the United States): Diego Catalan, Dept. of Literature, D-007, University of California, San Diego, La Jolla, CA 92093.

4.8.4. Logotheque, Göteborg

Logotheque is involved in various collaborative projects, one of which is the creation of a reference corpus of two million running spanish words from newspaper material from which a concordance has also been prepared. This is in collaboration with Per Rosengren and Mr. David Mighetto of the Department of Romance Languages at the University of Göteborg. The information concerning access and dissemination of the data is the same as in the following section.

4.9. Swedish

4.9.1. Språkdata, Göteborg

Språkdata is the abbreviated form for *Institutionen för språkvetenskaplig databehandling* which is the Department of Computional Linguistics at the Göteborg University. Logotheque, the Swedish language bank, a service branch of the linguistics department, was established by the Swedish government in 1975. It is responsible for the collecting, storing, processing, and providing of linguistic material in machine-readable form. The language material is basically Swedish, although data on other languages are also included. Data at the Institute are encoded on the spot, acquired from other researchers, or, when the text is available in printed form, converted into machine-readable form by an optical character reader.

The language corpora at Logotheque is synchronic, and except for a corpus of Strindberg's works, dates from post 1965 and is principally of the written language. The language corpus contains some 30 million running words: novels (69 published in 1976 and 60 in 1981, comprising approximately nine million words), legal texts (about 500 000 words), reports of the proceedings of the Swedish Parliament (1978—79), about four million), daily newspapers (1965, 1976, about 2.3 million), and weekly magazines.

An author corpus of between six and seven million words of the complete works of August Strindberg is also being encoded at Logotheque. — Word corpora within Logotheque comprise about 200 000 entries, including the vocabularies of the Word-list of the Swedish Academy and the Frequency Dictionary of Present-Day Swedish based on newspaper material. A project at Logotheque called "Lexin" is creating dictionaries for various immigrant groups. Along these lines a Swedish dictionary that serves as basis for translation has been developed and will be published by Språkdata. It contains about 15 000 words — copies of data in machine-readable form, as well as of concordances of encoded texts are available for academic research. For information, write to: Logotheque, Språkdata, Göteborg Universitet, Norra Allégatan 6, 3—413101 Göteborg, Sweden.

5. Literature (selected)

S. Hockey 1980 · IDS 1982 · W. Lenders 1985 · D. Rood 1980.

Robert F. Allen, Piscataway, New Jersey (USA)

14. Segmentierung in der Computerlinguistik

1. Problemstellung

Das Problem der Segmentierung in der Computerlinguistik hat seinen Ursprung darin, daß

erstens sprachliche Phänomene, besonders in der gesprochenen Form, nicht von vorneherein hinreichend deutlich in ihre Segmente zerlegt und in ihrer Funktion bestimmt sind, daß

aber *zweitens* für die Arbeit mit dem Computer auf diese Segmente, nämlich auf die Laute, die Silben, die Wörter, die Sätze, die Sinneinheiten und ihre Beziehungen zueinander zugegriffen werden muß.

Zwar wird in der geschriebenen Form mancher Sprachen eine Segmentierung be-

reits vorgegeben, z. B. durch die Buchstabenschrift, die Wortzwischenräume und Interpunktionszeichen. Für eine linguistische Struktur- und Funktionsbeschreibung reichen diese jedoch nicht aus. Ebenso sind in der gesprochenen Sprache zweifellos hörbare Segmentgrenzen vorhanden; die 'Hörbarkeit' allein aber liefert noch kein hinreichendes Abgrenzungskriterium für die Verarbeitung durch Computer.

Einige Beispiele mögen verdeutlichen, auf welche Weise fortlaufender Text (hier der erste Satz aus R. Musils „Mann ohne Eigenschaften") in Segmente zerlegt werden kann (vgl. auch die Beispiele in Art. 17 in diesem Handbuch):

1) Segmentierung in Buchstaben:
Ü-b-e-r-d-e-m-A-t-l-a-n-t-i-k-b-e-f-a-n-d-s-i-c-h-e-i-n-b-a-r-o-m-e-t-r-i-s-c-h-e-s-M-i-n-i-m-u-m

2) Segmentierung in Lautzeichen (hier durch eine phonetische Schrift wiedergegeben):
y:-b-ə-r-d-e:-m-α-t-l-α-n-t-i-K-b-ə-f-α-n-t-z-I-ç-α-e-n-b-a-r-om-e-t-r-I-ʃ-ə-s-m-i-n-I-m-U-m

3) Segmentierung in Silben:
über-dem-At-lan-tik-be-fand-sich-ein-ba-ro-me-tri-sches-Mi-ni-mum.

4) Segmentierung in Morphe:
über-dem-Atlant-ik-be-fand-sich-ein-bar-o-metr-isch-es-Minim-um

5) Segmentierung in Wörter (genauer: Wortformen):
über-dem-Atlantik-befand-sich-ein-baro-metrisches-Minimum.

6) Segmentierung in Satzteile:
Über dem Atlantik — befand sich — ein barometrisches Minimum.

Wenn man einen Satz in dieser Weise zerlegt, zieht man in der Regel unbewußtes Wissen über die Struktur der Sprache heran. Dieses Wissen ist — als Ergebnis eines langandauernden Sprachlernprozesses — im Gedächtnis eines Menschen als Wissen über sein Sprachsystem gespeichert. Es ist also begrifflich zu unterscheiden zwischen den Einheiten des Sprachsystems, den Phonemen, Morphemen, Lexemen auf der einen, und den beobachtbaren Einheiten des Sprachverhaltens, den Phonen, Morphen, Wortformen auf der anderen Seite (vgl. Lyons 1981, 1.3). Für die Entwicklung von maschinellen Verfahren muß das in jeder sprachlichen Erscheinung enthaltene Wissen explizit gemacht werden, d. h. daß anhand von exakt definierten Segmentgrenzen die Segmente aus der umgebenden größeren Einheit isoliert werden müssen. Die isolierten Segmente lassen sich miteinander vergleichen und klassifizieren.

Segmentierung erfolgt also auf allen linguistischen Ebenen und in allen Systemen, in denen Sprache, geschriebene oder gesprochene, maschinell verarbeitet wird.

In der Computerlinguistik wird das Problem der Segmentierung auf zwei Weisen gelöst,

zum einen durch Erkennung der physikalischen Segmentgrenzen in akustischen Sprachsignalen bzw. in Buchstabenfolgen; die physikalischen Segmentgrenzen sind hierbei durch phonetische Merkmale wie Pausen, Formanten, Intonation etc. und durch graphische Merkmale wie Wortzwischenräume, Satzzeichen etc. gegeben;

zum anderen durch besondere sprachwissenschaftliche Verfahren der funktionalen Beschreibung gleicher sprachlicher Einheiten und deren Klassifikation, womit vor allem die Methoden der Minimalpaaranalyse und des Parsing gemeint sind.

2. Physikalische Verfahren der Segmentierung

Im folgenden wird der Prozeß der Segmentierung anhand expliziter physikalischer Segmentgrenzen an einigen Beispielen aus verschiedenen linguistischen Beschreibungsebenen erläutert, und zwar an den Beispielen der Isolierung von Lauten und Buchstaben, der Silbentrennung, der Isolierung von Wörtern in Sätzen und schließlich der Zerlegung von Texten in Sätze.

2.1. Isolierung von Lauten

In vielen wissenschaftlichen Fragestellungen (z. B. in der Phonemanalyse, vgl. Art. 16) und in praktischen Anwendungen (z. B. der automatischen Spracherkennung, vgl. Art. 47) kann es erforderlich werden, die Laute eines gesprochenen Textes zu isolieren. Physikalisch gesehen stellt die Kette der Laute ein zeitliches Kontinuum akustischer Signale dar. Das Problem der Segmentierung besteht in der Umwandlung dieses Kontinuums in eine diskrete Zeichenkette.

Es liegt nahe, dieses Problem zu lösen, indem man sich zunutze macht, daß sich „die Eigenschaften des Sprachsignals ... in der Regel von Segment zu Segment" ändern, „während sie innerhalb eines Segmentes weitgehend konstant bleiben" (Ney 1983,

117). Es geht also darum, die über eine gewisse Dauer konstanten Eigenschaften eines Sprachsignals, z. B. die Formantstruktur, zu ermitteln und die aufgrund dieser Eigenschaften isolierbaren Segmente zu Klassen zu ordnen.

Abb. 14.1 zeigt einen Ausschnitt aus der Aufzeichnung eines gesprochenen Satzes in einem Sonagramm. Die kontinuierliche Signalkette ist nach der Signalfrequenz (y-Achse), der Intensität (Schwärzung) und dem Zeitverlauf (x-Achse) dargestellt. Der Fachmann erkennt aus dieser Aufzeichnung, an welcher Stelle ein bestimmter Laut anzunehmen ist, d. h. wo eine Segmentgrenze liegt. In unserem Beispiel wurde der jeweilige Laut explizit in phonetischer Umschrift unter dem Sonagramm angegeben; jeweils unter der phonetischen Umschrift befindet sich die Darstellung des Wortlauts in Schreibschrift. Die phonetischen Zeichen und die Zeichen der Schreibschrift stehen als Vertreter eines lautlichen Segments.

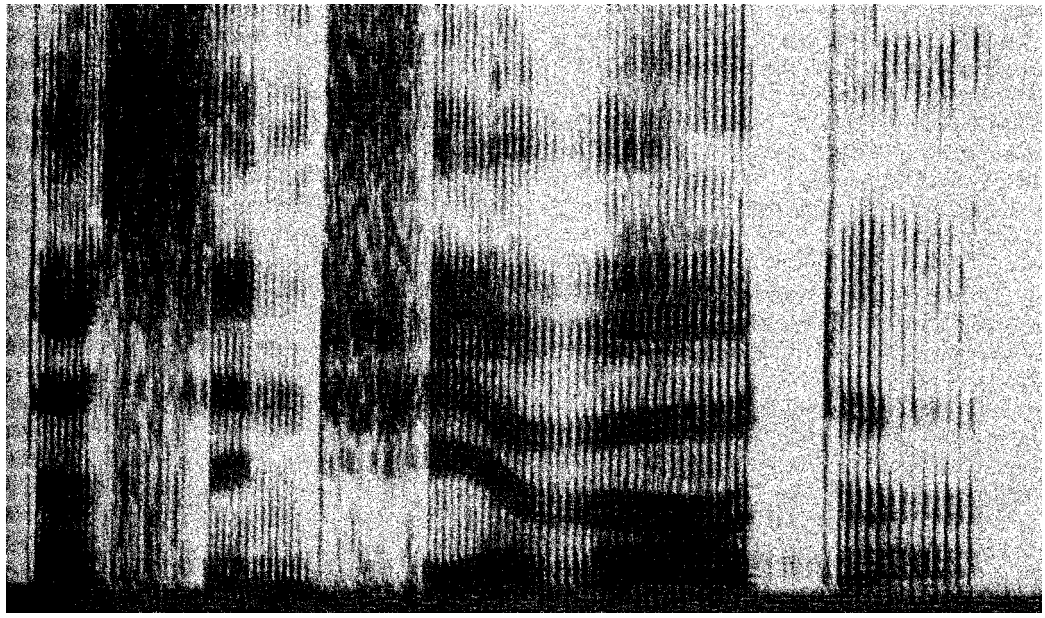

Abb. 14.1: Sonagramm

Die Segmentierung des akustischen Kontinuums in Laute wurde in diesem Beispiel durch das Gehör unterstützt. Es ist bis heute fraglich, ob man für die Erkennung von Lauten in fließender Rede überhaupt zuverlässige automatische Phonem-Erkenner entwikkeln kann; bisherigen Arbeiten auf diesem Gebiet war offenbar nur ein begrenzter Erfolg beschieden (vgl. Marcus 1983 a, 25).

Trotz dieses offenkundigen Problems, die Grenzen der Laute und Wörter im Kontinuum gesprochener Sprache explizit zu beschreiben, sind Menschen in der Lage, Einzellaute und Wörter zu erkennen. Sie verfügen über Muster der Einzellaute und Wörter (bzw. über die Art und Weise ihrer Erzeugung; vgl. Marcus 1983 a, 25), die Phoneme und Lexeme, die sie aufgrund eines langandauernden Spracherlernungsprozesses gespeichert haben, und verwenden dieses Wissen beständig bei der Erkennung der einzelnen Segmente eines Textes.

2.2. Silbensegmentierung

Die Segmentierung von Texten in Silben ist eine übliche Fähigkeit des Sprecher/Hörers, die er zur intonatorischen Gliederung der Rede einsetzt. Sie Silbe ist Träger so wichtiger Eigenschaften der gesprochenen Sprache wie Tonhöhe und Akzent. Im geschriebenen Deutsch macht man sich diese intonatorischen Gliederungsmarken für die Trennung der Wörter am Zeilenende zunutze: Man trennt „nach Sprechsilben, die sich beim langsamen Sprechen von selbst ergeben“ (Duden 1986, R 178, S. 58).

In der Automatisierung der Silbentren-

nung z. B. für Textverarbeitungsprogramme liegt die praktische, außerwissenschaftliche Bedeutung dieses Themas der Computerlinguistik. In wissenschaftlicher Hinsicht ist das Problem der Silbentrennung interessant für prosodische Untersuchungen an Texten, z. B. für Untersuchungen über den rhytmischen Aufbau von Verstexten (vgl. Chisholm 1980).

Wie im Falle der Laute, so muß jedoch auch für die Silbe festgestellt werden, daß eine einheitliche phonetische Definition der Silbe immer noch fehlt, d. h. daß es noch kein Verfahren gibt, aus den physikalischen Eigenschaften eines Sprachkontinuums die Grenzen zwischen Silben zu erkennen (vgl. Kohler 1977, S. 82; Marcus 1983 a, 26 f.).

Physikalisch gesehen sind die Grenzen zwischen Silben schlecht definiert. Doch macht man sich in neueren Systemen zur automatischen Sprachsynthese die Tatsache zunutze, daß es an den vokalischen *Silbenkernen* Energiemaxima gibt, an welchen man die Silbe in *Halbsilben* zerlegen kann. Aus diesen können dann „im Deutschen etwa 50 Konsonantengruppen für den Silbenanfang und etwa 150 für das Silbenende“ abgeleitet werden (vgl. Marcus 1983 a, 26; unter linguistischen Gesichtspunkten auch Bátori 1975, 321). Durch Kombination der Segmente dieses Inventars von Halbsilben lassen sich Wörter erzeugen und so Verfahren zur Sprachsynthese konstruieren.

2.3. Segmentierung von Sätzen in Wörter

Doch auch die algorithmische Erkennung einer auf den ersten Blick unproblematisch erscheinenden Einheit, des Wortes nämlich, erweist sich bei näherer Betrachung als nicht so einfach. Zu unterscheiden sind zwei Fälle,

a) die Segmentierung des lautlichen Kontinuums eines Satzes in Wörter als lautlich-akustische Einheiten und

b) die Segmentierung der schriftsprachlichen Zeichenkette eines Satzes in schriftsprachliche Einheiten.

Zu a)

Wie im oben schon beschriebenen Fall der Erkennung von Einzellauten, so muß es auch im Falle der akustischen *Worterkennung* darum gehen, die konstanten Eigenschaften des Sprachsignals festzustellen, die innerhalb eines Wortes für dessen Erkennung, d. h. Zuordnung zu einem Muster (Lexem), maßgebend sind.

Man kann grob zwischen Einzelworterkennung und Worterkennung in fließender Rede unterscheiden. Im ersten Bereich liegen heute ausgereifte und anwendungsbezogene Verfahren vor. In diesen wird durch eine akustische Analyse des gesprochenen Wortes ein Merkmalsmuster erstellt, das mit einer vorgegebenen Menge an Referenzmustern verglichen wird (zu den Verfahren im einzelnen vgl. z. B. Geppert/Kuhn/Ney 1983, 313 ff.; ferner Art. 47). Für die Erkennung von Wörtern in fließend gesprochener Rede dagegen reichen rein akustische Parameter nicht aus. Zur Erkennung der Wortgrenzen müssen hier andere ’Wissensquellen’, auch als ’höhere Wissensquellen’ bezeichnet, herangezogen werden (Geppert/Kuhn/Ney 1983, 372 ff., Brietzmann 1984).

Zu b)

In der Praxis des schriftlichen Sprachgebrauchs verwendet man beständig ein bestimmtes Merkmal, mit dem man die Wörter eines Textes voneinander trennt, den Wortzwischenraum. Es gibt aber auch kontrahierte Formen, die Informationen aus zwei Wortformen enthalten (z. B. *im*), und Verteilung des Wortinhalts auf mehrere Teile eines Wortes (z. B. in *kommt an* vs. *ankommen* und *ab und zu* vs. *gelegentlich*). Das Merkmal des Wortzwischenraums führt jedenfalls in vielen Fällen nicht zur eindeutigen Abgrenzung von Wörtern, es bedarf dazu vielmehr zweier zusätzlicher und für die Computerlinguistik grundlegender komplexer Bearbeitungsschritte, der Lemmatisierung und der Auflösung von Mehrdeutigkeiten (für Einzelheiten dieser beiden Themen vgl. Artikel 17).

Man erkennt aus diesen Ausführungen, daß zur Segmentierung eines Textes in Wörter nicht problemlos der Wortzwischenraum herangezogen werden kann. Zwar handelt es sich beim Wortzwischenraum um eine linguistische Information, die der Sprachbenutzer explizit in seinen geschriebenen Text einträgt; in vielen Fällen muß er aber, um Wörter eindeutig zu erkennen, auf den größeren Textzusammenhang zurückgreifen, also die betreffende Wortform hinsichtlich ihrer Funktion in der übergeordneten sprachlichen Einheit sehen (vgl. unten 3.).

2.4. Segmentierung von Texten in Sätze

In den meisten maschinellen Analysesystemen (so in den bisherigen Expertensystemen mit natürlich-sprachlicher Schnittstelle; vgl. Art. 57, 58, sowie in maschinellen Übersetzungsverfahren, vgl. Art. 53) wird die Einheit des Satzes als Analysegegenstand vorausgesetzt. Ein fortlaufender Text wird dabei meist

anhand von vorkodierten Identifikatoren in Sätze zerlegt. Im Deutschen können dazu die Interpunktionszeichen herangezogen werden. So dienten z. B. die Zeichen „.“ und „;“ im Saarbrücker maschinellen Analysesystem SUSY für das Deutsche als Satztrennungszeichen; “,“ wurde als Trennzeichen für Nebensätze verwendet. Bei der Übertragung von SUSY auf die Englische Sprache „erwies sich der für das Deutsche entwickelte Operator“ zur Satzsegmentierung „als untauglich, da dieser vor allem die strikte Zeichensetzung des Deutschen ausnutzt“. Das Englische dagegen zeichnet sich „durch eine Zeichensetzung auf, die sich nicht an den Teilsätzen, sondern eher an der Betonung orientiert“. Als Konsequenz aus diesem Sachverhalt mußte für das Englische ein Programm geschrieben werden, „bei dem zunächst auf der Basis der Wortklassen Wortgruppen zusammengefaßt und typisiert werden, innerhalb derer sich keine Teilsatzgrenze befinden kann“ (Blatt 1987, 302/303; auch Schmitz, K. D. 1986).

3. Linguistisch-klassifikatorische Verfahren der Segmentierung

Die explizite Angabe der Segmentinventare auf allen linguistischen Ebenen ist identisch mit der Explikation des menschlichen sprachlichen Wissens. Es ist daher nicht verwunderlich, daß das Problem der Segmentierung unter methodischem und inhaltlichem Aspekt Gegenstand mancher linguistischen Theorie ist. Für die strukturelle Linguistik wurde die Segmentierung sogar zur grundlegenden Methode der Linguistik überhaupt.

In strukturell-linguistischer Hinsicht ist Segmentierung Voraussetzung einer jeden Klassifikation von Segmenten nach ihrer Funktion; einer jeden Zuordnung struktureller und lexikalischer Information zu Texten geht die Segmentierung als methodischer Schritt voraus, für den formalisierbare Prozeduren gesucht werden, wenn sie auch praktisch nicht von der Klassifikation zu trennen ist. In der Praxis wurde das Problem der Segmentierung unter zwei verschiedenen Gesichtspunkten angegangen (vgl. auch Weber, H. 1973, 167): zum einen mit dem Ziel, in erster Linie die Inventare der funktional verschiedenen Segmente auf den einzelnen linguistischen Ebenen zu ermitteln, zum anderen mit der Absicht, die linearen und hierarchischen Beziehungen zwischen den Segmenten von Texten aufzuzeigen. Die hier bereitgestellten algorithmisierbaren und programmierbaren Verfahren sind einerseits die Minimalpaar- und Distributionsanalyse und andererseits das Parsing. Auf beide Themen wird in diesem Handbuch in besonderen Artikeln ausführlich eingegangen, so daß an dieser Stelle nur verdeutlicht werden soll, daß es sich um Verfahren der Segmentierung handelt (vgl. Artikel 6, 18, 31 und 32; zu einzelnen Parsing-Methoden vgl. auch King (ed.) 1983).

3.1. Minimalpaaranalyse

Minimalpaaranalysen wurden im wesentlichen auf den Ebenen der phonologischen und morphologischen Beschreibung verwendet. So führt Z. S. Harris „Segmentation“ ein als einen ersten Schritt, „toward obtaining phonemes“, als eine Prozedur, die

„represents the continuous flow of a unique occurence of speech as a succession of segmental elements, each representing some feature of a unique speech sound” (Harris 1951, 25).

Eine erste ’Zerteilung’ des kontinuierlichen Stromes der Rede wird aufgrund wahrgenommener Pausen, also physikalischer Phänomene gemäß Abb. 14.1, vorgenommen. Die so ermittelten „vorläufigen“ Segmente werden der „Minimalpaaranalyse“ unterzogen; d. h., daß Wortpaare gebildet werden, die sich minimal unterscheiden. Diese werden miteinander hinsichtlich des Kriteriums der Bedeutungsgleichheit verglichen, so daß abschließend die sie unterscheidenden Laute ggf. zu einer (Phonem)-Klasse geordnet werden können.

Die Segmentierung erfolgt im Falle der Minimalpaaranalyse in Verbindung mit der Klassifikation der Einheiten nach ihrer Funktion. Beide sind untrennbar miteinander verbunden. Ein umfangreiches Computerprogramm, in welchem eine Minimalpaaranalyse zur ’Entdeckung’ des Phonemsystems beliebiger Sprachen dargestellt wurde, ist TOPAS (Wothke 1983; vgl. auch Artikel 16).

Wie im Falle der Laute, so kann man auch für *Silben* eine linguistisch-funktionale Definition vornehmen, und zwar aus der Kombination bzw. Distribution der sie bildenden Phoneme (vgl. Kohler 1977, 112 ff.). Eine solche Definition läßt sich in Verfahren zur Silbentrennung und zur Erkennung von Silben einsetzen.

So sind die in der Duden-Grammatik (1966) der deutschen Gegenwartssprache aufgeführten 14 Regeln der Silbentrennung

wohl aus distributionellen Überlegungen entstanden. Einige dieser Regeln definieren die Silbengrenze aufgrund von Kriterien der lautlichen Umgebung des Vokals:

„Zwischen einem stimmlosen Verschlußlaut und folgendem b d g v z z liegt im Wortinnern eine Silbengrenze." (Duden 1966, 165).

Geht es nicht um Silbentrennung, sondern bloß um die *Erkennung* von Silben, z. B. zur Berechnung von deren statistischer Verteilung in einem Korpus, so kann eine einfachere funktionale Definition der Silbe angewendet werden (vgl. z. B. Krallmann 1966). Als Träger der Silbe wird hier, in Übereinstimmung mit der phonologischen Silbendefinition, ein Vokal angenommen, dem ein Konsonant vorausgehen und/oder folgen kann. Bei dieser Definition wird eine gewisse Fehlerquote in Kauf genommen, die jedoch für statistische Zwecke außer Betracht bleiben kann.

In ähnlicher Weise lassen sich Minimalpaaranalysen auch auf anderen Ebenen anwenden, besonders in der Morphologie. So liegt es z. B. nahe, auf der Grundlage von Nida (1946) und Harris (1951) die Morphemsysteme von Sprachen und die Klassifikation der Morpheme aufgrund von Minimalpaaranalysen zu gewinnen. Maschinelle Verfahren, welche die Prozeduren solcher morphologischen Minimalpaaranalysen ausführen, sind jedoch bis heute nicht bekannt.

3.2. Parsing

Unter Parsing versteht man allgemein den Prozeß der Zerteilung einer komplexen Einheit in ihre Segmente, nach Maßgabe eines Regelsystems, wobei die innere Struktur der komplexeren Einheit zum Vorschein kommt. Im Sinne des Strukturalismus wäre Parsing eine Methode zur Ermittlung der hierarchischen Ordnung der Morpheme im Wort und der Wörter im Satz mittels einer IC-Analyse (immediate constitutent analysis) (Weber 1973; Karttunen/Zwicky 1985).

3.2.1. Ermittlung von Morphen

Parsing-Strategien zur Erkennung der Wortstruktur ('wordparser') sind heutzutage Teil vieler komplexer Spracherkennungssysteme (vgl. z. B. Kay 1977; Pounder/Kommenda 1986; Vergne/Pagès 1986; Bear 1986; Russel/Pulman/Ritchie et al. 1986). Die Zahl der möglichen Zerlegung wird dabei von vornherein reduziert, indem Wörterbücher eingesetzt werden, welche die zulässigen Morpheme, nämlich Stämme, Präfixe, Suffixe, Infixe und Endungen enthalten. Im Prinzip lassen sich die Parsing-Verfahren, die aus der Syntax bekannt sind, auch zur Segmentierung von Wörtern verwenden, z. B. das 'chart parsing' in Kay 1977 (vgl. Artikel 17 und 31).

Im Unterschied zur Syntax muß bei einer eindeutigen Erkennung der Wortsegmente nicht nur mit einer internen Struktur, sondern auch mit morphologischen Veränderungen der Wörter auf graphematischer und phonematischer Ebene gerechnet werden. Zu diesen Erscheinungen gehören im Deutschen vor allem der Umlaut sowie Veränderungen im Auslaut, z. B. der Wechsel von 'ss' zu 'ß' in ERKENNTNISSE — ERKENNTNIß.

Weitere Probleme für morphologisches Parsing ergeben sich aus der morphologischen Mehrdeutigkeit mancher Wörter. So gibt es z. B. in folgenden Fällen mehrere Zerlegungsmöglichkeiten:

BEINHALTUNG 3	-->	BE-IN-HALT-UNG	BEIN-HALT-UNG
MÄDCHENHANDELSSCHULE	-->	MÄDCHEN-HANDEL-S-SCHULE	
		MADCHENHANDEL-S-SCHULE	
NASCHEN	-->	NASCH-EN	NAS-CHEN

In diesen Fällen muß die zutreffende Zerlegung aus dem größeren Zusammenhang erschlossen werden.

3.2.2. Isolierung von Satzkonstituenten

Die den Wörtern übergeordnete Einheit sprachlicher Äußerungen ist aus der traditionellen Grammatik als „Satzglied" oder „Satzteil" bekannt, in der modernen Grammatiktheorie auch „Konstituente" genannt.

Formal betrachtet besteht ein Satzglied aus einer Kette von Wortformen, die nach syntaktischen Regeln zusammengefügt werden.

Über die Art syntaktischer Regeln sowie über Parser vgl. im Einzelnen Art. 6, 31 und 32. Für den vorliegenden Zusammenhang sei hier nur festgestellt, daß der wichtigste Typ syntaktischer Regeln offenbar die Ersetzungsregel der Form

a + b ---> c

ist. In diesen Regeln stehen links und rechts

vom Pfeil Namen von Segmenten. Die Regeln geben Auskunft über die Zusammensetzung eines Segments (z. B. S, NP, VP ...) aus Segmenten der nächst niedrigeren sprachlichen Ebene. Es handelt sich hier also um Segmentierungsregeln, und Parser, mit denen Sätze in Konstituenten zerlegt werden, können als Segmentierungsalgorithmen betrachtet werden. Allerdings bilden die Segmente nicht eine lineare Abfolge, sie stehen vielmehr, da es sich um Segmente verschiedener Komplexitätsebenen handelt, in einem Abhängigkeitsverhältnis zueinander, durch das einem Satz ein bestimmtes strukturelles Muster, die Strukturbeschreibung, zugeordnet wird.

3.2.3. Segmentierung von Texten

In der wissenschaftlichen Untersuchung von Texten und in Anwendungen der Computerlinguistik, z. B. in der maschinellen Sprachübersetzung oder im Story Understanding, ist es häufig notwendig, die strukturellen Muster eines Textes, nach denen alle Einheiten aus jeweils kleineren oder größeren zusammengesetzt sind, zu ermitteln. Eine solche Beschreibung wird als 'Strukturbeschreibung' eines Textes bezeichnet.

Will man also die Struktur eines Textes beschreiben, so hat man ihn zunächst in Einheiten zu segmentieren. Sodann wird man die Einheiten klassifizieren, d. h. diejenigen Einheiten, die hinsichtlich bestimmter Kriterien gleich funktionieren, mit entsprechenden Angaben versehen und schließlich die Beziehungen zwischen den Einheiten, die regelmäßig auftreten, beschreiben wollen.

4. Klassifikation von Segmenten

Die durch Segmentierung isolierten Einheiten eines Textes werden unter definierten Gesichtspunkten zu Klassen zusammengeschlossen.

Typische Beispiele für Klassifikation sind die Zuordnung von Wortklassen zu Textwortformen oder die Angabe einer Satzteilbezeichnung zu einer Kette von Wörtern.

Nach der strukturalistischen Methode werden Klassen durch die Untersuchung der Distribution (Verteilung) eines potentiellen Segments in den Umgebungen, in denen es in einem Korpus vorkommt, gewonnen. Man bezeichnet diese Methode als 'Distributionsanalyse' oder 'Minimalpaaranalyse', wie sie in 2.1. behandelt wurden. In der sprachwissenschaftlichen Praxis hat man diese Verfahren jedoch kaum angewandt, sondern sich fast ausschließlich von pragmatischen oder plausiblen Segmentdefinitionen und Klassifikationen leiten lassen.

— So hat man z. B. als „Substantive" diejenigen Zeichenfolgen klassifiziert, die einen Gegenstand bezeichnet, als „Adjektive" diejenigen, die eine Eigenschaft zum Ausdruck bringen.

— Als „Laute" und „Silben" werden vom Phonetiker auditiv identifizierbare Signalfolgen bezeichnet, die ganz bestimmten auditiv wahrnehmbaren Merkmalen entsprechen müssen.

— Als „Wortform" wird die in Texten von Zwischenräumen umgebene Zeichenfolge bezeichnet, als „Wort" die abstrakte Größe, die verschiedenen Wortformen mit gleicher Bedeutung gemeinsam ist.

In allen drei Fällen werden Sprachsegmente zu Klassen geordnet.

Klassennamen sind Angaben über sprachliche Einheiten; sie können Funktionen bezeichnen, die sprachliche Einheiten in konkreten Texten ausüben können. Abb. 14.2 enthält einige Beispiele möglicher Klassen.

Segment-Ebene	mögliche Klassen
Phon/Phonem	stimmhafte, stimmlose, Plosive, Frikative
Morph/Morphem	Präfix, Suffix, Flexiv
Wortform/Wort	Substantiv, Adjektiv
Satzteil	Nominalgruppe, Verbalgruppe

Abb. 14.2: Beispiele von Klassen sprachlicher Ein- heiten

Beschreibt man Wortformen explizit durch Nennung der Klasse, so ergibt sich z. B. folgendes Bild:

Eisbrecher	*lockte*	*Wale*	*mit*	*klassischer*	*Musik.*
.	.	.	.	.	.
.	.	.	.	.	.
SUBST	VERB	SUBST	PRÄP	ADJ	SUBST

Das Beispiel zeigt, daß Segmentierung und Klassifikation nicht voneinander zu trennen sind. Denn um eine Klasse zuordnen zu können, muß zunächst das Segment ermittelt worden sein. Ob aber eine Zeichenfolge in einem Korpus als Textsegment aufgefaßt werden kann, hängt davon ab, ob es einer Klasse zugeordnet werden kann, d. h., ob es

in Texten eine bestimmte Funktion ausübt. Die Funktion eines Segment wird jedoch nicht von außersprachlichen Kriterien bestimmt, sondern aus der Sprache selbst, nämlich durch die Beziehungen, in denen das betreffende Segment im Sprachsystem anzutreffen ist.

Man kann die potentiellen Funktionen einer Einheit durch Angabe von Klassen in einem Lexikon verzeichnen, wie es z. B. in jedem konventionellen Wörterbuch durch Angabe von Wort- und Flexionsklasse geschieht.

5. Literatur (in Auswahl)

R. Geppert/M. H. Kuhn/H. Ney 1983 · Harris 1951 · Karttunen/Zwicky 1985 · K. Kohler 1977 · St. Marcus 1983 a · St. Marcus 1983 b · H. Ney 1983 · H. Weber 1973 · Wothke 1983.

Winfried Lenders, Bonn
(Bundesrepublik Deutschland)