor.

Der Fehlerantizipation liegt die Annahme zugrunde, dass Fehler gewissen Mustern folgen. Fehler werden dadurch gefunden, dass ein Abgleich des zu korrigierenden Textes mit zuvor definierten Fehlermustern vorgenommen wird. Somit werden aber nur diejenigen Fehler gefunden, die zuvor im Fehlermuster erfasst wurden. Darüber hinausgehende Fehler bleiben unbemerkt.<sup>18</sup>

Allerdings ist die Präzision derzeitiger verfügbarer Grammatikkorrekturprogramme noch unzureichend. Heutige Systeme erzielen normalerweise eine Präzision von höchstens 50 %, d. h., dass auf einen korrekterweise angezeigten Fehler mindestens ein Fall kommt, bei dem korrekte sprachliche Strukturen fälschlicherweise als Fehler interpretiert werden. Zudem werden bei Weitem nicht alle tatsächlichen Grammatikfehler gefunden, so dass Grammatikkorrekturprogramme gegenwärtig noch nicht zur Endkorrektur von Texten verwendet werden können.<sup>19</sup>

### **2.2.2. Volltextsuche**

Die Volltextsuche ermöglicht es, Texte oder Textpassagen aus einer großen Menge von elektronisch gespeicherten Texten herauszufiltern. Dies geschieht mithilfe eines Indexes, also eines meist alphabetisch sortierten Wörterverzeichnisses. Zu jedem dieser Indexterme genannten Wörter sind Verweise auf Textstellen aufgelistet, an denen der Indexterm erwähnt wird. Somit können die Texte oder Textstellen, die im Zusammenhang mit einem bestimmten Indexterm von Belang sind, schnell gefunden werden. Die Indexterme werden in der Regel automatisch aus den Texten extrahiert und in Form einer Wortliste organisiert. Dabei werden umfassende Wortlisten erstellt, die sich im Gegensatz zu Schlagwortregistern einer Bibliothek in Papierform nicht auf einige wenige Schlagwörter beschränken. Funktionswörter wie Artikel, Präposition usw. werden nicht indexiert, sondern als sogenannte Stoppwörter zuvor herausgefiltert.<sup>20</sup>

---

<sup>18</sup> Vgl. Flidner 2001: 414-415

<sup>19</sup> Vgl. Flidner 2001: 415

<sup>20</sup> Vgl. Dörre 2001: 427-428

In Suchanfragen können verschiedene Operationen ausgeführt werden, nämlich:

- Unschärfe Suche: Es wird nicht nur nach dem eingegebenen Wort gesucht, sondern auch nach Wörtern, die diesem ähneln. Dieses Verfahren bietet den Vorteil, dass auch Wörter in falscher Rechtschreibung gefunden werden.
- Phonetische Suche: Es wird neben dem eingegebenen Wort auch nach Wörtern gesucht, die ebenso wie das eingegebene Wort ausgesprochen werden (z. B. Maier, Meier, Meyer).
- Phrasensuche: Es wird nach Dokumenten gesucht, in denen nicht einzelne Wörter, sondern festgelegte Wortfolgen vorkommen.
- Suche in Feldern: Es werden Felder wie Autor, Titel, Erscheinungsjahr usw. festgelegt, so dass eine gezielte Suche im Rahmen bestimmter Kategorien durchgeführt werden kann.<sup>21</sup>

Eine große Leistungssteigerung bei der Suche nach Dokumenten mithilfe von Schlagwörtern kann durch die verknüpfte Suche nach mehreren Einzelwörtern erzielt werden, da dadurch die durchsuchten Dokumente viel stärker gefiltert werden. Auf welche Weise in einem Suchsystem einzelne Indexterme miteinander verknüpft werden können, hängt von dem Retrievalmodell (engl.: *retrieve* = dt.: *wiederfinden*) ab, mit dem das Volltextsuchsystem arbeitet. Ein Beispiel für ein solches Retrievalmodell ist das sogenannte Boole'sche oder auch mengentheoretische Retrievalmodell. Bei diesem Modell können aus mehreren Termen bestehende Anfragen mit den Boole'schen Operatoren UND, ODER und NICHT verknüpft werden. So hat eine Suchanfrage wie

Hotel UND Berlin UND NICHT Potsdam

die Bedeutung, dass Dokumente als Suchtreffer angezeigt werden, die sowohl den Begriff *Hotel* als auch den Begriff *Berlin*, nicht aber den Begriff *Potsdam* enthalten.<sup>22</sup>

---

<sup>21</sup> Vgl. Dörre 2001: 428

Zwar bietet das Boole'sche Retrievalmodell den Vorteil, dass die zu verwendende Anfragesprache für den Benutzer ohne Schwierigkeiten nachzuvollziehen ist, doch ist die Einfachheit dieses Systems in der Praxis zugleich ein Nachteil. Bei Suchanfragen wird nur zwischen relevanten und irrelevanten Dokumenten unterschieden. Das System toleriert die teilweise Erfüllung der in einer Anfrage genannten Kriterien nicht. Somit erhält der Benutzer auf seine Suchanfrage oft entweder eine Vielzahl von gefundenen Dokumenten oder nicht ein einziges. Um dieses Problem zu beheben, werden sogenannte erweiterte Boole'sche Retrievalmodelle verwendet. Dabei wird bei einer Anfrage jedem Dokument ein Relevanzwert zugeordnet. Dazu wird für jeden Boole'schen Operator eine Relevanzfunktion festgelegt. Aus den einzelnen Relevanzwerten der Teilausdrücke der Suchanfrage wird ein Relevanzwert für die gesamte Suchanfrage errechnet, so dass dem Benutzer bei komplexen, d. h. aus mehreren Termen bestehenden Anfragen zahlreiche Dokumente angezeigt werden, die nach dem Grad ihrer Übereinstimmung mit der Suchanfrage geordnet sind.<sup>23</sup>

### **2.2.3. Textzusammenfassung**

Das Ziel einer Zusammenfassung eines Textes ist es, einen komplexen Textzusammenhang je nach Zweck der Zusammenfassung auf die wichtigsten Aspekte zu reduzieren. Dafür ist es unerlässlich, den Text, der zusammengefasst werden soll, zu verstehen.

Dennoch hat man bei der Forschung und Entwicklung auf dem Gebiet der maschinellen Textzusammenfassung den Aspekt des Verstehens zunächst vernachlässigt und Systeme entwickelt, bei denen die Zusammenfassung ausschließlich auf der Grundlage der Auswertung statistischer Daten geschah. So beruhten die ersten Systeme der automatischen Zusammenfassung zur Generierung von Abstracts auf der Zählung von Worthäufigkeiten und der daraus abgeleiteten Wichtigkeit einzelner Wörter. Die Wichtigkeit von Sätzen leitete sich dabei aus der Anzahl ihrer wichtigen Wörter im Verhältnis zur Anzahl der weniger oder nicht wichtigen Wörter im Satz ab. Sodann wurden die Sätze nach ih-

---

<sup>22</sup> Vgl. Dörre 2001: 428, 430

<sup>23</sup> Vgl. Dörre 2001: 431



rer ermittelten Wichtigkeit geordnet, wobei nur die wichtigsten Sätze zur Generierung des Abstracts herangezogen wurden.<sup>24</sup>

Ein Beispiel für einen neueren statistischen Ansatz zur Textzusammenfassung ist der trainierbare Zusammenfasser. Dieser entscheidet mithilfe eines vom Menschen erstellten Korpus von Dokumenten und dazugehörigen Abstracts, welche Sätze eines Dokuments zur Anfertigung einer Zusammenfassung herangezogen werden sollen. Um die für die Zusammenfassung relevanten Sätze zu bestimmen, analysiert er den Text im Hinblick auf die folgenden Merkmale:

- Satzlänge: Sätze, die nicht eine bestimmte Mindestlänge aufweisen (z. B. fünf Wörter), werden als in Abstracts unwahrscheinlich angesehen und somit negativ bewertet.
- Indikatorphrasen: Sätze, in denen Indikatorphrasen wie *In conclusion...* oder *We found...* auftreten, werden positiv bewertet.
- Absatzstruktur: Die ersten zehn Absätze und die letzten fünf Absätze eines Textes werden positiv bewertet. Es wird darüber hinaus davon ausgegangen, dass innerhalb eines Absatzes am Anfang und am Ende Sätze stehen, die den Textinhalt zusammenfassend wiedergeben.
- Schlüsselwörter: Sätze, die vergleichsweise viele Schlüsselwörter aufweisen, werden positiv bewertet. Als Schlüsselwörter gelten dabei die am häufigsten auftretenden Wörter, wobei Funktionswörter ignoriert werden.
- Akronyme: Sätze mit Akronymen werden positiv bewertet, da diese als Eigennamen angesehen werden und Eigennamen oft wichtig für den Inhalt eines Textes sind.<sup>25</sup>

Um 1980 wurden zur automatischen Zusammenfassung erstmals wissensbasierte Systeme verwendet, die sich am Prozess des menschlichen Zusammenfassens orientieren. So gliedert sich das Verfahren der wissensbasierten automatischen Zusammenfassung in drei Schritte. Zunächst wird der Inhalt des Textes in eine Bedeutungsrepräsentation umgewandelt. Sodann wird diese Bedeutungsreprä-

---

<sup>24</sup> Vgl. Endres-Niggemeyer 2001: 456, 460

<sup>25</sup> Vgl. Endres-Niggemeyer 2001: 460-461

sensation auf die wichtigsten Informationen reduziert. Schließlich wird auf der Grundlage der reduzierten Repräsentation die Zusammenfassung generiert.<sup>26</sup>

Allerdings durchlaufen nicht alle Systeme diese drei Schritte vom Originaldokument bis zur Zusammenfassung. So beschränken sich einige Systeme darauf, aus einem strukturierten Datenbestand eine Zusammenfassung zu generieren. Dies trifft beispielsweise auf das System STREAK zu, das aufbauend auf strukturierte Daten Kurzbeschreibungen von Basketballspielen erzeugt. Diese Kurzbeschreibungen werden mittels eines Faktengenerators erstellt, der sich auf zwei Quellen stützt, nämlich auf abgespeicherte Fakten zu früheren Basketballspielen und auf neue Fakten zu einem aktuellen Spiel. Auf dieser Grundlage wird eine Satzstruktur generiert, die in weiteren Schritten wieder revidiert und dabei um weitere Fakten ergänzt wird. Wie dies funktioniert, zeigt das folgende Beispiel:

1. Initial draft:

Hartford, CT – Karl Malone scored 39 points Friday night as the Utah Jazz defeated the Boston Celtics 118-94.

2. Adjunctization:

Hartford, CT – Karl Malone *tied a session high with* 39 points Friday night as the Utah Jazz defeated the Boston Celtics 118-94.

3. Conjoin:

Hartford, CT – Karl Malone tied a session high with 39 points *and Jay Humphries added 24* Friday night as the Utah Jazz defeated the Boston Celtics 118-94.

4. Absorb:

Hartford, CT – Karl Malone tied a session high with 39 points and Jay Humphries *came off the bench to add 24* Friday night as the Utah Jazz defeated the Boston Celtics 118-94.

---

<sup>26</sup> Vgl. Endres-Niggemeyer 2001: 456-457

#### 5. Nominalization:

Hartford, CT – Karl Malone tied a session high with 39 points and Jay Humphries came off the bench to add 24 Friday night as the Utah Jazz *handed* the Boston Celtics *their sixth straight home defeat* 118-94.

#### 6. Adjoin:

Hartford, CT – Karl Malone tied a session high with 39 points and Jay Humphries came off the bench to add 24 Friday night as the Utah Jazz handed the Boston Celtics their *franchise record* sixth straight home defeat 118-94.

Aufbauend auf einen anfänglichen Entwurf ist der Satz unter Zuhilfenahme von gespeicherten Texten zu früheren Basketballspielen nach und nach umformuliert und durch Hinzufügung völlig neuer Informationen ergänzt worden.

### 2.2.4. Sprachsynthese

Sprachsynthesysteme werden immer dann eingesetzt, wenn es im Rahmen der Kommunikation zwischen Mensch und Maschine notwendig oder zumindest vorteilhaft ist, dass der Austausch von Informationen mündlich vonstatten geht. Die Anwendungen sind zahlreich. So kommt die Sprachsynthese beispielsweise nicht nur bei Navigationssystemen, Verkehrsmeldungen, Reiseauskünften, Kinoprogrammen und Börsenkursen, sondern auch am Computerarbeitsplatz für Blinde und Sehbehinderte oder zur Generierung einer künstlichen Stimme für Sprechbehinderte zum Einsatz.<sup>27</sup>

Die Sprachsynthese vollzieht sich in zwei Schritten: Zunächst wird der eingegebene Text linguistisch analysiert, sodann wird die linguistische Repräsentation, die sich aus der Textanalyse ergibt, in ein synthetisches akustisches Signal umgewandelt. Im Folgenden werden diejenigen Arbeitsschritte beschrieben, die auf der Grundlage des schriftlich eingegebenen Textes eine linguistische Repräsentation erstellen, die dann als Basis für die akustische Synthese dient.<sup>28</sup>

---

<sup>27</sup> Vgl. Möbius 2001: 462

<sup>28</sup> Vgl. Möbius 2001: 462-463

Die Komplexität der linguistischen Analyse eines Satzes wird am folgenden Beispiel verdeutlicht:

*Bei der Wahl am 12.3.1998 gewann Tony Blair ca. 52 % der Wählerstimmen.*

Um diesen Satz korrekt vorlesen zu können, muss der Sprecher über vielfältige Informationen verfügen. Er muss aus der schriftlichen Form der Wörter ihre Aussprache herleiten. Dazu sind unter anderem Kenntnisse der internen Struktur von Wörtern notwendig. So muss das Kompositum *Wählerstimmen* korrekt in seine Komponenten *Wähler* und *Stimmen* zerlegt werden, damit die Buchstabenfolgen *st* korrekt ausgesprochen werden kann. *Tony Blair* muss als englischer Name erkannt und somit gemäß den englischen Ausspracheregeln ausgesprochen werden. Die Abkürzung *ca.*, das Symbol *%*, die Zahl und das Datum müssen als Wortformen wiedergegeben werden. Dabei bereitet die richtige Interpretation der Punkte – einmal als Kennzeichen einer Abkürzung, zweimal als Teil der Datumsangabe, einmal als Satzendmarkierung – zusätzliche Schwierigkeiten. Es sind Kenntnisse über die korrekte Betonung von Wörtern und Silben erforderlich. Zudem muss der eingegebene Text in einzelne Wörter zerlegt werden. In den meisten Fällen ist im Deutschen diese Zerlegung durch die Leerzeichen zwischen einzelnen Wörtern vorgegeben, doch wie der Beispielsatz zeigt, ist dies keineswegs immer der Fall. So muss das Datum in Form von mehreren Wörtern dargeboten werden, obwohl es sich dabei um eine nicht durch Leerzeichen unterbrochene Zeichenkette handelt. Zudem ist für die korrekte Aussprache einzelner Wörter mitunter eine Analyse des Kontexts erforderlich, beispielsweise im Falle der Datumsangabe. So müssen der Tag und der Monat als Ordinalzahlen ausgedrückt werden und in Abhängigkeit davon, ob eine Präposition vorausgeht oder nicht, und davon, welche Präposition es ist, muss der richtige Kasus gewählt werden.<sup>29</sup>

Die Textanalyse gliedert sich in mehrere Schritte, die im Folgenden dargestellt werden.

Zunächst wird eine lexikalische Analyse vorgenommen. Bei den meisten Sprachsynthesystemen wird ein Lexikon verwendet, das zu jedem Eintrag grammatische Information (z. B. Wortart) und eine phonetische Transkription

---

<sup>29</sup> Vgl. Möbius 2001: 463-464

enthält. Im Hinblick auf die Erfassung aller flektierten und abgeleiteten Wörter einer Sprache lassen sich zwei Verfahren unterscheiden. Zum einen kann mit einem Vollformenlexikon gearbeitet werden, d. h. neben der Grundform eines Wortes sind auch sämtliche daraus abgeleitete Wortformen aufgeführt, zum anderen können lediglich die Wortstämme in das Lexikon aufgenommen werden. In diesem Fall muss bei allen Wortstämmen, die Wortarten angehören, welche morphologische Veränderungen erfahren können (im Deutschen Nomina, Adjektive und Verben), markiert werden, welches zuvor festgelegte Flexionsparadigma auf den jeweiligen Wortstamm anzuwenden ist. Darüber hinaus werden für die Expansion von Abkürzungen, für Eigennamen, für geographische Namen usw. besondere Wortlisten verwendet.

In einigen Sprachen können durch Zusammensetzung zweier oder mehrerer Wörter neue Wörter entstehen. So ist dieses Verfahren der Bildung von Komposita im Deutschen sehr produktiv. Da jederzeit spontan neue Komposita gebildet werden können, reicht es nicht aus, die vermeintlich gebräuchlichsten Komposita ins Lexikon aufzunehmen. Vielmehr muss ein linguistisches Analyseverfahren angewendet werden, das es ermöglicht, Komposita als aus mehreren Teilen bestehende Wörter zu interpretieren. Nun ist es keineswegs immer der Fall, dass ein Kompositum nur eine Möglichkeit der Segmentierung zulässt. So kann das Kompositum *Wählerstimmen* potentiell in unterschiedliche Morpheme segmentiert werden:

wähl [Verbstamm] + erst [Adjektivstamm] + imme [Nominalstamm] + n  
[Plural]

wähler [Nominalstamm] + st [Verbendung] + imme [Nominalstamm] + n  
[Plural]

wähler [Nominalstamm] + stimme [Nominalstamm] + n [Plural]

Um zu entscheiden, welche der drei möglichen Interpretationen die korrekte ist, werden Informationen über die Kombinierbarkeit von Morphemen und über die Auftretenshäufigkeiten in großen Korpora verwendet.

Wie das Beispiel zeigt, kann im Rahmen der lexikalischen und morphologischen Analyse einzelner Wörter nicht immer zuverlässig entschieden werden, wie ein Kompositum zu interpretieren ist. Daher ist es für die korrekte Disambiguierung

derartiger Mehrdeutigkeiten mitunter erforderlich, die syntaktische Struktur des Satzes zu analysieren, in dem ein Kompositum verwendet wird, das mehr als eine Segmentierung zulässt. Eine Syntaxanalyse ermöglicht es festzustellen, ob zusammengehörige Wörter syntaktisch kongruieren.

Die Aussprache eines Wortes wird durch seine Transkription im Lexikon festgelegt, zumindest dann, wenn ein Vollformenlexikon verwendet wird. Tritt im eingegebenen schriftlichen Text ein Wort auf, das nicht im Lexikon verzeichnet ist, so wird dieses mithilfe von Ausspracheregeln transkribiert. Derartige Systeme arbeiten in der Regel mit zahlreichen Ausnahmeregeln. Bei einem System, das nur Wortstämme im Lexikon verzeichnet, die durch Informationen über das anzuwendende Flexionsparadigma ergänzt werden, beruht die korrekte Aussprache einer Zeichenfolge darauf, dass jedes Wort mit den erforderlichen morphologischen Angaben versehen wird. Mit diesen Angaben können die Wörter aufgrund von Ausspracheregeln korrekt transkribiert werden. Bei im Lexikon verzeichneten Wortstämmen ist diese Transkription bereits vorhanden. Bei aus mehreren Morphemen bestehenden Wörtern und unbekanntem Wörtern wird die korrekte Aussprache im Rahmen einer Analyse der Morpheme des zusammengesetzten Wortes festgelegt. Die Transkription erfolgt in Analogie zu bereits bekannten Wörtern. Bei der Transkription und der darauf aufbauenden Aussprache ist logischerweise nicht nur die Phonemfolge, sondern auch die Betonung der Silben zu berücksichtigen. Dieses Verfahren erlaubt es, auf Ausnahmeregeln weitgehend zu verzichten.<sup>30</sup>

### **2.2.5. Mensch-Maschine-Dialog**

Durch natürlichsprachliche Dialogsysteme kann der Mensch mithilfe einer sprachlichen Ein- und Ausgabe mit der Maschine kommunizieren. Bei der ersten Generation von Dialogsystemen vollzog sich der Mensch-Maschine-Dialog noch ausschließlich in schriftlicher Form über Tastatur und Bildschirm. Moderne Dialogsysteme verfügen jedoch über Sprachsynthese- und Spracherkennungskomponenten, so dass der Benutzer mündliche Anfragen an ein Dialogsystem richten kann und von diesem ebenso in akustischer Form Antworten bekommt. Ein wichtiges Anwendungsgebiet sind maschinelle Auskunftssys-

---

<sup>30</sup> Vgl. Möbius 2001: 464-466