Erkennung und Beschreibung des prosodischen Fokus

Inauguraldissertation
zur
Erlangung der Doktorwürde

vorgelegt der
Philosophischen Fakultät
der
Rheinischen Friedrich-Wilhelms-
Universität
zu Bonn

von
Anja Elsner
aus
Bremen

Bonn
2000

Gedruckt mit Genehmigung der
Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Berichterstatter: Professor Dr. W. Hess

2. Berichterstatter: Professor Dr. W. Lenders

Tag der mündlichen Prüfung: 8. Dezember 1999

Inhaltsverzeichnis

1 Einleitung **1**

1.1 Kommunikation und Information 1

1.1.1 Verschiedene Klassifizierungen von Information 1

1.1.2 Entwurf eines Informationsmodells 3

1.2 Ziele dieser Arbeit 4

1.3 Gliederung 5

2 Prosodie in der Sprachproduktion **7**

2.1 Spracherzeugung 7

2.1.1 Sprachanregung (Phonation) 8

2.1.2 Sprachformung (Artikulation) 10

2.2 Prosodische Parameter 10

2.3 Eigenschaften von akustischen Parametern 11

2.3.1 Produktion und Messung 11

2.3.2 Eigenschaften im globalen Verlauf 13

2.4 Mikroprosodie 14

2.4.1 Intrinsische Eigenschaften 14

2.4.2 Koartikulatorische Zusammenhänge 15

2.5 Makroprosodie 16

2.5.1 Hervorhebung und Prominenz 16

2.5.2 Gliederung und Markierung von Äußerungen 17

2.6 Linguistische Funktion und prosodische Korrelate 19

2.6.1 Akzentuierung 19

2.6.2 Phrasierung 20

2.6.3 Satzmodus . . . 20
2.7 Para- und extralinguistische Parameter . . . 21
2.8 Zusammenfassung . . . 21

3 Fokusbegriff **23**
3.1 Linguistische Konzepte . . . 23
3.1.1 Funktion und Satzstruktur . . . 23
3.1.2 Fokus als semantisch-pragmatischer Begriff . . . 24
3.1.3 Akzent und Syntax . . . 25
3.1.4 Fokus und Informationsstruktur . . . 26
3.1.5 Unterscheidung zwischen 'normalem Akzent' und Kontrastakzent . 27
3.2 Automatische Zuweisung von Akzenten . . . 29
3.3 Fokusmarkierung im Sprachvergleich . . . 30
3.3.1 Deakzentuierung . . . 30
3.3.2 Optionaler Gebrauch von verschiedenen Satzakzenttypen . . . 31
3.4 Fokusrealisierung von Konzepten . . . 32
3.4.1 Bekannte und neue Information . . . 32
3.4.2 Kontrastfokus . . . 33
3.4.3 Markierung von Fokusakzenten in der Synthese . . . 34
3.5 Fokusrealisierung und Satzmodus/Satzposition . . . 35
3.5.1 Enger Fokus in unterschiedlicher Satzposition . . . 36
3.5.2 Position des Fokus in Aussagen und Fragen . . . 37
3.5.3 Enger und weiter Fokus, einfacher und doppelter Fokus . . . 37
3.5.4 Fokustypen in verschiedenen Satzstrukturen . . . 38
3.6 Automatische Erkennung . . . 39
3.6.1 Das MAFID-System . . . 39
3.6.2 Das Projekt Modus-Fokus-Intonation . . . 41
3.7 Zusammenfassung . . . 42
3.7.1 Definition des prosodischen Fokus . . . 42
3.7.2 Eigenschaften von Fokusakzenten . . . 43
3.7.3 Kontrastfokus . . . 44

4 Anwendungen in der automatischen Sprachverarbeitung **45**
4.1 Das Projekt Verbmobil . . . 45
4.2 Das Architekturteilprojekt Intarc . . . 47
4.2.1 Anforderungen an das System . . . 47
4.2.2 Module des Systems . . . 48
4.3 Prosodische Information in der automatischen Sprachverarbeitung . . . 50
4.3.1 Funktionen der Prosodie in Spontansprache . . . 50
4.3.2 Verwendung der Prosodie in Verbmobil und Intarc . . . 51

5 Sprachdaten **54**
5.1 Aufnahme der Sprachdaten . . . 54
5.2 Aufbau der Verbmobildialoge . . . 54
5.3 Eigenschaften von Spontansprache . . . 56
5.4 Prosodische Etikettierung . . . 56
5.5 Testdaten für die Fokuserkennung . . . 58

6 Verfahren zur Fokuserkennung **60**
6.1 Idee des Verfahrens . . . 60
6.2 Bearbeitung der Grundfrequenzkontur . . . 61
6.2.1 Störungen des F_0-Verlaufs . . . 61
6.2.2 Beschreibung des Verfahrens zur Grundfrequenznachverarbeitung . . . 62
6.3 Berechnung der Erkennungsraten . . . 64
6.4 Berechnung einer Referenzgerade . . . 65
6.4.1 Problemstellung . . . 65
6.4.2 Referenzgerade aus F_0-Minima und F_0-Maxima . . . 66
6.4.3 Auswertung der Parametereinstellungen . . . 67
6.5 Bestimmung der Fokusakzente . . . 68
6.5.1 Nutzung von Referenzgerade und F_0-Maxima . . . 69
6.5.2 Korrelationen der Referenzgerade . . . 69
6.6 Ergebnisse . . . 71
6.7 Satzmodus . . . 72
6.8 Klassifizierung von Fokusakzenten . . . 74

7 Weitere Untersuchungen **77**
7.1 Energie . . . 77
7.1.1 Experiment . . . 78
7.1.2 Auswertung für Silben und Wörter . . . 78
7.1.3 Auswertung für Vokale . . . 79
7.1.4 Diskussion . . . 79
7.2 Phrasengrenzen . . . 81
7.3 Emphase und Kontrast . . . 84
7.3.1 Verschiedene Messungen . . . 84
7.3.2 Diskussion . . . 88
7.4 Sprecherabhängigkeit . . . 88
7.4.1 Untersuchungen für alle Sprecher . . . 89
7.4.2 Dialogverhalten eines Sprechers im Vergleich . . . 90
7.5 Perzeption . . . 92
7.5.1 Experimente . . . 92
7.5.2 Unterschiede zwischen Prädiktion und Perzeption . . . 93
7.5.3 Zuverlässigkeit der Etikettierung . . . 93
7.5.4 Akustische Auffälligkeit und Hörerurteile . . . 94

8 Anwendung der Fokusinformation für linguistische Module **96**
8.1 Grundsätzliche Überlegungen . . . 96
8.2 Transfer in Verbmobil und Intarc . . . 97
8.2.1 Übersetzungsstrategien . . . 97
8.2.2 Transfer in Intarc . . . 99
8.2.3 Flache Übersetzung in Verbmobil . . . 100
8.3 Fokus und Dialogaktbasierter Transfer in Intarc . . . 100
8.3.1 Experimente mit Transfer . . . 100
8.3.2 Anwendungsbeispiele . . . 101
8.4 Fokus und Semantik . . . 104
8.4.1 Desambiguierung von Diskurspartikeln . . . 105
8.4.2 Prosodische Information zum Aufbau der semantischen Repräsentation in Verbmobil . . . 106
8.4.3 Einbindung und Korrektur von Prosodieinformation in Intarc . . . 106

9 Abschließende Diskussion **109**

9.1 Zusatzinformationen für die Fokuserkennung 109

9.2 Beitrag der Fokusinformation zur Unterstützung anderer Module 111

9.2.1 Worterkennung 111

9.2.2 Semantik 112

9.2.3 Transfer 112

9.2.4 Synthese 113

9.3 Zusammenfassung und Ausblick 114

Literaturverzeichnis **115**

1. Einleitung

Die vorliegende Arbeit bewegt sich im Rahmen der akustischen Phonetik, unter besonderer Berücksichtigung der Prosodie. Grundlage von prosodischen Untersuchungen ist die Information über akustische Parameter, die prosodische Phänomene markieren. Dies sind in erster Linie Sprachgrundfrequenz, Intensität und Dauer.

Ziel dieser Arbeit ist die Untersuchung von prosodischen Markierungen von Äußerungen, die eine linguistische Funktion signalisieren. Einerseits wird versucht, diese prosodische Markierung akustisch zu detektieren (Kapitel 6 und 7), andererseits wird die Interpretation der linguistischen Funktion betrachtet (Kapitel 8). Die Ergebnisse wurden in ein automatisches sprachverstehendes System integriert und getestet.

Im folgenden werden zunächst die grundlegenden Begriffe Kommunikation und Information diskutiert. Es wird ein Informationsmodell entworfen, das auf die Inhalte dieser Arbeit abgestimmt ist. Im Anschluß werden entsprechende Zielsetzungen definiert. Abschließend wird eine kurze Übersicht über die Gliederung dieser Arbeit gegeben.

1.1 Kommunikation und Information

Nach den Erkenntnissen der Sprechakttheorie, die wesentlich von Austin (1962) und Searle (1969) geprägt wurde, werden nicht einzelne Wörter oder Sätze als Grundelemente der menschlichen Kommunikation angesehen, sondern Sprechhandlungen, die durch ihre Äußerung vollzogen werden, die sog. *illokutiven Akte*. Illokutive Akte bezeichnen die Redeabsicht oder Intention eines Sprechers, eine kommunikative Wirkung auf Hörer auszuüben. Der Ausdruck der Intention eines Sprechers wird im allgemeinen durch die prosodische Markierung wesentlich unterstützt. Der Inhalt bzw. die Bedeutung dessen, was ausgesagt wird, wird als *Proposition* bezeichnet, die intendierte Wirkung des Sprechakts auf den Hörer wird von Searle (1969) als *Perlokution* definiert (siehe auch Bußmann, 1990).

1.1.1 Verschiedene Klassifizierungen von Information

Laver (1994) definiert 3 Arten von Information: semantisch, evidentell und regulativ. Die *semantische Information* meint die direkte Bedeutung einer Aussage, also den propositionalen Gehalt des kommunikativen Akts. Darunter fällt auch die indirekte Information, die eine pragmatische Bedeutung hat ("es ist kalt hier" - Aufforderung, das Fenster zu schlie-

ßen). Die *evidentelle Information* kann der Hörer aus der Sprechsituation direkt ableiten. Sie umfaßt Sprecherattribute wie Geschlecht, Alter, soziale Herkunft, Gesundheitszustand, Persönlichkeit, etc. Die *regulative Information* bezieht sich auf die Dialogführung an sich. Die Sprecher geben sich während des Dialogs Signale, wie lange sie jeweils sprechen wollen ('floor holding') und wann sie dem Dialogpartner das Wort überlassen wollen. Abgesehen von entsprechenden nonverbalen Gesten wird dies im wesentlichen durch die Prosodie übermittelt.

Die Übertragung von Information kann man nach Lyons (1977) von zwei Standpunkten aus betrachten: dem des Sprechers (kommunikativ) und dem des Hörers (informativ). Die *kommunikative Form* geht von der Intention des Sprechers aus, dem Hörer etwas mitzuteilen, was ihm vorher nicht bekannt war. Die *informative Form* geht vom Hörer aus, d. h., er bekommt vom Sprecher eine neue Information, egal ob dies vom Sprecher beabsichtigt war oder nicht. Im wesentlichen ist linguistische Aktivität kommunikativ und informativ zugleich.

Lyons (1977) unterscheidet außerdem zwischen *Signalinformation* und *semantischer Information*. Das Sprachsignal ist die physikalische Codierung der Mitteilung, die bestimmte akustische Eigenschaften hat. Die semantische Information muß aus dieser akustischen Information interpretiert werden. Hörer, die ein Sprachsignal wahrnehmen, können dieses mit Hilfe ihrer Sprachkompetenz und ihres pragmatischen Situationswissens dekodieren. Sie haben bezüglich der Mitteilung eines Sprechers bestimmte Erwartungen, d. h., sie kennen die Wahrscheinlichkeiten von Äußerungen, die in einem bestimmten Kontext auftreten. Wenn außerdem der Sprecher und seine Gewohnheiten gut bekannt sind, können diese Erwartungen noch weiter eingeschränkt werden.

Nach Lyons (1977) dient Sprache nicht nur der Kommunikation von tatsächlicher Information. Sie dient darüber hinaus dem Errichten und Aufrechterhalten von sozialen Beziehungen und dem persönlichen Ausdruck der Sprecher. Diese 3 Funktionen können mit den Begriffen *deskriptiv* (Ausdruck einer Proposition), *sozial* (Beziehung der Sprecher untereinander) und *expressiv* (charakteristischer Ausdruck eines Sprechers) bezeichnet werden.

Eine ähnliche Dreiteilung findet sich bereits im Sprachmodell von Bühler (1934). Bühler definiert die Funktionen von Sprache mit den Begriffen Darstellung, Ausdruck und Appell. Die *Darstellung* entspricht hier der deskriptiven Funktion (es wird über einen Gegenstand oder eine Person gesprochen), der *Ausdruck* der expressiven Funktion (der Sprecher drückt etwas über sich selbst aus), und der *Appell* hat eine vokative Funktion (eine andere Person wird direkt angesprochen). In Assoziation zu diesen 3 Funktionen definiert Bühler die Begriffe *Symbol* (Bedeutung eines Sachverhalts), *Symptom* (Vorstellungen eines Sprechers) und *Signal* (auf einen Empfänger bezogen).

Eine wesentliche Modifikation wurde von Jakobson (1960) vorgenommen: die vokative Funktion heißt bei ihm *konativ*. Es bleibt zwar die Implikation, daß ein Hörer direkt angesprochen wird (eben vokativ), doch es überwiegt hier der Aspekt der *Intention* des Sprechers, der eine Erfüllung seiner Bedürfnisse erwartet. Hier ist also eher eine *instrumentelle Funktion* gemeint, d. h., Sprache wird gebraucht, um eine praktische Wirkung hervorzubringen.

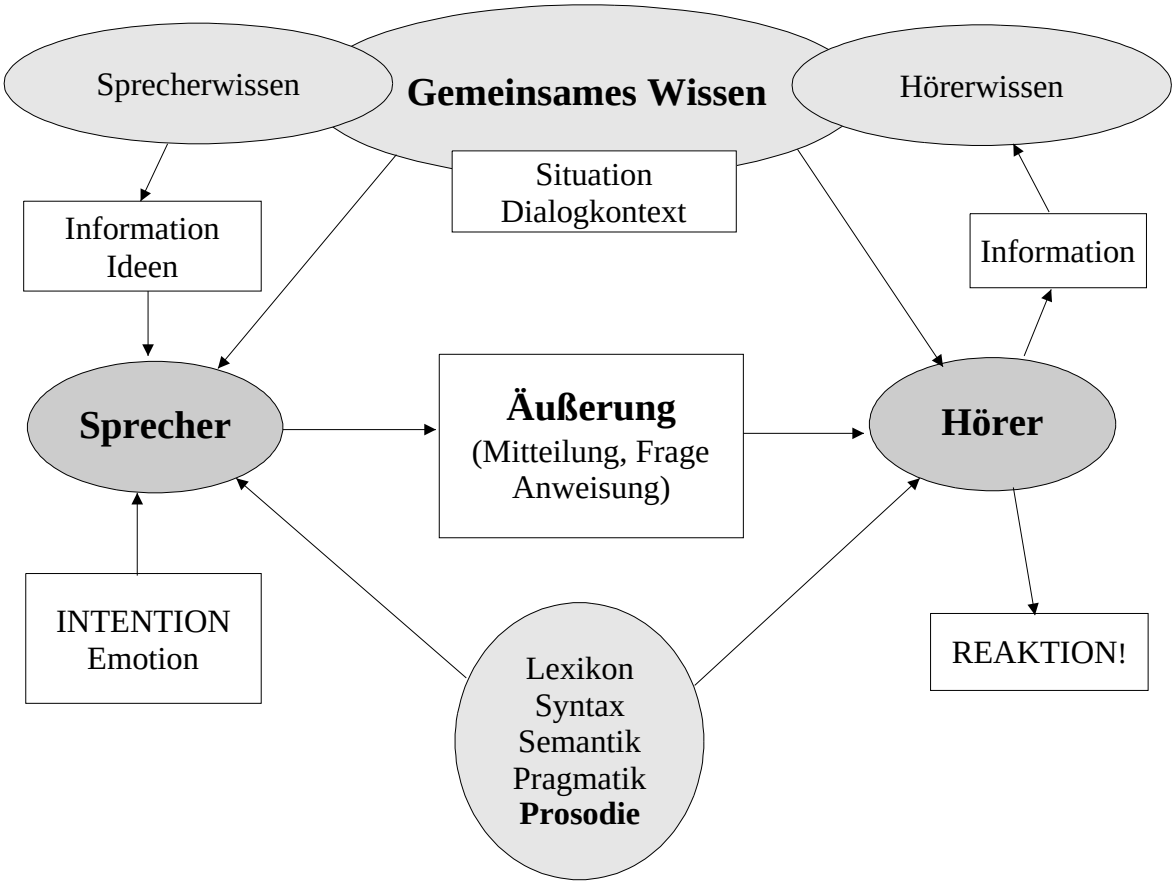

Abbildung 1.1: *Informationsmodell*

1.1.2 Entwurf eines Informationsmodells

Nach diesem Exkurs zu verschiedenen Klassifizierungen von Information sollen nun die hier behandelten Arten von Information aufgezeigt werden. Nach der Klassifikation von Laver (1994) beschäftigt sich diese Arbeit hauptsächlich mit *semantischer Information*, also dem eigentlichen Inhalt von Äußerungen. Bei der Beziehung Sprecher-Hörer nach Lyons (1977) liegt hier der Schwerpunkt auf der Intention des Sprechers, also auf der kommunikativen Form. Die Unterscheidung zwischen *Signalinformation* und *semantischer Information* ist wesentlich für diese Arbeit, da hier ja gerade versucht werden soll, aus akustischen Informationen die wesentlichen Inhalte für einen Hörer zu detektieren.

Im folgenden wird ein Informationsmodell entwickelt, das die für diese Arbeit relevanten Aspekte zusammenfaßt (siehe Abbildung 1.1): Zwischen Teilnehmern eines Dialoges herrscht ein kontinuierlicher Informationsfluß. Es soll hier von einem kooperativen Dialog ausgegangen werden, wo beide Partner ein gemeinsames Ziel verfolgen. Sowohl der Sprecher hat die Absicht, seinem Partner Information erfolgreich zu vermitteln, wie es ebenso im Interesse des Hörers ist, die gewünschte Information auch zu erhalten und zu verstehen.

Eine Information ist im Sprecher mental repräsentiert. Er hat den Wunsch, zu kommuni-

zieren und Informationen oder seine Ideen dazu mitzuteilen. Mit der Mitteilung verfolgt er bestimmte Intentionen. Er will den Hörer informieren, zusätzlich erwartet er oft auch eine angemessene Reaktion. Daraus ergibt sich die Struktur der realisierten Mitteilung (akustisch, prosodisch, linguistisch). Die Kommunikationsabsicht des Sprechers beeinflußt Wahl und Anordnung der Wörter (Lexikon, Syntax, Semantik) und ist abhängig vom jeweiligen Situationskontext (Pragmatik). Zusätzlich wird die Mitteilung noch durch die Prosodie unterstützt; Intention, aber vor allem auch Emotion und Einstellung des Sprechers werden wesentlich durch die Prosodie markiert.

Die Äußerung wird vom Hörer perzipiert. Mit Hilfe bereits vorhandenen Wissens sollte er idealerweise die neue Information des Sprechers im Situationskontext einordnen können. Mit seinem linguistischen Wissen kann er auch den Inhalt der Äußerung analysieren. Vor allem aber sollte er die Intention des Sprechers korrekt interpretieren; diese kann im wesentlichen aus der Prosodie erschlossen werden. Dies kann den Hörer möglicherweise zu einer Reaktion veranlassen (z. B. seine Meinung zu äußern, eine Frage zu beantworten oder eine Handlung auszuführen). Nach der Perzeption wird die Information wieder in eine mentale Repräsentation, diesmal die des Hörers, überführt (siehe Abbildung 1.1).

1.2 Ziele dieser Arbeit

Ein wesentlicher Punkt zum Verstehen eines Dialoges ist es, die Intention der Sprecher und damit die kommunikativen Akte (Sprechakte) zu erkennen. Es stellt sich zum einen die Frage, ob und ggf. wie diese Intention akustisch markiert wird. Zum anderen muß untersucht werden, welche Intention sich direkt aus der akustischen Realisierung ableiten läßt und welche nur mit Hilfe einer linguistischen Analyse interpretiert werden kann.

In dieser Arbeit geht es darum, die wesentlichen Teile einer gesprochenen Äußerung herauszufiltern. Die Arbeitshypothese geht davon aus, daß wesentliche Information auffälliger im Sprachstrom kodiert ist als unwesentliche. Diese wesentliche und zugleich akustisch auffällige Information wird in dieser Arbeit als *prosodischer Fokus* definiert.

Ziel ist zum einen, den prosodischen Fokus auf mehreren Ebenen zu beschreiben. Zunächst gilt es, die verschiedenen linguistischen und akustischen Definitionen zu untersuchen und möglichst eine Quintessenz daraus zu ziehen. Welche Arten von prosodischem Fokus gibt es, wie äußern sie sich?

Ein zweites wesentliches Ziel besteht darin, den prosodischen Fokus auf akustischer Ebene zu erkennen. Es wird davon ausgegangen, daß akustische Parameter darüber Aufschluß geben, welche Teile einer Äußerung auch perzeptiv stärker hervortreten. Diese Parameter äußern sich unterschiedlich in ihrem Beitrag zur Erkennung; die Sprachgrundfrequenz liefert hier offensichtlich die wesentliche Information.

Drittes Ziel ist die Anwendung der Fokuserkennung in einem System zur Erkennung, Übersetzung und Synthese von spontansprachlichen Dialogen (siehe Kapitel 4). Die Fokusinformation soll anderen linguistischen Erkennungsmodulen Hilfen zur weiteren Verarbeitung geben. Als 'Abnehmermodule' kommen dafür im wesentlichen Semantik und Transfer, aber auch Dialog und Synthese in Frage.

1.3 Gliederung

Im *zweiten Kapitel* wird zunächst ein Überblick zur Prosodie in der Sprachproduktion gegeben. Das Gebiet der Prosodie verfügt über eine große Begriffsvielfalt; daraus folgt leider auch, daß viele Begriffe nicht einheitlich definiert werden und daß auch und gerade bei so zentralen Begriffen wie Intonation und Betonung sehr genau beschrieben werden muß, was darunter zu verstehen ist. Die verwendeten Definitionen für prosodische Parameter und die entsprechenden akustischen Parameter werden vorgestellt, außerdem werden ihre verschiedenen Eigenschaften in bezug auf Produktion und Messung beschrieben.

Weiterhin wird die Mikroprosodie der akustischen prosodischen Parameter erläutert. Aufgrund der Rückkopplung zwischen Vokaltrakt und Artikulation kommt es zu bestimmten lokalen intrinsischen und koartikulatorischen Phänomenen. Auf globaler Ebene bewirken die prosodischen Parameter eine sog. Makroprosodie. Dies führt zu Hervorhebungen, Gliederung und Markierungen einer Äußerung. Diese können jeweils linguistisch als Akzentuierung, Phrasierung und Satzmodus interpretiert werden. Abschließend werden para- und extralinguistische Parameter erwähnt; die Zusammenhänge werden in einem Schaubild zusammengefaßt.

Im *dritten Kapitel* erfolgt eine intensive Analyse der Fokusliteratur. Das weite Feld der semantischen Definitionen wird nur am Rande gestreift, es werden lediglich einige wesentliche linguistische Konzepte vorgestellt. Einen Schwerpunkt bildet die Darstellung der phonetischen Untersuchungen, die sich mit den akustischen Korrelaten dieser semantischen Konzepte auseinandergesetzt haben. Weiterhin werden einige Arbeiten vorgestellt, die eine automatische Erkennung von Fokusakzenten bereits implementiert haben. Als Fazit wird eine eigene Definition erstellt, und die wesentlichen Eigenschaften von Fokusakzenten werden noch einmal zusammengefaßt.

Im *vierten Kapitel* wird eine Einführung in das Projekt Verbmobil gegeben, das den Rahmen für diese Arbeit bildet. Im Teilprojekt Intarc sollte eine spezielle Architektur für die Sprachverarbeitung getestet werden. Es wurde eine rege Kommunikation der Erkennungsmodule angestrebt, um die Erkennungsleistung aller Module (sowohl Schnelligkeit als auch Zuverlässigkeit) deutlich zu steigern.

Anschließend wird die Anwendung der Prosodie in der automatischen Spracherkennung diskutiert. Es wird begründet, warum die Prosodie eine unabhängige Hilfe für andere Erkennungsebenen wie Syntax und Semantik sein kann. Aktuelle Anwendungen in Verbmobil und Intarc werden vorgestellt. Die verwendeten Sprachdaten und ihre Etikettierung in Verbmobil werden im *fünften Kapitel* beschrieben.

Das *sechste Kapitel* beschreibt das entwickelte Verfahren zur Fokuserkennung. Das Verfahren versucht, die Fokuserkennung mit Hilfe einer globalen Verlaufsbeschreibung der Sprachäußerung zu lösen. Dabei wird zunächst nur die Sprachgrundfrequenz F_0 in Betracht gezogen. Die Arbeitshypothese war, daß der Abfall der F_0-Kontur durch den Fokus gesteuert wird: Nach dem Äußern des Fokus wird der F_0-Abfall signifikant steiler. Das regelbasierte Verfahren sucht daher anhand einer globalen Referenzgerade (die aus der F_0-Kontur berechnet wurde) entsprechende Zeitpunkte mit signifikant steilem Abfall.

Die ersten Erkennungsraten zeigten recht gute Ergebnisse, mit Verwendung von Zusatzinformationen wie Satzmodus konnten noch weitere Steigerungen erzielt werden. Abschließend folgt eine Klassifizierung der Fokusakzente nach akustischen Kriterien, um die unterschiedlichen Erkennungserfolge gezielt zu untersuchen

Im *siebten Kapitel* werden weitere Experimente dargestellt, die die Verwendung von zusätzlicher Information für die Fokuserkennung prüfen sollten. Im Hinblick auf die Energie wurde untersucht, ob entsprechende Energieinformationen für die Fokuserkennung nutzbar sind. Die Energieunterschiede zwischen fokussierten und nicht fokussierten Bereichen waren allerdings nur marginal und erwiesen sich als sehr unzuverlässig. Daher wurde dieser Ansatz nicht weiter verfolgt.

Vielversprechend waren dagegen Experimente mit der Einbeziehung von Phrasengrenzeninformation. Zunächst wurde mit manuell etikettierten Phrasengrenzen experimentiert, es ergab sich eine deutliche Reduktion der Fehlschläge. Untersuchungen mit 'detektierten' Phrasengrenzen aus einem Prosodiedetektor ergaben bisher nur geringe Steigerungen. Verbesserungen sind aber zu erwarten, wenn das Vorkommen von Doppelfokus angemessen berücksichtigt wird.

Des weiteren wurden die Eigenschaften von unterschiedlich akustisch auffälligen Fokusakzenten untersucht. Dazu wurde ihre Position in der Äußerung (Abstand zur nächsten Phrasengrenze), ihre Dauer und die Höhe des Grundfrequenzmaximums betrachtet. Für jede untersuchte Fokusakzentkategorie ergaben sich typische Formen.

Es wurden ebenfalls Untersuchungen zur Sprecherabhängigkeit durchgeführt. Eine wesentliche Frage war dabei, ob sich akustische Kriterien finden lassen, anhand derer vorauszusagen ist, ob sich Sprecher gut oder schlecht für die Erkennung eignen. Abschließende Untersuchungen zur Perzeption von Fokusakzenten zeigen deutliche Übereinstimmungen in der Etikettierung bei mehreren Versuchspersonen. Dies konnte zum einen die Etikettierung der Autorin bestätigen, zum anderen konnte eine hohe Übereinstimmung auch direkt auf akustische Parameter abgebildet werden.

Im *achten Kapitel* wird die Anwendung der Fokusakzentinformation diskutiert. Es wird angenommen, daß die akustischen Fokusmarkierungen eine Abbildung der linguistischen Konzepte darstellen und somit für die Verarbeitung in den linguistischen Modulen nützlich sind. Zunächst werden die Übersetzungsstrategien in Verbmobil und Intarc beschrieben; diese basieren im wesentlichen auf Sprechakten. Anschließend wird die Verwendung der detektierten Fokusakzente im Transfermodul detailliert vorgestellt. Weiterhin wird die Nutzung von prosodischer Information in den semantischen Modulen von Verbmobil und Intarc betrachtet.

Zum Abschluß werden im *neunten Kapitel* die wesentlichen Ergebnisse der Arbeit noch einmal kurz diskutiert und zusammengefaßt. Zum einen werden die Möglichkeiten zur Verbesserung der Fokusakzenterkennung beschrieben, zum anderen wird der Beitrag der Fokusinformation zur Unterstützung der anderen Module diskutiert. Abschließend folgt ein kurzer Ausblick.

2. Prosodie in der Sprachproduktion

In diesem Kapitel sollen vor allem die sprachlichen Vorgänge dargestellt werden, die einen Bezug zur Prosodie haben. Im wesentlichen werden aber nur die Bereiche skizziert, die für das Verständnis dieser Arbeit von Bedeutung sind. Dies beginnt bei der Spracherzeugung an sich, mit Schwerpunkt auf der Sprachanregung. Anschließend wird geklärt, was in dieser Arbeit unter prosodischer Information verstanden wird.

Wichtig ist die Unterscheidung zwischen objektiv meßbaren und subjektiv empfundenen Erscheinungen. Die prosodischen Parameter beziehen sich auf *perzeptive* Phänomene, es gibt keine einheitliche Korrespondenz zwischen ihnen und akustischen, meßbaren Parametern, sie haben aber jeweils ein akustisches Korrelat. Die akustischen Parameter, die für die Prosodie eine Rolle spielen, werden ausführlich diskutiert. Anschließend werden lokale (Mikroprosodie) und globale Aspekte (Makroprosodie) der prosodischen Parameter betrachtet.

Abschließend werden einige linguistische Funktionen der Prosodie dargestellt. Linguistische Phänomene werden durch *qualitative Änderungen* der prosodischen Parameter erzeugt. Es folgen ergänzende Anmerkungen über para- und extralinguistische Parameter. Die dargestellten Beziehungen zwischen Prosodie, Akustik und Linguistik werden in einem Schaubild zusammengefaßt.

2.1 Spracherzeugung

Eine mögliche Einteilung der Spracherzeugung ist die in Sprachanregung (Phonation) und Sprachformung (Artikulation). Unter Phonation versteht man nach Laver (1994) *"the use of the laryngeal system, with the help of an airstream provided by the respiratory system, to generate an audible source of acoustic energy which can then be modified by the articulatory actions of the rest of the vocal apparatus (S. 184)."* Die Phonation wird also im wesentlichen durch die Atmungsorgane des Brustraums (Lunge, Luftröhre etc.) gesteuert, während die Artikulation im Vokaltrakt (Mund-, Nasen- und Rachenraum, Kehlkopf) stattfindet. Diese strenge Trennung gilt genau genommen aber nur für Vokale, Liquide und Gleitlaute. Für die Produktion eines Frikativs beispielsweise findet auch im Vokaltrakt eine Anregung statt, nämlich die Erzeugung einer geräuschhaften Luftstromverwirbelung an einer Engstelle.

2.1.1 Sprachanregung (Phonation)

Bei der Phonation ist es Aufgabe des subglottalen Atmungssystems, einen Luftstrom mit relativ konstantem Luftdruck zur Verfügung zu stellen. Im Gegensatz zur normalen Atmung geschieht das Ausatmen bei der Phonationsatmung nicht passiv, sondern der Luftdruck muß aktiv erzeugt werden (Kohler, 1995, S. 42). Der Druckunterschied zwischen subglottalem Bereich und atmosphärischem Druck in der Außenluft ist ein wesentlicher Faktor, der das Öffnen und Schließen der Glottis als einen zyklischen Prozeß in Gang hält.

Ebenfalls von Bedeutung ist die elastische Muskulatur der Stimmlippen; nach jeder Veränderung der neutralen Position (dies ist bei normaler Atmung eine mittlere Öffnung der Glottis) sind Rückstellkräfte der Muskulatur bemüht, den alten Zustand wiederherzustellen. Die Druckkräfte des beim Ausatmen erzeugten Luftstroms arbeiten gegen die neutrale Position: Aerostatische Kräfte durch *subglottalen Überdruck* bewirken eine weite Öffnung der Glottis, während aerodynamische Kräfte durch *subglottalen Unterdruck* die Glottis zum Schließen veranlassen (Lieberman, 1967).

Eine weithin anerkannte Theorie für die *stimmhafte* Sprachanregung ist die *myoelastisch-aerodynamische Theorie der Phonation* (van den Berg, 1958). Die Anregung geschieht nach dieser Theorie durch den Druck der Atmungsluft mit Unterstützung der elastischen Stimmlippenmuskulatur im Bereich der Glottis (Stimmritze).

Ein Anregungszyklus läßt sich folgendermaßen beschreiben: Die Glottis wird zunächst als geschlossen angenommen. Beim Ausatmen wird ein Luftstrom erzeugt, der beim Passieren des Kehlkopfes durch den Glottisverschluß am Weiterströmen gehindert wird. Dadurch entsteht ein Überdruck; durch ständig steigenden Druck wird schließlich der Glottisverschluß gelöst. Die Luft strömt nun schnell durch die Glottisenge, um einen Druckausgleich zu erreichen. Durch die hohe Strömungsgeschwindigkeit der Luft entsteht ein Unterdruck, in diesem Zusammenhang auch als *Bernoulli-Effekt* bekannt. Die dadurch hervorgerufene Sogwirkung, unterstützt durch die elastische Stimmlippenmuskulatur, bewirkt, daß die Stimmlippen wieder zusammengepreßt werden.

Die Fortsetzung dieses Zyklus läßt ein mehr oder weniger starkes Vibrieren der Stimmlippen entstehen, welches die eigentliche Phonation ausmacht. Das Signal an sich entsteht dabei durch die periodische Unterbrechung des Luftstroms. Die Zahl der Stimmlippenschwingungen pro Zeiteinheit wird durch den akustischen Parameter Grundfrequenz (F_0) beschrieben.

Der Begriff Phonation wird gelegentlich auf die stimmhafte Sprachanregung beschränkt, viele Autoren fassen den Begriff aber etwas weiter und beziehen auch die stimmlose Anregung mit ein (z. B. Laver, 1994). Weitere Formen der Anregung entstehen auch durch *Art* und *Ort* der Glottisverengung. Zusätzlich spielt die Stimmlippenspannung und die Höhe des subglottalen Luftdrucks eine Rolle. Eine ausführliche Beschreibung findet sich bei Catford (1988) und Laver (1994); die wichtigsten Phonationsarten sollen im folgenden kurz vorgestellt werden.

Zu den primär stimmlosen Formen zählt Laver (1994) *Nil phonation, Breath phonation (Behauchung)* und *Whisper phonation (Flüstern)*. Bei *Nil phonation* und *Behauchung* ist die Glottis weit geöffnet (siehe Abbildung 2.1 a), während sie beim *Flüstern* stark verengt

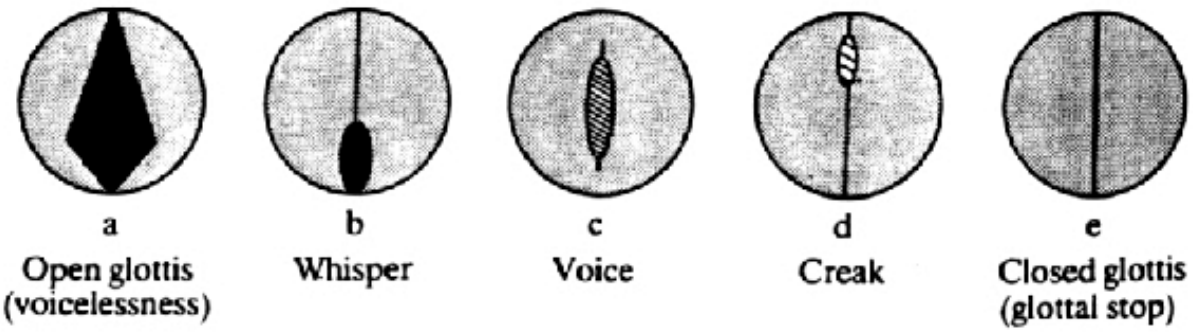

Abbildung 2.1: *Blick auf die Glottis bei verschiedenen Phonationsarten [aus Catford, 1988]*

ist (Abbildung 2.1 b). Diese drei Phonationsarten unterscheiden sich außerdem durch die Stärke des durchgeführten Luftstroms: Bei der *Nil phonation* ist die Luftstromrate sehr niedrig und ohne Turbulenzen. Der akustische Input für den Vokaltrakt ist praktisch Null. Bei der *Behauchung* gibt es einen leicht turbulenten Luftstrom (die Artikulation wirkt stark beatmet/behaucht), während beim *Flüstern* der Luftstrom stark turbulent ist. Dadurch entsteht ein geräuschhaftes, starkes Hauchen. Laver (1994) ordnet der Nil phonation auch den Zustand der geschlossenen Glottis zu; hier wird ebenfalls keine akustische Energie in den Vokaltrakt übertragen (siehe Abbildung 2.1 e).

Bei der stimmhaften Anregung *Voice* (Abbildung 2.1 c) wird der Luftstrom pulsartig durch die Glottis geführt, es entsteht eine periodische Vibration der Stimmlippen. Dabei kann man sog. *Register* wie Bruststimme (Modalregister) und Kopfstimme (Falsett) unterscheiden. Bei der Kopfstimme werden die Stimmlippen stärker gespannt, die Atmungskräfte werden verstärkt, und die Bernoullikraft wird kleiner. Durch die Spannung werden die Stimmlippen 'dünner', so daß höhere Tonlagen mit der Kopfstimme erreicht werden können.

Bei der *Creak phonation* (Abbildung 2.1 d) gibt es ebenfalls eine pulshafte Anregung, aber die Luftstromimpulse sind von niedriger Frequenz und sehr unregelmäßig. Außerdem vibriert nur ein kleiner Bereich der Stimmlippen. Die Öffnung der Glottis geschieht kurz und schnell, während die Schließungsphase relativ lange andauert. Es werden also weniger Anregungspulse produziert, was eine starke Erniedrigung der Grundfrequenz zur Folge hat; sie bildet eine *Subharmonische* der normalen Grundfrequenz eines Sprechers. Ein kurzfristiger, unbewußter Übergang in diese Phonationsart wird von Lehiste (1970) auch als *Laryngalisierung* bezeichnet.

Als vereinfachte Version der stimmhaften Phonationsarten hat sich vielfach der Vorschlag von Hollien (1974) durchgesetzt. Er schlägt die Zusammenfassung in drei Register vor, die durch ihren Tonumfang voneinander abgegrenzt sind: Die niedrigste Lage nimmt das *pulse register* ein (andere Bezeichnungen sind Strohbaßregister, vocal fry, creaky, Laryngalisierung). Für die mittlere, normale Stimmlage wird das *modal register* (Bruststimme bei normaler Phonation) verwendet, die höchsten Frequenzen lassen sich mit dem *loft register* (Falsett, Kopfstimme) erzielen.

Einige dieser Phonationsarten können auch kombiniert werden (eine Übersicht findet sich

bei Laver, 1994, S. 199). Die eigentlich stimmlosen Phonationsarten *Behauchung* und *Flüstern* können mit Stimmton produziert werden; dazu werden die Stimmlippen etwas näher zusammengeführt, so daß eine leichte Stimmlippenvibration entstehen kann. Linguistisch bedeutsam sind in erster Linie die Kombinationen "whispery voice" und "creaky voice". Erstere ist im Hindi und Urdu kontrastiv markiert, letztere findet sich als phonologische Differenzierung im Dänischen. Phonationsarten können außerdem pathologisch bedingt sein oder paralinguistisch eingesetzt werden (siehe Abschnitt 2.7).

2.1.2 Sprachformung (Artikulation)

Nach der Anregung im subglottalen Bereich ist es Aufgabe der supraglottalen Elemente, das Anregungssignal weiter zu formen. Der Vokaltrakt kann dabei als zeitlich variables akustisches Rohr gesehen werden, das durch Artikulationsorgane wie Zunge, Lippen und Velum geformt wird. Durch Heben und Senken des Velums beispielsweise kann der Nasenraum ab- und zugeschaltet werden. Wesentlich ist aber die Gestaltung des Mund- und Rachenraums durch Zunge und Lippen, unterstützt durch den Unterkiefer. Jede Vokaltraktstellung hat Resonanzbereiche, an welchen das Anregungssignal verstärkt wird. Dadurch entstehen als Resonanzfrequenzen die sog. *Formanten*.

Eine ausführliche Darstellung dieser Vorgänge ist in der *Akustischen Theorie der Vokalartikulation* formuliert worden (Fant, 1960Ungeheuer, 1962).

2.2 Prosodische Parameter

Im Bereich der Prosodie herrscht eine große Begriffsvielfalt und damit auch Begriffsmehrdeutigkeit. So erscheint es zweckmäßig, die in dieser Arbeit verwendeten Bezeichnungen zu definieren und sie in diesem Zusammenhang konsistent zu verwenden.

In älteren Publikationen wurden prosodische Elemente oft als unabhängig von den Segmenten gesehen, sie sind zusätzliche Merkmale, die von den Lauten nur getragen werden (z. B. von Essen (1953, S. 169), Trubetzkoy (1939, S. 166)). Nach Lehiste (1970) dagegen bedeutet Prosodie eine *sekundäre überlagernde Funktion* von Merkmalen, die den Lauten bereits *inhärent* sind. Die prosodischen Einheiten sind nicht an einzelne Segmente gebunden, sondern können sich über mehrere Lauteinheiten erstrecken.

Lehiste (1970) teilt die Prosodie zunächst einmal grob in Parameter mit linguistischer Funktion (*Suprasegmentalia*) und in para- und extralinguistische Parameter ein. Zu den beiden letzteren gehören beispielsweise Sprechtempo, Rhythmus, Stimmqualität und Pausen (siehe Abschnitt 2.7). Die Suprasegmentalia bestehen aus *tonal features, stress* und *quantity*, denen die jeweiligen akustischen Korrelate Sprachgrundfrequenz (F_0), Intensität und Dauer zugeordnet werden. Die *tonal features* werden bei Lehiste in *pitch* (perzeptive Tonhöhe), *tone* (distinktiv auf Wortebene, z. B. in Tonsprachen) und *intonation* (Funktion auf Satzebene) eingeteilt.

Eine ähnliche Einteilung mit etwas anderen Bezeichnungen findet sich bei Cruttenden (1986): Die drei *perzeptiven* Parameter mit linguistischer Funktion heißen bei ihm *pitch,*

loudness und *length*. *Pitch* und *loudness* bezeichnen bei ihm die *Variation* des jeweiligen Parameters über eine oder mehrere Einheiten; *length* ist die *relative Dauer* einer oder mehrerer Einheiten in einem bestimmten Kontext im Vergleich zu anderen Kontexten. Eine Einheit kann dabei aus einem Laut, einer Silbe oder einem Wort bestehen; es können aber auch größere Einheiten betrachtet werden, wie Phrasen, Sätze, Abschnitte oder ganze Diskurse.

Die Übertragung dieser Begriffe ins Deutsche ist nicht immer einheitlich. Der deutsche Begriff *Intonation* wird beispielsweise manchmal in erweitertem Sinn ausgelegt (Bußmann, 1990Nöth, 1991). Nöth versteht unter Intonation die *"distinktive Verwendung prosodischer Eigenschaften zur Bedeutungsdifferenzierung" (Nöth, 1991, S. 23)*. Dies geschah offensichtlich unter dem Kritikpunkt, daß in vielen Untersuchungen nur der Grundfrequenzverlauf in bezug auf linguistische Funktionen erforscht wird.

In dieser Arbeit soll der Begriff *Intonation* im ursprünglichen Sinne von Lehiste verwendet werden. Für den durch die Grundfrequenz hervorgerufenen Effekt auf Segmentebene findet der Begriff *Tonhöhe* (hier nicht ausschließlich perzeptiv verstanden) Verwendung.

Für das perzeptive Korrelat der Intensität wird vielfach der Begriff *Lautheit* gebraucht (Cruttenden, 1986Nöth, 1991Möbius, 1993). Aus Gründen der Eindeutigkeit soll er auch in dieser Arbeit verwendet werden. Der Begriff *stress* und seine deutsche Übertragung *Betonung* sind insofern problematisch, als daß sie oft mit der Hervorhebung von Silben (Akzentuierung) gleichgesetzt werden (siehe auch Diskussion in Cutler und Ladd, 1983, S. 141). Wie viele Untersuchungen belegen, ist aber das, was allgemein unter Akzentuierung verstanden wird, weit weniger durch die Intensität verursacht als vielmehr durch Grundfrequenz und Dauer (Abschnitt 2.5.1).

Der Begriff der *Quantität* kann hingegen ohne Probleme übernommen werden. Sein akustisches Korrelat ist die zeitliche Ausdehnung einer Spracheinheit, also die Dauer.

2.3 Eigenschaften von akustischen Parametern

Im folgenden sollen die drei akustischen Parameter F_0, Intensität und Dauer betrachtet werden. Die Darstellung lehnt sich an Lehiste (1970) und Vaissière (1983) an. Diese drei Parameter werden sowohl im Hinblick auf physiologische und meßtechnische Aspekte als auch auf ihre Eigenschaften im globalen Verlauf in einer Äußerung beschrieben.

2.3.1 Produktion und Messung

Die Grundfrequenz ist über das Anregungssignal, d. h. die Vibration der Stimmlippen bei der stimmhaften Phonation (Abschnitt 2.1.1), definiert. Für stimmlose Laute gibt es daher keine Grundfrequenz. Die Abstände zwischen den einzelnen Anregungsimpulsen (also zwischen zwei Glottisverschlüssen) ergeben die *Grundperiode* T_0. Die *Grundfrequenz* F_0 ist der reziproke Wert der Grundperiode.

Die Grundfrequenz ist von der Länge der Stimmlippen und ihrer Spannung, außerdem

von der Höhe des subglottalen Luftdrucks abhängig. Erhöhung von Grundfrequenz wird durch größere Länge und Spannung der Stimmlippen und durch höheren Luftdruck verursacht. Allerdings ist die Schwingung der Stimmlippen nicht immer regelmäßig, z. B. in Verschlußpausen von stimmhaften Plosiven oder beim Phänomen der Laryngalisierung (Abschnitt 2.1.1). Vor allem für diese unregelmäßigen Signalabschnitte ist die Bestimmung der Grundfrequenz nichttrivial und stark fehlerbehaftet; die Problematik der Grundfrequenzmessung wird in Hess (1983) ausführlich erläutert.

Die Intensität ist physiologisch durch die Höhe des respiratorischen Aufwands repräsentiert, der ein Ansteigen des subglottalen Drucks verursacht. In der Akustik ist die Intensität einer laufenden Welle als Leistung pro Flächeneinheit (W/m^2) definiert. Zum Vergleich verschiedener Intensitäten wird im allgemeinen ein Verhältnismaß (Pegel) verwendet: Die *Pegelstärke* L ist definiert als logarithmierte Intensität I im Verhältnis zu einem festgelegten Grundmaß I_0; die Maßeinheit heißt *Dezibel* (dB):

$$L = 10 \cdot log \, \frac{I}{I_0} \, dB \tag{2.1}$$

Im Unterschied zur physikalisch meßbaren Intensität gibt es außerdem die psychoakustische Einheit der *Lautstärke*, die von der Wahrnehmung durch das Gehör abhängt. Die subjektive Empfindung der Lautstärke ist extrem frequenzabhängig: Schallwellen mit gleicher Pegelstärke, aber mit unterschiedlicher Frequenz, werden im allgemeinen nicht als gleich laut wahrgenommen.

Die Intensität ist proportional zum Quadrat der Amplitude. Die *Amplitude* bezeichnet die maximale Auslenkung eines Teildrucks (hier der Luftdruck) bei einer Oszillation um eine Ruhelage. Zur praktischen Berechnung der Energie eines Sprachsignalabschnitts der Fensterlänge N wird beispielsweise der Kurzzeiteffektivwert verwendet (*root mean square, rms-Pegel*):

$$E_S = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x^2(n)} \tag{2.2}$$

Die Dauer beschreibt die zeitliche Ausdehnung eines akustischen Ereignisses, sei es segmental oder suprasegmental. Ein Problem bei der Messung der Dauer ist die Segmentierung. Benachbarte Lautsegmente und prosodische Funktionen werden nicht isoliert produziert, sondern gehen ineinander über; so ist es mitunter schwierig zu entscheiden, wo ein Ereignis beginnt oder aufhört.

Für alle drei Parameter sollten bei der Messung lautspezifische Werte nicht vernachlässigt werden (siehe Abschnitt 2.4); außerdem müssen globale Aspekte (z. B. *final lengthening* für die Dauer; siehe Abschnitt 2.3.2) beachtet werden.

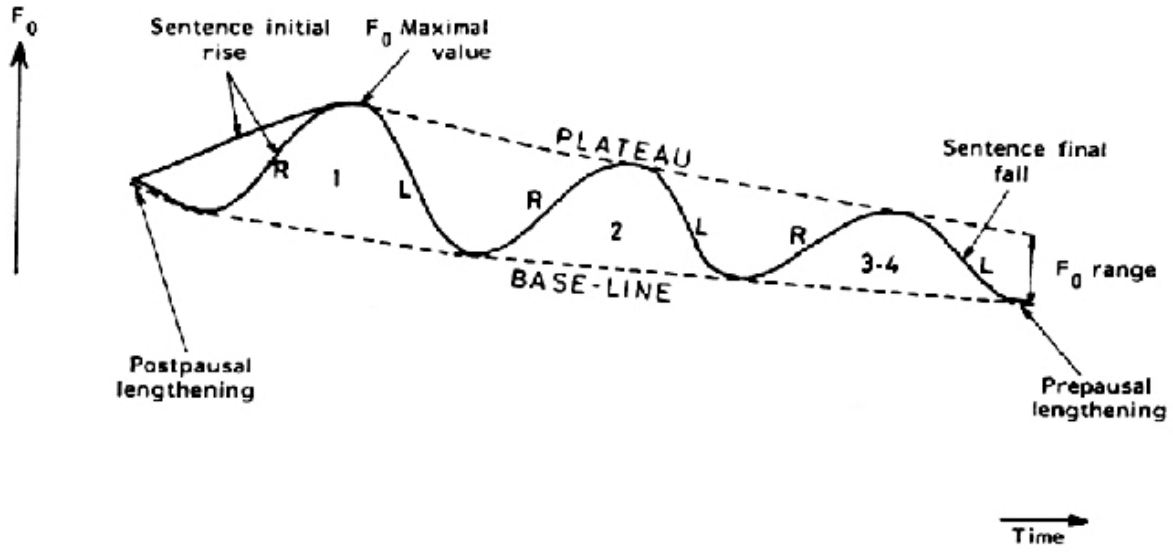

Abbildung 2.2: *Eigenschaften einer F_0-Kontur in einer unmarkierten Äußerung (R = Anstieg, L = Abfall) [aus Vaissière, 1983]*

2.3.2 Eigenschaften im globalen Verlauf

Bezogen auf den Gesamtverlauf lassen sich gewisse Tendenzen für die Grundfrequenz angeben, die in vielen bisher untersuchten Sprachen gültig sind (Vaissière, 1983). Eine schematische Darstellung findet sich in Abbildung 2.2.

Die Grundfrequenz hat die Tendenz, sich zwischen zwei imaginären Linien zu bewegen (*plateau* oder *topline* und *baseline*). Die 'baseline' wird im allgemeinen als sprecherabhängig eingestuft (siehe auch Pierrehumbert, 1980, S. 124). Der Wertebereich, der dadurch beschrieben wird, verengt sich im Verlauf des Äußerungsabschnitts. Dies gilt zumindest für lineare Darstellungsweise. Bei logarithmischer Darstellung, die stärker den Aspekt der Wahrnehmung betont, verlaufen 'topline' und 'baseline' in den meisten Fällen parallel; siehe z. B. Cohen et al. (1982). Die genannten Autoren verneinen die Unabhängigkeit der beiden Beschreibungslinien und verwenden in erster Linie die als stabiler angesehene 'baseline', aus der die parallele 'topline' leicht abgeleitet werden kann (siehe aber Pierrehumbert, 1979).

Zu Äußerungsbeginn wird ein starker F_0-Anstieg produziert, während gegen Ende die F_0-Werte stark absinken (sog. *final fall*); dies gilt in erster Linie für Aussagesätze. Der globale Verlauf ist charakterisiert durch eine Wellenbewegung von wechselndem Anstieg und Fall der Grundfrequenz. Dabei sinken nach Vaissière (1983) die lokalen F_0-Maxima schneller ab als die lokalen F_0-Minima. Die F_0-Extrema korrelieren mit den akzentuierten Silben, in sehr vielen Sprachen sind dies die Maxima, in einigen aber auch die Minima.

Eine wichtige Aufgabe der Intensität ist es, zwischen Sprache und Pause zu unterscheiden (Vaissière, 1983). Weiterhin kann sie als Indiz für die Silbeneinteilung gelten (obwohl dies durch segmentale Einflüsse gestört werden kann; siehe Abschnitt 2.4.2). Der Verlauf

der Intensität wird von der Grundfrequenz beeinflußt, d. h. Ansteigen oder Fallen der F_0-Kontur bewirkt eine ebensolche Tendenz beim Intensitätsverlauf. Dies läßt sich durch die gleichen physiologischen Mechanismen bei der Produktion erklären: Erhöhung dieser beiden Parameter wird durch Verstärkung der Atemanstrengung, des subglottalen Drucks und der Spannung der Stimmlippen erzeugt. Eine Ausnahme von dieser Abhängigkeit stellt eine final stark ansteigende Grundfrequenzkontur dar, die in vielen Sprachen zur Markierung einer Frage verwendet wird (Abschnitt 2.5.2).

Bei der Dauer läßt sich eine Längung der ersten und letzten Elemente (meist Vokale) einer Äußerung feststellen (siehe auch Abbildung 2.2). Offensichtlich ist es ein universelles Phänomen, daß die Terminierung von motorischen Sequenzen oder Planungseinheiten durch Verlängerung der Dauer (*final lengthening*) markiert wird (Vaissière, 1983).

2.4 Mikroprosodie

In diesem Abschnitt soll die Rückwirkung des Vokaltrakts auf die Phonation dargestellt werden. Die Artikulation der Laute wird von der Phonation beeinflußt; umkehrt beeinflussen lautspezifische Eigenschaften die Messung der Anregungsparameter und damit auch die Messung der akustischen prosodischen Korrelate Grundfrequenz, Intensität und Dauer. Der gegenseitige Einfluß von segmentalen und suprasegmentalen Faktoren wird auch als *Mikroprosodie* bezeichnet. Der Großteil der Untersuchungen beschäftigt sich allerdings nur mit der Grundfrequenz.

Die Auswirkungen äußern sich lautspezifisch (intrinsisch: IF_0) und koartikulationsbedingt (CF_0). Sie unterliegen nicht der aktiven Kontrolle des Sprechers (Petersen, 1986, S. 31) und spielen für die Wahrnehmung, zumindest von linguistischen Kategorien, keine Rolle (Daniloff et al., 1980, S. 207). Die Mikroprosodie hat also keine linguistische Funktion in der normalen Sprechsituation. Eine ausführliche Darstellung von mikroprosodischen Phänomenen findet sich bei di Cristo (1985) und Gartenberg (1987).

2.4.1 Intrinsische Eigenschaften

In bezug auf den Parameter F_0 haben hohe Vokale wie [i] oder [u] eine höhere intrinsische Grundfrequenz (Mohr, 1971) als der Tiefzungenvokalvokal [a]. Dies läßt sich physiologisch so erklären (nach Lehiste, 1970): Die Zungenhebung bei hohen Vokalen streckt den Teil der Kehlkopfmuskeln, der Position und Spannung der Stimmlippen kontrolliert und verstärkt damit die Stimmlippenspannung. Durch diese höhere Spannung wird die Grundfrequenz angehoben. Dieses Phänomen wird auch als *tongue-pull* -Theorie bezeichnet (z. B. Ohala und Ewan, 1973).

Der intrinsische Effekt wirkt sich auf akzentuierte Silben und zu Beginn einer Äußerung sehr viel stärker aus (Petersen, 1979). Hier ist also wiederum der Einfluß des globalen Tonhöhenverlaufs sichtbar. IF_0 ist weiterhin durch eine hohe Sprecherabhängigkeit geprägt (di Cristo und Hirst, 1986).

Für Vokale gibt es ebenfalls eine intrinsische Intensität. Kurzvokale sind im allgemeinen

energiereicher als Langvokale. Besonders hoch ist die Intensität, wenn die Grundfrequenz mit dem 1. Formanten zusammenfällt (dieser besitzt die meiste Energie).

Stimmhafte Konsonanten machen sich durch einen lokalen Abfall von F_0 und Intensität im globalen Verlauf bemerkbar. Dies ist besonders deutlich für stimmhafte Obstruenten. Bei der Produktion von Plosiven und Frikativen entsteht durch den mehr oder weniger starken Verschluß im Vokaltrakt ein Abfall des transglottalen Drucks (Vaissière, 1988). Dies bewirkt das Absinken von F_0 und Intensität.

Als intrinsische Charakteristik für die Dauer läßt sich angeben, daß Hochzungenvokale eher kürzer sind. Die tiefer gelegenen Vokale dagegen benötigen weitreichendere artikulatorische Bewegungen und sind daher etwas länger (Lehiste, 1970).

2.4.2 Koartikulatorische Zusammenhänge

Unter Koartikulation wird die zeitliche Überlappung von artikulatorischen Gesten verstanden. Das Phänomen der Koartikulation wurde bereits von Menzerath und de Lacerda (1933) untersucht. Lautsegmente im Kontext werden nicht isoliert produziert, sondern die einzelnen Bewegungen gehen ineinander über. Das macht die automatische Segmentierung und Erkennung von gesprochener Sprache besonders schwierig.

Verschiedene Untersuchungen haben gezeigt, daß gewisse Zusammenhänge zwischen Koartikulation und der Auswirkung auf den Tonhöhenverlauf bestehen. Lehiste und Peterson (1961) fanden Gesetzmäßigkeiten bei Vokalen in stimmlosem bzw. stimmhaftem Konsonantenkontext: Vokale in Verbindung mit stimmlosen Konsonanten ergaben eine relativ höhere F_0 mit schnellem Anstieg zu einem Maximum. Vokale in stimmhafter Umgebung hingegen waren zu Beginn etwas niedriger in F_0, stiegen dann langsam an und hatten ihren Spitzenwert erst in der Mitte des Vokals. Ähnliche Ergebnisse wurden bereits von House und Fairbanks (1953) ermittelt.

Bei Silverman (1986) werden die vorliegenden Ergebnisse allerdings nicht voll bestätigt: Er kritisiert, daß beim Aufbau der vorherigen Experimente der globale Verlauf der prosodischen Kontur, der sich ja auch wieder auf die Lautsegmente auswirkt, nicht genug berücksichtigt worden sei. In seinen Untersuchungen wird der Effekt dahingehend abgeschwächt, daß nach stimmlosen Konsonanten nicht notwendigerweise ein steiler Anstieg im nachfolgenden Vokal stattfindet. Es gibt dagegen die Tendenz eines Vokals, nach stimmlosen Konsonanten relativ höher zu beginnen als nach stimmhaften Konsonanten.

Als physiologische Ursache für den CF_0-Effekt wurden sowohl aerodynamische Einflüsse der supra- und subglottalen Luftdruckverhältnisse (Ladefoged, 1973) als auch die Möglichkeiten verschiedener Kehlkopfmuskeln zur Stimmlippenspannung (Halle und Stevens, 1971) betrachtet. Es erscheint aber sinnvoll, nicht nur eine Ursache zu favorisieren, sondern vielmehr eine Kombination aus verschiedenen Quellen anzunehmen (di Cristo und Hirst, 1986).

Die Intensität wechselt mit den artikulatorischen Konfigurationen des Vokaltrakts. So kann es vorkommen, daß sie an Transitionen zwischen Konsonant und Vokal höher ist als im Vokal selbst. Dies ist andererseits nicht zwingend; nichtsdestoweniger ist die Intensität

als Indikator für hervorgehobene Silben ein eher unzuverlässiger Parameter (siehe auch Abschnitt 7.1).

Die Auswirkung der Koartikulation auf die Dauer ist abhängig vom Ausmaß der artikulatorischen Bewegungen. Die Lautkombination [ib] ist z. B. kürzer als [ig], während [ug] wiederum kürzer als [ub] ist. Ansonsten ist die Dauer aber eher sprachenabhängig. Im Englischen z. B. ist die Länge eines Vokals ein wichtiges Indiz für die Stimmhaftigkeit des folgenden Konsonanten.

2.5 Makroprosodie

In diesem Abschnitt soll auf die globale Funktion der prosodischen Parameter eingegangen werden. In Anlehnung an die Mikroprosodie (Abschnitt 2.4) wird dies oft als *Makroprosodie* bezeichnet. Die Makroprosodie umfaßt im Gegensatz zur Mikroprosodie die *bewußte Steuerung* der prosodischen Parameter, um bestimmte wahrnehmbare Phänomene zu erzeugen. Dieser gezielte Einsatz der prosodischen Parameter bewirkt im wesentlichen die Hervorhebung von Einheiten (Silben, Wörter) und die Gliederung bzw. Markierung von Äußerungen. Aus diesen Erscheinungen kann eine *linguistische Funktion* (Abschnitt 2.6) interpretiert werden.

2.5.1 Hervorhebung und Prominenz

Es ist vielfach umstritten, in welchem Verhältnis prosodische Parameter zur wahrnehmbaren Hervorhebung von Silben oder Wörtern und damit zur linguistischen Funktion der *Akzentuierung* (siehe Abschnitt 2.6.1) beitragen. Dies ist in hohem Maße auch sprachenabhängig (Vaissière, 1983). Am weitesten verbreitet ist die Ansicht, daß die *Grundfrequenz* den größten Anteil daran hat (z. B. Vaissière, 1983Bolinger, 1958Kohler, 1989). Der zweitwichtigste Parameter ist nach Untersuchungen von Lieberman (1960) die Intensität, während Fry (1955) die Dauer favorisiert (beide für die englische Sprache). In anderen Untersuchungen für das Englische ist Beckman (1986) allerdings der Ansicht, daß die *Intensität* den größten Einfluß auf die Hervorhebung hat, während Adams und Munro (1978) wiederum zu einem anderen Ergebnis kamen: bei ihnen hatte die *Dauer* von Silben die stärkste Bedeutung bei der Hervorhebung. Für eine ausführliche Diskussion siehe auch Möbius (1993, S. 12 - 16).

Neppert und Pétursson (1986) differenzieren diese Problematik für verschiedene Sprachen. Die *Erhöhung der Grundfrequenz* wird als wichtigstes Akzentuierungskorrelat für Deutsch, Englisch, Französisch und die slawischen Sprachen angegeben bzw. als ausschließliche Ursache für Spanisch, Italienisch und Japanisch. Im Dänischen korreliert die Akzentuierung mit einer *Erniedrigung der Grundfrequenz*. Die *Intensität* steht bei germanischen Sprachen wie Deutsch, Englisch und Isländisch in ihrer Bedeutung an zweiter Position, während sie für die romanischen Sprachen im allgemeinen keine Rolle spielt. Die *Dauer* ist als zusätzliches Korrelat für das Französische wichtig; in sog. Quantitätssprachen wie dem Ungarischen (dort ist die Quantität eine phonologische Größe, die entscheidend

zur Unterscheidung von Wortbedeutungen beiträgt) spielt die Dauer in bezug auf die Akzentuierung überhaupt keine Rolle.

In manchen Sprachen wie Deutsch, Russisch oder Englisch kann auch die *Klangqualität* von Vokalen eine Bedeutung für die Akzentuierung haben, es gibt dann gewisse Vokalinventare für akzentuierte und nichtakzentuierte Silben. Als Tendenz läßt sich angeben, daß für nicht hervorgehobene Silben häufig Zentralvokale verwendet werden; diese werden in akzentuierten Silben dagegen nie verwendet.

Ein weiterer wichtiger Begriff in diesem Zusammenhang ist die *Prominenz*. Der Begriff Hervorhebung betont eher die aktive Beteiligung des Sprechers, Prominenz betont eher die Wahrnehmung des Hörers. Laver (1994) definiert Prominenz so, daß manche Silben perzeptiv stärker herausragen als andere; dies wird durch größere muskuläre Anstrengung erzeugt:

> *"Other things being equal, one syllable is more prominent than another to the extent that its constituent segments display higher pitch, greater loudness, longer duration or greater articulatory excursion from the neutral disposition of the vocal tract (Laver, 1994:450). "*

Untersuchungen zur quantitativen Messung der Prominenz wurden erstmals von Fant und Kruckenberg (1989) durchgeführt. In verschiedenen Hörerexperimenten ließen sie nicht nur die Prominenz von Silben beurteilen, sondern auch die von Grenzen. Die Wahrnehmung von Grenzen innerhalb einer Äußerung, und damit ihre Gliederung, kann demnach ebenfalls mit Hilfe der Prominenz beschrieben werden (vgl. auch Heuft, 1999).

2.5.2 Gliederung und Markierung von Äußerungen

Der Verlauf einer Äußerung wird in Lieberman (1967) mit einer sog. *Breath Group* beschrieben. Damit soll ein physiologisches Phänomen bei der Ausatmung erklärt werden: Durch sinkende Atemkraft fällt der subglottale Druck am Ende einer Äußerung, so daß ohne weitere Anstrengung des Sprechers ebenfalls Intensität und F_0 absinken. Eine markierte Form der Breath Group wird durch zusätzliche Aktivitäten der Kehlkopfmuskeln des Sprechers erreicht. Der subglottale Druck sinkt zwar trotzdem am Ende der Äußerung, durch bewußte Anspannung der Stimmlippen steigt aber die Grundfrequenz. Diese markierte Form der Breath Group wird in vielen Sprachen zur Darstellung einer Frage verwendet (Lieberman, 1967, S. 132).

Die Untersuchungen von Collier (1975) können die Ergebnisse von Lieberman (1967) teilweise bestätigen. Collier stellte als wichtigste Ursache für die Kontrolle der F_0-Variationen die Aktivität eines Kehlkopfmuskels (Crycothyroid) fest, während der subglottale Luftdruck in erster Linie die fallende *baseline* (siehe Abschnitt 2.3.2) der F_0-Bewegung bewirkt. Im Gegensatz zu Lieberman, der annimmt, daß hervorgehobene Silben sowohl durch Anstieg des subglottalen Luftdrucks als auch durch stärkere Kehlkopfmuskelspannung charakterisiert sind, macht Collier dafür allein die Anspannung der Kehlkopfmuskeln verantwortlich.

Das Fallen der F_0-Kontur am Ende einer Äußerung wird auch als *Deklination* bezeichnet (Pierrehumbert, 1979). Als physiologische Ursache wird der Abfall des subglottalen Drucks (Lieberman, 1967), die sog. *trachea-pull* -Theorie (Maeda, 1976) oder ein Trägheitsprinzip (Produktion eines F_0-Anstiegs kostet mehr Aufwand als das Absenken; Ohala und Ewan (1973)) angegeben. Diese drei Erklärungen sind nach Ansicht von Vaissière (1983) kompatibel und müssen nicht im Widerspruch zueinander stehen.

Nach Vaissière (1983) gibt es die Möglichkeit, die Deklinationsrate als konstant zu betrachten und mit einer *Deklinationsgerade* zu beschreiben; es gibt auch die Ansicht, daß die Grundfrequenz zu Beginn einer Äußerung schneller abfällt und dann langsamer absinkt (exponentielle Deklination). Die Steigung der Deklination ist nicht nur physiologisch bedingt, sondern wird vom Sprecher aktiv kontrolliert, um syntaktische Markierungen zu setzen (Pierrehumbert, 1979). Experimente von Pierrehumbert (1979) belegen außerdem, daß die Deklination in längeren Intonationsgruppen geringer ist; sie ist dafür um so stärker, je weiter der F_0-Bereich des Sprechers ist. In Pierrehumbert (1979) kommt im übrigen der 'topline' höhere Bedeutung zu, sie erscheint perzeptuell wichtiger und auch zuverlässiger in der Produktion, außerdem ist sie steiler als die 'baseline' (siehe auch Abschnitt 2.3.2). Das Phänomen der Deklination ist für gelesene Sprache und Spontansprache unterschiedlich ausgeprägt (Umeda, 1982).

Abgesehen von der Endemarkierung durch die Deklination (*declination reset*) kann die *Gliederung einer Äußerung* (siehe Abschnitt 2.6.2) mit weiteren Mitteln erzeugt werden. Ein wichtiges Mittel ist die Setzung von *Pausen*. Weiterhin kann eine intonatorische Markierung eingesetzt werden: Der Anstieg von F_0 markiert im allgemeinen den Beginn, ein Fallen das Ende eines zusammenhängenden Abschnitts bzw. einer Phrase. Eine Endemarke setzt außerdem das sog. *final lengthening*. Das bedeutet, daß Elemente am Ende einer Phrase oder einer Äußerung eine höhere zeitliche Dauer haben (siehe Abbildung 2.2). Die Erscheinung des *final lengthening* kann sowohl auf Wort-, Phrasen- oder Satzebene wirksam sein.

Eine weitere wichtige Grenzmarkierung kann auch durch die Stimmqualität, also das Einsetzen eines bestimmten phonatorischen Registers (siehe Abschnitt 2.1.1), erfolgen: Hedelin und Huber (1990) führten umfangreiche Untersuchungen über unregelmäßige Sprachanregungen wie verschiedene Formen der Laryngalisierung durch. In einem Großteil der Fälle wurde dadurch eine Endemarkierung gesetzt. Dies läßt sich auch physiologisch damit erklären, daß die Stimmlippenspannung und der Atmungsaufwand am Ende einer Äußerung nachlassen, so daß die Stimme leichter in ein tieferes Register absinken kann.

Zur Gliederung einer Äußerung können auch rhythmische Elemente beitragen, die weitestgehend sprachenabhängig sind. Vaissière (1988) unterscheidet zwischen *isosyllabilicity* und *isochrony*. Zum ersten Typ (auch *silbenzählend* genannt) gehört z. B. das Französische, wo die zeitlichen Abstände zwischen den einzelnen Silben relativ gleichmäßig verteilt sind. Die Isochronie besagt dagegen, daß gleiche zeitliche Abstände zwischen *akzentuierten* Silben angestrebt werden (daher auch die Bezeichnung *akzentzählend*). Im Deutschen, das zu den akzentzählenden Sprachen gehört, führt dies Phänomen zu einer starken Dauerreduktion der nichtakzentuierten Silben.

2.6 Linguistische Funktion und prosodische Korrelate

Die prosodischen Parameter können, einzeln oder auch in Kombination, eine linguistische Funktion markieren. Diese Funktionen können durch unterschiedliche Anordnungen und Ausprägungen erzeugt werden. Ohne Prosodie können linguistische Funktionen auch durch grammatikalische Markierungen dargestellt werden.

In der Schriftsprache gewinnt die grammatikalische Markierung stärker an Bedeutung. Dort ersetzen außerdem die Satzzeichen einige prosodische Markierungen (Komma und Punkt für Grenzen, Fragezeichen und Ausrufezeichen für Phrasenmarkierung). Akzentuierung wird eher selten schriftlich fixiert, dann finden die Möglichkeiten Unterstreichung, Fett- bzw. Kursivdruck oder Großschreibung Verwendung.

2.6.1 Akzentuierung

In Kohler (1995) wird die Bezeichnung *Akzent* für die Hervorhebung von Silben verwendet; damit wird eine linguistische Funktion bewirkt. Zu unterscheiden ist dabei zwischen Wort-, Ton und Satzakzent. Der *Wortakzent* ist sprachenabhängig: Er kann auf eine bestimmte Silbenposition festgelegt sein (z. B. die 1. Silbe im Ungarischen) und hat dann meist nur eine abgrenzende, rhythmische Funktion. Ein frei beweglicher Akzent kann dagegen auf jeder Silbe stehen und ist oft durch lexikalische Regeln (z. B. im Deutschen) für jedes Wort definiert.

Der *Tonakzent* spielt nur für die sog. Tonsprachen (wie z. B. Chinesisch) eine Rolle. Dort hat die *Tonhöhe* phonologische Relevanz, d. h. bedeutungsunterscheidende Funktion. Die Tonhöhe bezieht sich dabei auf morphologisch definierte Segmente (Morphe, Wörter).

Der *Satzakzent* ist der Akzent, der innerhalb der Akzente einer Äußerung am weitesten herausragt (durch Maximalwerte der akustischen Parameter). Durch ihn wird meist die Stelle einer Äußerung markiert, die den höchsten Grad an Information enthält (siehe Abschnitt 3.1.1). In diesem Fall wird die Akzentuierung inhaltlich (nicht nur lexikalisch/grammatisch) motiviert, so daß der Akzent auf jede Silbe fallen kann, z. B. auch zum Ausdrücken von Kontrast oder Emphase (Abschnitt 3.1.5).

Kohler (1991) favorisiert eine scharfe Trennung zwischen dem Akzent auf lexikalischer Ebene (Wortakzent) und dem Akzent auf Satzebene (Satzakzent). Der Wortakzent ist lexikalisch definiert und hat keine zusätzliche linguistische Funktion. Er dient lediglich zur Unterscheidung von Wortbedeutungen und ist im Deutschen auf wenige Beispiele beschränkt (z. B. übersetzen, umfahren; August, Tenor). Im Englischen können auch Wortarten unterschieden werden, z. B. *con*flict (Substantiv) vs. con*flict* (Verb). Der Satzakzent korreliert dagegen mit der Informationsstruktur einer ganzen Äußerung und damit mit den 'höheren' linguistischen Ebenen der Semantik und Pragmatik. Da der Schwerpunkt dieser Arbeit auf der semantisch-pragmatischen Funktion der Prosodie liegen soll, werde ich mich im weiteren auf die Akzentuierung auf Satzebene beschränken.

2.6.2 Phrasierung

Eine prosodische Phrase wird in erster Linie durch die Änderung des Intonationsverlaufs und durch Pausen wahrgenommen (Abschnitt 2.5.2). In einer Phrase werden in der Regel logisch zusammenhängende Elemente einer Äußerung zusammengefaßt. Die Markierung einer Phrasengrenze trennt verschiedene Phrasen voneinander ab. Die Grenzen zwischen Phrasen können unterschiedlich stark sein (de Pijper und Sanderman, 1994).

Prosodische Phrasen geben Hinweise auf die *syntaktische Gliederung* eines Satzes, sie sind aber nicht mit syntaktischen Phrasen identisch (vergleiche auch mit Abschnitt 4.3.2). Syntaktische Phrasen bilden eine Konstituente (Wortgruppe oder Satzteil von relativer Selbständigkeit) (Bußmann, 1990). Der Zusammenhang zwischen syntaktischen und prosodischen Phrasen wurde z. B. in Price et al. (1991) untersucht.

Prosodische Phrasen können grammatisch unvollständig sein, z. B. durch Abbrüche oder bestimmte Auswirkungen der Spontansprache. Die Gliederung von größeren Äußerungsabschnitten (Absätze, Dialoge, etc.) wird in Swerts (1993Swerts (1997Swerts et al. (1994) beschrieben.

2.6.3 Satzmodus

Der Satzmodus wird verstanden *"als komplexe syntaktische Struktur (z. B. Aussagesatz, Entscheidungsfragesatz, Wunschsatz), der regelhaft bestimmte abstrakte Formtypen zugeordnet sind"* (Altmann et al., 1989). Die Markierung des Satzmodus kann lexikalisch (W-Frage), syntaktisch (Verbstellung), morphologisch (Flexion des Verbs: z. B. Konjunktiv in Wunschsätzen) oder durch den Verlauf der prosodischen Parameter erfolgen.

Der linguistische Satzmodus kann aus der prosodischen Phrasenmarkierung interpretiert werden. Bei der Markierung einer Phrase spielt die Intonation eine wesentliche Rolle. Eine fallende Kontur markiert eine Phrase als *terminal*, eine ansteigende Kontur wird als *progredient* (weiterführend) oder *interrogativ* (fragend) und damit als 'nicht-terminal' wahrgenommen. Es werden allerdings nicht alle Fragesätze durch eine steigende Kontur markiert (z. B. W-Fragen in der Regel nicht).

Cruttenden (1981) beschreibt universelle Tendenzen für fallende und steigende Intonationskonturen. In den meisten Sprachen wird eine fallende Kontur für eine 'geschlossene Bedeutung', also für Aussagen, Bestätigung oder Feststellung verwendet. Ansteigende Konturen vermitteln eine 'offene Bedeutung', also Fragen, Zweifel, Höflichkeit, etc. Große Abweichungen gibt es in den verschiedenen Dialekten; tendenziell nimmt die Verwendung von steigenden Intonationskonturen zu. Oftmals wird eine bestimmte intonatorische Markierung als extralinguistische Markierung der Sprecherattitude (Abschnitt 2.7) wahrgenommen, auch wenn dies vom Dialektsprecher nicht beabsichtigt wurde.

Im Deutschen gilt es als Indiz für einen Fragesatz, wenn am Äußerungsende die Grundfrequenz ein fallend-steigendes Muster zeigt (Altmann et al., 1989). Nöth (1991) gibt als mögliche Markierung für einen Aussagesatz das Mittel der Laryngalisierung (siehe Abschnitt 2.1.1) am Ende einer Äußerung an.

2.7 Para- und extralinguistische Parameter

In den vorhergehenden Abschnitten wurden die drei wichtigsten prosodischen Parameter und ihre linguistischen Funktionen behandelt. Daneben gibt es noch eine weitere Reihe von Parametern, die zur *Unterstützung* bzw. *Variation* von linguistischer Funktion beitragen können. Zu diesen paralinguistischen Parametern zählen *Sprechtempo, Rhythmus, Stimmqualität, Phonationstyp, Pausen* und *Häsitationen*.

Pausen, Häsitationen und Sprechtempo können die zeitliche Strukturierung von Äußerungen beeinflussen. Rhythmus wirkt sich ebenfalls auf die Gliederung einer Äußerung aus und ist im allgemeinen sprachenspezifisch. Ein bestimmter Phonationstyp wie "creaky voice" (Laryngalisierung) kann Grenzen (Phrasen), aber auch Terminalität (Satzmodus) markieren.

Extralinguistische Parameter sind meist sprecherspezifisch und unterliegen oft nicht der aktiven Kontrolle eines Sprechers. Bestimmte permanente Eigenschaften (wie Alter, Geschlecht, Körperbau) und vorübergehende Zustände (wie Gesundheitszustand, Attitude, Emotionen) eines Sprechers können durch diverse Parameter markiert werden: Die Stimmqualität verrät oft etwas über den Gesundheitszustand oder die Emotionen des Sprechers, die Tonhöhe gibt Auskunft über Geschlecht und Alter, erhöhtes Sprechtempo und Lautheit können emotionale Erregung markieren. Die Intonation kann über die Einstellung/Attitude eines Sprechers zum Gesagten (Ironie, Zweifel) Auskunft geben.

Die Grenze zwischen para- und extralinguistischen Parametern ist fließend, alle prosodischen Parameter können sowohl linguistische Funktionen unterstützen als auch nur den situations- oder sprecherspezifischen Kontext markieren. Bestimmte Phonationstypen können in festgelegten Situationen verwendet werden, z. B. "Flüsterstimme" als aktives Mittel, um möglichst wenig zu stören oder um das Gesagte als 'geheim' zu markieren; oft zwingt aber auch allein der Gesundheitszustand einen Sprecher zum Flüstern. Extralinguistische Parameter können auch *aktiv* verwendet werden, um bestimmte Emotionen nur darzustellen (z. B. bei Schauspielern), um fremde Stimmen zu imitieren oder um die eigene Stimme zu verstellen.

2.8 Zusammenfassung

Die Beziehungen zwischen Prosodie, Akustik und Linguistik werden in Abbildung 2.3 zusammengefaßt. Abschließend dazu ein Zitat von Fujisaki (1994):

> *"The role of prosody in human speech communication cannot be overemphasized. The linguistic function of prosody covers lexical, syntactic, semantic and pragmatic information that supplements information carried by individual segments. Its function, however, extends beyond the scope of language in the narrow sense. A speaker expresses, either consciously or unconsciously, his/her intention, attitude, emotion, and even physical conditions that may or may not be related to the linguistic content of the utterance." [Vorwort]*

Abbildung 2.3: *Beziehungen zwischen prosodischen, akustischen und linguistischen Parametern*

3. Fokusbegriff

Es gibt verschiedene Betrachtungsweisen, sich dem Begriff des Fokus zu nähern. Unstrittig ist im allgemeinen, daß nur *wichtige* Elemente einer Äußerung im Fokus stehen. Die Fokusrealisierung bestimmt dabei die kommunikative Relevanz und Funktion einer Mitteilung im Dialog. Ein wichtiger Diskussionspunkt besteht darin, inwieweit die akustische Realisierung eines Fokus durch Syntax und/oder Informationsstruktur bestimmt ist. Dies läßt sich nicht universell festlegen; Untersuchungen mit unterschiedlichen Sprachen belegen dies (siehe Abschnitt 3.3).

Im folgenden werden einige Experimente vorgestellt, die untersuchen, wie sich linguistische Konzepte in der akustischen Realisierung manifestieren. Im ersten Berichtsteil wird eine eher semantische Sichtweise verfolgt, der zweite Teil konzentriert sich eher auf die Syntax: Dort wird versucht, akustische Korrelate für verschiedene Fokusarten in unterschiedlichen Satzpositionen und Satztypen zu finden.

Der überwiegende Teil der Untersuchungen wurde mit gelesener Sprache durchgeführt. Für die Spezialfälle von Fokus- und Kontrastakzenten hat das seine Gründe: Es ist sehr schwierig, in Spontansprache systematisch alle Positionen von möglichen Fokusakzenten zu untersuchen, da bestimmte Konstellationen häufiger auftreten als andere. Deswegen wurden meist Sätze konstruiert, die alle zu untersuchenden Phänomene abdecken sollten. Selbst dann ist nicht garantiert, daß alle Fokusakzente auch wie gewünscht realisiert werden, d. h. alle Testsätze müssen noch einmal perzeptiv überprüft werden.

Anschließend werden noch einige Arbeiten vorgestellt, die sich mit der automatischen Detektion von Fokusakzenten befassen. Zum Abschluß dieses Kapitels wird eine eigene Klassifikation entwickelt, die als Grundlage für diese Arbeit dient. Die ‘typischen’ Eigenschaften von Fokus und seiner akustischen Realisierung werden für verschiedene Ebenen zusammengefaßt.

3.1 Linguistische Konzepte

3.1.1 Funktion und Satzstruktur

Bereits in der Prager Schule wurde der Begriff der *Funktionalen Satzperspektive* eingeführt (Mathesius, 1929). Diese betrachtet die Gliederung eines Satzes unter dem Aspekt seiner *Mitteilungsfunktion*. In diesem Zusammenhang wurde auch das Begriffspaar *Thema-Rhema* entwickelt:

> *"Das Thema und Rhema stellen zwei komplementäre Mitteilungsfunktionen von verschiedenen semantischen Bestandteilen einer Äußerung dar; in fast jeder Aussage unterscheidet man das, worüber etwas mitgeteilt wird (Thema) und das, was darüber mitgeteilt wird (das Rhema, die Aussage im eigenen, engeren Sinn)." (Daneš, 1970)*

Synonym dazu werden vom gleichen Autor auch die Begriffspaare *Topic-Comment* (Daneš, 1967) oder *thèse-propos* (Daneš, 1960) verwendet. Als ähnlich oder synonym gelten auch die Begriffe *Fokus-Hintergrund* (Fokus entspricht dem Rhema, Hintergrund dem Thema):

> *"Die Fokus-Hintergrund-Gliederung gliedert die semantische Struktur von Sätzen zum Zwecke der Herstellung eines Alternativenbezugs in hervorgehobene und nicht-hervorgehobene Teile." (Jacobs, 1988)*

In der Prager Schule wurden dazu einige Universalien formuliert, deren Gültigkeit zumindest für die indo-europäische Sprachgruppe angenommen wird (Daneš, 1967Daneš, 1960):

1. In einer unmarkierten Äußerung wird das Thema vor dem Rhema angeordnet.
2. Das Intonationszentrum befindet sich im Bereich des Rhema.
3. Folglich befindet sich das Intonationszentrum eher im hinteren Bereich einer Äußerung.

Diese Universalien geben zunächst einmal nur Tendenzen für unmarkierte Äußerungen an. Wenn in markierten Äußerungen das Rhema im vorderen Bereich liegt, ist es aber sprachenabhängig, ob das Rhema eher durch Wortstellung oder durch Intonation markiert wird. In Sprachen mit relativ fester Wortstellung (wie Deutsch und Englisch) wird bevorzugt die Intonation zur Markierung des Rhemas verwendet, die Position ist weniger wichtig. In den slawischen Sprachen (mit relativ freier Wortstellung) wird es dagegen oft bevorzugt, ein Rhema im vorderen Bereich an das Ende einer Äußerung zu verschieben; die Position des Rhemas und damit auch des Intonationszentrums hat dort also entsprechend höhere Priorität (Daneš, 1967, S. 509).

3.1.2 Fokus als semantisch-pragmatischer Begriff

Das zu Beginn dieser Arbeit entwickelte Informationsmodell (Abschnitt 1.1) geht von einem *kooperativen* Dialog aus. Ziel ist die erfolgreiche Vermittlung von Information. Daraus ergibt sich auch, daß Sprecher ihre Mitteilungen effizient realisieren müssen, d. h. Wichtiges muß hervorgehoben und Unwichtiges in den Hintergrund geschoben werden.

Der *Fokus* ist in diesem Sinne sowohl ein semantischer als auch ein pragmatischer Begriff. Er enthält die wichtigsten Aspekte einer Mitteilung, bildet also das Informationszentrum des Satzes. Zugleich hat er eine Funktion im Diskurs; durch Herausstellen der wichtigsten Information wird dem Hörer die korrekte Interpretation des Satzes erleichtert bzw. erst ermöglicht.

In der akustischen Realisierung wird der Fokus durch den Satzakzent markiert. Der semantisch fokussierte Bereich (Skopus) kann im allgemeinen nur mit Hilfe des Kontextes

ermittelt werden. Der Fokusbereich umfaßt dann möglicherweise eine ganze Konstituente (weiter Fokus), kann aber auch auf einem Wort oder nur einer Silbe (eng) liegen (Ladd, 1980).

Wesentlich für diese Arbeit ist es, den Fokus in bezug auf seine aktuelle Realisierung im Diskurs zu betrachten. Die Markierung des Fokus geschieht durch Wort- und Satzgliedstellung und/oder intonatorisch durch den Satzakzent. Es ist sprachenabhängig, welche hervorhebenden Mittel dabei bevorzugt werden (Cruttenden, 1986). Hier soll in erster Linie die *intonatorische Markierung* von Fokus untersucht werden.

In Nooteboom und Kruyt (1987) findet sich dazu eine anschauliche Gegenüberstellung der hier verwendeten Grundbegriffe *Wortakzent (lexical stress), Akzent und Fokus*:

> *"Note, however, that accent should not be confused with lexical stress. Accent is superimposed on a word in the act of speaking. Lexical stress position is, of course, a permanent lexical property of the word. We also emphasize that, in the present view, only words can be accented, not word groups or whole sentences.*
> *... A constituent, which can be a single word but also a group of words, can be presented by the speaker as in focus by means of an accent on a single word that we will call the prosodic head of the constituent." (Nooteboom und Kruyt, 1987, S. 1513)*

Abschließend stellen sich folgende Fragen: Wenn der semantisch-pragmatische Fokus bekannt ist, kann dann die Silbe, die akustisch fokussiert wird, vorhergesagt werden? Und umgekehrt: Wenn der akustische Fokusakzent bestimmt ist, kann dann der semantisch-pragmatische Fokus abgeleitet werden? Einige grundlegende Ansätze zur Klärung der ersten Frage werden in den folgenden Abschnitten kurz vorgestellt. Wichtiger für diese Arbeit ist allerdings die zweite Frage, nämlich die praktische Nutzung von Fokusakzentinformation in einem sprachverarbeitenden System. Dieser Frage ist das 8. Kapitel gewidmet.

3.1.3 Akzent und Syntax

In der Generativen Grammatik wurden feste Regeln für die Akzentzuweisung im Englischen definiert (Chomsky und Halle, 1968): Innerhalb eines Wortes weist die *Compound stress rule* den Hauptakzent dem am weitesten *links* stehenden akzentuierten Vokal zu (1a). Innerhalb einer Phrase weist die *Nuclear Stress Rule* den Hauptakzent dem am weitesten *rechts* stehenden akzentuierten Vokal zu (1b). Diese Regeln gelten nur für unmarkierte, 'normale' Intonation (siehe auch Abschnitt 3.1.5).

(1a) There is a [blackbird.]$_{Fokus}$ → There is a **black**bird.
(1b) There is a [black bird.]$_{Fokus}$ → There is a black **bird**.

Die Anwendung dieser Regeln scheint für Chomsky und Halle (1968) obligatorisch zu sein:

> *"Once the speaker has selected a sentence with a particular syntactic structure and certain lexical items ..., the choice of stress contour is not a matter subject to further independent decision. (S. 25)"*

Ähnliche Ansätze finden sich auch in Culicover und Rochemont (1983). Ganz im Gegensatz dazu steht die Auffassung von Bolinger. In seinem bekannten Artikel "Accent is predictable (if you're a mind-reader)" äußert sich Bolinger (1972, S. 633) folgendermaßen:

> *"... Instead, accent should be viewed as independent, directly reflecting the speaker's intent and only indirectly the syntax. Accented words are points of information focus ... Location of sentence accents is not explainable by syntax or morphology."*

Nach Bolingers Ansicht ist die Positionierung des Akzents also nicht von der Syntax, sondern von der Intention des Sprechers bestimmt. Die Entscheidung des Sprechers drückt sich auch in der Wortwahl aus. Wenn ein Sprecher beispielsweise den Informationsschwerpunkt auf eine Tätigkeit legen will, wählt er wahrscheinlich eher ein 'semantisch reiches' Verb. Die Akzentuierung dieses Verbs ist dann eine weitere Konsequenz der Sprecherentscheidung:

I have a point to **emphasize** - I have a **point** to make. (Bolinger, 1972, S. 633)

In den Ausdrücken "books to write", "works to do" ist das Verb aber in hohem Grade vorhersagbar und bekommt daher keinen Akzent. Weniger gebräuchliche Verben werden dagegen seltener deakzentuiert. Bei einfachen Ausdrücken mit Verben wie "have", "there is", "do", "make", etc. ist die Tendenz höher, daß das zugehörige Substantiv akzentuiert wird. Entscheidend ist aber das relative semantische Gewicht im Kontext; je nach Kontext können Sprecher jede beliebige Stelle im Satz akzentuieren.

Nach Bolinger (1972, S. 644) ist die Verteilung von Satzakzenten nicht in erster Linie von der syntaktischen Struktur, sondern von "semantic und emotional highlighting" bestimmt. Syntax spielt nur indirekt eine Rolle: für manche Strukturen ist es wahrscheinlicher, hervorgehoben zu werden als für andere. Eine Beschreibung ist daher nur statistisch möglich, nicht als Regel oder Vorschrift.

3.1.4 Fokus und Informationsstruktur

Eine weitere Möglichkeit, die Fokussierung einer Äußerung zu beschreiben, beruht auf der Informationsstruktur eines Diskurses. In der folgenden Definition des *Informationsfokus* spielt, ebenso wie bei Bolinger (1972), die Sprecherentscheidung eine tragende Rolle:

> *"Information focus reflects the speaker's decision as to where the main burden of the message lies.... It is one kind of emphasis, that whereby the speaker marks out a part (which may be the whole) of a message block as that which he wishes to be interpreted as informative (Halliday, 1967, S. 204)."*

Der Teil einer Äußerung, der den Informationsfokus trägt, beinhaltet im allgemeinen *neue* Information für den Hörer. *Neue Information* in dieser Definition ist nicht immer völlig unbekannt; sie kann auch im Sinne von *"the speaker presents it as not being recoverable from discourse"* (Halliday, 1967, S. 204) oder als *"what the speaker assumes he is introducing into the addressee's consciousness by what he says"* (Chafe, 1976, S. 30) verstanden werden. Im Gegensatz zu *neuer* Information steht *bekannte ('given')* Information, die aus der Situation oder dem Diskurs zu erschließen oder nach Annahme des Sprechers bereits im Bewußtsein des Hörers präsent ist. Es gibt eine allgemeine Tendenz, neue Information grundsätzlich zu akzentuieren, während bekannte Information (oder eine Referenz darauf) nur in Ausnahmefällen akzentuiert wird (Chafe, 1974Nooteboom und Kruyt, 1987).

Es ist nicht eindeutig geklärt, wie weit eine Information im Diskurs zurückliegen muß, um wieder als *neu* zu gelten (Chafe, 1976). Im Sinne von Bolinger (1972) spielt wohl auch hier wieder die Sprecherentscheidung eine Rolle: Sprecher entscheiden, ob die Information neu für den Hörer ist; eine Wiederholung einer bereits erwähnten Information kann auch aus didaktischen Gründen geschehen, oder die Information wird für besonders ungewöhnlich erachtet. Akzentuierung von bereits erwähntem Material kann auch aus diskursregulatorischen Gründen vorgenommen werden, z. B. um das Rederecht weiter für sich zu beanspruchen (Couper-Kuhlen, 1986, S. 45).

Wie im vorhergehenden Abschnitt (3.1.3) beschrieben wurde, äußert sich bei Bolinger (1972) die Informationsstruktur eines Satzes auch in der Wortwahl eines Sprechers: Häufig auftretende Wörter tragen eher weniger Information, während Wörter mit geringer Auftretenswahrscheinlichkeit "points of information focus" sind und meist akzentuiert werden. Wichtig ist auch das semantische Gewicht von Wörtern; dies ist aber nicht von vornherein lexikalisch festgelegt, sondern läßt sich nur im Kontext ermitteln (in einem Bahnauskunftssystem haben beispielsweise die Wörter "Zug" und "fahren" nur ein relativ geringes semantisches Gewicht).

Das Begriffspaar *bekannte/neue Information* wird bei vielen Autoren mit *Thema-Rhema* (siehe Abschnitt 3.1.1) gleichgesetzt (z. B. Daneš, 1967). Nach Halliday (1967) dagegen ist die Einteilung in Thema/Rhema unabhängig vom vorherigen Diskursverlauf, während bekannt/neu nur vom vorher Gesagten her zu bestimmen ist: *"given: what I was talking before; theme: what I am talking about now"* (Halliday, 1967, S. 212). Die genaue Festlegung dieser Begriffe ist für diese Arbeit aber nicht entscheidend; zur weiteren Diskussion sei z. B. auf Daneš (1974) verwiesen.

3.1.5 Unterscheidung zwischen 'normalem Akzent' und Kontrastakzent

Die Vertreter der Funktionalen Satzperspektive (siehe Abschnitt 3.1.1) bezeichnen die unmarkierte Satzstruktur (Thema vor dem Rhema angeordnet) als 'normal' oder 'neutral' und entsprechend die terminale Position des Intonationszentrums als 'automatisch' (Daneš, 1960, S. 46). Dagegen spricht die Ansicht von Bolinger (1961), nach der es keine 'Normalintonation' gibt; für ihn ist im Prinzip jede Äußerung 'kontrastiv' im Sinne zu einer anderen, möglichen Aussage. Für Cruttenden (1986) gibt es eine normale Intonation

nur für isolierte, kontextfreie Sätze ('all-new' oder 'out-of-the-blue').

Nach Ladd (1980) bildet der Begriff der *Normalintonation* zumindest einen Ausgangspunkt, um überhaupt eine Beschreibung für die Intonation zu entwickeln. Er geht von der Regel aus, daß der Hauptakzent 'normalerweise' auf die letzte betonbare Konstituente gesetzt wird. Für den Akzent wird dann jeweils das letzte *akzentuierbare* Wort dieser Konstituente ausgewählt. Akzentuierbar sind im allgemeinen eher Inhaltswörter (Substantive, Hauptverben, Adjektive, Adverben) als Funktionswörter (Artikel, Präpositionen, Pronomina, etc.). Dies ist als graduelle Abstufung zu verstehen.

Weiterhin gilt aber, daß bekannte Information im allgemeinen deakzentuiert wird. Für diesen Fall formuliert Ladd (1980) die Regel, daß der 'normale Akzent' eines Satzes innerhalb eines bereits erwähnten Diskurskontextes jeweils auf das nächste benachbarte Wort (Priorität hat das rechte Nachbarwort) verschoben wird. Eine Akzentverschiebung zur Vermeidung eines Akzents auf bekannter Information zählt bei Ladd (1980) also noch zur Normalintonation; für andere Autoren scheint dieser Punkt aber nicht eindeutig geklärt zu sein (z. B. Cruttenden, 1986, S. 95).

Von diesen Regeln abweichende Akzente werden meist als Kontrastakzente bezeichnet. Couper-Kuhlen (1986) weist aber darauf hin, daß 'kontrastiv' hier auf zwei Arten verstanden werden kann: Kontrastiv kann zum einen eine Abweichung von 'normaler' Akzentuierung bedeuten, indem z. B. nicht ein Wort aus der letzten Konstituente oder ein normalerweise nicht akzentuierbares Wort akzentuiert wird. Zum anderen kann Kontrast auch als semantischer Kontrast interpretiert werden, also wenn ein Begriff einem gegensätzlichen aus einer begrenzten Menge gegenübergestellt wird, z. B. bei einer Korrekturantwort.

Dies führt oft zu Mißverständnissen, denn ein semantischer Kontrast kann auch mit 'normaler' Akzentuierung dargestellt werden, und ein von der Norm abweichender 'Kontrastakzent' ist auch ohne semantische Begründung möglich (Beispiele in Bolinger, 1961). Darüber hinaus kann semantischer Kontrast auch mit rein lexikalischen oder grammatischen Mitteln ohne besondere Akzentuierung dargestellt werden.

Ein weiteres Problem ist die Unterscheidung von *Kontrast* und *Emphase*, die in Fuchs (1976) ausführlich diskutiert wird. Oft werden die beiden Begriffe gleichgesetzt. Eine weitere Variante besteht darin, Emphase als Oberbegriff für Hervorhebung im Sinne von 'Satzakzent' im Vergleich zum Wortakzent zu sehen. Weiterhin gibt es die Möglichkeit, *Kontrast* nur für semantisch motivierten Kontrast zu verwenden, während der Begriff *Emphase* eine besonders auffällige akustische Realisierung (d. h.maximaler Einsatz der akustischen Parameter) impliziert, z. B. bei Ausrufen mit hohem emotionalem Gehalt.

Zur Vereinfachung dieser Begrifflichkeit wurden von Ladd (1980) die Begriffe *enger* und *weiter* Fokus eingeführt. Die 'normale' Akzentuierung steht für einen weiten Fokus, der auf der gesamten Konstituente liegt. Eine kontrastive Akzentuierung schränkt den Fokusbereich auf ein Wort oder eine bestimmte Silbe ein - dies wird als enger Fokus definiert. In dieser Definition wird akustische Auffälligkeit nicht zwingend mit dem semantischen Hintergrund des Akzents verknüpft. Nach Ladd (1980, S. 213) ist allein die *Position* des Akzents entscheidend für die Markierung von engem Fokus. *Emphase* ist bei Ladd die Verstärkung von akustisch-prosodischen Mitteln; sie tritt zwar sehr häufig zusammen mit engem Fokus auf, dies ist aber nicht obligatorisch.

3.2 Automatische Zuweisung von Akzenten

An dieser Stelle sollen noch einige neuere Verfahren vorgestellt werden, die sich mit der automatischen Zuweisung von Akzenten befassen. Bei den verwendeten Strategien werden unterschiedliche Schwerpunkte gesetzt; Grundlage sind aber jeweils die bereits diskutierten Faktoren, die verantwortlich für eine Akzentrealisierung sind:

1. Syntaktische Struktur (Wortart, Position im Satz)
2. Informationsstruktur (bekannte/neue Information für den Hörer)
3. Sprecherentscheidung (was ist für den Sprecher wichtig?)

Ein eher syntaktisch orientiertes Verfahren verwenden Batliner et al. (1998b). Das Verfahren dient zur automatischen Etikettierung von deutscher Spontansprache, um die Menge der verfügbaren Trainingsdaten für eine automatische Akzenterkennung zu erhöhen (siehe Abschnitt 5.4). Eingabeinformationen sind die jeweils aktuelle Äußerung, syntaktisch-prosodische Phrasengrenzen und automatisch annotierte Wortarten. Ein Schwerpunkt liegt auf der gesonderten Behandlung von Fokuspartikeln und Fragewörtern. Mit diesem Ansatz können keine kontrastiven oder emphatischen Akzente etikettiert werden, ebenfalls kann ohne Kontext auch keine Deakzentuierung von bekannter Information berücksichtigt werden (siehe auch Abschnitt 3.1.5). Dennoch werden 88 % der Akzente korrekt (im Vergleich zur manuellen Etikettierung; siehe Abschnitt 5.4) etikettiert (Batliner et al., 1998b).

Einen erweiterten Ansatz verfolgen Delin und Zacharski (1997). Ihr System dient zur Modellierung von Intonationskonturen in einem Dialogsystem. Zusätzlich zu syntaktischen Informationen (Wortart und Satzposition) integrieren die Autoren pragmatische Konzepte. Zum einen wird der *kognitive Status* innerhalb einer Äußerung berücksichtigt. Dies bezieht sich direkt auf das Konzept von given/new (siehe Abschnitt 3.1.4). Mit Hilfe einer *Givenness Hierarchy* (Gundel et al., 1993) wird für jedes Wort ein kognitiver Status bestimmt: Neu eingeführte Begriffe bekommen dann einen relativ hohen Wert, während soeben erwähnte Begriffe einen sehr niedrigen Wert bekommen. Dieser kognitive Status wird ständig aktualisiert. Eine dem kognitiven Status angepaßte Information nimmt Rücksicht auf den bereits vorhandenen Informationsstand des Hörers und erleichtert somit das Verständnis.

Zum anderen wird versucht, den *Informationswert* einer Äußerung zu berücksichtigen. Damit ist gemeint, daß ein Sprecher eine Äußerung so effizient gestaltet, daß Hörer wichtige Informationen gut identifizieren können. Daher spielt auch die Sprecherentscheidung eine Rolle (welche Information ist für den Sprecher wichtig?). Diese ist allerdings schwierig zu bestimmen, wenn die Ziele des Sprechers nicht bekannt sind. Zur adäquaten Umsetzung müßte ausreichend Sprecherinformation zur Verfügung stehen (Interessen, bevorzugte Stilmittel zur Markierung, emotionaler Zustand, aktuellen Sprecherwissen, etc.). Der hier verfolgte Ansatz geht zunächst nur davon aus, daß ein Sprecher *kooperativ* ist: Dieser markiert beispielsweise nur die Teile einer Äußerung, die im aktuellen Situationskontext

den Hörer auf entscheidende Unterschiede aufmerksam machen (sog. 'contrastive properties' bei Prevost und Steedman, 1994). Ähnliche Ansätze finden sich auch in Hiyakumoto et al. (1997) und Terken et al. (1997).

In van Deemter (1998) werden ebenfalls syntaktische und semantisch/pragmatische Informationen zur automatischen Akzentzuweisung verwendet. Der Ansatz beruht auf einer Einteilung in 'Akzent hervorrufende' und 'Akzent verhindernde' Faktoren. Zur ersten Kategorie gehören Neuheit einer Information oder ein Kontrast zu einer vorhergehenden Information, während bekannte Information zur zweiten Kategorie tendiert. Zur Verhinderung von unerwünschten Akzenten (zu viele Akzente führen zu mangelnder Akzeptanz des Systems) werden nicht nur wörtliche Wiederholungen, sondern auch Referenten darauf (Synonyme, Pronomina) deakzentuiert.

3.3 Fokusmarkierung im Sprachvergleich

Der Schwerpunkt der veröffentlichten Arbeiten zum Thema Fokus liegt im englischsprachigen Raum. Darüber hinaus gibt es viele Untersuchungen für Deutsch, Niederländisch und Schwedisch. Diese germanischen Sprachen haben gewisse Ähnlichkeiten in bezug auf ihre Fokusmarkierung. Im Gegensatz dazu befinden sich romanische Sprachen, in denen Fokusakzente in deutlich anderem Maße verwendet werden als in der germanischen Sprachgruppe.

In Hirst und di Cristo (1998) werden die Intonationsmuster für 20 verschiedene Sprachen vorgestellt. Hirst und di Cristo (1998) definieren hier eine 'normale' Akzentuierung als *basic neutral unmarked intonation pattern*. Für die dargestellten Sprachen wird dieses neutrale Intonationsmuster und ein davon abweichendes Emphase/Kontrast-Muster beschrieben. Für die indo-europäischen Sprachen beobachten die Autoren grundsätzliche Unterschiede in der rhythmischen Gliederung von akzentuierten und nicht-akzentuierten Silben (siehe auch Abschnitt 2.5.2): In der germanischen Sprachgruppe wird eine Intonationsgruppe von einer akzentuierten Silbe angeführt, der unakzentuierte Silben folgen (*left-headed foot*). In den romanischen Sprachen gibt es dagegen eher die Tendenz, daß eine Intonationsgruppe von einer akzentuierten Silbe beendet wird (*right-headed foot*).

3.3.1 Deakzentuierung

Eine verbreitete Annahme aus 'germanozentrischer' Sicht ist, daß die Positionierung von Akzenten in erster Linie vom informativen Gehalt eines Wortes oder einer Konstituente beeinflußt wird. Akzente werden daher im allgemeinen nicht auf wiederholte Wörter in einem Dialogkontext (siehe Abschnitt 3.1.4) gesetzt. Die Wörter, die bereits Bekanntes beinhalten, werden jeweils deakzentuiert. Ganz anders sieht dies aber in den romanischen Sprachen aus.

In Ladd (1996) werden Beispiele aus dem Rumänischen oder Italienischen gegeben, in denen Deakzentuierung nicht akzeptabel wäre. Weitere Beispiele für Spanisch finden sich in Delin und Zacharski (1997). Auch eine Verlagerung des lexikalischen Wortakzents zur Darstellung eines Korrekturkontrastes ist in den romanischen Sprachen nicht möglich:
"This whisky wasn't EXported, it was DEported." (Bolinger, 1961, S. 83)

Kontraste können in den romanischen Sprachen mit veränderter Wortstellung dargestellt werden, z. B. mit sog. *right-dislocation* (Ladd, 1996). Dabei wird eine Konstituente nach rechts ans Satzende verlagert und durch ein Pronomen ersetzt:
"Adesso faccio scorrere il TUO, di bagnetto" (Ladd, 1996, S. 179)
(Nun werde ich DIR ein Bad einlassen.)

Es gibt also Sprachen, in denen der relative Informationsgehalt für die Akzentuierung relevant ist, in anderen gilt dies nicht. Ladd (1996) definiert entsprechend zwei Sprachgruppen: Es gibt *deakzentuierende* Sprachen, die es erlauben oder sogar vorziehen, wiederholtes oder bekanntes Material zu deakzentuieren. Dazu zählen in erster Linie die germanischen Sprachen wie Deutsch, Englisch, Niederländisch oder Schwedisch. In *nicht-deakzentuierenden* Sprachen dagegen wird eine Deakzentuierung oft als unakzeptabel empfunden; diese verwenden zur Darstellung von Korrekturakzenten oder Kontrast eine andere Wortstellung. Dies ist bei den romanischen Sprachen der Fall, gilt aber beispielsweise auch für Sprachen wie Türkisch oder Ungarisch.

3.3.2 Optionaler Gebrauch von verschiedenen Satzakzenttypen

Auch zwischen sehr verwandten Sprachen bzw. Dialekten kann es starke Unterschiede in der Intonationsmarkierung geben. Eine Untersuchung für Dänisch (verschiedene Dialekte), Schwedisch und Deutsch (Norddeutsch) wurde von Grønnum (1989Grønnum (1990) durchgeführt. Es handelte sich dabei um gelesene Sätze, die von der Syntax und Wortstruktur her in allen Sprachen ähnlich waren.

Grønnum definiert als *Satzakzent* die akzentuierten Wörter eines Satzes (eins oder mehrere), die *perzeptiv* aus den anderen akzentuierten Wörtern herausragen. Dies geschieht im wesentlich durch Veränderungen der Intonation. Grønnum (1990) unterscheidet dabei zwei Formen des Satzakzentes:

"default accent": prosodisch/syntaktisch bedingt, nur äußerungsfinal
"focal accent": semantisch/pragmatisch, kontextuell bedingt, an allen Positionen möglich

Sie äußert die Hypothese, daß diese beiden Satzakzenttypen in äußerungsfinaler Position perzeptiv unterscheidbar sind; sie bezweifelt aber, daß diese Unterschiede ohne Kontext wahrnehmbar sind (Grønnum, 1990, S. 190). Insgesamt ist der Satzakzent in initialer oder medialer Stellung erheblich auffälliger als in finaler Position. Grønnum (1990) macht dafür den folgenden Kontext verantwortlich, der nach dem Satzakzent intonatorisch stark abgeschwächt wird. Nur in wenigen Fällen wird in ihren Sprachdaten auch der Kontext *vor* dem Satzakzent intonatorisch abgeschwächt.

Grønnum (1990) stellt aber erhebliche Unterschiede in der Anwendung bzw. Existenz von "default" und "focal" Akzenten in den untersuchten Sprachvarianten fest. In ihren deut-

schen Sprachdaten findet sie keine wahrnehmbaren Unterschiede zwischen “default” und finalen “focal” Akzenten, beide Formen sind intonatorisch sehr schwach ausgeprägt. Die nichtfinalen fokalen Akzente im Deutschen zeichnen sich durch eine verstärkt akzentuierte Silbe aus, der verbliebene Äußerungsrest wird deutlich abgesenkt/deakzentuiert.

Die von Grønnum (1990) untersuchten Sprachgruppen sind unterschiedlich stark ‘prosodisch expressiv’; ihre Gruppierungen sind eher durch geographische Nähe als durch die zugrunde liegende Sprache begründet. Die stärkste prosodische Variation findet sich in Stockholm-Schwedisch und Bornholm-Dänisch. Hier gibt es auch stark wahrnehmbare Unterschiede in der Ausprägung von finalen Satzakzenten: “focal” Akzente sind akustisch stärker markiert als “default” Akzente. Eine mittlere Gruppe bilden die norddeutschen Varianten, gewisse prosodische Ähnlichkeiten finden sich auch in Sønderborg-Dänisch. Eine eher ‘flache’ Prosodie herrscht in den dänischen Dialekten (dazu kann auch Malmö-Schwedisch gruppiert werden) vor. In den dänischen Sprachdaten von Grønnum (1990) fanden sich beispielsweise zum großen Teil überhaupt keine finalen Satzakzente.

3.4 Fokusrealisierung von Konzepten

In diesem Abschnitt soll auf Untersuchungen eingegangen werden, die sich mit den semantisch-pragmatischen Konzepten *given/new* und *Kontrast* beschäftigen. Dabei wurde sowohl die Produktion als auch die Perzeption untersucht. Einen Sonderbereich stellt die Synthese von Fokusakzenten dar; dort wurden verschiedene akustische Realisierungen daraufhin getestet, inwieweit sie als akzeptabel von Hörern wahrgenommen werden.

3.4.1 Bekannte und neue Information

Wie bereits in Abschnitt 3.1.4 erwähnt wurde, kann sich die Informationsstruktur auf die Fokussierung einer Äußerung und damit auf die Position des Fokusakzents auswirken. *Neue Information* innerhalb eines Diskurses steht im allgemeinen im Fokus und wird dementsprechend akzentuiert. Bereits erwähnte Information wird dagegen meist deakzentuiert. Dies kann nur innerhalb eines Kontextes bestimmt werden, bei isolierten Sätzen wird immer von *neuer* Information ausgegangen.

Nooteboom und Kruyt (1987) stellten eine Diskrepanz zwischen diesen von Linguisten allgemein anerkannten Regeln und der aktuellen Realisierung beispielsweise von Sprechern im Radio fest. Offenbar führt eine Akzentuierung, die die Informationsstruktur nicht beachtet, noch nicht unbedingt dazu, daß Sprecher mißverstanden werden. Daher wollten die Autoren gezielt untersuchen, inwieweit sich die Informationsstruktur eines Diskurskontextes auf die Akzentuierung überhaupt auswirkt. Die wesentlichen Fragen waren:

1. Wie wichtig sind Akzente für die Wahrnehmung von ‘given/new’?
2. Unter welchen Bedingungen werden ‘Akzentfehler’ (also akzentuiertes ‘given’ oder deakzentuiertes ‘new’) akzeptiert?

Nooteboom und Kruyt (1987) wandten folgende Methode zur Untersuchung ihrer Hypothesen an: Sie entwarfen kurze Diskurse mit zwei aufeinander folgenden Äußerungen

im Nachrichtenstil. Diese wurden von einem Sprecher neutral gelesen (in niederländischer Sprache), alle Sätze wurden resynthetisiert und dabei mit verschiedenen Akzentkombinationen versehen. Versuchspersonen sollten dann die Akzeptanz der Akzentrealisierungen beurteilen.

Als Ergebnis stellten Nooteboom und Kruyt (1987) fest, daß ein Fokusakzent auf bekannter Information im allgemeinen eher akzeptiert wird als eine Deakzentuierung von neuer Information. Im ersteren Fall gilt dies aber nur für bestimmte Satzstrukturen. Eine weitere Annahme war unter anderem, daß Referenzwörter (Umschreibungen oder Pronomina für einen bereits genannten Begriff) nicht akzentuiert werden (z. B. Oxford (Original) - the city (Referenz)). Diese wurde widerlegt: Ein Referenzwort auf bekannter Information wurde in akzentuierter Form akzeptiert, die bekannte Information im gleichen Wortlaut galt mit Fokusakzent als weniger akzeptabel.

3.4.2 Kontrastfokus

Krahmer und Swerts (1998) untersuchten den Kontrastfokus für Niederländisch. Die Autoren gehen von folgenden Definitionen aus: Akzente können entweder *neue* Information markieren oder einen Kontrast zwischen dem akzentuierten Element und einer beschränkten Menge von Alternativen aufzeigen. ('Kontrastiv' wird hier also semantisch, d. h. inhaltsbezogen, definiert). Grundlage ist ebenfalls die These von Ladd (1980), nach der enger Fokus oft semantisch kontrastiv interpretiert wird; dies ist aber auch abhängig davon, ob sich der Fokusakzent in 'default'-Position (also definitionsgemäß eher am Ende eines Satzes; siehe auch Abschnitt 3.1.5) befindet oder nicht. Krahmer und Swerts (1998) wollten darüber hinaus noch folgende Hypothesen untersuchen:

1. Sind kontrastive Akzente akustisch auffälliger als 'Neuheits'-Akzente?,
2. Haben kontrastive Akzente eine spezielle Intonationskontur?

Die Untersuchungen wurden für eine semi-spontane Dialogsituation durchgeführt. Die Dialogpartner mußten einander unterschiedlich farbige geometrische Figuren benennen. Im Dialogkontext können also neue Farben und Formen ('new') auftauchen bzw. bereits bekannt sein ('given'). *Kontrastiv* war eine Äußerung, wenn ein beschriebenes Objekt mit einem vorher erwähnten in einer Eigenschaft nicht übereinstimmte, also 'kontrastierte'. Akzente konnten auf einem Adjektiv (Farbe) und/oder einem Substantiv (geometrische Form) liegen. Es ließ sich nicht für jeden Satz aufgrund seiner Syntax entscheiden, ob es sich um engen (Kontrastfokus) oder weiten Fokus ('Neuheits'-Fokus) handelt. Diese 'mehrdeutigen' Sätze wurden speziell untersucht.

Die Ergebnisse konnten beide Hypothesen bestätigen. Kontrastive Akzente waren tatsächlich akustisch auffälliger als Neuheits-Akzente. Dies galt aber nicht für eine isolierte Darbietung der akzentuierten Wörter, Hörer konnten in einem entsprechenden Perzeptionsexperiment keine Unterschiede erkennen. Der prosodische Kontext, vor allem die Deakzentuierung der folgenden Wörter, scheint also wesentlich für eine kontrastive Interpretation zu sein. Die Form der Akzente im Grundfrequenzverlauf war interessanterweise für die Substantive bei beiden Akzenttypen (kontrastiv und neu) ähnlich, unterschied sich aber für die Adjektive. Die kontrastive Interpretation wird hier also durch die Position

des Akzents begünstigt, d. h. bei einem Adjektiv liegt er nicht auf einer 'default'-Position und fällt dadurch deutlich heraus.

Die erzielten Ergebnisse gelten in erster Linie für *deakzentuierende* Sprachen mit fester Wortstellung (siehe auch Abschnitt 3.3.1). Um die Unterschiede zu einer nicht-deakzentuierenden Sprache näher zu untersuchen, führten die Autoren ein weiteres Experiment mit italienischen und niederländischen Sprechern durch (Swerts et al., 1999).

Der Versuchsaufbau war sehr ähnlich, wieder mußten geometrische Formen benannt werden, verschiedene Akzentkombinationen auf Adjektiv und/oder Substantiv waren möglich. Gemeinsamkeiten in der Verteilung der Akzente gab es nur zu Beginn der Dialoge: Im Stadium 'all new' wurden in beiden Sprachen Akzente auf Adjektiv und Substantiv gesetzt. Im weiteren Dialogverlauf markierten die niederländischen Sprecher die Informationsstruktur durch Deakzentuieren von gegebener Information. Bei den italienischen Sprechern dagegen wurde in jedem Kontext die gegebene Information akzentuiert.

Perzeptionsexperimente zeigten allerdings, daß italienische Hörer die Akzente auf gegebener Information von denen auf neuer Information unterscheiden konnten: Es gab eine graduelle Abstufung, in der Form, daß 'given'-Akzente akustisch schwächer augeprägt waren als 'new'-Akzente. Offensichtlich werden im Italienischen die prosodischen Mittel zur Markierung der Informationsstruktur weniger stark genutzt als im Niederländischen. Für diese Markierung wird im Italienischen eher die freie Wortstellung ausgenutzt (siehe auch Abschnitt 3.3.1). Das konnte in diesem Experiment durch die notwendige Beschränkung des Versuchsaufbaus nicht gezeigt werden, bietet aber Ansätze zu weiteren Untersuchungen (Swerts et al., 1999).

3.4.3 Markierung von Fokusakzenten in der Synthese

Mit Hilfe einer prominenzbasierten Synthese (Portele und Heuft, 1997) wurde für deutsche Sprachdaten untersucht, ob sich verschiedene Fokustypen perzeptiv unterscheidbar synthetisieren lassen. Die Prominenz dient hier als quantitative Beschreibung von linguistischen Konzepten. Die von einer linguistischen Komponente vorgegebenen Prominenzwerte werden von der Synthese in die entsprechenden akustischen Korrelate umgesetzt.

Eine erste Untersuchung (Wolters und Wagner, 1998) befaßte sich mit verschiedenen Arten von sog. 'Antwortfokus'. Entsprechend gestellte Fragen bestimmen einen weiten Fokus (Fokus liegt auf ganzem Satz) oder einen engen Fokus (Fokus liegt auf Subjekt oder Objekt). Der enge Fokus wurde als kontrastiv definiert, wenn die Antwort eine direkte Korrektur der Frage darstellte. Im Versuch wurden verschiedene Sätze synthetisiert, denen von Hörern passende Fragen zugeordnet werden sollten. Als Grundprominenzwerte wurden die Werte einer vorher gesprochenen natürlichen Äußerung verwendet, nur die Prominenzwerte der zu fokussierenden Wörter wurden in verschiedenen Stufen manipuliert.

Insgesamt zeigte es sich, daß enger Fokus mit dieser Methode erfolgreicher zu erzeugen war als weiter Fokus, denn ein beabsichtigter enger Fokus wurde sehr viel besser erkannt. Ein weiteres Ergebnis war eine starke Asymmetrie beim Erkennen der Fokusposition: ein

beabsichtigter Fokus auf einem Subjekt wurde sehr viel schlechter erkannt als ein Objektfokus, selbst bei besonders hohen Prominenzwerten auf dem Subjekt. Ein zusätzlich kontrastiver Fokus wurde von den Hörern nur bei den höchsten Prominenzwerten wahrgenommen.

Im nächsten Experiment von Wagner (1999) wurde explizit versucht, einen Kontrastfokus perzeptiv wahrnehmbar zu synthetisieren. Aufgrund des vorherigen Experiments (Wolters und Wagner, 1998) sollte der Einfluß der Prominenz nicht nur lokal, sondern auch im globalen Zusammenhang untersucht werden. Als Kontrastfokus wurde hier in erster Linie ein Korrekturkontrast untersucht, also Fokussierung einer gegensätzlichen Information im Diskursverlauf.

Im Experiment wurden synthetisierte Sätze mit Kontrastfokus an verschiedenen Positionen mit unterschiedlichen Manipulationsmethoden erzeugt. Diesen wurden entsprechende Fragen vorangestellt, die einen Kontrast evozieren sollten. Hörer sollten daraufhin beurteilen, ob die Fokussierung (also die Korrektur) zur Frage paßte oder nicht. Es wurden verschiedene Strategien zur Synthese des intendierten Kontrastes eingesetzt. Als Grundlage dienten immer Prominenzwerte, dem intendierten Kontrast wurde jeweils der höchste Prominenzwert zugeordnet. Im Ergebnis schnitt die Strategie am besten ab, die mehrere Manipulationen kombinierte:

Der fokussierten Silbe wurde die höchste Prominenz zugewiesen, die Dauer wurde zusätzlich verlängert, und die anderen akzentuierten Silben wurden deakzentuiert. Dabei zeigte sich kein Unterschied, ob alle Silben der Phrase oder nur die nachfolgenden deakzentuiert wurden. Am schlechtesten waren die Ergebnisse, wenn jeweils auf die Deakzentuierung oder auf die Dauermanipulation verzichtet wurde. Es waren aber wieder Unterschiede in der Position des intendierten Kontrastes auszumachen: Satz-initiale Kontraste wurden weniger erfolgreich synthetisiert als satz-finale. Ähnliche Ergebnisse in dieser Hinsicht finden sich auch in Portele (1999).

3.5 Fokusrealisierung und Satzmodus/Satzposition

Offensichtlich hat die syntaktische Umgebung einen gewissen Einfluß auf die akustische Realisierung von Fokusakzenten. Dazu gehört vor allem die Position des Fokusakzents im Satz oder in einer Phrase, aber auch der Satzmodus. Diese Abhängigkeiten wurden von der Forschungsgruppe um Eady und Cooper ausführlich experimentell untersucht (für amerikanisches Englisch). Ein Problem ihres Ansatzes scheint allerdings zu sein, daß sie von direkten akustischen Korrelaten für linguistische Variablen ausgehen. Eine grundsätzliche Kritik zu ihrer Vorgehensweise findet sich beispielsweise in Ladd (1996, S. 20 ff.).

Die Gewinnung der Daten lief stets nach einem einheitlichen Schema ab. Es wurden einfache Sätze konstruiert, die von verschiedenen Sprechern gelesen wurden. Die Fokusmarkierung wurde durch entsprechende Fragen evoziert, so daß Fokusakzente entweder Subjekt, Verb oder Objekt hervorhoben (oder eine Kombination bei Doppelfokus). Alle Sätze wurden perzeptiv auf korrekte Fokusrealisierung überprüft. Die Sätze hatten eine einheitliche syntaktische Struktur, nur bestimmte Schlüsselwörter wurden ausgewechselt. Dies waren

stets Inhaltswörter (Substantiv, Verb oder Adjektiv). Für jedes dieser Schlüsselwörter wurde eine Messung der Dauer und der Grundfrequenz vorgenommen.

3.5.1 Enger Fokus in unterschiedlicher Satzposition

In einer ersten Untersuchungsreihe dieser Art (Cooper et al., 1985) sollten Fokusakzente an verschiedenen Positionen im Satz untersucht werden. Es wurde speziell enger Fokus auf genau einem Wort (hier von den Autoren als 'kontrastiv' definiert) untersucht; dieser wurde durch eine Alternativfrage evoziert. Die verwendeten Satzstrukturen sahen beispielsweise so aus (Schlüsselwörter erscheinen kursiv):

Chuck liked the *present* that *Shirley* sent to her *sister*.
The *fish* will be *fresh* and *cheap* at this *restaurant*.
Kate went to *Kansas* with *Jack* and *Peter*.

Es wurden Grundfrequenz und Dauer für jedes Schlüsselwort gemessen. Bei der Grundfrequenz wurde der Gipfelwert für jedes Schlüsselwort ermittelt. Es scheint allerdings nicht ganz unproblematisch, Sätze mit so unterschiedlicher Struktur miteinander zu vergleichen; es ist anzunehmen, daß die deutliche syntaktische Grenze in der ersten Satzstruktur (bei "that") die Intonationskontur stark beeinflußt.

Meßergebnisse für die Dauer

Insgesamt waren die fokussierten Wörter in jeder Position länger. Diese Verlängerung war besonders stark für initiale und mediale Position, für finale Satzposition aber eher schwach. Cooper et al. (1985) nehmen an, daß dies an der *finalen Längung* liegt, die jeweils am Satzende anzutreffen ist; bei zusätzlicher Fokussierung wird dann die Dauer nur geringfügig erhöht. Fokusmarkierung und satzfinale Längung wirken sich also nicht unbedingt additiv aus. Der Prozentsatz, um den die Dauer erhöht wird, ist bei mehrsilbigen Wörtern geringer; offensichtlich wird nur die Dauer der akzentuierten Silbe verlängert.

Meßergebnisse für die Grundfrequenz

Nach dem fokussierten Wort sinkt die Grundfrequenz stark ab (sog. *post-focal drop*). Eine Deklination findet vor und nach dem Fokusakzent statt. Der Grundfrequenzgipfel des Akzents unterbricht deutlich den Gesamtverlauf der Kontur. Der satzfinale Fokus hat einen weniger deutlichen F_0-Gipfel. Der satzinitiale Fokus im Subjekt unterscheidet sich in der F_0-Höhe nicht wesentlich von anderen initialen Subjekten; der Unterschied liegt hier wohl eher in der Dauer.

In einem Nachfolgeexperiment wurden die gleichen Untersuchungen für deutlich längere Sätze durchgeführt, außerdem wurde noch ein *neutraler* Fokus (im Sinne von weitem Fokus) untersucht. Die Messungen für die Dauer waren ähnlich wie im vorherigen Experiment, auch für den neutralen (weiten) Fokus änderte sich nichts. Kontrastiver (enger) Fokus schien hier eher lokal zu sein, die anderen Wörter wurden nicht davon beeinflußt.

Bei der Grundfrequenz ließen sich allerdings deutliche Unterschiede erkennen: Nach dem (eng) fokussierten Wort fiel die Grundfrequenz sehr stark ab. Ein fokussiertes Wort am Satzanfang bewirkte niedrigere Werte für alle folgenden F_0-Gipfel.

3.5.2 Position des Fokus in Aussagen und Fragen

In dieser Experimentreihe untersuchten Eady und Cooper (1986) Fokusakzente in unterschiedlichen Satzpositionen für Fragen und Aussagen. Als Fragentyp wurden Echofragen verwendet, die jeweils syntaktisch identisch mit der Aussagen waren und sich nur durch die Frageintonation von ihnen unterschieden. Sogenannnte W-Fragen haben dagegen ein ähnliches Intonationsmuster wie Aussagen. Die Autoren entwarfen verschiedene Satztypen für folgende Konstellationen:

Aussage/Frage mit weitem Fokus (final)
Aussage/Frage mit engem Initial-Fokus
Aussage/Frage mit engem Final-Fokus

Für alle Inhaltswörter wurden wieder Dauer und Grundfrequenz gemessen. Bezüglich der Dauern zeigten sich keine grundsätzlichen Unterschiede zwischen Aussagen und Fragen: Fokussierte Wörter waren im allgemeinen von längerer Dauer. Dies zeigte sich deutlicher in initialer Satzposition als in finaler (ähnlich wie in Cooper et al., 1985).

Bei der Grundfrequenz zeigten sich erwartungsgemäß Unterschiede zwischen Aussagen und Fragen, insbesondere im hinteren Bereich der Äußerung (weiter und enger Final-Fokus). Bei den Fragen war für alle Fokustypen der letzte Grundfrequenzgipfel deutlich höher als bei den Aussagen. Bei engem Initial-Fokus waren keine Unterschiede im fokussierten Wort selbst zu erkennen, im weiteren Verlauf fiel jedoch bei den Aussagen die Grundfrequenzkontur auf eine niedrigeres Niveau ab, während bei den Fragen die auf den Fokusakzent folgenden Grundfrequenzwerte weiterhin auf einem relativ hohen Niveau blieben.

Insgesamt wird also die Realisierung von Fokusakzenten von der Position des Fokus im Satz und vom Satzmodus beeinflußt. Die Dauer wirkt sich im allgemeinen nur auf das fokussierte Wort selbst aus, wohingegen die Grundfrequenzkontur vom Fokus global modifiziert wird.

3.5.3 Enger und weiter Fokus, einfacher und doppelter Fokus

Eady et al. (1986) gingen von der Annahme aus, daß es unterschiedliche Intonationsmuster für engen und weiten Fokus gibt (die perzeptive Unterscheidbarkeit spielte hier keine Rolle). Akustische Differenzen zwischen engem und sog. neutralem Fokus wurden bereits in einer vorherigen Untersuchung (Cooper et al., 1985) festgestellt. Die Begriffe ‘weit’ und ‘neutral’ werden bei den Autoren für unterschiedliche Abstufungen im Skopus des Fokus verwendet.

Im ersten Experimentteil untersuchten Eady et al. (1986) neutralen, weiten und engen Fokus. Alle Fokusvarianten wurden nach ihrer Definition durch den Skopus (Geltungs-

bereich) bestimmt. Die Sätze, die als Untersuchungsmaterial dienten, hatten wieder eine feste Struktur, an den Inhaltswörtern wurden jeweils Messungen für Grundfrequenz und Dauer vorgenommen:

Subjekt + Verb + direktes Objekt + Präpositionalobjekt
Die Struktur der evozierenden Fragen sah dann folgendermaßen aus:
neutraler Fokus: Was passierte? (Fokus auf ganzem Satz)
weiter Fokus: Was tat das Subjekt? (Fokus auf Verbalphrase)
enger Fokus: direkte Frage nach dem Präpositionalobjekt (Fokus auf Präp.Obj.)

Die Vergleiche zwischen den verschiedenen Fokusrealisierungen auf dem letzten Inhaltswort der Sätze zeigten folgende Ergebnisse: Erwartungsgemäß war der enge Fokus deutlich durch einen höheren F_0-Gipfel und eine längere Dauer akustisch markiert. Weiter Fokus im Vergleich zu neutralem Fokus tendierte zu einer Verlängerung der Dauer im gesamten fokussierten Bereich, zeigte aber keine Unterschiede in der Grundfrequenz.

Im zweiten Experimentteil zum Unterschied zwischen 'neutralem Einzelfokus' und 'doppeltem engem Fokus' wurden wieder entsprechende Sätze konstruiert:

Subjekt + Objekt + Präpositionalobjekt
Die Fragen wurden so konstruiert, daß neutraler Fokus (finaler *default accent*), Initialfokus (Akzent auf Subjekt), Finalfokus (Akzent auf Präp.Obj.) und Doppelfokus (Akzente auf Subjekt und Präp.Obj.) evoziert wurden.

Beim Doppelfokus verhielten sich die beiden fokussierten Wörter wie die jeweils einzeln fokussierten Wörter, es fand offensichtlich keine gegenseitige Beeinflussung statt. Der finale Einzelfokus hatte dabei die gleichen Werte wie das zweite Element des Doppelfokus; ebenso verhielt sich der Initialfokus ähnlich wie das erste Element des Doppelfokus. Die Unterschiede zeigten sich eher an nichtfokussierten Wörtern: der Initialfokus bewirkte eine starke Deakzentuierung des letzten Wortes (geringere Dauer und Grundfrequenz), während nach dem ersten Element des Doppelfokus keine typische Grundfrequenzabsenkung stattfand.

3.5.4 Fokustypen in verschiedenen Satzstrukturen

Bei weitem Fokus fällt der Akzent im allgemeinen auf das letzte Inhaltswort einer Intonationsgruppe; eine Ausnahme davon stellen sog *Ereignissätze* dar (Cruttenden, 1986). Diese enthalten ein intransitives Verb, das Subjekt ist oft keine menschliche Person: *"Der Zug kommt."* Hier bekommt zur Umsetzung des weiten Fokus das Subjekt 'normalerweise' den Akzent.

Eine Untersuchung zu dieser Problematik wurde von Hoskins (1996) durchgeführt. Im Gegensatz zu den oben beschriebenen Experimenten von Eady und Cooper untersuchte er engen und weiten Fokus in intransitiven Sätzen für 3 verschiedene Satztypen. Seine Annahme war, daß die Markierung für weiten Fokus davon abhängig ist, ob es sich beim Subjekt um ein Agens (bewußter Verursacher eines Geschehens) handelt:

passiv:	Subjekt ist kein Agens	“the letter was mailed”
intransitiv:	Subjekt ist kein Agens	“the water boiled”
intransitiv:	Subjekt ist Agens	“the daughter jogged”

Alle Sätze wurden von Versuchspersonen jeweils mit weitem Fokus (Frage: “was passierte?”) und engem Fokus auf Subjekt oder Verb gelesen. Für alle drei Bedingungen wurden Dauer und F_0 für Subjekt und Verb gemessen. Für engen Fokus entsprachen die Ergebnisse den Erwartungen: beide akustischen Parameter waren hier maximal für das Subjekt bzw. das Verb im engen Fokus.

Für weiten Fokus waren folgende Ergebnisse festzuhalten: Für die Satztypen *passiv* und *intransitiv-nicht-agens* waren die akustischen Parameter Dauer und F_0 für das Subjekt maximal, die Verben wurden dagegen deakzentuiert. Beim Typ *intransitiv-agens* bekam dagegen das Verb den Hauptakzent.

Eine weitere Untersuchung von Hoskins (1997) beschäftigt sich ebenfalls mit engem und weitem Fokus in Sätzen mit unterschiedlicher Objektstruktur. Die Annahme war hier, daß es keine akustischen Unterschiede zwischen engem und weitem Fokus bei Verbalphrasen mit direktem Objekt gibt, wohl aber bei Verbalphrasen mit Präpositionalobjekt. Das beinhaltete auch den Kritikpunkt, daß der Objekttyp in der Untersuchung von Eady et al. (1986) nicht berücksichtigt wurde.

Beispiel: *He was eating a banana.* vs. *He was eating at the dinner.* (Hoskins, 1997)

Die Sätze wurden wieder mit 3 Variationen gelesen: weiter Fokus (Default-Fokus auf Objekt), enger Fokus auf Verb und enger Fokus auf Objekt. Die statistischen Auswertungen zeigten, daß enger und weiter Fokus immer akustisch zu unterscheiden waren, unabhängig vom Objekttyp.

3.6 Automatische Erkennung

Frühere Arbeiten zur Fokuserkennung befaßten sich in erster Linie mit gelesener Sprache, um bestimmte Phänomene gezielt zu untersuchen. Im folgenden soll von zwei größeren Projekten berichtet werden, die sich mit der automatischen Erkennung von Fokusakzenten für das Deutsche befaßten (Bannert, 1991Batliner, 1989b).

3.6.1 Das MAFID-System

Im Rahmen eines DFG-Projekts wurde ein Modell für ein sprachverstehendes Dialogsystem entwickelt (Hoepelman und Machate, 1994). Der Schwerpunkt lag auf der Untersuchung der Fokusintonation im gesprochenen Dialog. Die Fokussierung wurde sowohl auf prosodischer Ebene als auch im Hinblick auf die semantische Steuerung des Dialogs untersucht. Es wurde ein Verfahren zu Erkennung der Fokusakzente im gesprochenen

Satz entwickelt und implementiert (Bannert, 1991). Außerdem wurde ein Modell zur Interpretation der Fokusintonation entworfen und implementiert. Das daraus entstandene MAFID-System (Hoepelman und Machate, 1994) ist eine prototypische Implementierung eines Sprachdialogsystems, das Information aus der intonatorischen Fokussierung im Dialog verarbeiten kann.

Die Arbeiten gehen im wesentlichen auf das Intonationsmodell von Bannert (1985aBannert (1985b) zurück. Der Fokusakzent ist bei ihm das "bedeutungswichtigste Wort einer Äußerung" und stellt einen "semantischen Stützpfeiler sprachlicher Kommunikation" dar. Eine Äußerung kann mehrere Fokusakzente enthalten. Fokusakzente werden im wesentlichen durch Bewegungen der Sprachgrundfrequenz markiert. Nach dem letzten Fokusakzent einer Äußerung findet kaum noch Bewegung von F_0 statt, sie sinkt bis zum Ende der Äußerung einfach ab. Eine Äußerung ist in prosodische Phrasen aufgeteilt, die nicht unmittelbar identisch mit den syntaktischen Phrasen sind. Eine prosodische Phrase wird dadurch definiert, daß sie rhythmisch und tonal abgeschlossen ist.

Die untersuchten Sprachkorpora enthielten einfache Äußerungen (eine prosodische Phrase) und komplexere (zwei bis drei prosodische Phrasen). Es handelte sich um gelesene Sprache (142 Satzmuster, von 5 verschiedenen Sprechern bis zu sieben Mal gelesen). Aufgrund der Probleme, die aus der Grundfrequenzbestimmung (siehe auch 6.1) resultieren, bestand ein Teil der Korpora hauptsächlich aus Sonoranten, die einen weitgehend lückenlosen F_0-Verlauf liefern.

Der Verfahren zur Fokuserkennung arbeitete im wesentlichen regelbasiert. Parameterwerte wurden experimentell bestimmt, dabei wurden einige Schwellwerte nur absolut und speziell auf die verwendeten Sprachkorpora bezogen verwendet. Als primäre Informationsquelle für das Verfahren diente die Grundfrequenz, in Kombination mit absoluten Zeitwerten als Dauerinformation. Es wurde gezielt nach sog. Akzentmaxima gesucht: Ein F_0-Wert in einem festgelegten Zeitintervall (ca. 300 ms) mußte höher als die Umgebungswerte und außerdem um mindestens 20 Hz höher als die F_0-Werte am linken und rechten Intervallrand sein.

Es werden keine Erkennungsraten angegeben, es heißt lediglich "die meisten Fokusakzente werden erkannt" (Hoepelman und Machate, 1994, S. 46). Probleme entstanden in erster Linie durch Fehler in der Grundfrequenzkontur, die auch durch gezielte Nachverarbeitung nicht vollständig behoben werden konnten. Beim Testen des Verfahrens an unbekannten Sprachdaten werden allerdings deutliche Einbrüche in der Erkennung zu erwarten sein:

1. Die Menge der verwendeten Sprachdaten mit komplexerer Struktur war sehr gering,
2. Die Sprachdaten waren akustisch 'gutartig' konstruiert (z. T. nur stimmhafte Laute),
3. Die festgelegten Schwellwerte waren zu sehr auf die Sprachdaten abgestimmt (Verwendung von absoluten Werten).

Die Gesamtkonzeption des MAFID-Systems ist aber durchaus positiv zu bewerten. Es wurde erstmals versucht, die Erkennung von akustischer Fokusinformation mit einer semantischen Interpretation zu verbinden. Von einem Dialogsystem, wie wir es heute kennen, war MAFID allerdings noch weit entfernt.

3.6.2 Das Projekt Modus-Fokus-Intonation

Ebenfalls in einer größeren Projektstudie ("Modus-Fokus-Intonation") sollten Prototypen für Fokusintonation in verschiedenen Satztypen gefunden werden (Altmann et al., 1989). Es wurden keine Fokusakzenttypen wie *Normalakzent* vs. *Kontrastakzent* unterschieden. Die Autoren gehen von folgender Definition aus:

> *" Fokus und Fokussierung sind semantische Begriffe, die die informationelle Gliederung von Äußerungen betreffen: Vorausgesetzter Hintergrundsinformation steht fokussierte Information gegenüber, die aus einer kontextuell festgelegten Menge von Alternativen ausgewählt wird. Gekennzeichnet wird der Fokus durch die akzentuelle Hervorhebung eines Ausdrucks, des Fokusexponenten. Dieser Ausdruck trägt aber auch einen Teil der oben erwähnten intonatorischen Satzmodusmarkierung, so daß jede Untersuchung die gegenseitige Beeinflussung der intonatorischen Kennzeichnung von Satzmodus und Fokus berücksichtigen muß. " (Altmann et al., 1989, S. 1)*

Zur Gewinnung eines Prototyps, der auf repräsentativen Daten (und nicht nur auf einzelnen Beispielen) beruhen sollte, wurde ein von 6 verschiedenen Sprechern gelesenes Korpus erzeugt (insgesamt 360 Sätze). Dieses enthielt 'Intonations-Minimalpaarsätze', also Sätze mit gleichem Wortlaut, die sich jeweils in ihrer Intonation bezüglich der Fokusmarkierung und/oder Satzmodusmarkierung unterschieden. Die Intonationsmuster wurden durch entsprechende Lesekontexte vorgegeben, die gelesenen Sätze wurden auf die gewünschte Intonationskontur hin perzeptiv überprüft. Auch hier wurden im Korpus weitgehend sonorante Laute verwendet, um die aus der Grundfrequenzbestimmung resultierenden Probleme zu minimieren.

Batliner (1989b) nimmt an, daß es prototypische Realisierungen bestimmter Akzentstrukturen gibt, die möglicherweise nicht ausschließlich mit Regeln darstellbar sind. Zur Prototypgewinnung wendet er zwei Methoden an: 1. Mittelwertbildung aller relevanten Parameter anhand einer großen Datenmenge und 2. Betrachten von 'typischen' Exemplaren, die von Hörern als besonders gut bewertet wurden.

Zur Bestimmung des ersten Prototyps wurden akustische Merkmale aus den Korpussätzen extrahiert, 12 verschiedene Energie-, Dauer- und F_0-Merkmale wurden zu einem Merkmalsvektor zusammengestellt (Batliner, 1989b). Die Relevanz dieser akustischen Parameter wurde statistisch mit Hilfe einer Diskriminanzanalyse ausgewertet. Die Durchschnittswerte ergaben ein Muster für 'normale, prototypische' Intonation. Daneben gab es noch diverse sprecherspezifische Abweichungen.

Ein weiteres Ziel war die Vorhersage von Fokus bzw. welcher Teil der Äußerung den Fokusakzent trägt (Batliner, 1991). Mit Hilfe der akustischen Merkmale wurde ein Erkenner trainiert; dabei sollte entschieden werden, ob in einfachen Sätzen mit mehreren Objektphrasen jeweils die 2. oder die 3. Objektphrase fokussiert wurde. Als entscheidende Merkmalsvariablen für alle Sätze stellten sich die F_0-Maxima und -Minima der Objektphrasen sowie deren Position zueinander heraus. Der Beitrag von Dauer und Intensität war für die Entscheidung von eher untergeordneter Bedeutung; bei Nichtberücksichtigung dieser Parameter verschlechterten sich aber die Ergebnisse.

Entscheidend für eine gute Klassifikation war die Trennung der Daten in Fragen und Nicht-Fragen (Aussagen und Imperativsätze). Für beide Satzgruppen waren unterschiedliche Merkmale für die Fokusklassifikation relevant. Für Nicht-Fragen waren dies F_0-Maximum und Dauer der 3. Objektphrase, für Fragen waren eher Merkmale der 2. Objektphrase relevant (F_0-Maximum und Positionsparameter). Insgesamt werden Fokusakzente in Nicht-Fragen besser klassifiziert. Offensichtlich überlagern sich in den Fragen Fokus- und Satzmodusmarkierung, so daß diese beiden Phänomene schlecht voneinander getrennt werden können.

Die Klassifikation (Lernstichprobe $\neq$ Prüfstichprobe, 5 Sprecher) war für Fragen und Nicht-Fragen insgesamt in 77.5 % aller Fälle korrekt. Für Fragen allein waren dies 75.7 %, für Nicht-Fragen 81.6 %.

Die Methode der statistischen Klassifikation wird auch in Verbmobil und Intarc (siehe 4. Kapitel) zur Akzenterkennung verwendet (Kompe, 1997Kießling, 1997Strom, 1998). Die Klassifikatoren (oder auch neuronale Netze) wurden auf allen etikettierten Akzenten (ohne Unterscheidung verschiedener Akzenttypen) trainiert (siehe Abschnitt 5.4). Die dabei erzielte mittlere Erkennungsrate für Akzente beträgt zwischen 71.5 % (ohne Dauermerkmale; Strom, 1998) und 82.8 % (Kießling, 1997).

3.7 Zusammenfassung

3.7.1 Definition des prosodischen Fokus

Der in dieser Arbeit verwendete Fokusbegriff soll hier auf einen *prosodischen Fokus* beschränkt werden. Dieser kann allein aus der aktuellen akustischen Realisierung bestimmt werden. Auf der Grundlage eines schriftlichen Textes kann er dagegen nur bedingt vorhergesagt werden (siehe Abschnitt 7.5). Auch der Diskurskontext gibt nicht immer einen Hinweis darauf, ob ein prosodischer Fokus zu erwarten ist oder nicht.

Der prosodische Fokus bezeichnet Wörter oder Satzteile, die stärker intonatorisch hervorgehoben sind als andere. Er ist aber andererseits nicht unabhängig vom *linguistischen Fokus*. Die prosodische Auffälligkeit hängt im allgemeinen unmittelbar mit der Wichtigkeit zusammen: Die stärker prosodisch markierten Wörter reflektieren die Absicht von Sprechern, die Abschnitte eines Satzes besonders hervorzuheben, die für sie wichtig sind. Dies kann sowohl *neue* Information beinhalten als auch Bekanntes, auf das noch einmal besonders hingewiesen wird.

Der prosodische Fokus ist also auch als linguistischer Fokus interpretierbar, während ein linguistischer Fokus nicht immer prosodisch markiert sein muß. Der linguistische Fokus kann auch mit anderen Mitteln ausgedrückt werden, z. B. durch lexikalische (hier gibt es Fokuspartikeln wie “nur”, “auch” und “sogar”, die den Fokus assoziieren) und grammatische Mittel (hier insbesondere Passivkonstruktionen und Spaltsätze, wie z. B. “Es war der *Mittwoch*, an dem wir uns treffen wollten.”). Im Sprachvergleich gibt es große Unterschiede, inwieweit grammatische oder prosodische Mittel zur Markierung des Fokus eingesetzt werden (siehe Abschnitt 3.3.1). Dies gilt auch für Sprechstile: Im Deutschen werden gram-

matische Mittel in Spontansprache eher selten verwendet; diese Ausdrucksform wird oft als umständlich und formal empfunden.

Bei der Klassifizierung von verschiedenen Arten des prosodischen Fokus spielt vor allem der Kontrastfokus eine Rolle (weitere Einzelheiten in Abschnitt 6.8). In der Definition des *Kontrastfokus* lege ich den Schwerpunkt auf die *akustische Auffälligkeit*. Ein kontrastiver Fokusakzent ragt akustisch besonders aus seiner Umgebung heraus. Dies schließt natürlich nicht aus, daß diese Auffälligkeit durch eine Abweichung von der 'normalen', erwarteten Intonationskontur erzeugt wurde, die wiederum durch die Nichtanwendung der Syntaxstrukturregeln bedingt wird. Das Gegenstück dazu ist ein *Defaultfokus*, der an syntaktisch definierter Stelle am Ende einer Phrase plaziert wird und nur dadurch als Fokus wahrgenommen wird, weil kein anderes Wort akustisch herausragt.

3.7.2 Eigenschaften von Fokusakzenten

Fokusakzente haben bestimmte Eigenschaften, diese sollen hier abschließend zusammengefaßt werden. Für 'normale' Fokusakzente in den germanischen Sprachen (mit mehr oder weniger großen Abweichungen) gelten folgende Beobachtungen; dabei sind die ersten fünf Eigenschaften als obligatorisch zu verstehen, während die restlichen eher als grobe Tendenzen zu sehen sind:

- Akustik
  - perzeptiv auffällig, erhöhter Einsatz der akustischen Parameter
  - alle folgenden Akzente werden deakzentuiert
  - satzinitial werden akustische Parameter stärker eingesetzt
  - unterschiedlicher Einsatz der akustischen Parameter in Fragen und Aussagen
- Lexikon/Syntax
  - Akzent fällt auf Silbe mit lexikalisch festgelegtem Wortakzent
  - Akzent fällt eher auf Inhaltswörter (Substantiv, Verb, etc.)
  - Akzent befindet sich eher am Ende einer syntaktischen Phrase/Konstituente (*Default*-Position)
- Semantik/Pragmatik:
  - Akzent markiert semantisch wichtigem Abschnitt
  - Akzent liegt auf neuer Information im Diskurs
  - Bekannte Information wird deakzentuiert

3.7.3 Kontrastfokus

Ein *Kontrastfokus* ist relativ unabhängig von den allgemeinen Regeln für Fokusakzente. Die Grundfunktion (Markierung von wichtiger Information) bleibt gleich, aber die *Realisierung* kann anders erfolgen: Der Fokusakzent kann auf jeder beliebigen Silbe liegen, und/oder der Akzent kann *zusätzlich* akustisch verstärkt werden.

Kontrastfokus kann allerdings auch den rein semantischen Kontrast (ohne Bezug zur akustischen Realisierung) bezeichnen (siehe auch Abschnitt 3.1.5). Dabei gilt für alle Betrachtungsweisen, daß ein Kontrast immer ein Vergleichselement braucht, also nicht in Isolierung festzustellen ist. Abschließend folgt eine Zusammenstellung der unterschiedlichen *Sichtweisen* zur Definition eines Kontrastfokus. Bei Zutreffen von mindestens einer Eigenschaft ist - je nach Sichtweise - die Bedingung für einen Kontrastfokus erfüllt.

- akustisch/perzeptiv
  - perzeptiv **besonders** auffällig im Vergleich zur Umgebung
  - Begriff *Emphase* als Synonym verwendbar
- Akzentposition
  - Akzent liegt auf Funktionswort
  - Akzent liegt nicht auf lexikalisch definierter Silbe
  - Akzent ist nicht vorhersagbar anhand von syntaktischen Regeln
  - Akzent liegt auf bekannter Information
- semantisch/logisch (auch im geschriebenen Text entscheidbar)
  - Gegenüberstellung von semantischen Alternativen
  - Korrektur einer Behauptung

4. Anwendungen in der automatischen Sprachverarbeitung

Die vorliegende Arbeit entstand im Rahmen des BMBF-Verbundprojekts Verbmobil. Das Ziel von Verbmobil ist ein mobiles System zur automatischen Übersetzung spontansprachlicher Dialoge. In Verbmobil werden daher die Bereiche Spracherkennung, Sprachübersetzung und Sprachsynthese gefördert.

Ein Teilprojekt in Verbmobil ist das experimentelle System Intarc (integrated architecture). Bei ähnlicher Grundfunktionalität wie in Verbmobil war das Hauptziel die Erforschung neuer Architekturen, d. h. die Organisation und Abstimmung der Verarbeitungsmodule untereinander. In dieser Arbeit wurde die Einbindung prosodischer Informationen (hier speziell die Fokusinformation) zur Unterstützung der linguistischen Verarbeitungsmodule (Semantik und Transfer) untersucht. Die Nutzung von prosodischer Information in der automatischen Sprachverarbeitung, mit besonderer Berücksichtigung von Verbmobil und Intarc, wird im folgenden diskutiert.

4.1 Das Projekt Verbmobil

Verbmobil ist ein langfristig angelegtes Verbundprojekt zur automatischen Übersetzung spontansprachlicher Dialoge (Wahlster, 1993). Die Planung sieht 2 Phasen für jeweils einen Zeitraum von 4 Jahren vor. Das Szenario für die 1. Phase sah folgendermaßen aus: Zwei Geschäftspartner, ein Deutscher und ein Japaner, wollen einen Termin vereinbaren. Beide haben einen aktiven und passiven englischen Wortschatz, sprechen aber nicht fließend Englisch. Der Dialog wird also im wesentlichen in englischer Sprache geführt; wenn die englischen Sprachkenntnisse der Partner an ihre Grenzen stoßen, kann von Verbmobil eine Übersetzung angefordert werden. Die Verhandlungspartner können dann in ihrer Muttersprache weitersprechen, und Verbmobil erstellt jeweils eine Übersetzung Deutsch-Englisch oder Japanisch-Englisch. Ein langfristig angestrebtes Ziel ist die Entwicklung eines tragbaren (also *mobilen*) Übersetzungsgerätes für diverse Sprachen.

Die Aufgaben von Verbmobil betreffen verschiedene Bereiche:

- Verbmobil muß Spontansprache erkennen und das Wesentliche verstehen - auch bei schlechter Aussprache und ungrammatischen Äußerungen.
- Während der englischen Konversation muß Verbmobil der Verhandlung folgen, um

Kontextwissen für eine eventuelle Übersetzung aufzubauen.

- Bei Unklarheiten, die nicht mehr aus dem Kontext geklärt werden können, muß Verbmobil die Initiative ergreifen und mit dem Benutzer einen Klärungsdialog führen.

- Bei angefordertem Übersetzungswunsch muß ein Transfer Deutsch-Englisch oder Japanisch-Englisch erfolgen. Die Übersetzung muß nicht jedes Detail genau übersetzen, doch die wesentliche Intention des Sprechers soll auf jeden Fall enthalten sein.

- Bei der Ausgabe der Übersetzung sollte die Synthese möglichst natürlich klingen. Dabei ist wünschenswert, daß sich die ursprüngliche Intention der Sprecher auch hier in der Prosodie widerspiegelt. Außerdem soll der Stimmcharakter der jeweiligen Sprecher modelliert werden, d. h. Stimmtonhöhe und Geschlecht müssen in der Synthese dem Originalsprecher oder der Originalsprecherin angemessen sein.

Es kann davon ausgegangen werden, daß beide Dialogpartner kooperativ, d. h. an einer gemeinsamen Lösung eines Verhandlungsproblems interessiert sind. Für eine korrekte Übersetzung ist meist Weltwissen über den Gesprächsgegenstand nötig. Verbmobil muß daher bereits Wissen in einer bestimmten Repräsentation gespeichert haben (passend zur Domäne "Verhandlungsdialoge") und weiteres Wissen im Verlauf des Dialogs erwerben, um im richtigen Kontext zu übersetzen.

Bei der Verarbeitung in Verbmobil müssen die Faktoren Rechenzeit versus Qualität der Ergebnisse gegeneinander abgewogen werden. Einerseits sollen Resultate möglichst schnell geliefert werden, um die Akzeptanz des Systems generell zu erhöhen. Andererseits dürfen diese nicht so schlecht sein, daß sie vom Benutzer als unbrauchbar abgelehnt werden. Die Verarbeitungstiefe der Analyse sollte einstellbar sein, um wahlweise schnellere Resultate (mit geringerer Qualität) oder bessere Resultate (mit längerer Verarbeitungszeit) zu erhalten.

In der zweiten Phase von Verbmobil wurde das Szenario auf Reiseplanung und Fernwartung (Wahlster, 1997) erweitert. Es wird ein größerer Wortschatzumfang angestrebt, außerdem auch die bidirektionale Übersetzung weiterer Sprachpaare, wie z. B. Deutsch-Englisch und Deutsch-Japanisch. Das setzt weitgehend sprachenunabhängige und reversible Verarbeitungsverfahren voraus. Wünschenswert ist auch eine automatische Erkennung des Hauptgesprächsthemas oder eines Themenwechsels (Topic Spotting), um automatisch auf eine neue Domäne umschalten zu können. Es ist ebenfalls eine Protokollfunktion geplant, wobei Verbmobil die wesentlichen Inhalte eines Gesprächs protokollieren soll.

Außerdem soll die Spracheingabe nicht mehr nur über Nahbesprechungsmikrophone erfolgen, sondern es ist auch die Möglichkeit des Freisprechens, z. B. auch über Telefon oder Mobiltelefon geplant. Das setzt weitere Forschung in Bezug auf robuste Spracherkennung voraus. Auch die Qualität der Sprachausgabe soll deutlich gesteigert werden. Prosodie und satzübergreifende Phänomene sollen besser modelliert werden; dazu wird ein "Concept-to-Speech"-Ansatz (siehe auch Abschnitt 9.2.4) verfolgt, der eine enge Verknüpfung von

Generierung und Synthese vorsieht. Vom Sprecher intendierte Hervorhebungen sollen direkt an die Synthese weitergereicht werden, so daß aufwendige Reanalysen des Textes entfallen können. Ausführliche Beschreibungen über Verbmobil finden sich in (Wahlster, 1997Bub et al., 1997).

4.2 Das Architekturteilprojekt Intarc

Intarc (*integrated architecture*) ist ein Teilprojekt von Gesamt-Verbmobil. Es sollte ein experimentelles System aufgebaut werden, das dieselbe Grundfunktionalität wie Verbmobil hat (also Übersetzung spontansprachlicher Dialoge). Der Umfang von Intarc ist allerdings wesentlich geringer: Das System arbeitet mit 9 Modulen, während Gesamt-Verbmobil 43 funktionale Module hat (siehe Bub et al., 1997). Der Schwerpunkt lag auch nicht auf der Optimierung der einzelnen Module, sondern auf der *Interaktion* zwischen diesen. Es sollten neue Informationspfade, und damit auch neue Informationsquellen, für die einzelnen Module untersucht und getestet werden. Der Austausch von Nachrichten zwischen den Modulen sollte schon zu einem relativ frühen Verarbeitungszeitpunkt stattfinden, also beispielsweise auch, während die Eingabe noch anhält.

4.2.1 Anforderungen an das System

Das geplante Sprachverarbeitungssystem sollte mit möglichst geringer Verzögerung zur Echtzeit arbeiten. Das bedeutet, daß die Systemreaktion für einen Benutzer so schnell sein muß, daß sie mit seiner Eingabe Schritt hält. Dafür sind bestimmte Architekturvorgaben nötig. Es ist anzustreben, in allen Teilbereichen nur Algorithmen und Verfahren anzuwenden, die mit der Signaleingabe Schritt halten können und nicht aus ihrer Natur heraus eine lange Verarbeitungszeit haben. Das System sollte *inkrementell, interaktiv, parallel* und *robust* sein (Weber et al., 1997).

Die Anforderung der *Inkrementalität* verlangt, daß jedes Ereignis unverzüglich gemeldet wird, auch wenn es noch nicht abgeschlossen ist. Es werden also Hypothesen gebildet, ohne daß die dazu nötige Information vollständig zur Verfügung steht. Die Verarbeitungsrichtung geht grundsätzlich und bevorzugt von links nach rechts, also in Zeitrichtung. Algorithmen, die in Gegenrichtung arbeiten, wie *Inselstrategien* oder *Backtracking*, sind eher unerwünscht. Da die rechte Kontextinformation weitgehend wegfällt, müssen gegebenenfalls neue Algorithmen und Verfahren entworfen werden.

Interaktivität: Es findet ein massiver asynchroner Nachrichtenaustausch statt. Restriktionen sollen möglichst früh in die Analyse einfließen. Die Verarbeitungsrichtung geht nicht nur in der Richtung signalnah→linguistisch/symbolisch, sondern auch umgekehrt, d. h. Erkennungsmodule bekommen Rückmeldungen aus einer 'höheren' Verarbeitungsstufe (top-down-Erwartung). Dazu müssen alle Module gleichzeitig aktiv sein und parallel arbeiten.

Performanzerhaltung: Für das Treffen von Entscheidungen steht nur der linke Kontext zur Verfügung. Dadurch bedingt die schritthaltende Verarbeitung eine Erhöhung der Verar-

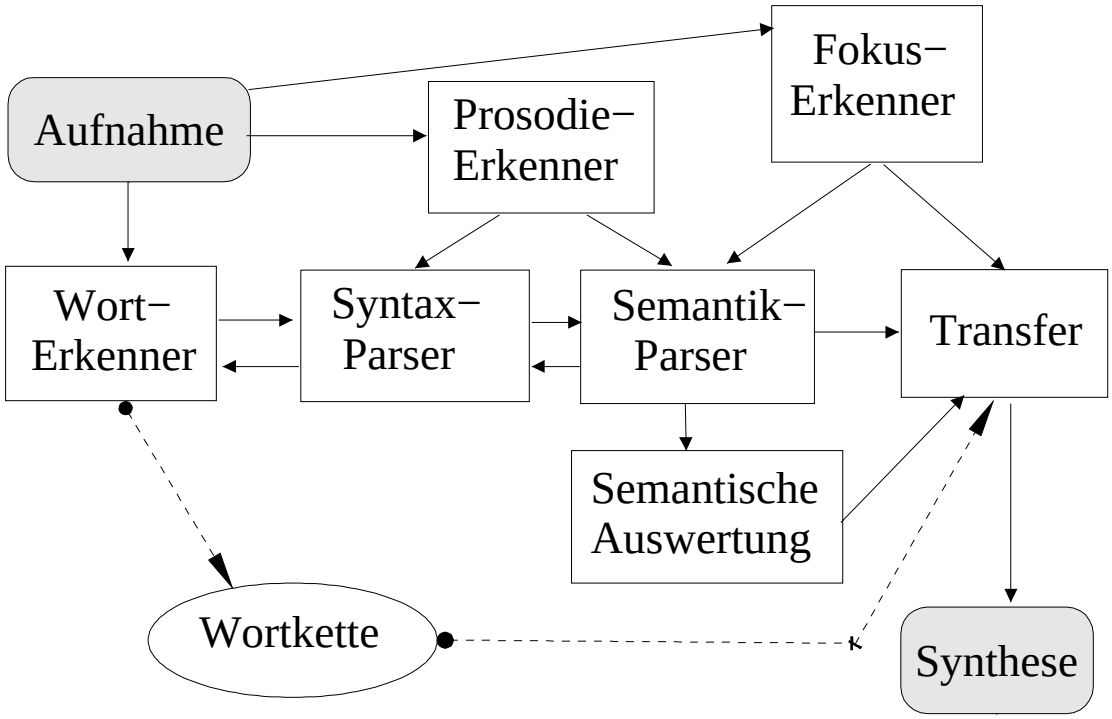

Abbildung 4.1: *Zusammenarbeit der Verarbeitungsmodule in Intarc; die gestrichelten Linien zeigen den Weg der flachen Analyse*

beitungszeit der Einzelmodule, außerdem kann die Korrektheit der Ergebnisse möglicherweise darunter leiden. Diese negativen Konsequenzen sollten möglichst gering gehalten werden.

Robustheit: Das System muß bei Fehlschlägen sinnvoll weiterlaufen, tiefe und flache Analyse sollten parallel und unabhängig voneinander stattfinden. Wenn die tiefe linguistische Analyse scheitert, muß durch eine flache Analyse ein approximativ korrektes Ergebnis geliefert werden.

4.2.2 Module des Systems

Alle Module des Intarc-Systems sind inkrementell angelegt, um eine möglichst geringe Zeitverzögerung in der Gesamtverarbeitung zu gewährleisten. Alle Teilergebnisse werden sofort an das nächste Modul weitergereicht, und alle Module arbeiten parallel (siehe Abschnitt 4.2.1). Ausführlichere Beschreibungen finden sich in Weber et al. (1997) und Strom (1998). Einen Überblick über das Gesamtsystem gibt Abbildung 4.1.

Die Signaleingabe und -verarbeitung erfolgt über ein Mikrophon und eine daran angeschlossene Gradient Box. Die darauf aufsetzende *Worterkennung* besteht aus einem HMM-basierten Dekodierer, der schritthaltend Wortgraphen erzeugt. Es werden potentiell erkannte Wörter ausgegeben; pro Sekunde Sprechzeit werden ca. 200 Worthypothesen erzeugt (Weber et al., 1997). Jede gefundene Worthypothese wird an den Syntaxparser

geschickt. Im Worthypothesengraph wird nur eine Vorwärtssuche durchgeführt; im Gegensatz zu anderen HMM-basierten Verfahren ist eine Rückwärtssuche in das spätere Parsing-Verfahren integriert. Die beste Wortkette (mit den höchsten Wahrscheinlichkeiten) wird außerdem noch an den Transfer weitergereicht, um eventuell eine flache Analyse durchführen zu können (falls die tiefe linguistische Verarbeitung scheitert).

Die traditionelle Rückwärtssuche zur Reduktion des Wortgraphen des Erkenners wird im *Syntaxparser* durchgeführt. Rechte Kontextinformation wird zur Bewahrung der Inkrementalität nur minimal verwendet; um diesen Mangel auszugleichen, wird die Suche als stochastisches Parsing ausgeführt. Dabei werden komplexere Modelle eingesetzt, die auch längere Kontexte berücksichtigen.

Das *Symbolische Parsing* (Semantik-Parser) untersucht konkurrierende Hypothesen von Wortketten. Dabei muß die Grammatik auf Spontansprache abgestimmt sein. Die eingegangenen Worthypothesen werden auf ihre Kompatibilität mit der Grammatik überprüft. Ausgegeben werden Merkmalsstrukturen, die den syntaktischen und semantischen Gehalt der Äußerungen beschreiben.

Die *Semantische Auswertung* überführt die Informationen in Dialogakte und führt ein Dialogaktgedächtnis. Beim *Transfer* wird aus dem übermittelten Dialogakt der propositionale Gehalt entnommen, und eine Übersetzung wird erstellt. Die gelieferten Teilstücke müssen übersetzbar sein, d. h. es handelt sich bereits um Sätze oder Turns. In den Transfer ist eine schemabasierte Generierung integriert; deren Ausgabe wird an eine kommerzielle Synthese weitergegeben.

Zusätzliches Wissen an die einzelnen Module wird von der *Prosodie* geliefert. Diese besteht aus einem Erkenner für Phrasengrenzen und Satzmodus (Strom, 1995) und der Fokuserkennung (siehe Kapitel 6). Die Phrasengrenzen tragen zur Reduktion des Suchraums für die beiden Parser bei: Beim syntaktischen Parser wird der Suchraum um 20 % reduziert bzw. der Suchbaum wird um 65 % verringert. Beim symbolischen Parser wird die Anzahl der Lesarten um 25 % reduziert (Strom et al., 1997). Satzmodus und Fokusinformation werden über den Semantik-Parser an die Semantische Auswertung geliefert (Kasper und Krieger, 1996).

In einer flachen Analyse nimmt der Transfer eine Übersetzung mit Hilfe der besten Wortkette und der fokussierten Wörter in dieser Kette vor. Anhand der fokussierten Wörter werden Dialogakte bestimmt, die in eine schablonenbasierte Übersetzung einfließen. Dies soll zur Erhöhung der Robustheit dienen. Bei Scheitern der tiefen Analyse liefert der fokusgesteuerte Transfer in ca. 30 % der Fälle akzeptable (aber möglicherweise reduzierte) Übersetzungen (siehe Kapitel 8).

4.3 Prosodische Information in der automatischen Sprachverarbeitung

Die Prosodie wurde in der automatischen Spracherkennung in früheren Jahren meist nicht berücksichtigt. Dabei ist es unbestritten, daß die Prosodie einen bedeutenden Faktor für die menschliche Sprachwahrnehmung darstellt. Vorreiter für die Forderung nach mehr Verwendung von prosodischer Information in Spracherkennungssystemen waren Lea (1980) und Vaissière (1988), erste Anwendungen finden sich bei Waibel (1988) und Nöth (1991). Im Rahmen von Verbmobil sind mittlerweile diverse Dissertationen entstanden, die sich mit der Nutzung der Prosodie in der automatischen Spracherkennung beschäftigen (Kießling, 1997Kompe, 1997Strom, 1998).

4.3.1 Funktionen der Prosodie in Spontansprache

In welcher Weise kann Prosodie also für die automatische Spracherkennung nützlich sein? Ein wichtiger Aspekt ist, daß die Prosodie eine *unabhängige* Hilfe für die anderen Erkennungsebenen darstellen kann. Die Module für Semantik, Syntax, Morphologie, Phonologie und Akustik-Phonetik bauen jeweils aufeinander auf; die prosodischen Parameter werden zwar auch auf der akustischen Ebene gewonnen, doch ihre Interpretation kann weitgehend unabhängig von anderen Modulen stattfinden. Natürlich kann die prosodische Analyse trotzdem von den anderen Ebenen unterstützt werden; in erster Linie soll aber die Prosodie zur Verbesserung der anderen Analysemodule beitragen. Dadurch kann sowohl die Geschwindigkeit der Erkennung verbessert als auch eine größere Zuverlässigkeit gewährleistet werden.

Wie bereits in Abschnitt 2.6 gezeigt wurde, kann die bewußte Steuerung der prosodischen Parameter zur Markierung von verschiedenen linguistischen Funktionen eingesetzt werden. Das Sprachsignal wird in einzelne Abschnitte aufgeteilt (Phrasen), besonders wichtige Elemente werden hervorgehoben (Akzentuierung). Phrasen können markiert werden (pragmatische Funktion → Satzmodus). Akzentuierung findet auf Wortebene (lexikalischer Akzent) und auf Satzebene (Satzakzent → Fokussierung) statt.

In Vergleich zu früheren Spracherkennungssystemen ist bei Verbmobil eine höhere Komplexität vorhanden: Die Sprachäußerungen sind länger und aufgrund der Verwendung von Spontansprache durch grammatische Unregelmäßigkeiten gekennzeichnet (siehe Abschnitt 5.3). Die wesentlichen Aufgaben der Prosodie bestehen hier in der *Strukturierung* der Sprache, um den Suchraum einzuschränken. Außerdem dient die Prosodie zur *Desambiguierung* von mehrdeutigen Äußerungen; andernfalls müßte dies aus dem Kontext abgeleitet werden, was wiederum zu einer Verlängerung der Suche führt.

Die Gliederung von syntaktisch und semantisch zusammenhängenden Elementen durch die Prosodie erleichtert es erheblich, die Satzstruktur durch eine Einteilung in ihre Konstituenten offenzulegen. Es folgt ein Beispiel für verschiedene Gliederungsmöglichkeiten einer Äußerung in Phrasen; eine zusätzliche Markierung des Satzmodus kann hier weitere Hinweise zur korrekten Aufteilung geben:

Wie sieht es aus am Dienstag um 17 Uhr geht es nicht

[Wie sieht es aus?] [Am Dienstag um 17 Uhr geht es nicht.]
[Wie sieht es aus am Dienstag?] [Um 17 Uhr geht es nicht.]
[Wie sieht es aus am Dienstag um 17 Uhr?] [Geht es nicht?]

Eine Desambiguierung kann mit Hilfe des Fokusakzentes vorgenommen werden. In den folgenden Beispielen markiert die Fokussierung eine jeweils unterschiedliche Bedeutung und erfordert damit eine andere Übersetzung (fokussierte Wörter sind durch Großbuchstaben markiert).

Wir brauchen noch einen TERMIN. – We still need a date.
Wir brauchen NOCH einen Termin. – We need another date.

In der WOCHE kann ich nicht. – During the week it's impossible for me.
In DER Woche kann ich nicht. – In that week it's impossible for me.

Für Übersetzung und Synthese ist es wichtig, über zusätzliche prosodische Information zu verfügen, da viele Äußerungen sonst nur durch aufwendige linguistische Analyse unter Einbeziehung des Kontextes korrekt interpretiert werden können. In vielen Fällen kann es sogar unmöglich sein, aus einer isolierten Äußerung ohne prosodische Informationen die von der Sprecherin/vom Sprecher intendierte besondere Emphase in bestimmten Teilen der Äußerung zu erkennen.

4.3.2 Verwendung der Prosodie in Verbmobil und Intarc

Im folgenden sollen einige Möglichkeiten für die Verwendung von prosodischer Information in einem realen System aufgezeigt werden; soweit vorhanden, werden Beispiele aus aktuellen Anwendungen in Verbmobil und Intarc gebracht. Die Prosodiemodule in beiden Systemen sind allerdings nicht unmittelbar vergleichbar, da sie in unterschiedlicher Weise mit den anderen Modulen verknüpft sind und damit auch über andere Informationsquellen verfügen (Das Prosodiemodul in Gesamt-Verbmobil verfügt neben der reinen Signalinformation noch zusätzlich über Wortgrapheninformation, siehe Nöth et al., 1997). Die Zusammenarbeit der Module in Verbmobil wird in Abbildung 4.2 dargestellt (vgl. auch mit Abbildung 4.1).

Zu bedenken ist, daß alle Prosodiemodule in Verbmobil und Intarc in erster Linie Wahrscheinlichkeiten bzw. Konfidenzen für ein detektiertes prosodisches Phänomen ausgeben. Diese entsprechen den Wahrnehmungskategorien in Abschnitt 2.5. Die linguistische Funktion (Abschnitt 2.6) muß dann jeweils von den linguistischen Verarbeitungsmodulen interpretiert werden. Zum einen dürfen Erkennungsfehler in der Prosodie nicht zum Mißerfolg der linguistischen Analyse führen (ein entsprechender Recovery-Mechanismus wird in Kasper und Krieger (1996) vorgestellt). Zum anderen muß jedes Modul für sich entscheiden, ab welchem Wahrscheinlichkeitswert beispielsweise eine Hervorhebung als Akzent interpretiert werden soll und damit in die Analyse einfließen darf. Ein anderes Problem ist, daß prosodische Grenzen nicht automatisch syntaktische Grenzen sind (eine ausführliche

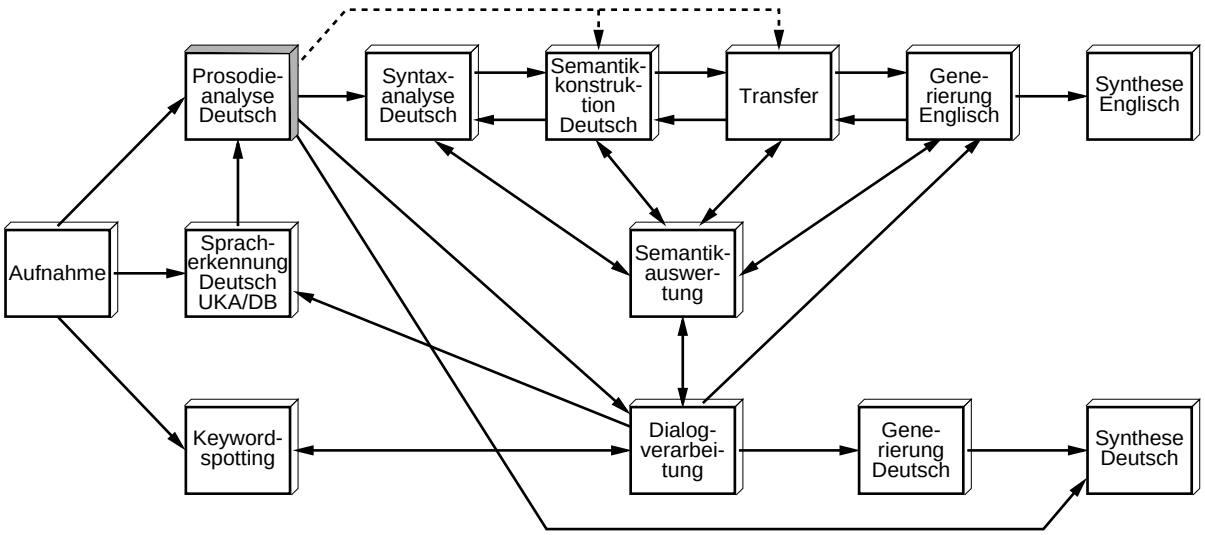

Abbildung 4.2: *Zusammenspiel der Module in Verbmobil: Das Prosodiemodul bekommt Information aus Aufnahme und Worterkennung, als Ausgabe werden Hypothesen und Wahrscheinlichkeiten über Akzente, Phrasengrenzen und Satzmodus direkt an Syntax und Dialog weitergeleitet. Diese Informationen fließen indirekt in Semantik und Transfer ein (gestrichelte Linie). Die Synthese bekommt Information über die Stimmlage des Sprechers oder der Sprecherin. [aus Nöth et al, 1997]*

Diskussion findet sich in Batliner et al., 1998a). Die Phrasenmarkierung (fallend oder steigend) muß ebenfalls auf einen linguistischen Satzmodus abgebildet werden.

Syntaktische und symbolische Parser: Auf der Parserebene dient die Prosodie in erster Linie zur Reduktion der Worthypothesengraphen, die in den Verbmobildialogen aufgrund der Spontansprache (Abschnitt 5.3) besonders umfangreich sind. Dies bedingt eine aufwendige Suche nach plausiblen (d. h. im Sinne der Grammatik korrekten) Wortketten. Wenn aber bereits Information über prosodische Phrasengrenzen vorliegt, kann die Anzahl der möglichen Lesarten deutlich reduziert werden.

Die Phrasengrenzenformation wird sowohl in Verbmobil als auch in Intarc verwendet. In Verbmobil verringert sich die Anzahl der Lesarten um 96 %, die Zeit für die Suche kann um 92 % reduziert werden (Nöth et al., 1997). In Intarc wird der Suchraum um 20 % eingeschränkt bzw. der Suchbaum wird um 65 % verringert. Beim symbolischen Parser (Semantik-Parser) wird die Anzahl der Lesarten um 25 % reduziert (Strom et al., 1997).

Semantische Konstruktion und Auswertung: In Verbmobil erhält das Semantikmodul einen Ableitungsbaum, die Wortkette und indirekt über das Syntaxmodul Akzentuierungsinformation. Aus diesen Informationen werden Diskurs-Repräsentationsstrukturen erzeugt. Diese Strukturen können mit Hilfe der Akzentinformation reduziert werden, indem nicht zur Akzentuierung passende Strukturen verworfen werden können. (siehe auch Bos et al., 1995). In Intarc werden Satzmodus und Fokusinformation indirekt über den Semantik-Parser an die Semantische Auswertung geliefert (siehe Abschnitt 8.4). Die prosodische Information dient hier ebenfalls zur Desambiguierung von semantischen Merkmalsstruk-

turen (Kasper und Krieger, 1996).

Dialog: In Verbmobil dienen Dialogakte zum einen als Übersetzungseinheiten, zum anderen werden sie als elementare Einheiten zur Planerkennung eingesetzt. In Jekat et al. (1995) wurden sämtliche in Verbmobil verwendeten Dialogakte definiert. Die grobe Dialogverfolgung (im inaktiven Modus von Verbmobil) geschieht durch Klassifikation der Dialogakte anhand von Schlüsselwörtern, die von einem Schlüsselworterkenner geliefert werden. Dazu werden semantische Klassifikationsbäume eingesetzt, die auch die Schlüsselwortlisten erzeugen (Mast, 1995).

Vor der Klassifikation müssen die Dialogakte jedoch segmentiert werden, denn die meisten Verbmobil-Äußerungen enthalten mehr als einen Dialogakt (siehe Beispiel in Abschnitt 5.2). Die Information über prosodische Phrasengrenzen wird vom Dialogaktmodul entsprechend ausgewertet: Jede Dialogaktgrenze entspricht auch einer Phrasengrenze, dies gilt aber nicht umgekehrt.

In Verbmobil gibt es ebenfalls eine flache linguistische Analyse, die vom Dialogmodul durchgeführt wird. Aus den erkannten Dialogakten der Äußerung wird mit Hilfe von entsprechenden Schablonen eine Übersetzung erstellt.

Transfer: Das Transfermodul in Verbmobil erhält von der Semantik Diskurs-Repräsentationsstrukturen. Diese müssen wiederum desambiguiert und pragmatisch analysiert werden. Dazu werden Satzmodus- und Akzentuierungsinformation herangezogen. Letztere erleichtert insbesondere die korrekte Interpretation von Partikeln (Ripplinger und Alexandersson, 1996Stede und Schmitz, 1998).

In Intarc wird bei der tiefen Analyse die prosodische Information indirekt genutzt, sie ist in den semantischen Merkmalsstrukturen enthalten, die von der Semantischen Auswertung geliefert werden. Bei der flachen Analyse in Intarc wird die beste Wortkette mit Hilfe der Fokusinformation übersetzt.

Synthese: Zur Zeit wird in Verbmobil nur Information über die Stimmhöhe direkt weitergegeben, um eine passende Synthesestimme (Männer- oder Frauenstimme) auszugeben. Es ist aber die Weitergabe von weiteren prosodischen Informationen vorgesehen, um einen "Concept-to-speech"-Ansatz (Alter et al., 1997) zu realisieren.

5. Sprachdaten

Für das Projekt Verbmobil wird eine sehr große Menge spontansprachlicher Daten benötigt. Schon allein für die Spracherkennung, die in erster Linie statistische Verfahren verwendet, steigt die Qualität der Verfahren mit der Anzahl der verfügbaren Trainingsdaten. Für das Szenario von Verbmobil 1 stehen inzwischen 8 CDs mit deutschen Sprachdaten zur Verfügung, weitere 9 CDs gibt es für Verbmobil 2. Außerdem wurden Sprachdaten für Englisch (Amerikanisch) und Japanisch bereitgestellt.

Im folgenden werden einige Eigenschaften der Sprachdaten beschrieben. Es wurden nur Dialoge aus dem Szenario von Verbmobil 1 verwendet. Ein geringer Teil der Dialoge ist prosodisch etikettiert (hauptsächlich von CD 1). Im Rahmen dieser Arbeit wurden zusätzlich Fokusakzente etikettiert, da die bereits vorhandenen Etiketten nicht den Vorstellungen des hier entwickelten Fokuskonzepts genügten.

5.1 Aufnahme der Sprachdaten

Im vereinbarten Szenario Terminvereinbarungsdialoge sollten jeweils 2 Gesprächspartner anhand eines vorliegenden Kalenders einen Dialog simulieren. Die Sprachäußerungen waren weitgehend spontan, also ohne Vorgabe des Dialogverlaufs. Es sollte aber ein Ziel, nämlich ein gemeinsamer Termin, erreicht werden. Sprecherwechsel wurde durch Knopfdruck simuliert, d. h. es kann jeweils nur ein Sprecher reden, wobei er einen Knopf gedrückt hält; hat er seinen Beitrag beendet, gibt er durch Loslassen des Knopfes den Kanal frei.

Die Aufnahmen sind von akustisch guter Qualität, sie wurden in einem ruhigen, aber nicht schalltoten Raum aufgenommen. Dabei wurden Nahbesprechungsmikrophone verwendet. Die Sprachdaten wurden mit einer Abtastfrequenz von 48 kHz digitalisiert und anschließend auf 16 kHz heruntergetastet. Die Dialoge wurden transliteriert, auch unter Einbeziehung der spontansprachlichen Phänomene (5.3). Eine ausführliche Beschreibung der Transliterationskonventionen findet sich in Kohler et al. (1994).

5.2 Aufbau der Verbmobildialoge

Die Verhandlungsdialoge in Verbmobil haben einen typischen Aufbau: Meist sind sie eingerahmt in eine Begrüßung und Verabschiedung. Einer der Gesprächspartner erwähnt dann das Thema, nämlich die Verabredung eines gemeinsamen Termins. In der darauffol-

genden Diskussion werden Terminvorschläge von beiden Partnern gemacht, es wird aber selten bereits der erste Vorschlag angenommen. Bei Ablehnung eines Terminvorschlags wird meist eine Begründung gegeben, die dienstlich oder privat sein kann. Bei Annahme des Termins erfolgt gelegentlich eine positive Bewertung ("prima") und eine Wiederholung der genauen Daten (Mast, 1995).

Jeder Dialog enthält eine oder mehrere Terminabsprachen. Für jeden Termin müssen im allgemeinen das Datum, die Uhrzeit, die Dauer und gegebenenfalls der Treffpunkt abgesprochen werden. Die fokussierten Bereiche (mit semantisch wichtiger Information) enthalten im wesentlichen Zeit- und Ortsangaben ("Donnerstag nachmittag", "14 Uhr 45", "bei mir im Büro"), aber auch Bewertungen ("das wird etwas knapp bei mir", "das ist schlecht", "hervorragend").

Im folgenden wird ein Verbmobildialog (n002kb, CD 1) vorgestellt. Ein Sprecher (nps1) und eine Sprecherin (nmw1) verhandeln über einen Termin. $< A >$ und $< P >$ bezeichnen jeweils Beginn und Ende des Knopfdrucks (d. h. Freischalten, bzw. Weitergeben des Sprecherkanals). Störungen der Aufnahmequalität (<#Stoergeraeusch> und '%' undeutlich) sind ebenfalls angegeben. Die Bezeichnung der Dialogakte ist mit Großbuchstaben und in runden Klammern kodiert.

```
nmw1k001
<A> der Termin den wir neulich abgesprochen haben am zehnten an dem
Samstag da kann ich doch nich' @(DIGRESS_SCENARIO) %wir sollten einen
anderen ausmachen <#St"orger"ausch> <P> @(INIT_DATE)
nps1k002
<A> wenn ich da so meinen Termin-Kalender anschaue @(DELIBERATE_EXPLICITE)
<A> <P> das %sieht schlecht aus <P> @(FEEDBACK_RESERVATION)
ich kann Ihnen den Dienstag sechsten April anbieten oder Freitag den
sechzehnten April <A> @(SUGGEST_SUPPORT_DATE) wo pa"st 's Ihnen denn am
besten <A> <#St"orger"ausch> @(REQUEST_COMMENT_DATE)
nmw1k003
alles vor sechzehn Uhr <P> <#St"orger"ausch> <A> @(SUGGEST_SUPPORT_DATE)
nps1k004
<P> dann w"urd‘ ich sagen am besten gleich nach dem Mittagessen um
vierzehn Uhr <P> auch hier in diesem Zimmer <#St"orger"ausch> <P>
@m(SUGGEST_SUPPORT_DATE) @m(SUGGEST_SUPPORT_LOCATION)
nmw1k005
<P> gut prima @m(ACCEPT_DATE) @m(ACCEPT_LOCATION)
vielen Dank @(THANK_INIT)
dann is' das ja kein Problem <P> @(FEEDBACK_ACKNOWLEDGEMENT)
nps1k006
<P> bis dann @(BYE) tsch"u"s <#St"orger"ausch> @(BYE)
```

5.3 Eigenschaften von Spontansprache

Bei den Daten handelt es sich um Spontansprache. Spontansprache ist frei formuliert und unterscheidet sich stark von vorbereiteten Texten, die abgelesen werden. Diese Form besteht häufig aus nicht wohlgeformten Äußerungen. Folgende Eigenschaften sind für die Spontansprache in Verbmobil charakteristisch (Wahlster, 1997):

- nicht-kanonische Aussprache (Dialekt, Verschleifungen)
- freie bis ungrammatische Syntax
- unvollständige Sätze (Ellipsen)
- Häsitationen (äh, hm)
- Pausen, Dehnungen (irregulär innerhalb der Äußerung)
- nonverbale Äußerungen: Husten, Lachen, Schmatzen, Atmen
- Wort- und Satzabbrüche, eventueller Neustart
- Versprecher und Korrekturen

5.4 Prosodische Etikettierung

Eine Teilmenge der Verbmobildialoge (36) wurde von den Projektpartnern in Braunschweig prosodisch etikettiert. Der überwiegende Teil davon befindet sich auf CD1 (29), der Rest verteilt sich auch CD2, CD3 und CD5. Es stehen Etiketten für Phrasengrenzen, Akzente und Satzmodus, dazu noch eine formale Intonationsbeschreibung zur Verfügung. Eine ausführliche Beschreibung des Etikettierungskonzepts findet sich in Batliner und Reyelt (1994).

Im Rahmen dieser Arbeit wurden nur die Etiketten für starke Phrasengrenzen (**B3**) verwendet. Diese sind deutlich durch einen eigenständigen Intonationsverlauf und/oder eine anschließende Pause markiert. Weiterhin gibt es noch schwache Phrasengrenzen (**B2**) und irreguläre Phrasengrenzen (**B9**), die die syntaktische Struktur unterbrechen. An einer B3-Grenze kann fakultativ das Etikett '**?**' angebracht werden, wenn eine Frage mit ansteigendem Intonationsverlauf realisiert wurde.

Die Etiketten für Akzente wurden nach der Prominenzstärke vergeben. Die höchste Prominenz hat ein Emphase/Kontrastakzent (**EK**). Dieser tritt relativ selten auf (siehe auch Abschnitt 7.3). Regelmäßig wurde dagegen für jede Phrase, die mit einer B3-Grenze abschließt, ein Phrasenakzent (**PA**) etikettiert. Dieser entspricht dem prominentesten Wort der Phrase. Im allgemeinen sollte ein Phrasenakzent pro Phrase etikettiert werden. Daneben wurden noch weitere, schwächer akzentuierte Silben mit einem Nebenakzent (**NA**) etikettiert.

Außerdem wurde der Tonhöhenverlauf nach einem leicht modifizierten ToBI-System etikettiert (Silverman et al., 1992). Der Verlauf wird dabei mit einer Folge von hohen und tiefen Tönen formal beschrieben. Für jedes intonatorische Phänomen sind bestimmte Töne oder Tonkombinationen definiert. Phrasengrenzen werden mit einem Phrasenton beschrieben, für B3-Grenzen gibt es zusätzlich einen Grenzton. Akzente werden mit einem entsprechenden Akzentton etikettiert. Alle Etikettierungen wurden perzeptiv und manuell ausgeführt (zur Konsistenz der Etikettierungen siehe z. B. Reyelt, 1995).

Die Etikettierung auf perzeptiver Basis ist sehr zeitaufwendig. Für die statistischen Verfahren war aber eine größere Datenmenge nötig als die bisher prosodisch etikettierten Dialoge. Für die Prosodiemodule in Verbmobil wurden daher zwei Lösungen realisiert: die Etikettierung aufgrund von syntaktischen Regeln anhand der schriftlichen Transkription und die automatische Etikettierung eines gelesenen Korpus mit bekannten Satzstrukturen. In Verbmobil 2 wird mittlerweile eine Kombination dieser beiden Verfahren untersucht (Batliner et al., 1998b).

Die syntaktische Etikettierung von Phrasengrenzen erfolgte in Absprache mit dem Modul für Syntax. Es hatte sich herausgestellt, daß das Training mit den perzeptiv etikettierten Phrasengrenzen nicht die Ergebnisse brachte, die von der Syntax gewünscht wurden (Batliner et al., 1996). Daher wurde ein Etikettierungssystem entworfen, das prosodische und syntaktische Phrasengrenzen gleichermaßen berücksichtigen konnte. Diese Etikettierung wurde für die 8 CDs mit deutschen Sprachdaten aus Verbmobil 1 anhand der schriftlichen Transkription vorgenommen (Batliner et al., 1998a).

Für die automatische prosodische Etikettierung wurde auf ein gelesenes Korpus für Bahnanfragen (ERBA-Korpus) zurückgegriffen (Kießling, 1997). Die Textstruktur war mit Hilfe eines Satzgenerators automatisch erzeugt worden. Aufgrund dieser vorhersagbaren Satzstruktur konnten Phrasengrenzen und Akzente nach bestimmten Regeln automatisch etikettiert werden (Kießling et al., 1994Kießling, 1997).

Eine Weiterentwicklung dieser automatischen Etikettierung wird inzwischen auch auf die spontansprachlichen Daten von Verbmobil angewendet (Batliner et al., 1998b). Das Verfahren verwendet die syntaktisch-prosodische Phrasengrenzenetikettierung (Batliner et al., 1998a) und eine automatische Annotation der Wortarten. Die Etikettierung basiert auf der Grundregel, daß Akzente eher am Ende einer Phrase zu finden sind; außerdem sind bestimmte Wortarten unterschiedlich *akzentuierbar*: Es werden z. B. eher Inhaltswörter als Funktionswörter akzentuiert, Substantive eher als Verben, etc. (siehe auch die Abschnitte 3.1.5 und 3.2). Spezielle Wortklassen wie Fokuspartikeln, Fragewörter und Partikelverben werden mit gesonderten Regeln behandelt. Die Emphase/Kontrast-Akzente werden mit diesen Regeln allerdings nicht erfaßt, da sie auf jedes Wort fallen können.

Weiterhin gibt es Annotierungen für Dialogakte (Abschnitt 4.3.2). Dialogakte können stark durch die Prosodie geprägt sein: Durch den Satzmodus oder eine bestimmte Akzentuierung können mit dem gleichen Wortlaut verschiedene Dialogakte realisiert werden. Allerdings wurden die Dialogaktetiketten ebenfalls nur anhand des Textes vergeben und nicht perzeptiv überprüft. Dies bedeutete eine große Zeitersparnis (es wurden sehr viele Testdaten benötigt); in manchen Fällen müßte die Dialogaktannotation aber aufgrund der akustischen Realisierung revidiert werden (Jekat et al., 1995).

Dialognr.	Anzahl				Anteil FA in %	Wortlänge in ms	Dialogpartner	
	Turns	FA	PA	B3				
n001k	21	56	59	62	20.67	287	nps1(m)	nbs1(w)
n002ka	22	55	64	51	23.64	303	nps1(m)	nmw1(w)
n002kb	6	19	27	19	20.00	288	nps1(m)	nmw1(w)
n002kc	15	37	38	36	23.47	287	nps1(m)	nmw1(w)
n003k	16	65	70	63	21.88	304	nps1(m)	nsp2(w)
n008ka	30	63	73	73	19.03	315	nhk1(m)	njk2(m)
n008kb	10	25	31	23	22.60	285	nhk1(m)	njk2(m)
n009k	24	51	64	50	18.46	331	nps1(m)	nhm1(m)
n011k	15	45	61	53	21.20	324	nps1(m)	nhw3(m)
n017k	15	40	44	34	19.13	325	nps1(m)	nms5(m)
n019k	21	46	47	40	22.90	295	nhk1(m)	noh2(m)
Gesamt	195	502	578	504	21.18	304	-	-

Tabelle 5.1: *Inhaltsdaten für die Testdialoge: Anzahl der Turns und der Fokusakzente (FA), Anzahl der Phrasenakzente (PA) und Phrasengrenzen (B3), prozentualer Anteil der fokussierten Wörter, die durchschnittliche Wortlänge (aller Wörter) und die Dialogpartner (männlich/weiblich).*

5.5 Testdaten für die Fokuserkennung

Zum Testen der Fokuserkennung wurde eine Teilmenge der prosodisch etikettierten Verbmobildaten ausgewählt. Diese Teilmenge wurde als Testmenge unter den Intarc-Projektpartnern abgesprochen. Es handelt sich um 11 Dialoge der CD 1. Die Dialoge wurden von 7 Sprechern und 3 Sprecherinnen gesprochen. Im Durchschnitt sind 21 % der Wörter fokussiert (siehe Tabelle 5.1).

Nach Prüfung der vorhandenen prosodischen Etiketten wurde entschieden, eine eigene Fokusetikettierung durchzuführen. In einer anschließenden Auswertung zeigte es sich, daß die etikettierten Phrasenakzente dem Fokuskonzept nur zu etwa 70 % entsprachen (siehe Tabelle 5.2). Die Fokusetiketten wurden für die 11 Testdialoge von der Autorin nach akustischer Wahrnehmung gesetzt. Die Grenzen der Fokusakzente wurden dabei auf ein Wort festgelegt.

Tabelle 5.2 zeigt das Verhältnis der Fokusakzente zu den anderen prosodischen Etiketten. Etwa 73 % der Fokusakzente befinden sich direkt (hier definiert auf dem letzten Wort der Phrase) an einer starken Phrasengrenze (B3). 99 % der Fokusakzente (Summe der Spalten 2 bis 4) sind ebenfalls entweder als Phrasenakzent (PA), Nebenakzent (NA) oder Emphase/Kontrast (EK) etikettiert worden. Damit ist garantiert, daß diese Fokusakzente auch von einer anderen Person als perzeptiv auffällig eingestuft wurden. Umgekehrt sieht das Verhältnis etwas ungünstiger aus: Nur 70 % der Phrasenakzente sind gleichzeitig Fokusakzente. Immerhin gilt dies aber für über 90 % der Emphase/Kontrast-Akzente, so daß diese Etiketten für die Fokuserkennung genutzt werden könnten.

Dialognr.	FA-B3	FA-PA	FA-NA	FA-EK	PA-FA	NA-FA	EK-FA
n001k	89.29	78.57	5.36	16.07	74.58	5.56	90.00
n002ka	78.18	74.55	10.91	14.55	64.06	15.00	80.00
n002kb	63.16	94.74	5.26	–	66.67	9.09	–
n002kc	62.16	81.08	5.41	8.11	78.95	6.45	75.00
n003k	70.77	72.31	12.31	10.77	67.14	16.67	100.00
n008ka	73.02	77.78	9.52	12.70	67.12	8.57	88.89
n008kb	84.00	80.00	16.00	–	64.52	19.05	–
n009k	68.63	88.24	5.88	5.88	70.31	5.45	100.00
n011k	88.89	91.11	4.44	4.44	67.21	6.90	100.00
n017k	60.00	80.00	17.50	2.50	72.73	15.91	100.00
n019k	69.57	80.43	10.87	6.52	78.72	16.67	100.00
Gesamt	73.42	81.71	9.41	7.41	70.18	11.39	92.65

Tabelle 5.2: *Verhältnis der Fokusakzente (FA) zu den anderen prosodischen Etiketten: starke Phrasengrenze (B3), Phrasenakzent (PA), Nebenakzent (NA), Emphase/Kontrast (EK) (Angaben in %).*

In Tabelle 5.3 sind die Anzahlen der Fokus- und Phrasenakzente pro Phrase aufgeführt. Im Gegensatz zur ursprünglichen Forderung, im Regelfall nur einen Phrasenakzent pro Phrase zu vergeben (siehe Abschnitt 5.4), wurden 20 % der Phrasen mit mehr als einem Phrasenakzent etikettiert. Fokusakzente wurden insgesamt weniger vergeben; zu 75.8 % befindet sich in jeder Phrase ein Fokusakzent. Ein Mehrfachfokus kommt nur in 11.7 % der Phrasen vor.

Zur weiteren Absicherung der etikettierten Fokusakzente, die ja nur auf der Wahrnehmung von einer Person basierten, wurde eine Teilmenge dieser Daten in einer späteren Untersuchung von 5 Mitarbeitern des Instituts (phonetisch geschult) noch einmal etikettiert (siehe Abschnitt 7.5). Nach Auswertung dieses Perzeptionstests wurden die ursprünglichen Fokusetiketten noch einmal überprüft und gegebenenfalls revidiert. Alle angegebenen Vergleichsdaten und Erkennungsraten sind mit den neu durchgesehenen Fokusetiketten berechnet worden.

	Anzahl Akzente pro Phrase				
	0	1	2	3	4
Anzahl Phrasen mit FA	63	382	58	1	0
Prozentanteil	12.5	75.8	11.5	0.2	0.0
Anzahl Phrasen mit PA	47	356	82	18	1
Prozentanteil	9.3	70.6	16.3	3.6	0.2

Tabelle 5.3: *Verhältnis der Fokusakzente und Phrasenakzente zu den mit B3-Grenze etikettierten Phrasen (Anzahl und prozentualer Anteil).*

6. Verfahren zur Fokuserkennung

Das folgende Kapitel beschäftigt sich mit der Entwicklung und Implementierung eines Verfahrens zur Fokuserkennung. Die Grundidee, die zu der Entwicklung führte, wird beschrieben, und die Umsetzung in ein automatisches Verfahren wird ausführlich erläutert. Das Verfahren gliedert sich in drei Module: 1. Bearbeitung der Grundfrequenzkontur, 2. Erstellung einer Referenzgerade aus der Grundfrequenz, 3. Ermittlung der Fokusakzente. Die Ergebnisse und mögliche Verbesserungen mit Hilfe des Satzmodus werden ausführlich dargestellt.

Das Verfahren versucht, die Fokuserkennung mit Hilfe einer globalen Verlaufsbeschreibung der Sprachäußerung zu lösen. Es verwendet nur die Grundfrequenz als Informationsquelle. Die Energie erwies sich als unzuverlässig (siehe Abschnitt 7.1), eine direkte Dauerinformation war nicht vorhanden, da nur das Signal Eingabequelle war. Ziel war es, ohne Worthypotheseninformation möglichst viel mit alleiniger Hilfe der akustisch-prosodischen Parameter zu erkennen. Durch die Architekturvorgabe von Intarc (siehe Abschnitt 4.2) war es außerdem erforderlich, daß das Verfahren inkrementell arbeitete.

6.1 Idee des Verfahrens

Das Verfahren zu Fokusakzenterkennung wurde inspiriert durch Untersuchungen für schwedische Spontansprache (Bruce und Touati, 1990Bruce und Touati, 1992). In diesem Projekt ging es darum, die Prosodie von gesprochener und gelesener Sprache zu untersuchen. Ein wesentliches Ziel bestand darin, den Einsatz der prosodischen Mittel in einer Dialogstruktur zu erforschen.

Besonders interessant sind die Untersuchungen für Fokusakzente. Im Laufe einer Äußerung bleiben unfokussierte Akzente stets auf einem Prominenzlevel; tritt aber ein Fokusakzent auf, gibt es ein klares *Downstepping* aller folgenden Nicht-Fokusakzente. Ein Fokusakzent markiert also einen Wendepunkt in einer Äußerung, da danach die Intonationskontur signifikant steil abfällt. Der Verlauf der Intonation ist daher auch ‘semantisch gesteuert’. In dieser Beziehung traten keine Unterschiede zwischen spontaner und gelesener Sprache auf (Bruce und Touati, 1992).

Ausgehend von diesen Untersuchungen ergaben sich folgende Fragestellungen für diese Arbeit:

1. Läßt sich das Phänomen eines Wendepunkts in der Grundfrequenzkontur, gesteuert durch den Fokusakzent, ebenfalls im deutscher Spontansprache finden?

2. Ist diese Eigenschaft stark genug ausgeprägt, um sie für die automatische Erkennung zu nutzen?

3. Wie kann die F_0-Kontur in einer einfachen Referenzgerade abstrahiert werden, welche Ankerpunkte sind notwendig?

Wie in Abschnitt 3.3.2 gezeigt wurde, ist der Einsatz der prosodischen Mittel auch in nah verwandten Sprachen unterschiedlich intensiv. Schwedisch zählt dabei eher zu den 'prosodisch expressiven' Sprachen (Grønnum, 1990). Für deutsche Spontansprache ist zumindest zu erwarten, daß *finale* Fokusakzente sich nur schwach intonatorisch hervorheben (Grønnum, 1990) und somit schlechter erkannt werden können.

6.2 Bearbeitung der Grundfrequenzkontur

Eine Grundvoraussetzung für die Nutzung prosodischer Information ist, daß ein möglichst genauer und verläßlicher F_0-Verlauf der jeweiligen Äußerung vorliegt. Ein erster Schritt hierzu ist die Gewinnung der F_0-Werte durch ein Verfahren zur Grundfrequenzbestimmung. Wie im 2. Kapitel gezeigt wurde (Abschnitte 2.3 und 2.4), ist der Grundfrequenzverlauf einer Äußerung aber ein komplexes Gebilde, das vielen verschiedenen Einflüssen unterworfen ist. Hinzu kommt noch, daß gerade auch durch diese Einflüsse die Messung der Grundfrequenz nicht hundertprozentig verläßlich ausgeführt werden kann.

Um die Untersuchung von F_0-Verläufen im globalen Verlauf auf eine sicherere Grundlage zu stellen, wurde ein Verfahren angewendet, das sowohl Fehler der Grundfrequenzmessung als auch mikroprosodische Einflüsse (siehe Abschnitt 2.4) bereinigt. Dieses Verfahren war bereits für eine andere Anwendung entwickelt worden (Petzold, 1993); zur Verwendung in dieser Arbeit wurde es entsprechend angepaßt.

6.2.1 Störungen des F_0-Verlaufs

Unerwünschte Störungen des F_0-Verlaufs können in unserem Fall durch Fehler aus der Grundfrequenzmessung oder aus der Sprachproduktion an sich (mikroprosodische Einflüsse, unregelmäßige Anregung; siehe Abschnitte 2.1.1, 2.4) entstehen. Diese äußern sich folgendermaßen in den akustischen Parametern:

- Bei nur teilweiser Stimmhaftigkeit treten Streuungen im F_0-Verlauf auf.
- Überdurchschnittlich hohe F_0-Werte deuten auf einen Oktavsprung oder ganz allgemein auf einen Meßfehler hin.
- Unregelmäßige Anregung wie Laryngalisierung erzeugt sehr tiefe F_0-Werte.

- Stimmhafte Konsonanten, insbesondere Obstruenten, verursachen Absenkungen sowohl im F_0- als auch im Intensitätsverlauf.
- Bei koartikulatorischen Einflüssen von stimmlosen Obstruenten ist oft ein steiler Anstieg von F_0 im folgenden Vokal zu beobachten.
- Die Dauer der lokalen Abweichung ist im Vergleich zu makroprosodischen Erscheinungen eher kurz.

6.2.2 Beschreibung des Verfahrens zur Grundfrequenznachverarbeitung

Zur Grundfrequenzbestimmung wurde das kommerzielle Verfahren von ESPS [1] verwendet. Besonders häufig treten in diesem Verfahren Oktavsprünge als Fehler auf. Im Vergleich mit anderen Verfahren schneidet ESPS dennoch recht gut ab, es hat eine relativ geringe Grobfehlerrate (Kießling, 1997, S. 181). Ein wesentliches Ziel bestand darin, die Fehler im Grundfrequenzverlauf zu korrigieren und die Kontur allgemein zu glätten, ohne dabei ihren charakteristischen Verlauf zu verändern.

Einfache Filterlösungen wie Medianfilter sind leicht zu implementieren. Die Entfernung von Grobfehlern längerer Dauer ist allerdings problematisch, da ihre Entfernung von der Medianlänge abhängig ist. Je länger aber ein Medianfilter ist, desto stärker werden charakteristische Gipfel verflacht und damit die typischen Eigenschaften einer Kontur verwischt. Taylor (1992) beklagt ebenfalls dieses Problem: Bei zu kleiner Medianlänge werden segmentale Abweichungen nicht entfernt, bei zu großer Länge werden wichtige F_0-Maxima zu sehr verflacht. Nur mit Hilfe von segmentaler Information ist es möglich, die Medianlänge variabel einzustellen (z. B. längerer Median im Bereich von stimmhaften Obstruenten).

Für die angestrebte prosodische Analyse in dieser Arbeit wurde ein komplexeres Verfahren benötigt, das sowohl die Ursachen aus der Sprachproduktion als auch aus der Grundfrequenzmessung selbst einbezieht. Dies sind zum einen die relativ kurze Dauer der mikroprosodischen Ereignisse und zum anderen einige typische Verlaufsmerkmale. Ein wesentliches Ziel ist es also, die Bereiche, die wirklich ein Intonationsmuster darstellen, von denen zu trennen, die segmentale Einflüsse repräsentieren.

Die Grundidee des hier beschriebenen Verfahrens ist nun, Bereiche mit einer längerfristigen Tendenz (sowohl fallend als auch steigend) zu detektieren. Diese Bereiche müssen dann nicht weiter bearbeitet werden. Eine allgemeine Glättung entfernt alle Werte, die nur kurzfristige Richtungsänderungen (bis 20 ms) verursachen, oder paßt sie der aktuellen globalen Verlaufsrichtung an.

Das charakteristische Muster einer mikroprosodischen Abweichung durch stimmhafte Obstruenten äußert sich in einem langsamen, glockenartigen Abfall und einem anschließenden schnellen Wiederanstieg der Kontur (siehe Abschnitt 2.4.2). In der Kontur läßt sich dies also in einem *relativ starken lokalen Minimum* finden, das nicht zu einer allgemein abfal-

[1] Entropic Signal Processing System

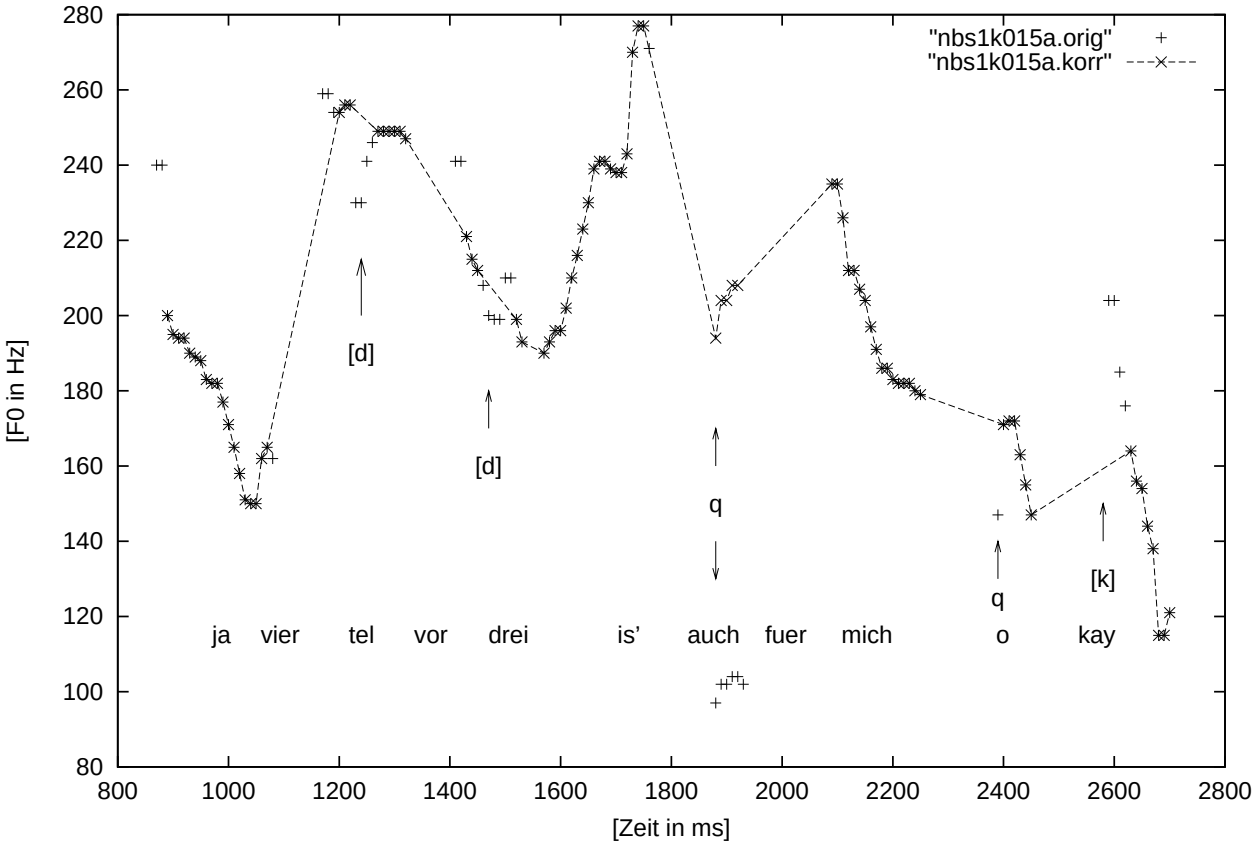

Abbildung 6.1: *Vergleich der Originalkontur (nbs1k015a.orig) mit der korrigierten Kontur (nbs1k015a.korr). Im Bereich der gestrichelten Linien befinden sich keine F_0-Werte, es handelt sich lediglich um Verbindungslinien. Die Markierung 'q' weist auf Laryngalisierung hin.*

lenden Tendenz wie z. B. einem bestimmten Intonationsmuster gehört (siehe Abbildung 6.1, bei den [d] - Lauten).

Kurzfristige Absenkungen durch stimmhafte Obstruenten werden vom Verfahren gelöscht. Ebenso werden stimmhafte Randbereiche abgeschnitten, wenn sie eine große Streuung aufweisen oder eine sehr starke Steigung (sowohl positiv als auch negativ) vorliegt (siehe Abbildung 6.1, beim [k]). Einzelne Ausreißer (z. B. Oktavsprünge) werden ebenfalls gelöscht, anschließend interpoliert und damit der allgemeinen Kontur angeglichen.

Besonders hohe oder niedrige Frequenzwerte müssen gesondert behandelt werden. Sehr hohe Werte treten besonders zu Beginn einer Äußerung auf, da das Grundfrequenzbestimmungsverfahren eine gewisse Einschwingphase benötigt; hier werden meist Werte geliefert, die in ihrer Höhe unwahrscheinlich sind. Diese Werte werden gelöscht. Sehr niedrige Werte können ein Hinweis für eine Laryngalisierung sein. Wenn die Werte um die Hälfte unter der allgemeinen Durchschnittsfrequenz liegen, werden diese durch einfaches Verdoppeln angeglichen (siehe Abbildung 6.1, beim Wort "auch").

Es stellt sich noch die Frage, was mit den stimmlosen Abschnitten geschehen soll. Für stimmlose Laute ist die Grundfrequenz nicht definiert, da auch keine Glottisschwingung

stattfindet. Trotzdem nehmen Hörer keine 'Löcher' in der Grundfrequenzkontur wahr, sie interpolieren über die stimmlosen Abschnitte (Cruttenden, 1986, S. 4). Für die weitere Analyse ist in unserem Fall eine komplizierte rechnerische Interpolation nicht nötig. Die stimmlosen Abschnitte werden durch einfache Geraden verknüpft. Weitere Details des Verfahrens sind in Petzold (1993Petzold (1994) dokumentiert.

Ein Beispiel für eine geglättete Grundfrequenzkontur findet sich in Abbildung 6.1. Die Werte der korrigierten Kontur wurden zur Verdeutlichung mit einer gestrichelten Linie verbunden. Die gelöschten Originalwerte sind noch als einzelne '+'-Zeichen im Bild enthalten. Sowohl die Absenkung durch stimmhafte Plosive (hier [d]) als auch ein Anstieg durch einen stimmlosen Plosiv ([k]) sind deutlich zu erkennen. (Man beachte, daß im Wort "viertel" das /t/ als [d] realisiert wurde!) In der Äußerung finden sich auch zwei Laryngalisierungen. Beim Wort "auch" realisiert die Sprecherin einen Teil der Werte in einer sehr tiefen Frequenzlage; dies wird vom Verfahren durch Verdopplung der Werte ausgeglichen. Im Bereich des Wortes "okay" äußert sich die Laryngalisierung nur in einer kurzen Streuung; Werte, die nicht zum Gesamtverlauf passen, werden vom Verfahren gelöscht.

Das Glättungsverfahren wurde zur Aufbereitung für die Fokuserkennung (Abschnitt 6.5) verwendet. In Tabelle 6.7 sind die Ergebnisse zu sehen: Ohne Glättung sinkt die Gesamterkennungsrate um 3 Prozentpunkte. Bei den ungeglätteten Daten werden sehr viel mehr Einfügungsfehler gemacht (Erkennungsrate für Nichtfokusbereiche), daher auch die sehr viel niedrigere Akkuratheit. Die Erkennungsrate für Fokusbereiche dagegen ist leicht höher. Weitere Erläuterungen finden sich in Abschnitt 6.6.

6.3 Berechnung der Erkennungsraten

Bevor nun die weiteren Erkennungsmodule vorgestellt werden, müssen zunächst einige Evaluationskriterien bestimmt werden. In jedem Einzelmodul muß ausgewertet werden, ob die jeweiligen Parametereinstellungen zum optimalen Ergebnis führen. Die Leistungsfähigkeit des Verfahrens insgesamt wird in Abschnitt 6.6 diskutiert.

Eine Forderung von Intarc bestand darin, daß die Verarbeitungsgeschwindigkeit sich an Echtzeit annähern sollte (siehe Abschnitt 4.2.1). Dies stellt kein Problem für das Verfahren dar; aufgrund der einfachen Regeln und der geringen Zeitverzögerung (etwa eine Phrase) kann die Berechnung in Echtzeit durchgeführt werden.

Am wichtigsten ist hier natürlich die Korrektheit der Ergebnisse. Die etikettierten Fokusakzente müssen detektiert werden, und es dürfen keine zusätzlichen Fokusakzente in nichtfokussierte Bereiche eingefügt werden. Dabei können verschiedene Prioritäten gesetzt werden: Ist es wichtiger, möglichst viele Fokusakzente zu finden, oder sollen möglichst wenig falsche Akzente eingefügt werden? Zur Verwendung in dieser Arbeit werden folgende Erkennungsraten definiert (in Anlehnung an Kießling, 1997).

Zunächst einmal wurden die Erkennungsraten für die einzelnen Klassen der fokussierten

Wörter und der nichtfokussierten Wörter berechnet:

$$\mathcal{ER}_F = \frac{\text{Anzahl korrekt erkannte fokussierte Wörter}}{\text{Anzahl fokussierte Wörter}} \tag{6.1}$$

$$\mathcal{ER}_{NF} = \frac{\text{Anzahl korrekt erkannte nichtfokussierte Wörter}}{\text{Anzahl nichtfokussierte Wörter}} \tag{6.2}$$

Die mittlere Erkennungsrate der beiden Klassen wird dann folgendermaßen berechnet:

$$\mathcal{ER}_M = \frac{\mathcal{ER}_F + \mathcal{ER}_{NF}}{2} \tag{6.3}$$

Da die Anteile der einzelnen Klassen sehr unterschiedlich sind, wurde auch eine Erkennungsrate mit entsprechender Gewichtung (die Gewichtungswerte 0.2/0.8 sind Durchschnittswerte, für jeden Dialog werden die ermittelten Fokusanteilwerte eingesetzt) berechnet:

$$\mathcal{ER}_{Gew} = (\mathcal{ER}_F \cdot 0.2) + (\mathcal{ER}_{NF} \cdot 0.8) \tag{6.4}$$

Außerdem wurde die Gesamterkennungsrate ermittelt:

$$\mathcal{ER}_{Ges} = \frac{\text{Anzahl korrekt erkannte Wörter (fokussiert + nichtfokussiert)}}{\text{Anzahl Wörter gesamt}} \tag{6.5}$$

Als sehr strenges Maß wird auch noch die Akkuratheit verwendet. Hier werden Einfügungsfehler besonders negativ bewertet: Von den richtig erkannten fokussierten Wörtern werden die fehlerhaft erkannten subtrahiert. Bei vielen Einfügungsfehlern kann die Akkuratheit negativ werden.

$$\mathcal{A}_F = \frac{\text{Anz. korrekt erkannte fok. Wörter - Anz. falsch erkannte nichtfok. Wörter}}{\text{Anzahl fokussierte Wörter}} \tag{6.6}$$

6.4 Berechnung einer Referenzgerade

Verschiedene Studien haben untersucht, wie Hörer die Prominenzen in einer Äußerung auf der Basis von F_0-Charakteristika und der allgemeinen Kontur beurteilen (Übersicht in Gussenhoven et al., 1997). Eine wesentliche Frage war, welches die relevanten Parameter der Kontur für die Wahrnehmung sind: Sind es benachbarte F_0-Maxima, F_0-Minima oder beide?

6.4.1 Problemstellung

Gussenhoven und Rietveld (1994) beschreiben das Problem, eine passende Referenzgerade zu finden, um die Prominenzen benachbarter F_0-Maxima vergleichen. Die Distanz eines Maximums zu einer Referenzgeraden soll dabei Aufschluß über die Prominenz geben. Wichtig ist die Frage, ob die Maxima oder die Minima stabiler sind, um daraus

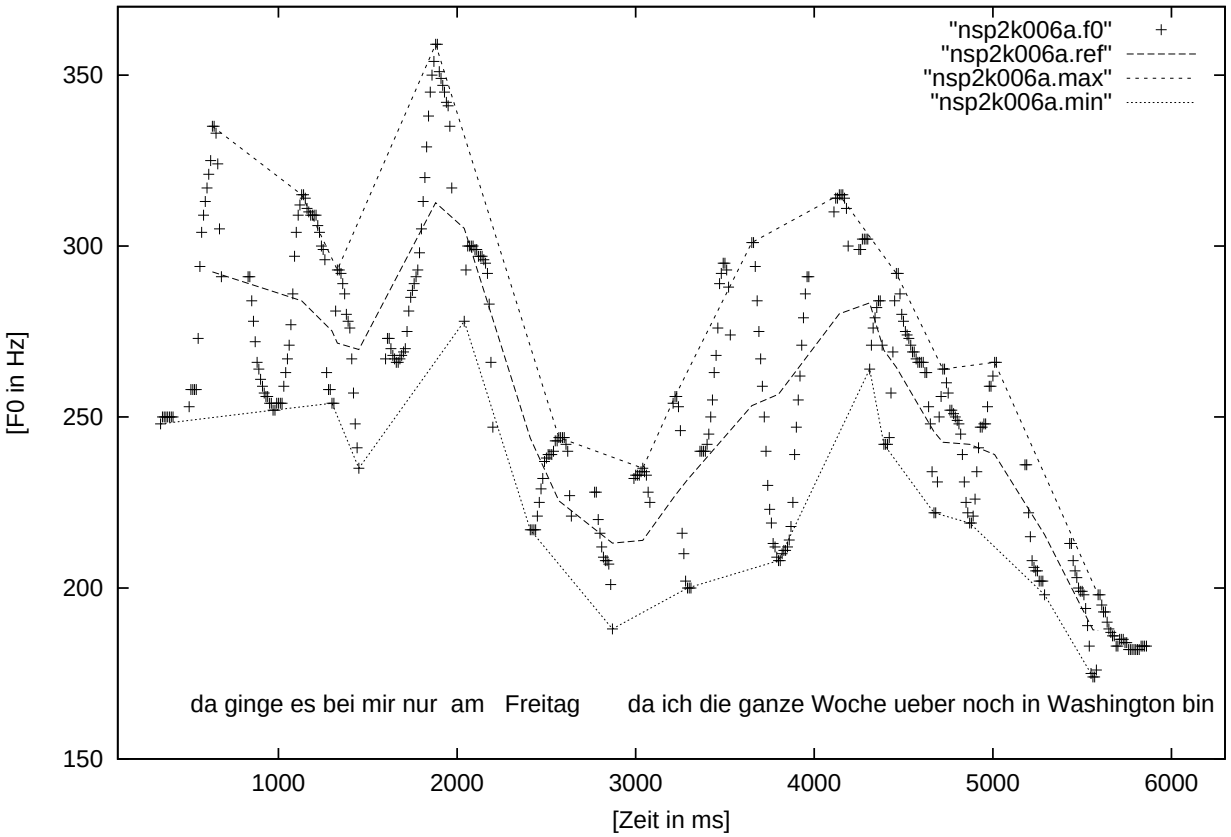

Abbildung 6.2: *F_0-Verlauf (nsp2k006a.f0) einer Äußerung mit Maximum- (nsp2k006a.max), Minimum- (nsp2k006a.min) und Referenzgerade (nsp2k006a.ref).*

eine Referenzgerade zu gewinnen. In der IPO-Tradition wird eine Basisgerade bevorzugt (t'Hart et al., 1990), die aus den Minima der F_0-Kontur gewonnen wird. Durch die Minimumpunkte wird eine Gerade gelegt, die in Richtung Äußerungsende fällt. Die Relevanz der Basisgeraden für die Wahrnehmung wurde in Terken (1991) gezeigt.

Bei Pierrehumbert (1979) kommt dagegen den F_0-Maxima höhere Bedeutung zu; ihrer Meinung nach sind sie perzeptuell wichtiger und auch zuverlässiger in der Produktion. Außerdem sei eine Basisgerade aus den Minima schwieriger zu bestimmen, da die Minima selten eine Gerade bildeten. Ihrer Meinung nach spielen die F_0-Minima keine oder nur eine kleine Rolle als Referenz.

6.4.2 Referenzgerade aus F_0-Minima und F_0-Maxima

Die Ergebnisse aus den verschiedenen Arbeiten legen nahe, daß sowohl die Maxima als auch die Minima der Grundfrequenzkontur als signifikant für eine Verlaufsbeschreibung angenommen werden müssen. Aus der nachbehandelten Grundfrequenzkontur wurden daher zunächst die Minima und Maxima als Stützpunkte für die Referenzgerade berechnet. Es wurde mit verschiedenen Parameterwerten experimentiert, um optimal die für dieses Problem signifikanten Minima und Maxima herauszufiltern. Eine anschließende Auswertung der Parameterwerte findet sich im nächsten Abschnitt.

Definition:
Innerhalb eines festgelegten Fensters (Fenstergröße **80 ms**) gelte ein Wert als signifikantes Maximum **sigMax** bzw. Minimum **sigMin**, wenn

$$1.\ \text{Bedingung} = \begin{cases} \text{sigMax} > \text{mind. } 2/3 \text{ der Fensterwerte} \\ \text{sigMin} < \text{mind. } 2/3 \text{ der Fensterwerte} \end{cases}$$

$$2.\ \text{Bedingung} = \begin{cases} \text{sigMax} > \text{vorheriges sigMin} + 10\ \% \\ \text{sigMin} < \text{vorheriges sigMax} - 10\ \% \end{cases}$$

Zusätzlich werden jeweils der erste und der letzte F_0-Wert der gesamten Äußerung als Maximum oder Minimum festgesetzt (abhängig von den folgenden bzw. vorhergehenden F_0-Werten).

Nach einigen Voruntersuchungen stellte sich heraus, daß weder die Maximumwerte noch die Minimumwerte allein eine verläßliche Referenzgerade darstellten. Da in dieser Arbeit vor allem längere Äußerungen mit mehreren Phrasen bearbeitet wurden, wurde außerdem eine lineare Deklinationsgerade verworfen. Als Lösung wurde ein Durchschnitt zwischen den Verbindungsgeraden der Maxima und der Minima gewählt. Ein Beispiel ist in Abbildung 6.2 zu sehen.

6.4.3 Auswertung der Parametereinstellungen

Die Parameterwerte wurden zunächst empirisch für einen Teil der Testdialoge ermittelt. Eine vollständige Auswertung war allerdings nur durch Berechnung von Erkennungsraten möglich. Die Werte sollten so eingestellt werden, daß die mittlere Erkennungsrate $\mathcal{ER}_M$ maximal wird.

	Erkennungsraten in %		
Fenstergröße	$\mathcal{ER}_F$	$\mathcal{ER}_{NF}$	$\mathcal{ER}_M$
40 ms	46.96	86.78	66.87
60 ms	48.50	88.38	68.44
80 ms	47.00	90.21	68.61
100 ms	44.75	91.50	68.12
120 ms	39.56	91.90	65.73

Tabelle 6.1: *Erkennungsraten für verschiedene Fenstergrößen bei der Berechnung von signifikanten Maxima und Minima.*

Tabelle 6.1 zeigt eine Auswertung für verschiedene Fenstergrößen bei der Berechnung der signifikanten Maxima und Minima aus der F_0-Kontur. Erwartungsgemäß steigt bei zunehmender Fenstergröße die Korrektheit für Nichtfokusbereiche $\mathcal{ER}_{NF}$, denn es werden weniger Maxima und Minima als Stützpunkte ausgewählt, so daß auch weniger Fehler gemacht werden. Bei kleinerer Fenstergröße steigt die Korrektheit für Fokusbereiche $\mathcal{ER}_F$,

Schwelle	Erkennungsraten in %		
	$\mathcal{ER}_F$	$\mathcal{ER}_{NF}$	$\mathcal{ER}_M$
0 %	48.95	86.92	67.94
5 %	48.63	87.89	68.26
10 %	47.00	90.21	68.61
15 %	42.39	91.55	66.97
20 %	39.80	92.64	66.22

Tabelle 6.2: *Erkennungsraten für verschiedene Schwellen bei der Berechnung von signifikanten Maxima und Minima.*

Kombination	Erkennungsraten in %		
	$\mathcal{ER}_F$	$\mathcal{ER}_{NF}$	$\mathcal{ER}_M$
60 ms und 5 %	50.54	85.00	67.77
60 ms und 15 %	43.40	90.20	66.80
100 ms und 5 %	46.02	89.99	68.01
100 ms und 15 %	41.18	92.77	66.98

Tabelle 6.3: *Erkennungsraten für weitere Parameterkombinationen bei der Berechnung von signifikanten Maxima und Minima.*

dies gilt allerdings nicht für die Größe von 40 ms: Offensichtlich ist bei sehr vielen Maxima und Minima als Stützpunkten die Verwechslungsgefahr so groß, daß die korrekten Fokusmaxima übersehen werden. Als optimaler Wert hat sich eine Fenstergröße von 80 ms erwiesen, die eine maximale mittlere Erkennungsrate $\mathcal{ER}_M$ garantiert. Wenn eine höhere Fokuserkennungsrate bei zu vernachlässigenden Einfügungen bevorzugt wird, ist auch eine Fenstergröße von 60 ms noch akzeptabel.

In Tabelle 6.2 wird die zweite Bedingung zur Berechnung der signifikanten Minima und Maxima untersucht. Je höher die Schwelle ist, ab der ein Maximum oder Minimum akzeptiert wird (es muß eine bestimmte Minimal- bzw. Maximalhöhe zum vorhergehenden Minimum/Maximum haben; siehe auch Abschnitt 6.4.2), desto höher ist auch die Erkennungsrate für Nichtfokusbereiche $\mathcal{ER}_{NF}$. Umgekehrt sinkt dann die Erkennungsrate für Fokusbereiche $\mathcal{ER}_F$. Der optimale Wert für die mittlere Erkennungsrate $\mathcal{ER}_M$ liegt bei einer Schwelle von 10 %. Die Tabellen 6.1 und 6.2 wurden jeweils mit der optimalen Schwelle (10%) bzw. der optimalen Fenstergröße (80 ms) berechnet. Eine Kombination von geringfügig schlechteren Parameterwerten untereinander (siehe Tabelle 6.3) brachte keine weitere Verbesserung der Erkennungsraten.

6.5 Bestimmung der Fokusakzente

In Anlehnung an Bruce und Touati (1990) muß sich der Fokusakzent im Bereich des stärksten Abfalls des F_0-Verlaufs befinden. Dieser Bereich wird innerhalb der Referenz-

geraden ermittelt. Zur weiteren Auswertung des Verfahrens werden die Korrelationen der Referenzgerade mit verschiedenen Parametern untersucht.

6.5.1 Nutzung von Referenzgerade und F_0-Maxima

Es wird im allgemeinen davon ausgegangen, daß sich ein Fokusakzent im Deutschen auf einem F_0-Maximum befindet (im Gegensatz z. B. zum Dänischen). Eine Auswertung der gefundenen F_0-Maxima mit den besten Einstellungswerten ergab, daß sich 85 % von diesen innerhalb eines etikettierten Fokusbereichs befinden. Aus der Menge der gefundenen F_0-Maxima müssen nun diejenigen ermittelt werden, die einen Fokusakzent markieren. Dies geschieht mit Hilfe der Referenzgeraden.

Für diese werden zunächst alle Steigungen berechnet. Innerhalb der fallenden Abschnitte der Referenzgeraden wird jeweils der Bereich mit der maximalen negativen Steigung ermittelt. Ausgehend vom Mittelpunkt dieses maximal abfallenden Bereichs wird das nächstliegende, vorhergehende F_0-Maximum als Fokuspunkt festgelegt.

Jeweils bei einem Wiederansteigen der Referenzgeraden wird ein neuer Satzabschnitt angenommen, daher können mehrere Punkte mit steilstem Abfall in einer Äußerung gefunden werden. Es gibt also keine Beschränkung bezüglich der Anzahl der Fokusakzente in einer Äußerung; inwieweit die Referenzgerade mit Phrasengrenzen (etikettierte B3-Grenzen) korreliert, wird im nächsten Abschnitt untersucht.

Das Verfahren ist inkrementell: Nach Durcharbeiten einer Phrase kann bereits ein Fokus ausgegeben werden. Mit diesem Verfahren werden nur Fokuspunkte gesucht. Die Abbildung dieser Zeitpunkte auf Wörter oder Konstituenten muß von anderen Modulen durchgeführt werden. Benötigt werden dazu Worthypothesen und – sofern es nötig ist, engen und weiten Fokus zu unterscheiden – auch Kontextinformation (siehe Abschnitte 3.1.5 und 8.4).

6.5.2 Korrelationen der Referenzgerade

Eine Auswertung der Steigungswerte der Referenzgerade soll Aufschluß darüber geben, wie stark die Steigung mit den etikettierten Fokusbereichen korreliert. Die gemittelten Steigungswerte für alle Dialoge (positive und negative Steigung) ergeben eine Zahl von -0.06, d. h. Steigen und Fallen der Referenzgerade sind in etwa gleichverteilt. Die negativen Steigungswerte allein haben einen Mittelwert von -0.38. Die Stellen mit maximalem Abfall, die vom Fokusdetektor gefunden werden, haben einen mittleren Steigungswert von -0.58, wobei die korrekten Fokuspunkte einen Wert von -0.63 und die fälschlich erkannten einen Wert von -0.52 haben.

In Tabelle 6.4 sind die gemittelten minimalen Steigungswerte für bestimmte Bereiche zu sehen. Im etikettierten Fokusbereich wurde jeweils die geringste Steigung ermittelt. Dies wurde ebenfalls für die Abschnitte vor und nach dem etikettierten Fokuswort (im Abstand von 200 ms) durchgeführt. Die Mittelwerte wurden für alle Steigungswerte insgesamt und für die negativen Steigungswerte gesondert berechnet. Da jeweils die geringste Steigung

	Abstände zum Fokuswort in ms					
Mittel der minimalen Steigungswerte	200 - 400	0 - 200	0	0 - 200	200 - 400	400 - 600
	vor Fokus		Fokuswort	nach Fokus		
alle Werte	-0.10	-0.11	-0.18	-0.25	-0.17	-0.10
nur negative Werte	-0.37	-0.40	-0.46	- 0.44	-0.42	-0.38

Tabelle 6.4: *Mittlere Steigungswerte in der Referenzgeraden für etikettierte Fokusbereiche und für die Bereiche davor und danach*

		Abstände zu B3 in ms			
Verlaufsform	Anteil gesamt	0 - 100	100 - 200	200 - 300	> 300
fallend	35.0	49.4	14.9	7.1	28.6
steigend	33.0	35.9	18.2	12.6	33.3
fallend-steigend	18.0	25.6	22.1	22.1	30.2
steigend-fallend	14.0	61.2	22.4	8.9	7.5

Tabelle 6.5: *Korrelationen des Referenzgeradenverlaufs mit den etikettierten B3-Grenzen, sortiert nach Abständen, in Prozent*

ermittelt wurde, überwiegen die negativen Steigungswerte auch in der Gesamtberechnung.

Es läßt sich aber eine klare Tendenz erkennen: Die Referenzgerade fällt bereits im fokussierten Wort ab und hat ihre höchste negative Steigung direkt in den ersten 200 ms danach (-0.25). Da die durchschnittliche Wortlänge etwa 300 ms beträgt (siehe Tabelle 5.1), heißt das also, daß die Referenzgerade bereits im ersten Wort nach dem fokussierten Wort am stärksten abfällt. Die Zahlen für die rein negativen Steigungswerte sind weniger deutlich, bestätigen aber die Tendenz.

Weiterhin wurde untersucht, inwieweit die Referenzgerade mit den etikettierten B3-Grenzen korreliert (siehe Abschnitt 5.4). Im vorgestellten Erkennungsverfahren wird jeweils bei einem Wiederansteigen der Referenzgeraden (Wechsel von 'fallend' zu 'steigend') die Suche neu aufgenommen. Eine andere denkbare Möglichkeit wäre, jeweils pro etikettierter Phrase einen Fokusakzent zu suchen (siehe Abschnitt 7.2).

Zur Auswertung wurde jedes Geradenteilstück der Referenzgerade als fallend, steigend oder als Kombination davon automatisch etikettiert (abhängig vom direkt vorhergehen-

		Abstände zu B3 in ms				
Verlaufsform	Anteil gesamt	0 - 100	100 - 200	200 - 300	300 - 400	> 400
fallend-steigend	54.8	10.0	18.5	17.6	11.8	41.2
steigend-fallend	45.2	23.0	13.8	12.8	6.1	44.3

Tabelle 6.6: *Korrelationen des Referenzgeradenverlaufs (nur wechselnde Abschnitte) mit den etikettierten B3-Grenzen, sortiert nach Abständen, in Prozent*

Dialognr.	Fokusanteil	$\mathcal{ER}_{Gew}$	$\mathcal{ER}_{M}$	$\mathcal{ER}_{F}$	$\mathcal{ER}_{NF}$	$\mathcal{A}_{F}$	$\mathcal{ER}_{Ges}$
n001k	20.67	77.88	61.33	33.57	89.10	-28.52	75.43
n002ka	23.64	76.31	63.82	38.95	88.68	-17.05	73.59
n002kb	20.00	87.97	81.00	67.67	94.33	25.50	86.33
n002kc	23.47	82.74	72.43	52.27	92.60	18.00	80.87
n003k	21.88	79.06	68.62	48.88	88.38	-5.56	76.06
n008ka	19.03	78.79	66.40	43.23	89.57	-22.53	76.87
n008kb	22.60	80.75	67.55	44.10	91.00	3.40	78.30
n009k	18.46	82.61	71.73	51.29	92.17	0.38	80.92
n011k	21.20	79.48	71.30	55.13	87.47	-3.80	77.00
n017k	19.13	82.15	70.00	50.33	89.67	-5.60	80.07
n019k	22.90	76.54	60.50	31.62	89.38	-23.86	74.00
Gesamt	21.18	80.39	68.61	47.00	90.21	-5.42	78.13
ohne Glättung	21.18	78.20	67.84	48.59	87.08	-20.30	75.42

Tabelle 6.7: *Fokusanteile und Erkennungsraten in Prozenten.*

den Geradenteilstück). Zu jeder B3-Grenze wurde der nächstliegende Referenzgeradenabschnitt ermittelt, und der zeitliche Abstand wurde gemessen. Tabelle 6.5 zeigt die Ergebnisse: Nur ein kleinerer Anteil der B3-Grenzen korreliert direkt mit einem Wechsel im Verlauf (18 % bei 'fallend-steigend', 14 % bei 'steigend-fallend'). Die 'steigend-fallend'-Abschnitte liegen aber zu einem sehr hohen Prozentsatz (61.2 %) im Abstand von weniger als 100 ms an einer B3-Grenze.

Anschließend wurden nur die Korrelationen der B3-Grenzen mit den 'fallend-steigend' und 'steigend-fallend'-Abschnitten untersucht (siehe Tabelle 6.6). Die Gesamtanteile verteilen sich auf 54.8 % auf 'fallend-steigend' und 45.2 % auf 'steigend-fallend'. Auch hier liegt ein relativ hoher Prozentsatz der 'steigend-fallend'-Abschnitte sehr nah (Abstand < 100 ms) an einer B3-Grenze. Insgesamt kann man aber nicht von einem direkten Zusammenhang zwischen Referenzgerade und etikettierten Phrasengrenzen ausgehen; von daher können B3-Grenzen eine zusätzliche Informationsquelle für die Erkennung darstellen. Dies wird im nächsten Kapitel (Abschnitt 7.2) näher untersucht.

6.6 Ergebnisse

Das Verfahren wurde mit sämtlichen Dialogen der etikettierten Testmenge (Abschnitt 5.5) getestet. Der Anteil der fokussierten Wörter in den Äußerungen beträgt im Durchschnitt 21.2 %, er schwankt zwischen 18.5 % und 23.5 %, abhängig vom jeweiligen Dialog. Die Fokusanteile werden zur Berechnung der gewichteten Erkennungsrate $\mathcal{ER}_{Gew}$ benötigt (zur Berechnung der Erkennungsraten siehe Abschnitt 6.3). Die Erkennungsraten für die einzelnen Dialoge finden sich in Tabelle 6.7.

In Tabelle 6.7 ist zu erkennen, daß die Erkennungsraten für Fokusbereiche signifikant schlechter sind als für Nichtfokusbereiche, d. h. es gibt sehr viel mehr Auslassungen als

Einfügungen. Die Erfahrungen in der Zusammenarbeit mit anderen Modulen haben aber gezeigt (Abschnitt 4.2), daß es weniger gravierend ist, wenn ein Fokus nicht erkannt wird, als wenn in einem nichtfokussierten Bereich ein Fokus gefunden wird. Ein Hinweis auf einen Fokus soll für die linguistischen Module ja nur eine zusätzliche Stütze sein - bleibt dieser Hinweis aus, wird es die semantische Analyse nicht ablenken; wird allerdings ein falscher Fokus geliefert, kann dies zu mehr oder weniger schweren Fehlern führen. Von daher ist es durchaus sinnvoll, die 'leichter' zu erkennenden Nichtfokusbereiche in $\mathcal{ER}_{Gew}$ relativ hoch zu gewichten.

Die Akkuratheit $\mathcal{A}_F$ ist insgesamt noch recht niedrig. Die Unterschiede zwischen den einzelnen Dialogen zeigen sich aber recht deutlich. Manche Sprecher oder Sprecherinnen stellen offensichtlich noch ein Problem für das Verfahren dar. Die unterschiedlichen Erkennungsraten für die einzelnen Dialoge spiegeln möglicherweise den Grad der 'Lebhaftigkeit' wider. Bei einer engagierteren Diskussion werden auch die Fokusakzente deutlicher markiert. Wesentliche Unterschiede zwischen Männer- und Frauenstimmen in der Deutlichkeit der Fokusmarkierungen wurden aber nicht festgestellt (siehe auch 7.4).

Die ersten Ergebnisse mit dem Verfahren sind durchaus zufriedenstellend. Es läßt sich aber noch einiges verbessern: Die Erkennung ist stark abhängig von einem korrekten Grundfrequenzverlauf. Sowohl die Grundfrequenzberechnung als auch die Nachbearbeitung müssen mit möglichst wenig Fehlern stattfinden. Die Berechnung der Referenzgerade ist davon stark abhängig; bei falschen Extrema durch einen fehlerhaften Grundfrequenzverlauf kann die gesamte Erkennung fehlschlagen. Die Berechnung der Extrema ist weniger fehleranfällig; es kann außerdem eingestellt werden, wie streng die Auswahl der Extrema gehalten werden soll (Verhältnis Minima-Maxima und Anzahl der Umgebungswerte, die kleiner bzw. größer sein müssen; siehe auch Abschnitt 6.4.3).

6.7 Satzmodus

Bei Durchsicht der Daten fiel vor allem auf, daß die Erkennungsrate für Fragen deutlich niedriger war als für Aussagen. Außerdem gibt es eine starke Interaktion zwischen dem Satzmodus und der Position des Fokus im Satz. Wenn sich der Fokus am Beginn eines Satzes befindet, ist er meist sehr deutlich markiert. In satzfinaler Position dagegen werden nicht-terminaler Satzmodus und Fokus mit denselben intonatorischen Mitteln markiert (Anstieg der Grundfrequenz), so daß es schwierig ist, die beiden Phänomene zu trennen. Auch Batliner (1989a) vermutete, daß sich bei einer Trennung von Fragen und Nichtfragen die Erkennungsrate von Fokusakzenten deutlich steigern ließe.

Eady und Cooper (1986) untersuchten Fokusakzente in unterschiedlichen Satzpositionen für Fragen und Aussagen (siehe auch Abschnitt 3.5.2). Unterschiede zwischen Fragen und Aussagen wurden in erster Linie am Ende einer Äußerung festgestellt: Bei den Fragen war der letzte Grundfrequenzgipfel deutlich höher als bei den Aussagen. Bei Fokusakzenten zu Äußerungsbeginn fielen in einer Aussage die nachfolgenden Grundfrequenzwerte signifikant ab, während sie bei den Fragen weiterhin auf einem hohen Niveau blieben.

Wenn also der Fokus in einer ansteigenden F_0-Kontur (Frage oder progredienter Anstieg)

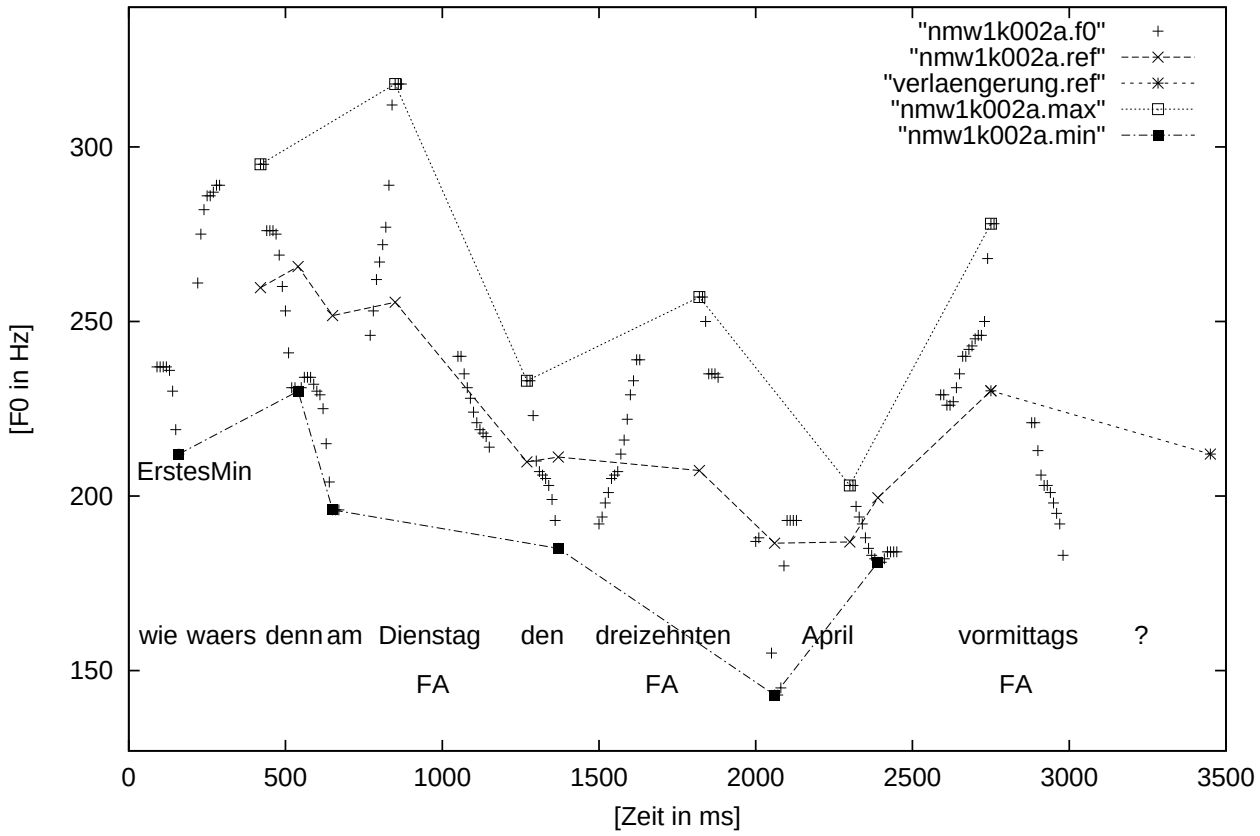

Abbildung 6.3: *F_0-Verlauf (nmw1k002a.f0) einer Äußerung mit Maximum- (nmw1k002a.max), Minimum- (nmw1k002a.min) und Referenzgerade (nmw1k002a.ref). Die Referenzgerade wurde mit den Koordinaten Erstes Minimum/ Ende der Äußerung verlängert (verlaengerung.ref). Dadurch kann der Fokusakzent in "vormittags" erkannt werden.*

Dialognr.	Fokusanteil	$\mathcal{ER}_{Gew}$	$\mathcal{ER}_M$	$\mathcal{ER}_F$	$\mathcal{ER}_{NF}$	$\mathcal{A}_F$	$\mathcal{ER}_{Ges}$
n001k	20.67	77.27	62.40	38.71	86.10	-40.05	74.57
n002ka	23.64	76.09	66.59	47.73	85.45	-22.55	72.68
n002kb	20.00	83.03	78.67	67.67	89.67	28.83	81.67
n002kc	23.47	79.72	73.10	56.67	89.53	5.20	77.27
n003k	21.88	76.65	69.62	54.31	84.94	-3.38	73.62
n008ka	19.03	76.35	67.40	49.33	85.47	-26.70	73.97
n008kb	22.60	84.84	78.05	65.70	90.40	22.50	82.30
n009k	18.46	81.16	73.67	58.58	88.75	-3.46	78.88
n011k	21.20	80.03	73.30	59.80	86.80	-4.67	77.53
n017k	19.13	82.45	72.27	55.93	88.60	-7.27	80.27
n019k	22.90	77.29	66.60	45.10	88.10	-15.10	74.43
Gesamt	21.18	79.53	71.06	54.50	87.62	-6.06	77.02

Tabelle 6.8: *Fokusanteile und Erkennungsraten in Prozenten für die Version mit integriertem Satzmodus.*

'versteckt' ist, kann er nicht mehr mit den gleichen Methoden bestimmt werden wie bei den Aussagen. Im hier vorgestellten Fokusverfahren steigt die Referenzgerade dann ebenfalls an (Abschnitt 6.5), so daß kein 'steilster Abfall' mehr gefunden werden kann. Daher wurde nach Möglichkeiten gesucht, den Satzmodus als zusätzliche Information miteinzubeziehen (Elsner, 1997a). Zur Lösung des Problems wurde eine 'grobe' Satzmodusinformation simuliert: Die Referenzgerade wurde künstlich verlängert, mit der Höhe des ersten gefundenen Minimums als y-Wert und dem Endezeitpunkt der Äußerung als x-Wert (siehe Beispiel in Abbildung 6.3).

Bei einer ansteigenden Kontur endet die Referenzgerade ebenfalls ansteigend. Die Höhe des ersten Minimums liegt meist unterhalb des letzten Referenzpunkts; bei einer entsprechenden Verlängerung der Referenzgerade wird also ein letzter 'steiler Abfall' konstruiert. Umgekehrt führt die Verlängerung der Referenzgerade bei einer fallenden Kontur (z. B. bei einer Aussage) zu einem letzten Anstieg der Referenzgerade, da die Höhe des ersten Minimums meist darüber liegt. In diesem Fall können dann keine Einfügungsfehler durch zusätzlich detektierte Fokusakzente entstehen.

Im Ergebnis verbesserte sich die Erkennungsrate für die Fokusbereiche deutlich. Die Erkennungsraten sind abhängig von den einzelnen Dialogen, die jeweils unterschiedlich Anteile an ansteigenden Konturen haben. Die Werte für die einzelnen Dialoge sind in Tabelle 6.8 zu sehen. Ein Problem ist hier aber wieder, daß durch mehr gefundene Fokusakzente auch die Anzahl der Einfügungen steigt, d. h. die Richtigkeit der Erkennung für Nichtfokusbereiche wird schlechter. Die Akkuratheit sinkt allerdings nur unwesentlich ab.

Es bleibt also schwierig, die Fokusakzente am Ende einer ansteigenden Kontur zu erkennen bzw. umgekehrt keine Erkennungsfehler zu machen, indem eine Frageintonation fälschlicherweise für einen Fokusakzent gehalten wird. Zur Vermeidung von Einfügungsfehlern könnte evtl. der akustische Parameter Energie nützlich sein, da die Energie bei einem Frageanstieg absinkt, im Gegensatz zur Grundfrequenz (siehe Abschnitt 2.5.2). Diese Idee wurde hier allerdings nicht weiterverfolgt; der Parameter Energie wird in bezug auf die Markierung von Fokusakzenten im nächsten Kapitel untersucht (Abschnitt 7.1).

6.8 Klassifizierung von Fokusakzenten

Um zu untersuchen, wie gut verschiedene Ausprägungen von Fokusakzenten erkannt werden, wurde die Etikettierung der Fokusakzente noch einmal verfeinert. Die Akzente wurden in 5 verschiedene Gruppen klassifiziert; die Kriterien betreffen im wesentlichen die akustische Auffälligkeit. Dies schließt aber nicht aus, daß diese Auffälligkeit durch bestimmte syntaktische oder semantische Gegebenheiten unterstützt wird.

1. **Fm** Diese Art tritt meist zu Anfang einer Äußerung auf. Sie besteht oft aus einem kurzen Ausruf mit hohem emotionalem Gehalt, aber von geringer semantischer Bedeutung ("oh", "hm", "gut").

2. **Fd** Dies soll einem Standardfokus (*Default*) von geringer Prominenz entsprechen. Er kommt in deklarativen Äußerungen am Ende einer Phrase vor. Keine Silbe in

Kategorien	Fm	Fd	Ff	Fq	Fk
Originalversion	62.2	18.3	48.5	35.3	59.5
mit Satzmodus	57.8	25.0	53.8	55.9	66.6

Tabelle 6.9: *Erkennungsraten für Fokusbereiche $\mathcal{ER}_F$ in Prozent für die einzelnen Fokusakzentkategorien, für beide Programmversionen; Abkürzungen siehe Abschnitt 6.8*

dieser Phrase ragt akustisch hervor, so daß im allgemeinen das letzte Inhaltswort einer Phrase als Fokusakzent wahrgenommen wird.

3. **Ff** Dies bezeichnet einen Fokusakzent ohne spezielle Eigenschaften, der die prominenteste Silbe in einem Äußerungsabschnitt markiert.

4. **Fq** Hiermit werden die Fokusakzente benannt, die sich in einer Phrase mit ansteigendem Grundfrequenzverlauf befinden. Fokusintonation und Satzmodusmarkierung überlagern sich, daher ist die Trennung oft schwierig.

5. **Fk** Dies entspricht einem Kontrastakzent, der akustisch besonders auffällig aus seiner Umgebung herausragt (auch Emphase genannt), also eine Steigerung der Kategorie 'Ff'.

Der Hauptanteil in den Dialogäußerungen besteht aus den unmarkierten Fokusakzenten *Ff*, er entspricht etwa 53 %. Die Kategorien *Fd* und *Fq* umfassen 16 % bzw. 13 %. Die übrigen *Fm* und *Fk* befinden sich zu je 9 % in den Dialogen. Interessant sind nun die Erkennungsraten im Fokusbereich $\mathcal{ER}_F$ für diese 5 Kategorien (siehe Tabelle 6.9):

Die Ergebnisse legen nahe, daß es offensichtlich nicht sinnvoll ist, die Akzente der Kategorie *Fd* mit Hilfe von akustischen Parametern zu detektieren. Da diese akustisch nur schwach markiert sind, scheint es erfolgversprechender, diese mit Kenntnis von Phrasengrenzeninformation in einem linguistischen Modul zu ermitteln. Die Kategorie *Fq* war stark verbesserungswürdig, konnte aber durch die Integration von Satzmodusinformation deutlich gesteigert werden. Erwartungsgemäß war die Erkennungsrate für die Kategorie *Fk* besonders hoch für beide Verfahrensversionen, da sich die Kontrastakzente hier definitionsgemäß durch akustische Auffälligkeit auszeichnen.

In der Kategorie *Fm* wurden ebenfalls recht hohe Erkennungswerte erzielt. Wie auch verschiedene Untersuchungen in Kapitel 3 zeigten, sind Fokusakzente am Anfang einer Äußerung offensichtlich besonders deutlich akustisch markiert, unabhängig vom semantischen Gehalt. Bei dieser Kategorie ist natürlich auch ein relativ hoher emotionaler Gehalt für die akustische Auffälligkeit verantwortlich. In der Originalversion ist die Erkennung sogar noch etwas höher als in der Satzmodusversion; bei letzterer werden durch die Verlängerung der Referenzgerade wohl eher die späteren als die früheren Fokusakzente erkannt. Abbildung 6.4 zeigt ein Beispiel einer Äußerung mit verschiedenen Arten von Fokusakzenten. Die Kategorien *Fd* und *Fq* werden hier nicht erkannt; die Satzmodusversion kann aber *Fq* erkennen.

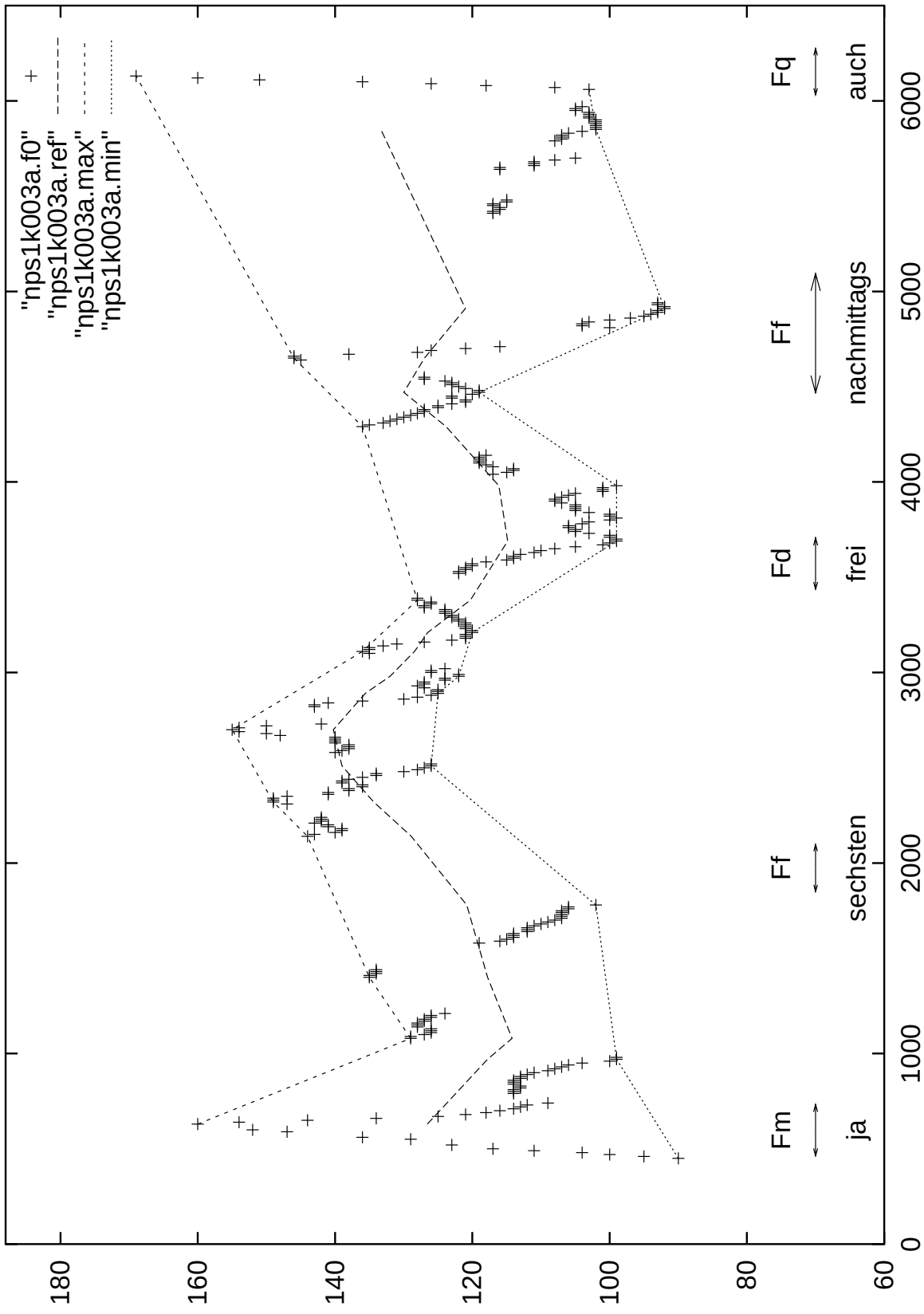

Abbildung 6.4: *Verschiedene Arten von Fokusakzenten in der Äußerung "Ja, am Dienstag, den 6. April hätt' ich noch frei, allerdings nur nachmittags. Geht es da bei Ihnen auch?". Erkannt wurden die Fokusakzente in "ja" und "nachmittags".*

7. Weitere Untersuchungen

Zur Ergänzung der Fokuserkennung wurden noch weitere Informationen geprüft. Zunächst erschien es vielversprechend, Energiewerte in die Bewertung miteinzubeziehen, zumal die Berechnung nicht aufwendig ist und auch inkrementell erfolgen kann. Außerdem wurden Experimente gemacht, um weitere syntaktische Information wie Phrasengrenzen zu nutzen. Phrasengrenzen könnten dazu beitragen, die Fokussuche gezielt auf einzelne Phrasen anzusetzen und somit das Problem in kleinere Zwischenschritte aufzuspalten.

Weitere Experimente beschäftigen sich mit der Unterscheidung von verschiedenen Fokusakzentformen. Untersucht wird die Frage, ob sich emphatisch/kontrastive Akzente in ihrer akustischen Form von den anderen Akzenten unterscheiden. Die untersuchten Parameter sind die Fokusmaxima im Grundfrequenzverlauf, außerdem ihre Verteilung bei unterschiedlichem Abstand zu einer Phrasengrenze.

Der überwiegende Teil der folgenden Untersuchungen wurde nicht mehr im Rahmen von Intarc durchgeführt, da das Projekt inzwischen abgeschlossen wurde. Deswegen wurden auch Experimente unternommen, die die Voraussetzungen von Intarc überschreiten. Abschließend folgt eine Auswertung der verschiedenen Sprecher in den untersuchten Verbmobil-Testdaten; die Erkennung wird auf ihre Sprecherabhängigkeit hin betrachtet. Weiterhin wurde die Fokusetikettierung ausgewertet, indem die Urteile von verschiedenen Hörern in einem Perzeptionstest untersucht werden.

7.1 Energie

Im allgemeinen gilt die Energie als ein wenig zuverlässiges Merkmal zur Bestimmung von Prominenzunterschieden. Verschiedene Untersuchungen (Campbell, 1995Sluijter und van Heuven, 1993) begründen dies aber damit, daß oft nur die Gesamtenergie getestet wurde - dagegen wurden insbesondere in den höheren Frequenzbändern durchaus signifikante Unterschiede zwischen prominenten und nichtprominenten Bereichen gefunden. Bei Campbell (1995) war dies im Bereich 2000 - 8000 Hz, bei Sluijter und van Heuven (1993) dagegen im Bereich 500 - 4000 Hz. In beiden Arbeiten wurde in erster Linie Laborsprache verwendet.

Dadurch ergaben sich folgende Fragestellungen: Sind die obengenannten Ergebnisse auf Spontansprache übertragbar? Ist der Parameter Energie für unsere Zwecke (zur Verbesserung der Fokuserkennung) brauchbar? Gibt es Unterschiede in den einzelnen Energiebändern, in denen sich fokussierte von nichtfokussierten Bereichen besonders deutlich abheben?

7.1.1 Experiment

In einer früheren Untersuchung (Petzold, 1996) wurde zunächst nur der erste Dialog (n001k) aus der Verbmobil-Testmenge (siehe Abschnitt 5.5) ausgewertet. Später stellte sich heraus, daß der Dialog n001k eher problematisch ist, da insbesondere die Erkennungsrate der beteiligten Sprecherin 'nbs1' extrem niedrig ist (siehe Tabelle 7.8). Da die Ergebnisse mit diesem Dialog nicht sehr befriedigend waren, wurde die Untersuchung noch einmal für andere Dialoge durchgeführt.

Für diese zweite Untersuchung wurden die 3 Dialoge n002ka, n002kb und n002kc verwendet. Sie enthalten insgesamt 43 Äußerungen, die Dialogpartner sind ein Sprecher (nps1) und eine Sprecherin (nmw1). Für die Dialoge stand eine Wortsegmentierung und neuerdings auch eine Silbensegmentierung zur Verfügung. Außerdem gab es eine inzwischen stark verbesserte Lautsegmentierung.

Nach Berechnung des Fourierspektrums wurde die Kurzzeitenergie für folgende Energiebänder gemessen: 0 - 500 Hz, 500 - 1000 Hz, 1000 - 2000 Hz, 2000 - 4000 Hz, 4000 - 8000 Hz, 0 - 8000 Hz. Die Energiewerte wurden jeweils für Wörter, Silben und Laute gemittelt. Die Energiewerte für die Laute wurden zusätzlich mit der Energie der jeweiligen Silbe normiert. Für jeden Laut bzw. Lautgruppe wurden Mittelwerte für fokussierte und nichtfokussierte Bereiche gemessen. Die Werte für die fokussierten Bereiche wurden als 100 % festgesetzt. die Werte für die nichtfokussierten Bereiche wurden als prozentualer Anteil davon berechnet.

7.1.2 Auswertung für Silben und Wörter

Die Energiewerte für einzelne Silben und Wörter zeigen sehr geringe Unterschiede für fokussierte Bereiche im Vergleich zu nichtfokussierten (siehe Tabelle 7.1). Dabei sind die Energiewerte für Wörter noch weniger aussagekräftig als die für Silben. Die Differenzen zeigen sich in erster Linie in den höheren Frequenzbändern, ab 2000 Hz für die Sprecherin, ab 4000 Hz für den Sprecher. Diese Differenzen erscheinen aber zu gering, als daß sie für eine Erkennung genutzt werden könnten.

Frequenzband (in Hz)	Sprecherin nmw1		Sprecher nps1	
	Silben	Wörter	Silben	Wörter
0 - 500	98.9	101.2	100.2	101.9
500 - 1000	97.6	100.7	99.1	100.1
1000 - 2000	97.5	100.1	99.2	99.7
2000 - 4000	96.0	99.5	98.7	99.9
4000 - 8000	97.3	99.3	98.2	99.3
0 - 8000	98.1	100.3	99.0	100.5

Tabelle 7.1: *Energieanteile (in %) der nichtfokussierten Wörter und Silben (Fokus = 100 %).*

7.1.3 Auswertung für Vokale

Da die Energiedifferenzen zwischen fokussierten und nichtfokussierten Wörtern oder Silben nicht ausreichend erscheinen, wurden die Berechnungen auch noch für einzelne Vokale durchgeführt. Die bereits vorliegende Lautsegmentierung wurde von der Autorin manuell nachkorrigiert. Dies war insbesondere für die Sprecherin nötig, um zuverlässige Energiewerte für die einzelnen Vokale zu bekommen.

In bezug auf die Energie haben Vokale bestimmte intrinsische Eigenschaften (siehe auch Abschnitt 2.4). Diese werden bestimmt durch den Öffnungsgrad des Vokals und damit die Höhe des 1. und 2. Formanten ([a] hat die höchste und [i] die niedrigste intrinsische Energie). Kurzvokale haben i. a. eine höhere Energiedifferenz als Langvokale. Berechnungen für intrinsische Vokalwerte finden sich für das Französische bei Rossi (1971) und für Englisch (Amerikanisch) bei Lehiste und Peterson (1959).

Die Energieanteile der nichtfokussierten Vokale im Verhältnis zu den jeweils fokussierten Vokalen sind in den Abbildungen 7.1 und 7.2 dargestellt. Es wurden nur die Vokale in die Auswertung mit aufgenommen, für die ausreichend Werte für beide Bereiche (fokussiert und nichtfokussiert) zur Verfügung standen.

Die Energiedifferenzen sind recht unterschiedlich für die einzelnen Vokale und für den Sprecher und die Sprecherin. Bei den o- und a-Lauten sind die Energiewerte für fokussierte und nichtfokussierte Vokale fast gleich. Aufgrund der allgemein höheren intrinsischen Energie dieser Vokale ist offensichtlich auch die Energie bei nichtfokussierten Vokalen noch relativ hoch. Etwas größere Energieunterschiede zeigen sich nur bei den e-Lauten. Es läßt sich allerdings kein Frequenzband ausmachen, für welches die Energiedifferenzen generell höher sind; dies ist stark abhängig vom Vokal und von den Sprechern.

Ähnliche Ergebnisse finden sich auch bei Portele (1998). Dort wurden die Korrelationen zwischen Energie und Prominenz untersucht (für amerikanisches Englisch, gelesene Sprache). Die Gesamtenergie steigt zwar mit höherer Prominenz, die Korrelation ist aber stark vokal- und sprecherabhängig.

7.1.4 Diskussion

Die Ergebnisse zeigen, daß die Energieunterschiede zwischen fokussierten und nichtfokussierten Bereichen zu gering sind, als daß sie für die Erkennung genutzt werden könnten. Es lassen sich einige Gründe denken, warum die Untersuchungen in der Literatur in dieser Hinsicht erfolgreicher waren:

In den anderen Untersuchungen wurden allgemein akzentuierte und nichtakzentuierte Silben miteinander verglichen. In dieser Arbeit ging es um eine Untermenge davon, es werden also fokussierte Silben mit Silben verglichen, die nichtfokussiert sind, aber dennoch leicht akzentuiert sein können. Deswegen sind die Durchschnittswerte für nichtfokussierte Silben auch relativ hoch. Zur Feststellung des Fokus scheint allgemein die Grundfrequenz ein besseres Mittel zu sein, vor allem durch das starke Absinken von F_0 nach dem Fokus.

Nicht zu vernachlässigen sind auch die Unterschiede zwischen Labor- und Spontansprache.

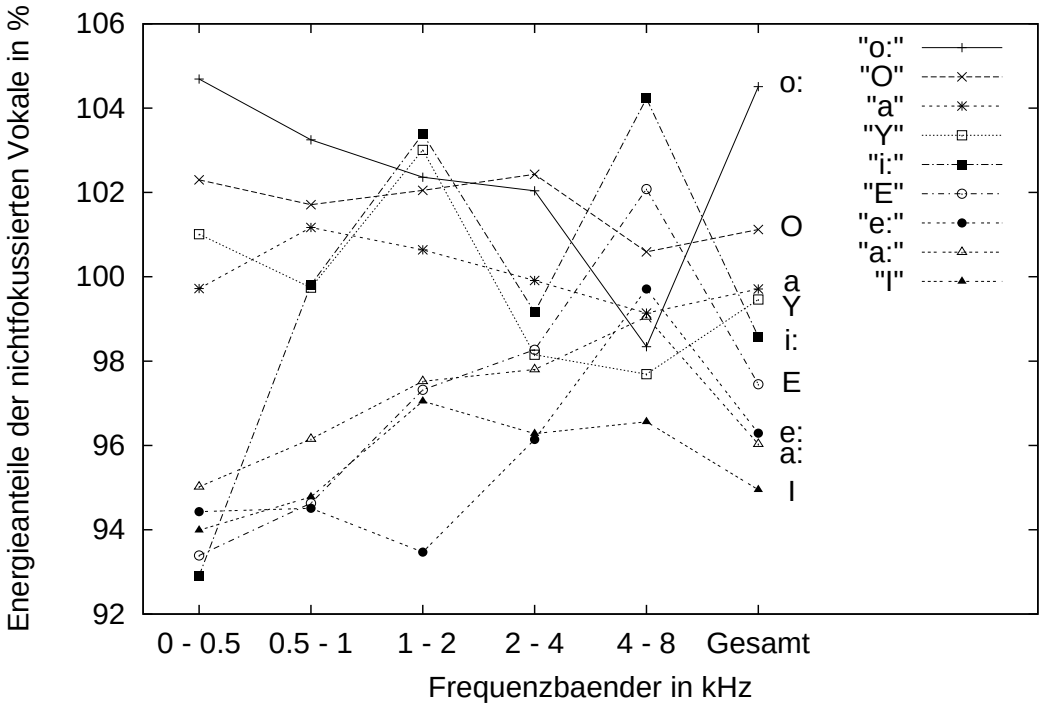

Abbildung 7.1: *Energieanteile der nichtfokussierten Vokale für Sprecher nps1*

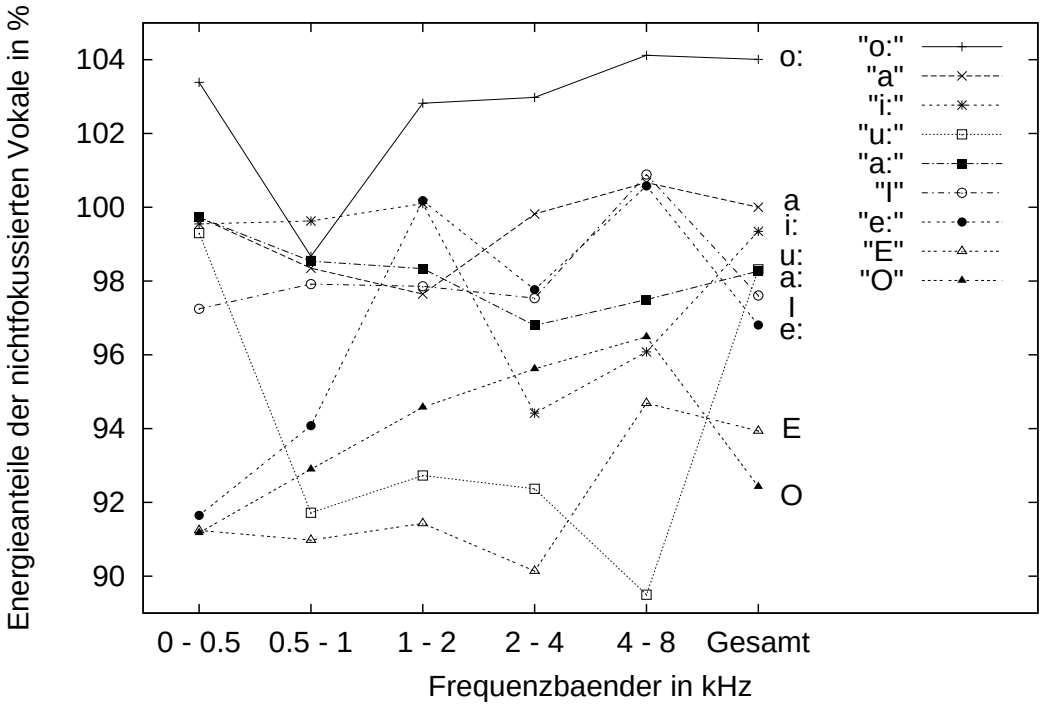

Abbildung 7.2: *Energieanteile der nichtfokussierten Vokale für Sprecherin nmw1*

In Spontansprache werden die nichtfokussierten Vokale zum Teil sehr stark reduziert. Deswegen sind entsprechende Vokale in fokussierter Form und nichtfokussierter Form oft in ihrer Qualität gar nicht mehr vergleichbar. Unterschiede zwischen Laborsprache und Spontansprache zeigten sich ebenfalls in Campbell (1995): bei Einbeziehung der gewonnenen Erkenntnisse über die Energie verbesserte sich seine Erkennung nur für die Laborsprache in deutlicher Weise.

Ein weiterer Störfaktor bei solchen Untersuchungen kann auch in der Lautsegmentierung liegen. Deswegen wurden die Segmentierungen für dieses Experiment manuell nachkorrigiert. Auch bei Campbell (1995) zeigte es sich, daß die Energiedifferenzen in den handsegmentierten Daten in seiner Untersuchung deutlich höher waren. Ein großer Nachteil von manueller Segmentierung ist natürlich der Zeitaufwand. Deswegen wurde diese Untersuchung auch nur für 3 Dialoge durchgeführt.

Zusammenfassend läßt sich sagen, daß die Energiedifferenzen sehr stark vokalabhängig sind. Aufgrund der intrinsischen Vokaleigenschaften ist es offensichtlich nicht möglich, einzelne Vokale automatisch als 'fokussiert' zu klassifizieren, wenn der Vokal nicht bekannt ist. Wenn man den Erkennungsaufbau in Intarc voraussetzt, liegt keine Wort- bzw. Segmentinformation vor. Bei anderem Aufbau wie in Gesamtverbmobil könnte die intrinsische Intensität der einzelnen Vokale besser berücksichtigt werden (siehe Abschnitt 4.3.2). Möglicherweise gilt dies aber nur für das Erkennen von akzentuierten Bereichen allgemein; für die spezielle Erkennung von Fokusakzenten sollten eindeutigere Parameter verwendet werden.

7.2 Phrasengrenzen

Als weitere Möglichkeit zur Verbesserung der Fokusakzenterkennung sollte die Integration von Phrasengrenzen getestet werden (siehe auch Elsner, 1997a). Eine Auswertung der Testdaten (Abschnitt 5.5) zeigte, daß etwa 73 % der Fokusakzente direkt mit einer Phrasengrenze korrespondieren, d. h. sie liegen auf dem letzten Wort einer Phrase. Außerdem liegen auch noch sehr viele Fokusakzente auf dem vorletzten Wort einer Phrase. Phrasengrenzen könnten also eine Möglichkeit sein, die Fokusakzenterkennung nur noch für einzelne Phrasen auszuführen und somit die Aufgabe in kleinere Unterprobleme aufzuteilen.

In der Originalversion des Fokuserkennungsverfahrens wurde die Anzahl der Fokusakzente pro Phrase nicht festgelegt, zumal die Phrasengrenzen als direkte Information ohnehin nicht einbezogen wurden. Für jeden fallenden Abschnitt der Referenzgerade wurde ein Fokusakzent bestimmt (Abschnitt 6.5). Die Zeitpunkte, zu denen in der Referenzgerade ein Wechsel von fallend zu steigend stattfindet, entsprechen nicht notwendigerweise einer Phrasengrenze (siehe auch Abschnitt 6.5.2). Deswegen stellen Phrasengrenzen möglicherweise eine zusätzliche nützliche Information für die Erkennung dar.

In einem ersten Experiment wurden die manuell etikettierten Grenzen aus Braunschweig (Abschnitt 5.4) verwendet. Die Phrasengrenzeninformation wurde folgendermaßen integriert: Für jede Phrase wurde innerhalb der Referenzgerade der Zeitpunkt des 'steilsten

Abfalls' berechnet, der dann auf ein Fokusmaximum abgebildet wurde. Es wurde damit festgelegt, nur einen Fokusakzent pro Phrase zu bestimmen. Enthielt die Referenzgerade im Phrasenbereich nur steigende Abschnitte, wurde kein Fokusakzent gefunden.

Die Ergebnisse sind in Tabelle 7.2 zu sehen, einen Vergleich mit der Originalversion liefert die letzte Zeile 'Gesamt orig'. Die Erkennungsrate stieg in erster Linie für die Nichtfokusbereiche, denn durch die Beschränkung auf einen Fokusakzent pro Phrase wurden insgesamt weniger Fokusakzente detektiert; dies reduzierte deutlich die Fehlschläge. Dies bewirkt insgesamt auch eine deutliche Verbesserung der Akkuratheit, die hier einen positiven Wert annimmt. Für manche Dialoge sinkt allerdings die Erkennungsrate für Fokusbereiche sehr stark ab, da offensichtlich mehr als ein Fokusakzent pro Phrase vorhanden ist.

Für Dialoge wie n008ka oder n001k ist es daher nicht angebracht, die Anzahl der Fokusakzente pro Phrase zu beschränken; die Erkennungsraten sinken hier sehr stark ab. Die Verbesserung der Erkennungsrate durch Phrasengrenzen hängt sehr davon ab, wieviele etikettierte Grenzen es im Vergleich zu den Fokusakzenten gibt (siehe letzte Spalte in Tabelle 7.2). Wo das Verhältnis ausgewogen ist (wie bei n002kb, n003k, n009k), steigt die Erkennungsrate; bei großen Differenzen (n001k, n008ka, n019k) sinkt die Erkennungsrate stark ab. Dabei ist besonders die Erkennungsrate für Fokusbereiche $\mathcal{ER}_F$ betroffen.

Im folgenden Experiment wurden 'detektierte' Phrasengrenzen aus dem Prosodiedetektor des Intarc-Systems verwendet (Strom, 1995). Die Erkennungsrate für die Phrasengrenzen beträgt 81 %; von daher war eine sehr viel geringere Verbesserung der Fokuserkennungsrate zu erwarten, wenn überhaupt. Die Ergebnisse finden sich in Tabelle 7.3. Im Vergleich mit der Originalversion gibt es leichte Verbesserungen für einige Dialoge, insgesamt schneidet dieses Experiment aber etwas schlechter ab.

In einem Dialog (n003k) ist die Erkennungsrate allerdings sogar höher als für die manuell etikettierten Daten. Dies könnte bedeuten, daß in manchen Fällen die prosodischen Grenzen, die automatisch auf akustischer Grundlage erkannt wurden, zuverlässiger sein können als manuell erstellte, die auch durch syntaktisches und linguistisches Wissen beeinflußt sein können. Andererseits ist die Erkennungsrate für manche Dialoge wie n002kc und n008kb sehr viel niedriger, die detektierten Phrasengrenzen sind hier offensichtlich nicht so hilfreich für die Fokuserkennung. Wie beim ersten Experiment entscheidet aber auch ein ausgewogenes Verhältnis von Fokusakzenten und Phrasengrenzen über den verbesserten Erfolg in der Erkennung.

Bei der Integration von Phrasengrenzeninformation muß offensichtlich das Vorkommen von *Doppelfokus* noch stärker berücksichtigt werden. Manche Sprecher haben einen sehr lebhaften Sprechstil, in den Verhandlungsdialogen sagen sie sehr betont, was sie wünschen, z. B. bei einer Aufzählung von möglichen Daten für ein Treffen. In den Verbmobildaten befindet sich beispielsweise folgender Satz:

"In der zweiten **Oktober-***Woche <PG> kann ich nur am* **Montag** *und* **Dienstag***".*

In dieser Äußerung sind die Wörter "Montag" und "Dienstag" von gleicher semantischer Wichtigkeit, und auch auf der akustischen Ebene ist es schwierig zu entscheiden, welches der Wörter einen stärkeren Akzent trägt. In diesem Fall ist also die Annahme eines Doppelfokus angebracht. Wie auch schon die statistische Auswertung der Testdaten ergab (siehe Tabelle 5.3), ist in 11.7 % der Phrasen mehr als ein Fokusakzent etikettiert.

Dialognr.	Fokusanteil	$\mathcal{ER}_{Gew}$	$\mathcal{ER}_{M}$	$\mathcal{ER}_{F}$	$\mathcal{ER}_{NF}$	$\mathcal{A}_{F}$	$\mathcal{ER}_{Ges}$	FA:B3
n001k	20.67	77.07	59.07	29.76	88.38	-29.57	74.24	56:62
n002ka	23.64	76.14	62.52	34.77	90.27	-12.23	73.68	55:51
n002kb	20.00	89.71	83.17	71.00	95.33	35.50	88.50	19:19
n002kc	23.47	83.26	71.53	50.00	93.07	17.07	81.60	37:36
n003k	21.88	79.32	66.50	42.19	90.81	0.69	76.75	65:63
n008ka	19.03	77.34	62.60	36.30	88.90	-28.03	75.03	63:73
n008kb	22.60	82.46	70.30	48.30	92.30	15.10	80.40	25:23
n009k	18.46	83.27	70.96	48.50	93.42	8.33	81.75	51:50
n011k	21.20	80.63	71.83	54.87	88.80	7.53	78.47	45:53
n017k	19.13	82.79	66.67	40.40	92.93	4.40	81.20	40:34
n019k	22.90	77.83	59.95	29.24	90.67	-17.52	75.57	46:40
Gesamt	21.18	80.89	67.74	44.12	91.35	0.12	78.84	502:504
Gesamt orig	21.18	80.39	68.61	47.00	90.21	-5.42	78.13	502:504

Tabelle 7.2: *Experimente mit Phrasengrenzen: Fokusanteile und Erkennungsraten in Prozenten, Verhältnis Anzahl Fokusakzente zu etikettierten B3-Grenzen (FA:B3).*

Dialognr.	Fokusanteil	$\mathcal{ER}_{Gew}$	$\mathcal{ER}_{M}$	$\mathcal{ER}_{F}$	$\mathcal{ER}_{NF}$	$\mathcal{A}_{F}$	$\mathcal{ER}_{Ges}$
n001k	20.67	76.54	55.76	21.14	90.38	-31.71	74.52
n002ka	23.64	76.39	62.93	36.18	89.68	-14.55	73.95
n002kb	20.00	86.00	78.00	62.17	93.83	16.67	84.50
n002kc	23.47	79.31	67.23	43.47	91.00	0.60	77.07
n003k	21.88	79.73	66.94	42.56	91.31	2.38	77.38
n008ka	19.03	78.39	63.40	36.30	90.50	-18.87	76.40
n008kb	22.60	77.50	61.35	30.80	91.90	-4.90	75.10
n009k	18.46	81.97	70.12	48.50	91.75	0.71	80.12
n011k	21.20	80.65	72.50	56.53	88.47	7.80	78.33
n017k	19.13	80.62	63.57	35.40	91.73	-8.93	78.67
n019k	22.90	79.44	59.98	26.86	93.10	-10.38	77.52
Gesamt	21.18	79.69	65.62	39.99	91.24	-5.56	77.60
Gesamt orig	21.18	80.39	68.61	47.00	90.21	-5.42	78.13

Tabelle 7.3: *Experimente mit Phrasengrenzen aus dem automatischen Prosodiedetektor, Fokusanteile und Erkennungsraten in Prozenten*

Für die Spontansprache in den Verbmobildaten scheint es also wünschenswert, den Doppelfokus in der Erkennung zu berücksichtigen. Untersuchungen zum Doppelfokus von Batliner et al. (1991) haben allerdings gezeigt, daß Doppelfokus nur gelegentlich intonatorisch markiert ist. Dies ist offensichtlich sprecher- und situationsabhängig. Im hier vorgestellten Erkennungsverfahren wird das Problem nur durch den Verlauf der Referenzgerade gelöst, die bei starkem Doppelfokus auch zweimal absinken kann. Wenn die Phrasengrenzeninformation optimal genutzt werden soll, müßte die Art und Weise der Integration noch in einigen Punkten verbessert werden.

7.3 Emphase und Kontrast

In diesem Abschnitt soll untersucht werden, inwieweit akustisch unterschiedlich auffällige Fokusakzente automatisch klassifiziert werden könnten (Elsner, 1996). Zum Auflösen von semantischen Ambiguitäten scheint es sinnvoll zu sein, Kontrastakzente von 'normalen' Akzenten zu unterscheiden (siehe auch Abschnitt 8.4). In dieser Untersuchung wurde die bereits vorgestellte Klassifizierung der Fokusakzente verwendet (siehe Abschnitt 6.8). Drei verschiedene Fokuskategorien (Ff, Fd, Fk) wurden in bezug auf ihren Abstand zur nächsten Phrasengrenze und auf ihre akustische Form hin untersucht.

Auch in der Untersuchung von Krahmer und Swerts (1998) spielte die Frage eine wesentliche Rolle, ob und wie sich kontrastive Akzente akustisch (in perzeptiver Auffälligkeit und Intonationskontur) von Akzenten für 'neue Information' unterscheiden. Ihre Ergebnisse zeigten, daß Kontrastakzente tatsächlich perzeptiv gegenüber den anderen Akzenten herausragten (allerdings nicht bei isolierter Darbietung). Eine veränderte Form in der Intonationskontur ließ sich nur erkennen, wenn der Fokusakzent nicht in einer syntaktischen *Default*-Position war (weitere Details in Abschnitt 3.4.2).

7.3.1 Verschiedene Messungen

Tabelle 7.4 zeigt die Verteilung für die 3 untersuchten Fokuskategorien Ff (Fokusakzent), Fd (Standardfokus/Default) und Fk (Kontrast-/Emphaseakzent). Die Kategorien Fm und Fq wurden nicht in die Untersuchung mit aufgenommen, da sie definitionsgemäß nur am Anfang oder Ende einer Äußerung auftreten. Für jeden etikettierten Fokusakzentbereich wurde der Zeitpunkt des Fokusmaximums ermittelt (aus der Menge der F_0-Maxima zur Berechnung der Referenzgerade) und der Abstand zur nächsten Phrasengrenze gemessen. Die Kategorie Fd tritt nur sehr nahe an einer Phrasengrenze auf, im Bereich bis zu 800 ms davor. Die Kategorien Ff und Fk treten bevorzugt im Bereich von 400 - 800 ms vor einer Phrasengrenze auf; die Kategorie Fk befindet sich dabei recht selten direkt vor einer Phrasengrenze (0 - 400 ms).

Weiterhin wurde die akustische Form für jede Fokusakzentkategorie ausgewertet. Die Kategorien wurden wieder nach dem Abstand der Fokusmaxima zur nächsten Phrasengrenze gruppiert. Für jedes Fokusmaximum wurde außerdem der Zeitabstand zum jeweils linken und rechten F_0-Minimum (ebenfalls aus der Menge zur Berechnung der Referenzgerade)

Fokus-Maximum - Phrasengrenze	Anteil Ff Fokus	Anteil Fk Kontrast	Anteil Fd Default
0.0 - 0.4	27.8	14.3	57.7
0.4 - 0.8	28.7	31.4	42.3
0.8 - 1.2	21.1	22.9	0
1.2 - 1.8	16.1	22.9	0
1.8 - 3.5	6.3	8.5	0
Anteil in der Testmenge	53	9	16

Tabelle 7.4: *Gemessene Zeitabstände (in Sekunden) und Verteilung der verschiedenen Fokusakzente Ff, Fk, Fd in Prozent*

gemessen. Danach wurde die relative Höhe des Fokusmaximums bestimmt, indem jeweils der Prozentanteil der F_0-Minima zum Fokusmaximum berechnet wurde (Fokusmaximum = 100 %). Die verschiedenen Kategorien finden sich in den Tabellen 7.5, 7.6 und 7.7.

Die Höhe der F_0-Minima ist für alle Kategorien recht symmetrisch, linkes und rechtes Minimum unterscheiden sich kaum. Es ist aber deutlich zu erkennen, daß die relative Höhe für ein Fokusmaximum für die Kategorie Fd deutlich niedriger ist, während sie für die Kategorie Fk am höchsten ist. Die Gesamtdauer ist für Fk-Akzente am längsten, Ff-Akzente sind geringfügig kürzer, und Fd-Akzente haben eine besondere Verkürzung im linken Bereich.

In bezug auf die Abstände zu einer Phrasengrenze gibt es bei der Kategorie Fk bedeutende Schwankungen. Direkt an einer Phrasengrenze (0 - 400 ms) ist der Akzent besonders lang im linken Zeitbereich, der Höhenabstand vom linken Minimum zum Fokusmaximum ist relativ gering. Mit zunehmender Entfernung von der Phrasengrenze wird die zeitliche Form etwas kürzer, bis auf einen Ausreißer im Bereich 800 - 1200 ms. Das linke Minimum ist im allgemeinen etwas niedriger als das rechte.

Die Abbildungen 7.3, 7.4 und 7.5 geben einen grafischen Überblick über die verschiedenen Akzentformen. Das Fokusmaximum ist jeweils im Nullpunkt und bei 100 %, die Zeitpunkte der linken F_0-Minima sind negativ, die der rechten F_0-Minima positiv aufgetragen. Die Fokusakzente aus der Kategorie Ff geben ein sehr einheitliches Bild ab, im Hinblick auf ihren Abstand zu einer Phrasengrenze verändert sich die Form nicht wesentlich. Bei der Kategorie Fk sind dagegen deutliche Unterschiede zu erkennen: Mit zunehmendem Abstand von einer Phrasengrenze erfolgt eine Rechtsdrehung des 'Akzentdreiecks', mit dem Fokusmaximum als Drehpunkt. Die Eigenschaften der Kategorie Fd sind in Abbildung 7.5 recht deutlich zu erkennen: diese Akzente sind von relativ kurzer Dauer und deutlich niedriger im Vergleich zu den anderen.

Ff-Akzent	Zeitabstand in sec		Anteil vom F_0-Gipfel in %	
FokMaximum-Phrasengrenze	Fokmax-linkes Min	Fokmax-rechtes Min	linkes Min/ Fokmax	rechtes Min/ Fokmax
0.0 - 0.4	0.318	0.252	73.48	72.83
0.4 - 0.8	0.293	0.304	74.08	70.27
0.8 - 1.2	0.248	0.276	76.16	74.57
1.2 - 1.8	0.361	0.295	72.06	75.89
1.8 - 3.5	0.343	0.262	73.66	76.06
Gesamt	0.304	0.280	74.05	73.24

Tabelle 7.5: *Gemessene Zeitabstände (in Sekunden) und Höhenunterschiede für Fokusakzente (Ff)*

Fk-Akzent	Zeitabstand in sec		Anteil vom F_0-Gipfel in %	
FokMaximum-Phrasengrenze	Fokmax-linkes Min	Fokmax-rechtes Min	linkes Min/ Fokmax	rechtes Min/ Fokmax
0.0 - 0.4	0.508	0.161	71.04	62.80
0.4 - 0.8	0.271	0.295	65.83	62.30
0.8 - 1.2	0.449	0.335	68.92	70.37
1.2 - 1.8	0.260	0.391	68.87	74.56
1.8 - 3.5	0.463	0.323	68.21	77.08
Gesamt	0.359	0.310	68.16	68.45

Tabelle 7.6: *Gemessene Zeitabstände (in Sekunden) und Höhenunterschiede für Fokuskontrastakzente (Fk)*

Fd-Akzent	Zeitabstand in sec		Anteil vom F_0-Gipfel in %	
FokMaximum-Phrasengrenze	Fokmax-linkes Min	Fokmax-rechtes Min	linkes Min/ Fokmax	rechtes Min/ Fokmax
0.0 - 0.4	0.185	0.302	79.98	80.62
0.4 - 0.8	0.181	0.217	82.32	76.50
Gesamt	0.183	0.266	80.97	78.70

Tabelle 7.7: *Gemessene Zeitabstände (in Sekunden) und Höhenunterschiede für Fokusakzente (Fd)*

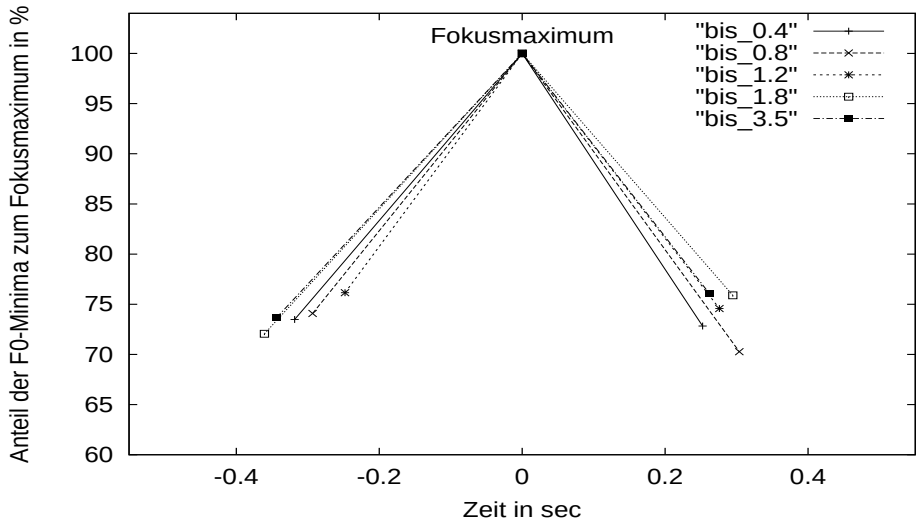

Abbildung 7.3: *Stilisierte Akzentform im F_0-Verlauf für Ff-Akzente*

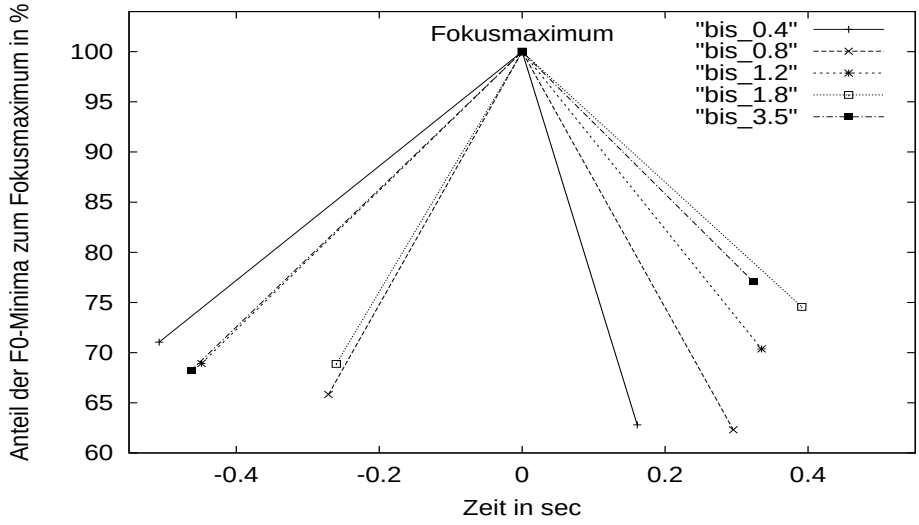

Abbildung 7.4: *Stilisierte Akzentform im F_0-Verlauf für Fk-Akzente*

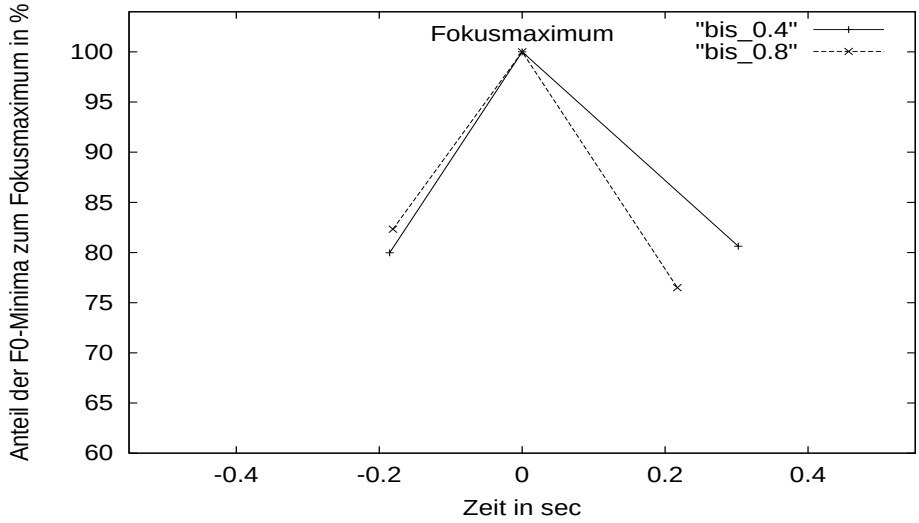

Abbildung 7.5: *Stilisierte Akzentform im F_0-Verlauf für Fd-Akzente*

7.3.2 Diskussion

Reichen die gemessenen Unterschiede aus, um die verschiedenen Akzentkategorien zu unterscheiden? Es ist relativ klar, daß im Bereich einer Phrasengrenze eher weniger mit einem Kontrastakzent zu rechnen ist. Kontrastakzente zeichnen sich außerdem durch eine größere Höhe des Fokusmaximums und durch eine längere Dauer aus. Die Festlegung der Schwellwerte erscheint aber schwierig; diese wird sicherlich auch stark sprecherabhängig sein. Mit Hilfe von Phrasengrenzen kann also nur eine grobe und teilweise Klassifizierung zwischen emphatischem/kontrastivem Fokus und 'normalem Fokus' vorgenommen werden.

Im Erkennungsverfahren innerhalb von Intarc könnten wieder die Phrasengrenzen des Prosodie-Erkennungsmoduls genutzt werden. (Strom, 1995). Zu bedenken ist aber, daß in Verbmobil emphatisch/kontrastive Akzente recht selten sind: Dies betrifft nur 1,2 % aller Wörter bzw. 3 % aller akzentuierten Wörter (Niemann et al., 1998). Es erscheint also fraglich, ob sich in diesem Szenario ein großer Erkennungsaufwand lohnt. Es wurde daher zunächst nicht weiterverfolgt, eine automatische Unterscheidung von 'normalen' und kontrastiven Akzenten zu implementieren.

Sprecher/in	$\mathcal{ER}_{Gew}$	$\mathcal{ER}_{M}$	$\mathcal{ER}_{F}$	$\mathcal{ER}_{NF}$	$\mathcal{A}_{F}$	$\mathcal{ER}_{Ges}$
nmw1	79.55	68.73	45.21	92.25	8.04	77.42
nsp2	78.66	70.12	54.62	85.62	-4.12	75.12
nbs1	76.56	55.77	23.45	88.09	-51.45	74.27
Frauen gesamt	78.26	64.87	41.09	88.65	-15.84	75.60
nhm1	83.65	74.00	55.50	92.50	2.83	82.17
nps1	82.22	71.52	52.00	91.02	2.00	80.15
njk2	81.68	68.79	46.47	91.11	-6.95	80.11
noh2	81.59	63.86	37.73	90.00	-25.82	79.64
nms5	77.36	63.19	39.25	87.12	-27.12	74.75
nhk1	74.05	60.80	32.80	88.78	-22.99	71.21
nhw3	73.46	64.00	45.57	82.43	-23.00	70.29
Männer gesamt	79.14	66.59	44.19	88.99	-14.44	76.90

Tabelle 7.8: *Erkennungsraten für die verschiedenen Sprecher und Sprecherinnen in Prozent*

7.4 Sprecherabhängigkeit

Es ist bekannt, daß die Realisierung des Fokusakzents unterschiedlich stark ausfällt, abhängig z. B. vom Satzmodus und der Position im Satz (Eady und Cooper, 1986). Ein weiterer wichtiger Faktor scheint die Dialoggestaltung der einzelnen Sprecher zu sein. Es fällt nämlich auf, daß die Fokuserkennungsraten für verschiedene Dialoge sehr stark differieren, sie bewegen sich zwischen 73 % und 83 %. Daraufhin wurden die Erkennungsdaten der Verbmobil-Testmenge (siehe Abschnitt 5.5) in bezug auf die Sprecherabhängigkeit untersucht (Elsner, 1997b).

Sprecher/in	Grundfrequenz			Maximum			F_0-Max./ Fokmax	$\mathcal{ER}_{Ges}$
	mittlere	Max.	Min.	davor	Fokus	danach		
nmw1	224.34	255.66	198.63	104.42	268.26	88.66	95.30	77.42
nsp2	256.32	287.32	223.07	101.75	308.44	95.68	93.15	75.12
nbs1	214.64	254.06	187.82	101.55	269.71	95.47	94.20	74.27
nhm1	101.28	105.73	85.30	94.29	121.68	90.81	86.90	82.17
nps1	123.03	137.83	106.20	96.57	149.79	92.39	92.06	80.15
njk2	136.83	154.74	123.11	95.47	166.92	86.15	92.70	80.11
noh2	130.66	142.81	118.88	111.86	152.26	95.36	93.80	79.64
nms5	142.43	157.52	121.32	96.66	159.50	96.73	98.76	74.75
nhk1	138.86	159.27	121.84	100.56	169.19	90.75	94.14	71.21
nhw3	131.79	149.04	116.61	97.92	156.35	91.09	95.30	70.29

Tabelle 7.9: *Weitere Daten für die Sprecher/innen aus Tabelle 7.8: mittlere F_0, F_0-Maxima und F_0-Minima in Hz, mittlere Höhe der Fokusmaxima in Hz mit vorhergehendem und nachfolgendem Maximum in Prozent, Verhältnis der F_0-Maxima zu den Fokusmaxima in Prozent, Gesamt-Erkennungsrate in Prozent.*

7.4.1 Untersuchungen für alle Sprecher

Zur Untersuchung der Sprecherabhängigkeit wurden für alle Sprecher und Sprecherinnen der Verbmobil-Testmenge (11 Dialoge, 3 Sprecherinnen und 7 Sprecher) die mittleren F_0-Werte berechnet, außerdem noch Mittelwerte für die F_0-Maxima und F_0-Minima (siehe Tabelle 7.9). Setzt man diese Daten mit den Erkennungsraten der Sprecher in Beziehung (Tabelle 7.8), lassen sich allein anhand der F_0-Werte kaum Korrelationen erkennen. Es gibt beispielsweise zwei Sprecher (nhk1 und njk2), deren F_0-Mittelwerte sehr ähnlich sind. Ihre Erkennungsraten unterscheiden sich jedoch stark (74.05 % vs. 81.68 %). Die F_0-Mittelwerte allein haben also kaum Aussagekraft bezüglich der zu erwartenden Erkennungsrate.

So wurde ein anderes Maß verwendet, das die Fokusmaxima der Sprecher mit den umgebenden Maxima in Beziehung setzt. Bei einem Fokusmaximum sollte zumindest das nachfolgende F0-Maximum deutlich niedriger sein; idealerweise ist auch das vorhergehende F0-Maximum etwas niedriger, damit eine optimale Fokusakzentuierung erreicht wird. Tabelle 7.9 zeigt die entsprechenden Werte: 'Fokmax' gibt den Mittelwert für Fokusmaxima an, der zu 100 % angenommen wird. 'vorMax' und 'nachMax' geben den prozentualen Anteil der umgebenden Maximumwerte an. Verglichen mit den Erkennungsraten lassen sich nun deutlichere Korrelate erkennen.

Die Sprecherin mit der besten Erkennungsrate (nmw1) hat z. B. einen sehr deutlichen Abfall nach den Fokusmaxima; die Maxima danach haben nur noch einen Prozentanteil von 88.66 %. Die unterschiedlichen Erkennungsraten der Sprecher nhk1 und njk2 werden jetzt ebenfalls deutlicher erkennbar. Sprecher nhm1 mit der besten Erkennungsrate hat auch sehr gute Werte für die Umgebungsmaxima. Nicht erklären läßt sich aber das schlechte Abschneiden des Sprechers nhw3, der trotz besserer Umgebungsmaxima-Verhältnisse

schlechter erkannt wird als Sprecher nhk1.

Eine weitere Korrelation ergibt sich zwischen den F_0-Maxima (F_0-Max) und den Fokusmaxima (Fokmax) in Tabelle 7.9. Wenn die Fokusmaxima signifikant höher sind als die F_0-Maxima (wie bei Sprecher nhm1), ergibt sich auch eine höhere Erkennungsrate. Dies gilt für alle Sprecher bis auf die beiden letztplazierten. Bei den Sprecherinnen gibt es dagegen keine Korrelation mit diesem Parameter.

Ein Vergleich mit einer anderen Untersuchung bezüglich der Verständlichkeit von Sprechern (allerdings ging es dort in erster Linie um segmentale Verständlichkeit) zeigt ähnliche Ergebnisse. In Bradlow et al. (1996) wurde festgestellt, daß die mittlere Grundfrequenz und auch die Sprechgeschwindigkeit nicht unbedingt mit der Verständlichkeit korreliert. Dagegen trägt eine höhere F_0-Variation zu einer besseren Verständlichkeit bei. Für Sprecherinnen wurde allgemein eine bessere Verständlichkeit festgestellt; dies lag in erster Linie an einer deutlicheren Aussprache mit weniger Reduktionen.

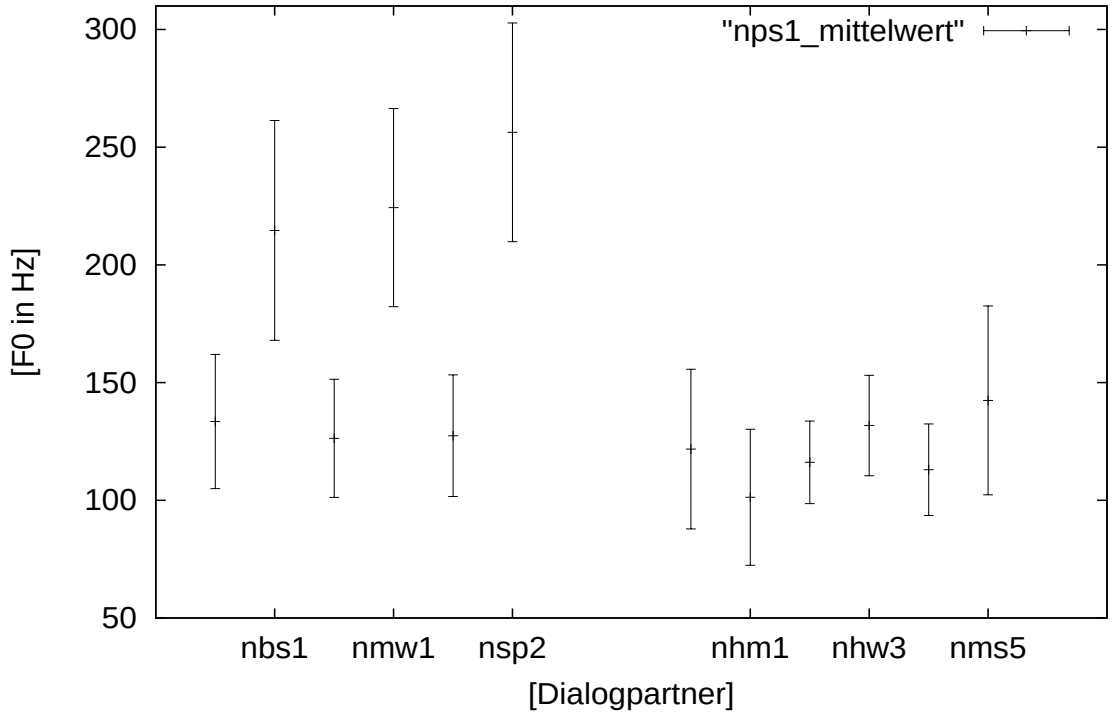

Abbildung 7.6: *Mittlere F_0-Werte (mit Standardabweichung) für Sprecher nps1 mit verschiedenen Dialogpartnern. Bei jedem Wertepaar gehört jeweils der linke Balken zu Sprecher nps1.*

7.4.2 Dialogverhalten eines Sprechers im Vergleich

Im untersuchten Verbmobil-Datenmaterial ließen sich bisher keine grundsätzlichen Unterschiede bezüglich der Fokuserkennungsrate von männlichen und weiblichen Dialogpartnern erkennen. Interessante Beobachtungen ließen sich allerdings mit einem Sprecher (nps1)

Dialogpartner/in	$\mathcal{ER}_{Gew}$	$\mathcal{ER}_{M}$	$\mathcal{ER}_{F}$	$\mathcal{ER}_{NF}$	$\mathcal{A}_{F}$	$\mathcal{ER}_{Ges}$
nmw1 (w)	80.98	69.84	50.63	89.05	-7.63	78.53
nsp2 (w)	79.45	67.12	43.12	91.12	-7.00	77.00
nbs1 (w)	78.94	67.20	44.70	89.70	-3.30	76.70
mit Frauen gesamt	79.79	68.05	46.15	89.96	-5.98	77.41
nms5 (m)	87.63	77.79	63.00	92.57	19.00	86.14
nhw3 (m)	84.75	77.69	63.50	91.88	13.00	82.88
nhm1 (m)	81.57	69.46	47.08	91.83	-2.08	79.67
mit Männern gesamt	84.65	74.98	57.86	92.09	9.97	82.90
Gesamt für Sprecher nps1	82.22	71.52	52.00	91.02	2.00	80.15

Tabelle 7.10: *Erkennungsraten für den Sprecher nps1 im Dialog mit Frauen und Männern, in Prozent*

machen, der an 6 Dialogen beteiligt war. Bei den Dialogen mit den Sprecherinnen hatte er eine deutlich höhere mittlere Grundfrequenz (129 Hz) als bei den tieferfrequenten Sprechern (117 Hz). Offensichtlich passen sich Dialogpartner bezüglich ihrer mittleren Grundfrequenz aneinander an (siehe Abbildung 7.6). Ähnliche Beobachtungen finden sich bei Collins (1998): In den Untersuchungen zeigte es sich, daß während einer längeren Konversation die mittleren F_0-Werte der beiden Dialogpartner an einigen Punkten konvergierten.

	Grundfrequenz			Maximum			F_0-Max./
Dialogpartner/in	mittlere	Max.	Min.	davor	Fokus	danach	Fokmax
nmw1 (w)	126.34	140.88	110.67	97.18	154.74	94.03	91.04
nsp2 (w)	127.45	143.63	110.77	97.41	153.91	94.28	93.32
nbs1 (w)	133.48	150.73	110.61	94.66	163.88	90.37	91.98
mit Frauen gesamt	129.09	145.08	110.68	96.42	157.51	92.89	92.11
nms5 (m)	113.03	127.36	98.30	93.75	136.71	91.81	93.16
nhw3 (m)	116.13	127.49	101.14	96.28	141.50	89.11	90.10
nhm1 (m)	121.73	136.89	105.73	100.12	148.00	94.76	92.49
mit Männern gesamt	116.96	130.58	101.72	96.72	142.07	91.89	91.91
Gesamt für nps1	123.03	137.83	106.20	96.57	149.79	92.39	92.06

Tabelle 7.11: *Daten für verschiedene Dialogpartner des Sprechers nps1: mittlere F_0, F_0-Maxima und F_0-Minima in Hz, mittlere Höhe der Fokusmaxima in Hz mit vorhergehendem und nachfolgendem Maximum in Prozent, Verhältnis der F_0-Maxima zu den Fokusmaxima in Prozent.*

Aufschlußreich sind auch die einzelnen Erkennungsraten für Sprecher nps1, die in Tabelle 7.10 dargestellt sind. Sein unterschiedliches Dialogverhalten im Gespräch mit Frauen und Männern wirkt sich offensichtlich auch stark auf die Erkennungsraten aus; die Erkennung ist für Gespräche mit Männern deutlich höher.

Beim Vergleich der F_0-Daten mit den einzelnen Dialogpartnern (siehe Tabelle 7.11) ragt

nur die mittlere Grundfrequenz als unterscheidendes Merkmal heraus. Die Verhältnisse der Umgebungsmaxima für weibliche und männliche Dialogpartner sind annähernd gleich, die starken Unterschiede in der Erkennung lassen sich hiermit nicht erklären. Für genauere Aussagen sollten aber weitere Untersuchungen mit mehr Sprechern gemacht werden.

7.5 Perzeption

Um die Fokusetikettierungen näher zu untersuchen (Elsner, 1999), wurden für einen Perzeptionstest 80 Dialogäußerungen aus der Verbmobil-Testmenge (Abschnitt 5.5) ausgewählt (ohne Kontext). Die Äußerungen enthalten 1039 Wörter und können aus mehreren Sätzen bestehen. Als Versuchspersonen dienten 5 Mitarbeiter des Instituts (ausgebildete Phonetiker).

In der Untersuchung ging es nicht um die Wahrnehmung von Akzenten an sich, sondern um die Zuverlässigkeit der Etikettierung. Es wurde bewußt miteinbezogen, daß die Hörer auch ihr linguistisches Wissen für die Etikettierung nutzen. Denn es ist das primäre Ziel der hier entwickelten Fokusakzenterkennung, in erster Linie die Fokusakzente zu erkennen, die für die linguistische Weiterverarbeitung verwertbar sind.

7.5.1 Experimente

In einem Vortest wurden den Versuchspersonen die schriftlichen Transkriptionen vorgelegt. Sie sollten alle Wörter unterstreichen, wo sie in freier spontaner Sprache einen Fokusakzent setzen würden. Die Annahme war, daß die Versuchspersonen in erster Linie Inhaltswörter und Wörter mit 'neuer' Information unterstreichen würden.

Die Ergebnisse finden sich in Tabelle 7.12 in der Zeile 'Prädiktion'. Insgesamt wurden 368 Wörter markiert, das entspricht 35.4 %. Die Kategorien entsprechen der Anzahl der Versuchspersonen, die ein Wort markiert hat. Die Kategorien 'III-V' und 'I-II' wurden zusammengefaßt; in der Auswertung zählen sie dann jeweils als 'markiert' bzw. 'unmarkiert'. Der Prozentsatz der 'markiert' Kategorie ist geringer als für die niedrigeren Urteile. Im geschriebenen Text sind offensichtlich mehr Möglichkeiten, um Fokusakzente zu markieren, deswegen ist die Übereinstimmung der Testpersonen noch nicht so hoch wie erwartet.

Im zweiten Experiment sollten dieselben Versuchspersonen sich die spontanen Originaldialoge anhören. Das zweite Experiment fand zeitlich etwa einen Monat später als das erste statt, um zu gewährleisten, daß die Testpersonen nicht durch ihre vorherigen Prädiktionen beeinflußt würden. Sie mußten im schriftlichen Text die Wörter unterstreichen, die sie als fokussiert wahrnahmen. Es wurde keine Begrenzung gegeben, wieviele Fokusakzente sie markieren sollten.

Die Ergebnisse sind wieder in Tabelle 7.12 zu sehen, in der Zeile 'Perzeption'. Diesmal wurden insgesamt etwas weniger Wörter markiert (350), dies entspricht einem Anteil von 33.6 %. Die Anteile in der 'markiert'-Kategorie III-V sind jetzt deutlich höher (19.6 %) als in der Kategorie mit sehr niedriger Übereinstimmung I-II (14.0 %). Die akustischen Daten reduzieren die Möglichkeiten für markierbare Fokusakzente, so daß die Variation zwischen

Kategorien	V	IV	III	II	I	0	III - V	I - II	I - V
Prädiktion	3.8	5.8	7.0	8.0	10.8	64.6	16.6	18.8	35.4
Perzeption	6.1	6.8	7.2	4.4	9.6	66.4	19.6	14.0	33.6

Tabelle 7.12: *Anteile (in Prozent) der Wörter mit entsprechenden Markierungen der Versuchspersonen (bezogen auf die Gesamtzahl)*

den Versuchspersonen abnimmt. Es ist aber bemerkenswert, daß sowohl für die Prädiktion als auch für die Perzeption etwa 2/3 der Wörter für die Testpersonen unmarkiert sind.

7.5.2 Unterschiede zwischen Prädiktion und Perzeption

Wahrgenommene und vorhergesagte Fokusakzente differieren aber in verschiedenen Aspekten. Zum Beispiel sagten einige Versuchspersonen voraus, daß sie Fragewörter wie "wann" und "welcher" fokussieren würden. Dies wurde in den Dialogen aber kein einziges Mal so wahrgenommen. Dies stimmt mit den Aussagen von Ladd (1996) überein: Er stellt fest, daß Fragewörter im Englischen (und in anderen germanischen Sprachen) normalerweise keinen Fokusakzent tragen, auch wenn es rein logisch gesehen sinnvoll wäre. In anderen Sprachen wie Türkisch, Rumänisch oder Ungarisch haben Fragewörter dagegen den prominentesten Akzent im Satz.

Für das Datum läßt sich eine weitere interessante Beobachtung machen. Alle Versuchspersonen sagten Fokusakzente für den Tag eines Monats voraus, in den Dialogen wurde aber meist der Monat als mit höherer Prominenz empfunden. Für die Hörer war es also sinnvoller, den Tag eines Monats stärker hervorzuheben, da sie davon ausgingen, daß der fragliche Monat im Dialog wahrscheinlich bereits bekannt sein dürfte.

Offensichtlich gibt es einen Konflikt mit der *Default-Regel*, also daß Fokusakzente dazu tendieren, möglichst am Ende der Phrase plaziert zu werden. Diese Regel wird nur verletzt, wenn ein besonderer Kontrast oder Emphase beabsichtigt ist. In den spontanen Dialogen wird für die meisten Diskussionen dieser zusätzliche 'vocal effort' vermieden, so daß der semantisch weniger bedeutende 'Monat' den Fokusakzent bekommt.

Ein anderer Grund liegt wohl auch darin, daß den Testpersonen die Dialogsätze ohne Kontext vorgelegt wurden. Normalerweise werden Wörter, die bereits erwähnt wurden ('given'), in den nächsten Sätzen deakzentuiert. Ohne Kontext war für die Testpersonen die meiste Information als 'neu' anzunehmen, so daß sie sehr viel mehr Fokusakzente vorhersagten, als im spontanen Dialog tatsächlich realisiert wurden.

7.5.3 Zuverlässigkeit der Etikettierung

Ein besonders wichtiger Grund, diese Untersuchung durchzuführen, war natürlich der Vergleich zwischen den Etiketten der Autorin und denen der Testpersonen. Es war bis dahin immer etwas kritisch, anzunehmen, daß die Etikettierung einer einzigen Person verläßlich genug ist, um darauf ein Erkennungsverfahren zu basieren. Die Ergebnisse der

Kategorien	V	IV	III	II	I	0	III - V	I - II	I - V
Anz. Wörter (Testhörer)	63	66	75	46	100	689	204	146	350
Anz. Wörter (Autorin)	63	62	64	11	14	825	189	25	214
Prozentanteile	100	93.9	85.3	23.9	14	-	92.7	17.1	61.1

Tabelle 7.13: *Anzahlen der markierten Wörter im Vergleich zwischen Testhörern und Etikettierung der Autorin; Prozentanteile Autorin vs. Testhörer*

Auswertung sind nun in Tabelle 7.13 zu sehen.

Im allgemeinen ist die Variation der Testpersonen untereinander größer: Sie markierten insgesamt 350 Wörter, während von der Autorin nur 214 markiert wurden. Aber die Prozentangaben zeigen, daß es eine sehr hohe Übereinstimmung für die höheren Kategorien gibt, für die 'markiert'- Kategorie 'III-V' sind es beispielsweise 92.7 % Übereinstimmung, für die Kategorie 'V' sind es sogar 100 %. So kann man wahrscheinlich schließen, daß der bei weitem überwiegende Anteil der Autorenetiketten von den phonetischen Experten unterstützt wird.

Als Konsequenz aus dieser Untersuchung wurden alle Etiketten der Autorin mit niedriger Übereinstimmung (I-II) noch einmal sorgfältig überprüft, die meisten wurden daraufhin entfernt. Alle in dieser Arbeit angegebenen Tabellen sind mit den revidierten Fokusetiketten berechnet worden.

7.5.4 Akustische Auffälligkeit und Hörerurteile

Eine weitere wichtige Frage ist, ob die Anzahl der Hörerurteile etwas über die akustische Auffälligkeit eines Wortes aussagt, d.h. ist ein Wort mit hoher Übereinstimmung markiert, ragt es dann auch akustisch aus den anderen heraus? Akustische Auffälligkeit soll hier folgendermaßen festgelegt werden (mindestens eine der Bedingungen sollte zutreffen):

- Die mittlere F_0 eines Wortes *vor* einem markierten Wort sollte niedriger sein als die mittlere F_0 des markierten Wortes.
- Die mittlere F_0 eines Wortes *nach* einem markierten Wort sollte niedriger sein als die mittlere F_0 des markierten Wortes.
- Die mittlere F_0 von 3 Wörtern (Wort davor, markiertes Wort, Wort danach) sollte niedriger sein als die mittlere F_0 des markierten Wortes.

Zur Auswertung wurde daher die mittlere F_0 für jedes Wort berechnet. Um den Unterschieden im Grundfrequenzumfang der einzelnen Sprecher und Sprecherinnen Rechnung zu tragen, wurden Prozentanteile als Vergleichsmaß verwendet. Für jedes Wort wurden daraufhin also die Prozentanteile der Wörter davor und danach in bezug auf die F_0 berechnet. Darüber hinaus wurde eine mittlere F_0 aus der des aktuellen Wortes, des Wortes davor und danach berechnet, anschließend wurde der Prozentanteil zum aktuellen Wort

Urteile	Wort davor	Wort danach	Mittelwert der 3 Wörter
einzelne Urteilsklassen			
V	96.53	94.15	96.89
VI	95.23	94.69	97.17
III	99.41	94.68	98.28
II	99.82	94.90	98.24
I	100.77	99.71	100.09
0	104.65	102.24	102.27
kombinierte Urteilsklassen (2)			
V - IV	95.84	94.42	97.04
IV - III	97.22	94.68	97.71
III - II	99.55	94.76	98.27
II - I	100.50	98.31	99.58
I - 0	104.19	101.95	101.99
kombinierte Urteilsklassen (3)			
V - III	97.00	94.50	97.46
II - 0	103.99	101.62	101.82

Tabelle 7.14: *Hörerurteile und prozentuale Anteile der mittleren F_0 der Umgebungswörter*

ermittelt. Alle Wörter mit der gleichen Anzahl von Hörerurteilen wurden, so wie in den vorhergehenden Abschnitten, in entsprechende Klassen zusammengefaßt.

Die Ergebnisse in Tabelle 7.14 zeigen eine deutliche Tendenz. Je mehr Testpersonen ein Wort markiert hatten, desto niedriger sind die prozentualen Anteile der Wörter davor und danach. Dies gilt besonders für das Wort nach einem markierten Wort. Dies bestätigt wiederum den Ausgangspunkt der in dieser Arbeit vorgestellten Fokuserkennung, die davon ausgeht, daß nach einem Fokusakzent die Grundfrequenz stark absinkt. Die Unterschiede zwischen den Prozentanteilen werden noch größer, wenn zwei oder drei Urteilsklassen zusammengefaßt werden. Die letzten beiden Zeilen in Tabelle 7.14 zeigen besonders deutliche Unterschiede zwischen der 'markiert'-Kategorie (III-V) und der 'unmarkiert'-Kategorie. Dies gilt wiederum besonders für das Wort nach dem markierten Wort.

Ähnliche Ergebnisse finden sich in Streefkerk et al. (1998). In ihrer Untersuchung mußten 10 Hörer 500 gelesene Sätze (Niederländisch) beurteilen. Sie fanden eine lineare Relation zwischen der Anzahl der Hörerurteile (hier mit Prominenz gleichgesetzt) und dem F_0-Umfang einer Silbe bzw. der Lautheit eines Vokals. Die Korrelation mit der Silbendauer war allerdings geringer, vermutlich beeinflußt durch Sprechgeschwindigkeit und finale Längung.

8. Anwendung der Fokusinformation für linguistische Module

In diesem Kapitel soll noch eine ungeklärte Frage aus dem 3. Kapitel (Abschnitt 3.1.2) untersucht werden, nämlich ob aus dem akustisch ermittelten Fokusakzent der semantische Fokus abgeleitet werden kann. Es gibt ja überhaupt unterschiedliche Vorstellungen davon, was ein Fokus ist, daher ist es eine wichtige Frage, inwieweit die akustische Fokusinformation für linguistische Module nützlich ist.

Zunächst werden einige Vorüberlegungen zur Nutzung der Fokusinformation angestellt. Daran schließt sich ein Exkurs zu den in Verbmobil verwendeten Übersetzungsstrategien an. Daraufhin wird die Anwendung in Intarc und Verbmobil beschrieben. Im Teilprojekt Intarc (siehe 4.2) wurden die erkannten Fokusakzente von den Modulen für Semantik und Transfer genutzt, das Transfermodul realisiert darüber hinaus noch eine flache Übersetzung mit Hilfe der Fokusinformation und der besten Wortkette des Erkennermoduls. Die Experimente in diesem Bereich werden ausführlich dargestellt. Abschließend wird noch auf spezielle Aspekte der Verwendung von Fokusinformation in den Semantikmodulen von Verbmobil und Intarc eingegangen.

8.1 Grundsätzliche Überlegungen

Die Prosodie kann als Brücke zwischen den akustischen Modulen und den linguistischen Modulen dienen. Auf beiden Seiten werden jeweils eigene Forderungen und Begrifflichkeiten vertreten. Die Prosodie beschreibt Präferenzen für eine bestimmte Strukturierung oder Hervorhebung (siehe auch Abschnitt 4.3.2). Für linguistische Module ist es schwierig, diese tendenziellen Aussagen als feste Regel in einen Parser oder in eine Grammatik einzubauen, es werden eher kategoriale Beschreibungen bevorzugt, z. B. akzentuiert vs. nicht-akzentuiert.

Die Prosodie kann als direkte Abbildung von Signaleigenschaften auf Konzepte gesehen werden. Sie dient der *Strukturierung von Informationen*: Dies kann zum einen durch eine Gliederung mit Hilfe von Phrasengrenzen geschehen, zum anderen kann durch Akzentuierung relevante von weniger relevanter Information abgegrenzt werden. Dabei ist die Relevanz oft wichtiger als die Aktualität der Information (siehe auch Diskussion in Abschnitt 3.1.4).

Bei der Übersetzung von Spontansprache kann in der Regel nicht auf wohlgeformte Äuße-

rungen zurückgegriffen werden (siehe Abschnitt 5.3). In längeren Diskursen gibt es eine große Variation, die nicht immer regelhaft zu erfassen ist. Der Dialogkontext beeinflußt die Form einer Äußerung besonders stark. Wichtig ist auch der emotionale Zustand eines Sprechers, der entsprechend seiner Einstellung zum Gesagten unterschiedlich starke Akzente setzt.

Vorüberlegungen und Tests mit dem Transfermodul aus Intarc (Elsner und Klein, 1996) ergaben, daß die akustisch bestimmten Fokusakzente in hohem Grad mit einem perzeptiven Fokus übereinstimmen. Dies war eine Grundvoraussetzung für eine Verwendung des prosodischen Fokus im Transfer. Weiterhin war es offensichtlich, daß eine enge Beziehung zwischen dem vom Algorithmus bestimmten und dem pragmatischen Fokus besteht. Im allgemeinen kann man also davon ausgehen, daß die prosodisch markierten Wörter auch die in kommunikativer Hinsicht wichtigsten Wörter in einem Dialogbeitrag sind.

Die prosodischen Fokusinformationen sind besonders auf der pragmatischen Ebene relevant, da sie den zentralen Äußerungsgehalt eines Turns wiedergeben. Sie markieren die gerade im Dialogkontext relevante Information. Fokussierte Äußerungsbereiche können für ein *concept-spotting* eingesetzt werden, das auch bei schlechten Erkennungshypothesen noch richtige Analysen liefern kann. Fokussierte Elemente mit lokalen semantischen Funktionen (wie Anaphernresolution, Kontrast) treten zwar auch auf, aber auch hier ist die Akzentuierung abhängig von der kommunikativen Funktion.

8.2 Transfer in Verbmobil und Intarc

Beim automatischen Übersetzen ist es wichtig, den zentralen Gehalt oder die pragmatisch-kommunikative Funktion einer Äußerung korrekt in die Zielsprache zu übertragen. Daher wurden Strategien entwickelt, diese wesentlichen Informationen auf Dialogakte abzubilden. Dabei war es auch wichtig, die Relevanz von Dialogbeiträgen korrekt zu beurteilen, um eine adäquate Übersetzung zu erstellen. Im folgenden werden die grundsätzlichen Übersetzungsstrategien und jeweils ihre Anwendungen in Intarc und Verbmobil beschrieben.

8.2.1 Übersetzungsstrategien

In Schmitz et al. (1994) wird ein Überblick zur Übersetzung von Dialogen gegeben. Grundsätzlich wird eine Übersetzungsäquivalenz in drei wichtigen Bereichen gefordert: strukturell, inhaltlich und bezüglich der kommunikativen Funktion. Dies entspricht den traditionellen Ebenen Syntax, Semantik und Pragmatik. Eine adäquate Übersetzung von Dialogen muß diese Funktionen von Sprache berücksichtigen.

Probleme der Übersetzungsäquivalenz treten aber beispielsweise bei sterotypen Formeln und Floskeln auf, denn es stellt sich die Frage, inwieweit diese Formeln in die Zielsprache übertragen werden können oder müssen. Stereotype Formeln sind dadurch charakterisiert, daß der propositionale Gehalt bei der Übersetzung keine Rolle spielt. Ausschlaggebender Faktor scheint hier die kommunikative Funktion zu sein.

Eine weitere wichtige Eigenschaft von Sprechakten ist es, daß sie eine wesentliche Information für eine erfolgreiche Desambiguierung darstellen. Für viele Wörter im Lexikon gibt es unterschiedliche Lesarten, die von ihrer Funktion im Kontext abhängen. Zur Auflösung von Mehrdeutigkeiten werden zusätzliche Informationen aus Semantik und Pragmatik benötigt, auch die Prosodie stellt ein wichtiges Hilfsmittel dar. Die Lesarten können oft nicht eindeutig festgelegt werden, es ergeben sich meist gewisse Präferenzen für Lesarten in einem bestimmten Kontext. Diese Präferenzen sind unterschiedlich relevant, sie können als gewichtete *Defaults* modelliert werden (Schmitz et al., 1994).

In der Schriftsprache gilt im allgemeinen ein Satz als Übersetzungseinheit (Schmitz und Quantz, 1996). Die spontanen Äußerungen in Verbmobil liefern dagegen oft nur Satzfragmente, die grammatisch unvollständig sind. Daher wurden in Verbmobil *Sprechakte* (siehe Abschnitt 1.1) als Übersetzungseinheiten gewählt . Um die Sprechakte automatisch identifizieren zu können, wurde versucht, einen systematischen Zusammenhang zwischen einem Sprechakt und seiner aktuellen lexikalischen Realisierung herzustellen, nämlich mit der Aufstellung von Schlüsselwörtern (Mast, 1995). Eine Dialogstruktur kann dann als Sequenz von Sprechakttypen beschrieben werden. Bestimmte Wahrscheinlichkeiten für das Auftreten eines Sprechakts im Dialogverlauf können ebenfalls zur Identifizierung des Sprechakts beitragen.

In Verbmobil wurde ein Inventar von sog. *Dialogakttypen* festgelegt, das für Terminabsprachen charakteristisch ist (Jekat et al., 1995). *Dialogakte* unterscheiden sich von der reinen Sprechhandlung (illokutiver Akt) dadurch, daß sie noch zusätzlich Inhaltsinformation (Proposition) enthalten und damit speziell auf die Domäne Terminvereinbarungen in Verbmobil abgestimmt sind (Schmitz und Quantz, 1996). Vom reinen Sprechakt *Vorschlag* können beispielsweise die Dialogakte *Vorschlag-Datum* und *Vorschlag-Ort* abgeleitet werden.

Die Dialogakte beschreiben also den zentralen Informationsgehalt und die kommunikative Funktion einer Äußerung. Prosodische Informationen können dazu dienen, Dialogakte zu segmentieren und zu identifizieren (Mast et al., 1996). Prosodische Indikatoren erlauben darüber hinaus eine *Gewichtung* von Dialogakten nach ihrer Relevanz im Kontext und damit auch nach ihrer Relevanz für die Übersetzung. Prosodisch fokussierte Wörter markieren dabei die Relevanz von Wörtern innerhalb einer Äußerung. Unter Umständen kann damit eine reduzierte bis vollständige Übersetzung erstellt werden (siehe Abschnitt 8.3.1).

Zur Bewertung der Korrektheit einer Übersetzung im Rahmen des Szenarios wurden folgende Kriterien aufgestellt (Schmitz und Quantz, 1996Bub et al., 1997):

- Alle zeitlichen Ausdrücke (Uhrzeit und Datum) müssen korrekt wiedergegeben werden.
- Der passende Dialogakt muß wiedergegeben sein.
- Eine angemessene Ebene der Höflichkeit muß beibehalten werden.
- Die Übersetzung muß verständlich sein.

Diese Kriterien sind natürlich teilweise recht subjektiv. Das Auswertungsproblem wurde in Verbmobil so gelöst, daß mehrere professionelle Dolmetscher die Angemessenheit der Übersetzungen beurteilen sollten. Eine umfangreiche Evaluation wurde für über 20.000 Verbmobil-Äußerungen via Internet durchgeführt (Bub et al., 1997Hauenschild et al., 1997).

8.2.2 Transfer in Intarc

Der Transfer in Intarc arbeitet zweigeteilt: In einer traditionellen 'tiefen Analyse' bekommt das Modul Information aus der Semantik-Auswertung (Dialogakt und Feature-Struktur). In einem zusätzlichen Pfad wird mit der besten Wortkette (aus der Worterkennung) und der Fokusinformation ein 'flacher Transfer' durchgeführt (siehe auch Abbildung 4.1).

Die Übersetzungsstrategie im Transfer ist dialogaktbasiert (Jekat et al., 1995Schmitz und Quantz, 1996), die benötigten Informationen für die Übersetzung werden aus verschiedenen Beschreibungsebenen und damit von verschiedenen Modulen bezogen (Jekat, 1996). Die größte Rolle spielen hierbei die Module für die Semantikkonstruktion, die semantische Auswertung und die Fokuserkennung. Sie liefern die Informationen, die zur Ermittlung des zentralen Äußerungsgehalts beitragen.

Das Kernstück der dialogaktbasierten Transferkomponente ist der Aufbau einer Repräsentation von Äußerungen oder Äußerungsteilen als Zusammenhang von illokutionärer Funktion (kommunikativ) und propositionalem Gehalt (Aussageinhalt), bis aus der Dialogaktstruktur eine Übersetzung generiert werden kann. Zu diesem Zweck wurde ein Repräsentationsformat gewählt, das neutral genug ist, um aus verschiedenen Informationsquellen Wissen für unterschiedliche Sprachen repräsentieren zu können.

Die Repräsentationsstruktur besteht aus einer Ebene für die illokutionäre Funktion und einer zweiten Ebene für den propositionalen Gehalt. Dialogakte (als Abbildung der illokutionären Funktion) und Konzepte (zur Darstellung des propositionalen Gehalts) werden in Merkmalsstrukturen repräsentiert, wobei die Konzepte in die Dialogaktstrukturen eingebettet werden. Die resultierende Struktur gibt den zentralen Gehalt einer Äußerung oder eines Segments wieder.

Beim flachen Transfer spielen die erkannten Fokusakzente eine steuernde Rolle. Da die Fokuserkennung Zeitpunkte und Konfidenzwerte liefert, die sich vom F_0-Verlauf ableiten, müssen noch Informationen der Worterkennung hinzugezogen werden, damit die Zeitpunkte auf eventuell fokussierte Wörter abgebildet werden können. Insgesamt hat sich die beste Wortkette als ausreichend aussagekräftig erwiesen, doch stützt sich die Bestimmung der fokussierten Wörter auf eine beste Kette, in der durch Pausen, Störungen oder Erkennungsfehler zertrennte Komposita in der Morphologiekomponente rekonstruiert wurden.

Die beste Wortkette mit den markierten Fokusakzenten wird dann folgendermaßen analysiert: Wenn der Fokus auf einem Inhaltswort liegt, wird ein probabilistisch bestimmter Dialogakt ausgewählt. Dieser wird dann mit der Information aus der Wortkette erweitert. Der flache Transfer wird nur verwendet, wenn die tiefe Analyse fehlschlägt (Strom et al., 1997).

8.2.3 Flache Übersetzung in Verbmobil

Wenn die tiefe Analyse scheitert, wird auch in Verbmobil eine flache Übersetzung angefertigt. Eine Analyse kann scheitern, wenn der Sprecher etwas äußert, wofür das System nicht angelegt ist (unbekannte Wörter, falsche Grammatik) oder wenn die akustische Qualität zu schlecht ist. Die flache Übersetzung wird folgendermaßen durchgeführt (Niemann et al., 1998Block, 1997):

1. Die Äußerung wird in semantisch-pragmatische Einheiten aufgeteilt. Dies geschieht mit Hilfe von prosodischer Information (in erster Linie Phrasengrenzen).
2. Für diese Einheiten wird der Dialogakt bestimmt (Klassifikator für 18 Klassen). 45 % der Dialogakte werden damit korrekt klassifiziert. Außerdem wird der propositionale Gehalt der Äußerung extrahiert (z. B. ein Datum).
3. Für jeden Dialogakt gibt es vorgefertigte Schablonen. Die entsprechenden Lücken werden dann mit dem propositionalen Gehalt gefüllt, und eine Übersetzung wird erstellt.

Die schablonenhafte Übersetzung entspricht natürlich nicht dem originalen Wortlaut, sie ist nur vom groben Inhalt her korrekt und gibt nicht jede Einzelheit wieder. Mit Hilfe dieser flachen Übersetzung können etwa 47 % der Verbmobil-Äußerungen approximativ korrekt ausgeführt werden. Die tiefe Analyse allein liefert zu 52 % approximativ korrekte Übersetzungen. Bei einer Kombination der beiden ergeben sich 74 % approximativ korrekte Übersetzungen (Bub et al., 1997).

8.3 Fokus und Dialogaktbasierter Transfer in Intarc

In den Experimenten mit Fokus und Transfer war es nicht das Ziel, jedes einzelne Wort korrekt zu übertragen, sondern anhand der fokussierten Wörter das Wesentliche einer Aussage zu erfassen. Es sollte die kommunikative Funktion der Äußerung und damit der Dialogakt bestimmt werden, außerdem sollte der propositionale Gehalt erfaßt werden. Für eine Übersetzung in diesem Zusammenhang reicht eine sinngemäße aus, eine wörtliche ist nicht unbedingt nötig.

8.3.1 Experimente mit Transfer

Zur Untersuchung der Verwertbarkeit des prosodischen Fokus wurde eine Übersetzung bzw. eine Dialogaktbestimmung *nur* mit fokussierten Wörtern vom Transfermodul in Intarc durchgeführt (Elsner und Klein, 1996). Zur Auswertung wurde eine Teilmenge der in Intarc abgesprochenen Testdaten verwendet (siehe Abschnitt 5.5). Die Ergebnisse für 49 Äußerungen aus den Dialogen n011k, n017k und n019k sahen folgendermaßen aus:

- 25 korrekte Übersetzungen (zwar reduziert, aber adäquat)
- 2 völlig falsche Übersetzungen
- für 22 Äußerungen reicht die Fokusinformation zu einer Übersetzung nicht aus
- Dialogaktzuweisungen sind in 27 Fällen möglich
- Aussagen über den propositionalen Gehalt (z. B. *date* wie bei suggest-support-date, accept-date, reject-date etc.) sind zusätzlich in 23 Fällen zutreffend

Manche Dialogakte lassen sich sehr gut bestimmen, z.B. *Vorschlag-Ort* oder *Vorschlag-Datum*, da die entsprechenden Schlüsselwörter (für *Ort* z. B. "Büro", "Mensa", "bei mir" (Mast, 1995)) sehr oft fokussiert sind. Ein Problem bereitet allerdings die *Bewertung* der Aussagen: Zustimmung oder Ablehnung eines Termins ist nur manchmal fokussiert: *ja, leider, schön, prima, ausgeschlossen*. Fokussierte Wörter allein reichen also oft nicht aus. Im Satz

Am **Montag**, *den* **15.** *kann ich nicht.*

ist z. B. nur der Termin fokussiert, aber nicht die Bewertung dieses Termins. Dieses Phänomen finden wir in den Dialogen häufiger: Datum und Ort sind fokussiert, die Bewertungen (*gut/schlecht; da habe ich (keine) Zeit*) sind dagegen eher selten markiert.

Das Problem für die Übersetzung ist also, das fokussierte Wort korrekt zu interpretieren. Handelt es sich dabei um ein Schlüsselwort für einen Dialogakt, kann zumindest die Dialogakterkennung ihren Suchraum deutlich einschränken. Bei manchen Dialogakten (z. B. *Begrüßung*) ist außerdem eine Bewertung nicht relevant, so daß die Übersetzung direkt erfolgen kann. In den meisten Fällen sind aber weitere Informationen, wie z. B. über den Satzmodus nötig.

8.3.2 Anwendungsbeispiele

Die folgenden Beispiele stammen aus einer Auswertung des Transfers zur Nutzung der Fokusinformation (Elsner und Klein, 1996). Es wurden 330 Äußerungen der CD 4 untersucht. In den Beispielen hat sich gezeigt, daß der prosodische Fokus oft eine pragmatische Funktion markiert. Der Fokus kann sich dabei sowohl auf ein linguistisches Konzept als auch auf einen Dialogakt beziehen. Ein wichtiges Konzept ist hier z. B. die Negation.

Es ist möglich, in einer Äußerung ein Konzept zu negieren:

Das ist **kein** guter Vorschlag.

Ebenso finden sich jedoch auch andere Negationen, deren Skopus über den ganzen Dialogakt (z.B. *Ablehnung*) reicht:

Das gefällt mir **nicht**.

Negation

Nach der Durchsicht der Daten scheint es, als sei in spontaner Sprache Negation häufig prosodisch markiert, wenn aus der Perspektive der Dialogakte der propositionale Gehalt relevant ist. Dagegen ist Negation eher nicht markiert, wenn die illokutionäre Funktion im Vordergrund steht. Beide Aspekte beziehen sich auf Negation, die auch syntaktisch realisiert ist. Im folgenden ein Beispiel für nicht markierte Negation (gefundene Fokusakzente des Erkenners sind fettgedruckt):

> g271a015 ja mein **Problem** ist daß ich am Dienstag **vormittag** was habe und am **Mittwoch** nachmittag äh also zwei **zusammenhängende** Tage sind bei **mir** immer nur Donnerstag und **Freitag** äh und dann **eventuell** am **Wochenende** wenn es Ihnen nichts ausmacht der **Montag** ist bei mir **auch** ganz frei

Das *"nichts"* in der Höflichkeitsfloskel *"wenn es Ihnen nichts ausmacht"* ist prosodisch nicht markiert. Aus pragmatischer Sicht ist dies ein typisches Beispiel für eine Negation in einer floskelhaften Verwendung, die für den Dialogverlauf nicht relevant ist. In der folgenden Äußerung ist dagegen *"nicht"* in zwei Fällen fokussiert:

> n107k003 **Montag** der dritte **Juli** da geht es leider bei mir **nicht** da äh habe ich nämlich **Urlaub** äh da bin ich nicht da und du hast den **Alternativvorschlag** äh gemacht am **fünften** äh da sieht es sehr **schlecht** aus da habe ich nämlich **morgens** äh einen Zahnarzttermin und da bin ich auch äh sonst den ganzen **Tag** beschäftigt also sieht es **nicht** gut aus ich würde als **Ersatztermin** würde ich den **zehnten** Juli **vorschlagen** das ist ein **Montag** wie sieht es da bei dir aus

Im ersten Satz *"da geht es leider bei mir nicht"* ist die Negierung eine wichtige Information zum vorher genannten Datum und hat daher eine wesentliche kommunikative Funktion. Die weitere Aussage *"da bin ich nicht da"* beinhaltet nur eine Konsequenz des bereits Gesagten (der Sprecher ist im Urlaub); dieser Satz trägt keine neue Information und ist daher auch nicht prosodisch markiert. Eine weitere Negierung eines Termins drückt sich im Wort *"schlecht"* aus, das dann auch fokussiert ist. Zum Abschluß der Termindiskussion folgt noch eine konklusive Negation: *"also sieht es nicht gut aus"*, in der *"nicht"* wiederum fokussiert ist.

In den Dialogteilen, die Abfolgen von Vorschlägen, Ablehnungen und Annahmen sind, finden sich verschiedene Arten von Negationen. In vielen Fällen ist die Negation dann markiert, wenn sie von konstituierender Bedeutung für den entsprechenden Dialogakt (vor allem bei Ablehnungen) ist. Dies scheint sowohl für Negation mit Skopus über ein Konzept des propositionalen Gehalts (aus semantisch-pragmatischer Perspektive) beziehungsweise über eine Nominalphrase (aus syntaktischer Sicht) wie auch für Negationen zu gelten, deren Skopus einem ganzen Dialogakt entspricht.

Nur in den Fällen, in denen eine Negation explizit zum Dialogverlauf gehört, muß sie auch übersetzt werden. Die Untersuchungen haben gezeigt, daß die Information zum prosodischen Fokus ein wichtiges Kriterium in dieser Unterscheidung ist, da Negation in beiden Fällen syntaktisch manifestiert sein kann. Problematisch wird es allerdings, wenn eine zentrale Negation nicht prosodisch markiert ist, da in diesen Fällen Kontextinformationen hinzugezogen werden müssen.

Routineformeln und Diskurspartikeln

Was beispielhaft für die Negation dargestellt wurde, gilt auch allgemeiner für Routineformeln und Diskurspartikeln. Wie in (Schmitz und Fischer, 1995) beschrieben, sollten für eine kommunikativ adäquate Übertragung in die Zielsprache nur die Routineformeln und Diskurspartikeln wörtlich übersetzt werden, die eine distinktive semantische Funktion haben. Für Formeln und Partikeln, die eher pragmatisch relevant sind und beispielsweise eine Abtönung oder allgemein eine Sprechereinstellung ausdrücken, ist ein geeignetes zielsprachliches Äquivalent auszuwählen, während einige Elemente in der Übersetzung besser gar nicht wiedergegeben werden.

Beispiele für solche Elemente sind *"eventuell"* und *"vielleicht"*, die beide sowohl eine echte Einschränkung als auch Bestandteil einer höflichen Floskel sein können. Auch hier scheint es für die Unterscheidung bedeutsam zu sein, ob die Elemente prosodisch fokussiert sind.

> m041n013 . . . also bis auf den **Mittwoch** äh bis auf den **fünfzehnten** und **sechzehnten** Februar **könnte** ich im Februar wenn Sie **da** vielleicht Zeit hätten

Im letzten Teil der Äußerung ist das *"vielleicht"* lediglich Bestandteil der Routineformel *"wenn Sie da vielleicht Zeit hätten"*. Die Routineformel markiert nur einen höflichen Abschluß eines Terminvorschlags. Sie hat keine eigenständige semantische Funktion und ist dementsprechend prosodisch nicht besonders markiert.

> m154d009 Guten Tag Frau **Luger** äh bei unsrer **Terminabsprache** vorhin da habe ich was übersehen äh wir **können** die **Absprache** so nicht **lassen** äh ich mache Ihnen einen **anderen Vorschlag** äh könnten **Sie eventuell** äh langsam äh **siebzehnter** bis äh **einundzwanzigster** Mai

In diesem Beispiel hat die Sprecherin offensichtlich einen Fehler begangen und bittet um eine Terminkorrektur. Mit der Fokussierung von *"eventuell"* wird noch einmal betont, daß es sich um eine sehr höfliche Bitte handelt, die die Dialogpartnerin nicht unter Druck setzen will. Hier hat die Diskurspartikel also eher eine abtönende Funktion, der Vorschlag wird 'vorsichtig' formuliert, um die Dialogpartnerin nicht zu bedrängen. Im Satz 'g271a015' aus dem vorhergehenden Abschnitt hat die Aussage *"eventuell am Wochenende"* dagegen eine deutlich einschränkende Funktion, da ein Termin an einem Wochentag im allgemeinen bevorzugt wird.

Probleme

Wie im Zusammenhang mit der Negation schon kurz angesprochen wurde, gibt es beim dialogaktbasierten Tranfer mit prosodischen Informationen dann Schwierigkeiten, wenn nicht alle relevanten Äußerungsteile prosodisch markiert sind. Dies ist der Fall, wenn eine Negation ausnahmsweise nicht prosodisch markiert ist. In anderen Situationen sind nur bestimmte Teile der Äußerung eines zusammengesetzten Zeitausdrucks prosodisch markiert, was wiederum zu unvollständigen und auch inkorrekten Übersetzungen führen kann. Dies wird im folgenden Beispiel deutlich:

> m157d004 es tut **mir** sehr leid aber da bin ich schon auf einem auf einer **Tagung** äh mir wäre ab **achten** Montag den achten März äh **besser** gedient

Hier wird nur der Tag fokussiert, der Monat ist nicht markiert. Das kann daran liegen, daß als selbstverständlich angenommen wird, daß März gemeint sein muß. In diesem Fall muß der Dialogkontext untersucht werden. Besondere, zusätzliche Schwierigkeiten treten auf, wenn die prosodische Fokussierung nicht nur unvollständig scheint, sondern wenn es auch auf der Worterkennungsebene Probleme gegeben hat, so daß aus der besten Kette mit der prosodischen Fokussierung kein sinnvolles Ergebnis herausgelesen werden kann:

> n114k012 Beste Kette: Dienstag vormittag äh habe ich schon ein Treffen um zehn das äh wären höchstens den zehnten und dann noch möglich

> n114k012 Tatsächlich gesprochen: **Dienstag** vormittag äh habe ich schon ein **Treffen** um zehn das äh wäre dann höchstens Dienstag **nachmittag** noch möglich

Es ist klar, daß nur das fokussierte Wort für eine Übersetzung in vielen Fällen nicht ausreicht. Es bildet aber einen wichtigen Ansatz, um den Satzinhalt zu verstehen. Eine weitere Hilfe für den flachen Transfer bieten natürlich auch weitere prosodische Informationen wie Satzmodus und Phrasengrenzen.

8.4 Fokus und Semantik

Zum Abschluß dieses Kapitels wird noch auf einige Sonderprobleme der Verarbeitung von Fokusinformation in den semantischen Modulen eingegangen. In Verbmobil wird die Information aus dem prosodischen Akzenterkenner schwerpunktmäßig zur Desambiguierung von Diskurspartikeln verwendet. Die erstellten *Diskurs-Repräsentationsstrukturen* (siehe auch Abschnitt 4.3.2) können mit Hilfe von Akzentinformation reduziert werden (Bos et al., 1995). Im anschließenden Abschnitt wird noch einmal detailliert die Einbindung der prosodischen Information im Semantik-Parser von Intarc beschrieben.

8.4.1 Desambiguierung von Diskurspartikeln

Ein zentrales Problem in der Semantik und beim Transfer ist die korrekte Verarbeitung von Diskurspartikeln. Diese haben eine recht hohe Frequenz in den Verbmobildaten, ihr Anteil liegt bei über 10 % (Schmitz et al., 1994). Der Gebrauch von Partikeln scheint damit typisch für deutsche Gespräche zu sein. Charakteristisch für gesprochene Sprache sind auch Abtönungspartikeln, die die Gesamtaussage des Satzes abschwächen, wie z. B. *"denn, doch, eigentlich, irgendwie"*. Häufige Partikeln sind außerdem *"dann, ja, noch, also, nämlich"*. Sie bringen zusätzliche Nuancen in eine Aussage ein; dies muß in der Zielsprache oft mit völlig anderen Konstruktionen wiedergegeben werden (sofern die Partikeln überhaupt mit übersetzt werden).

Die Diskurspartikeln können verschiedene semantische Lesarten haben, aber auch einer pragmatischen Funktion dienen, ohne die Wahrheitsbedingungen der Äußerung zu beeinflussen. Die direkte Übersetzung von Diskurspartikeln ist daher nur angebracht, wenn sie eine distinktive semantische Funktion haben. Partikeln mit pragmatischer Funktion wie Abtönung oder Sprechereinstellung müssen ein angemessenes Äquivalent in der Zielsprache erhalten, andere Partikeln sollten gar nicht mit übersetzt werden. Zur Unterscheidung der verschiedenen Funktionen von Partikeln kann die Prosodie einen wesentlichen Beitrag leisten: unterschiedliche Lesarten können auch daran festgemacht werden, ob die Partikel einen Fokusakzent trägt oder nicht.

In Bos et al. (1995) wurde versucht, einige Regeln zur Bestimmung des semantischen Fokus aufzustellen, in Abhängigkeit von der Akzentuierung von Diskurspartikeln. Am Beispiel der Partikel "auch" konnte im Verbmobilkorpus nachgewiesen werden, daß bei einem nichtakzentuierten "auch" der semantische Fokus rechts davon mit dem finalen Phrasenakzent (PA) zusammentrifft. Wenn "auch" akzentuiert ist, wandert der semantische Fokus nach links zum Subjekt.

Stede und Schmitz (1998) geben einen ausführlichen Überblick über die verschiedenen Diskursfunktionen der Partikeln in Verbmobil. Die Bestimmung des vom Sprecher intendierten Dialogakts ist abhängig von der jeweiligen Funktion der Partikel in einer Äußerung. Die rein semantischen Lesarten der deutschen Partikeln sind in (Bos und Schiehlen, 1999) aufgelistet.

Zur Desambiguierung von Partikeln werden in Verbmobil folgende Informationsquellen verwendet (Ripplinger und Alexandersson, 1996):

1. Prosodie: Phrasengrenzen und Akzente
2. Kontext: Dialogakt und Satzmodus der aktuellen und der vorhergehenden Äußerung
3. Kotext: Fokustyp, Skopus von Fokuspartikeln
4. Position: Syntaktische Position der Partikel (Anfang, Mitte oder Ende der Äußerung)

8.4.2 Prosodische Information zum Aufbau der semantischen Repräsentation in Verbmobil

Lieske et al. (1997) beschreiben die Einbindung der prosodischen Information in den linguistischen Modulen von Verbmobil. In den Modulen für Syntax und Semantik wird die Prosodie zur Ermittlung von syntaktischen Grenzen, zur Auswahl des Satzmodus und zur Bestimmung des semantischen Fokus verwendet. Im Transfermodul beeinflußt die Prosodie die lexikalische Auswahl zur adäquaten Übertragung der Äußerung in die Zielsprache.

Im Modul für Semantische Konstruktion werden sog. *VIT-Strukturen* (Verbmobil Interface Terms) aufgebaut, die syntaktische, semantische, prosodische und pragmatische Informationen enthalten. Damit können bereits Ambiguitäten wie relativer Skopus aufgelöst und lokale Anaphernresolutionen durchgeführt werden.

In der syntaktisch-semantischen Analyse wird die Prosodie zur Unterstützung herangezogen; steht sie im Widerspruch zur bisherigen Analyse, hat die syntaktisch-semantische Information Vorrang. Bei der Segmentierung der Dialogäußerungen verarbeitet die syntaktische Analyse die Information für prosodische Phrasengrenzen; dabei ist zu beachten, daß nicht jede prosodische Phrasengrenze einer syntaktischen Grenze entspricht und umgekehrt (Batliner et al., 1998a). Bei der Bestimmung des Satzmodus reicht vielfach syntaktisch-semantische Information aus (Stellung des Verbs, initiales Fragepronomen), der prosodische Satzmodus hilft der Analyse bei mehrdeutigen Sätzen. Die Information über prosodische Akzente wird bei der Beurteilung von fokussensitiven Adverbien verwendet.

Das Transfermodul verwendet die prosodische Information indirekt dann, wenn fokussensitive Adverbien auftreten. Die Wortwahl zur Übersetzung eines Adverbs ist davon abhängig, ob es fokussiert wurde oder nicht. Dies kann sich auch noch auf den Rest des Satzes auswirken: Wenn eine Diskurspartikel Abtönung oder Einschränkung ausdrückt, muß möglicherweise auch ein anderes Verb in der Zielsprache gewählt werden (Stede und Schmitz, 1998). Wenn Diskurspartikeln oder Fokuspartikeln nicht fokussiert sind, können sie oftmals in der Zielsprache auch entfallen (Ripplinger und Alexandersson, 1996).

8.4.3 Einbindung und Korrektur von Prosodieinformation in Intarc

In der Architektur von Intarc (siehe Abschnitt 4.2) sind die beiden linguistischen Parser (Syntax- und Semantikparser) aufeinander abgestimmt (sog. Tandem-Parser). Im Syntaxparser findet eine erste Reduktion des Wortgraphen mit einer reduzierten kontextfreien Grammatik statt. Der Semantikparser überprüft die verbleibenden Worthypothesen mit der vollständigen Grammatik (Kasper und Krieger, 1996).

In Abbildung 8.1 ist der Ablauf zwischen den linguistischen Modulen noch einmal detailliert dargestellt (vergleiche auch mit Abbildung 4.1). An den Schnittstellen reichen die Module ihre Informationen in bestimmten Kodierungen weiter:

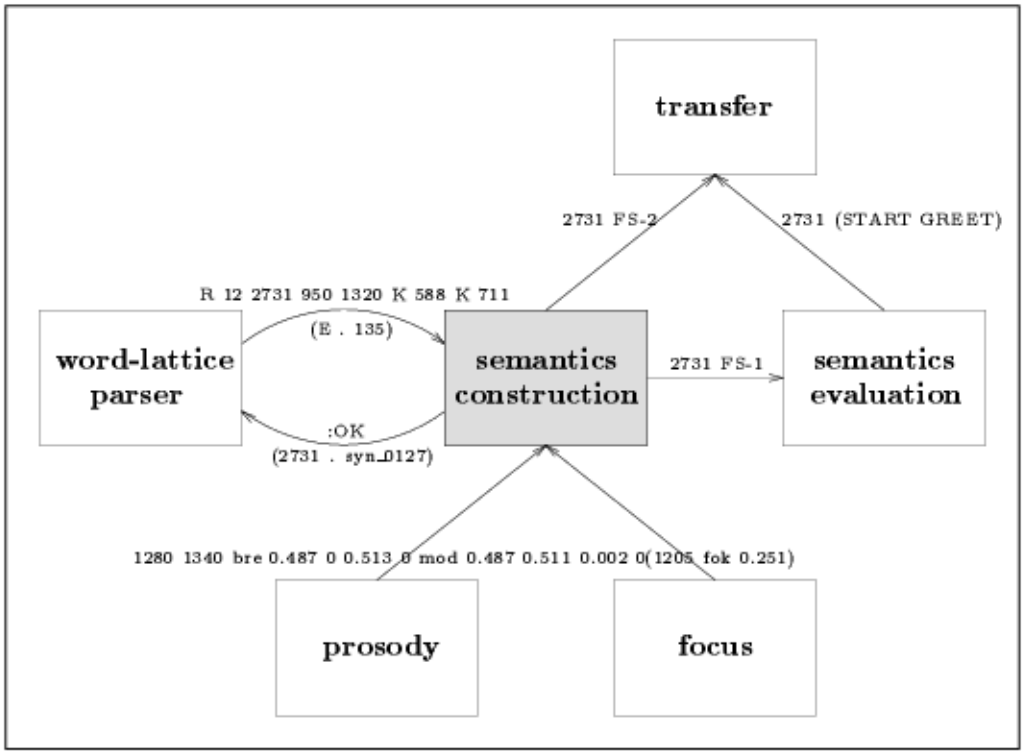

Abbildung 8.1: *Verarbeitung in den linguistischen Modulen von Intarc: Der semantische Parser (semantics construction) bekommt Hypothesen vom syntaktischen Parser (word-lattice parser) und aus den beiden prosodischen Modulen (prosody und focus). Die Beispiele für die Hypothesen sind in der entsprechenden Kodierung dargestellt (Erklärung im Text). Semantisch erlaubte Lesarten werden an die Semantische Auswertung und an den Transfer geschickt. [aus Kasper und Krieger, 1996]*

Fokuserkennung → Semantik-Parser
`(1205 fok 0.251)` =
Fokus zum Zeitpunkt 1205 ms, Konfidenzwert 0.251 (unsicherer Fokus)

Prosodie-Erkennung → Semantik-Parser
`1280 1340` = Phrasengrenze im Bereich zwischen 1280 ms und 1340 ms

`bre 0.487 0 0.513 0` = Wahrscheinlichkeiten für Grenztypen:
B0 (keine Grenze), B2 (schwache Grenze), B3 (starke Grenze) und B9 (irreguläre Grenze) (siehe auch Abschnitt 5.4)

`mod 0.487 0.511 0.002 0` = Moduswahrscheinlichkeiten für eine B3-Grenze:
keine Grenze (0.487), progrediente B3 (0.511), terminale B3 (0.002), interrogative B3 (0) (siehe auch Abschnitt 2.6.3)

Beide Parser machen starken Gebrauch von prosodischer Phrasengrenzeninformation (Strom et al., 1997). Die Grammatiken sind auf Sätze ausgerichtet; deswegen ist eine korrekte Segmentierung der Dialogbeiträge in Sätze oder Phrasen notwendig. Diese Segmentierung kann nicht immer allein mit linguistischen Kriterien durchgeführt werden; die

Information über prosodische Grenzen bedeutet außerdem eine starke Zeitreduktion für die Parser (siehe Abschnitt 4.3.2).

Im Semantikparser werden Merkmalsstrukturen aufgebaut, die den syntaktischen und semantischen Gehalt der Äußerungen beschreiben. Die prosodische Information wird direkt in diese Merkmalsstrukturen eingebaut. Die Grammatik wurde entsprechend angepaßt, so daß prosodische Ereignisse zusammen mit syntaktisch-semantischen Beschränkungen abgeglichen werden können. In Intarc werden Satzmodus und Fokusinformation indirekt über den Semantik-Parser an das Modul für Semantische Auswertung geliefert. Die Semantische Auswertung dient der Auflösung von Referenzen, der Erkennung von Dialogakten; sie verwaltet darüber hinaus das Dialogaktgedächtnis.

Die Fokusinformation kann dazu dienen, die korrekte Lesart eines Satzes und damit auch den entsprechenden Dialogakt zu bestimmen. Da wichtige Information eher akustisch markiert wird, kann damit leichter die Diskursfunktion ermittelt werden. Der erkannte Fokusakzent wird direkt in die lexikalische Struktur der Wortbeschreibung in der Grammatik integriert (Kasper und Krieger, 1996).

Für den Semantikparser ist es problematisch, wenn die prosodische Information fehlerhaft ist. Da die Reduktion der Worthypothesen durch die prosodischen Grenzen stark gesteuert wird, können zusätzliche oder fehlende Grenzen die Segmentierung in Phrasen und damit die Analyse durch die Grammatik stark beeinträchtigen oder sogar ganz unmöglich machen. Zur Lösung dieses Problems wurde ein *Recovery-Mechanismus* entwickelt (Kasper und Krieger, 1996): Dieser erlaubt es, bereits verworfene Hypothesen zu reaktivieren. Alle Hypothesen, die aufgrund von prosodischer Information verworfen wurden, werden in einem Pool für verzögerte Hypothesen gesammelt. Wenn die semantische Analyse zu keinem akzeptablen Ergebnis kommt, kann auf diesem Pool wieder aufgesetzt werden.

Eine weitere Möglichkeit, fehlerhafte prosodische Information zu verwerfen, ist die Nutzung von Schwellwerten. Sowohl der Prosodie-Erkenner als auch die Fokuserkennung liefern einen Konfidenzwert für die erkannten prosodischen Ereignisse. Anhand dieses Konfidenzwertes kann die prosodische Information entsprechend gewichtet werden: bei niedrigen Werten wird sie ignoriert, bei sehr hohen Werten darf sie eine wichtige Steuerungsfunktion einnehmen.

Für die korrekte semantische Analyse ist es besonders wichtig zu wissen, ob bestimmte Diskurspartikeln akzentuiert sind oder nicht, oder ob ein Kontrastfokus vorliegt. Wenn die Fokusakzenterkennung zu tolerant eingestellt ist, werden zu viele Akzente erkannt, so daß eine sinnvolle Gewichtung in relevante und weniger relevante Information nicht mehr stattfinden kann. Es erscheint daher auch problematisch, daß vom Akzenterkenner in Verbmobil Haupt- und Nebenakzente mit gleicher Gewichtung trainiert werden. Eine Unterscheidung dieser beiden Akzentstärken erscheint für die semantischen Module durchaus wünschenswert (Batliner, 1999). Daher wurde inzwischen beschlossen, die Akzenterkennung besser zu differenzieren, entweder durch getrenntes Training der einzelnen Akzentklassen oder auch durch Weglassen der Nebenakzente beim Training (Batliner, 1999).

9. Abschließende Diskussion

In diesem letzten Kapitel sollen die verschiedenen Ergebnisse abschließend diskutiert werden. Im Rahmen dieser Arbeit wurde ein Verfahren zur Erkennung des *prosodischen Fokus* entwickelt. Es wird zunächst noch einmal die Frage untersucht, mit welchen Hilfsmitteln diese Erkennung verbessert werden kann. Weiterhin werden die Anwendungen der Fokusinformation diskutiert. Anschließend zu diesen Betrachtungen folgt ein kurzer Ausblick.

Diese Arbeit wurde unter anderem wesentlich durch die Architekturvorgaben innerhalb des Teilprojektes Intarc geprägt (Abschnitt 4.2). Eine wichtige Rolle spielte dabei die Forderung nach Inkrementalität, die einige Verfahrensmöglichkeiten von vornherein ausschloß. Einige Untersuchungen gehen dennoch über Intarc hinaus; diese konnten allerdings nicht mehr im Gesamtsystem getestet werden.

9.1 Zusatzinformationen für die Fokuserkennung

Die in dieser Arbeit vorgestellte automatische Erkennung von Fokusakzenten stützt sich in erster Linie auf den Grundfrequenzverlauf. Wenn zusätzliche Information zu Hilfe genommen werden soll, muß die Zuverlässigkeit entsprechend gewichtet werden: Leistet die Information nur einen geringen Beitrag, dient sie nur der Bestätigung von bereits gefundenen Akzenten? Oder steuert sie die Fokusanalyse und führt zu völlig neuen Erkennungsergebnissen?

Als zusätzliche Information wurden zunächst die weiteren prosodischen Parameter wie Energie und Dauer in Betracht gezogen. Die Energieunterschiede zwischen fokussierten und nichtfokussierten Bereichen schienen allerdings zu gering, um für die Erkennung nützlich zu sein. Die Messungen für einzelne Vokale zeigten jeweils unterschiedliche Energiedifferenzen. Ohne Kenntnis des gesprochenen Vokals lassen sich diese Informationen also kaum sinnvoll nutzen. Auch der Parameter Dauer ist nur mit Hilfe von Segmentinformation interpretierbar. Wenn Worthypotheseninformation zur Verfügung steht (Nöth et al., 1997), kann die verlängerte Dauer eines Wortes eine sehr nützliche Unterstützung für die Fokusakzenterkennung sein. Am Ende eines Satzes ist allerdings das Phänomen der finalen Längung zu beachten; für einen finalen Fokus ist die Dauerinformation nicht so aussagekräftig (Cooper et al., 1985).

Weiterhin wurde die Verwertbarkeit von komplexeren prosodischen Ereignissen wie Satzmodus oder Phrasengrenzen untersucht. Im 3. Kapitel wurden bereits einige Experimente vorgestellt, die die Abhängigkeiten von Fokusakzenten in bezug auf den Satzmodus und

ihre Stellung im Satz belegen (Abschnitte 3.5.1 und 3.5.2). Das hier entwickelte Verfahren für die Fokuserkennung zeigte beispielsweise schlechtere Ergebnisse für Äußerungen mit final ansteigender Intonationskontur. Mit Hilfe einer leichten Modifikation des Verfahrens konnte dieses Problem gebessert werden. Optimal wäre allerdings eine unterschiedliche Erkennung von Fragen und Nicht-Fragen, wie es auch schon von Batliner (1989b) vorgeschlagen wurde. Dies bringt allerdings Probleme mit der Forderung der Inkrementalität mit sich, da erst am Ende einer Äußerung ein finaler Anstieg bestimmt werden kann.

Die Einbindung von prosodischen Phrasengrenzen zeigte ebenfalls recht gute Ansätze. Ziel war es, die Fokuserkennung in sinnvolle Erkennungsabschnitte zu unterteilen. Damit konnten die Einfügungsfehler noch einmal stark reduziert werden, da die Anzahl der zu erkennenden Akzente begrenzt wurde. In 11.7 % der untersuchten Daten gibt es allerdings einen Doppelfokus, der in der Erkennung bisher noch nicht berücksichtigt werden konnte. Zur besseren Nutzung der Phrasengrenzeninformation müßte die Möglichkeit eines Doppelfokus in die Erkennung integriert werden.

In einer ersten Phase von Intarc gab es eine Akzenterkennung, die Akzentinformation an einen linguistischen Worterkenner lieferte. Diese Arbeiten wurden allerdings vorzeitig eingestellt (Strom, 1998). Die Akzenterkennung wurde auf den Verbmobil-Etiketten trainiert (siehe Abschnitt 5.4), d. h., es sollten Phrasenakzente, Nebenakzente und Emphase/Kontrast-Akzente erkannt werden. Wie im Abschnitt 5.5 gezeigt wurde, sind aber nur etwa 70 % der Phrasenakzente gleichzeitig Fokusakzente, so daß es zweifelhaft scheint, ob erkannte Akzente für das Fokusmodul nutzbar wären.

Denkbar wäre eine Nutzung eher als Negativbestätigung: wo kein Akzent erkannt wurde, liegt mit sehr hoher Wahrscheinlichkeit auch kein Fokusakzent vor. Die Stärke des Fokusmoduls liegt aber ohnehin schon darin, daß es relativ wenig Einfügungsfehler macht. Es erscheint recht kritisch, die Akzentinformation für eine Bestätigung von Fokusakzenten zu nutzen: Da Fokuserkennung und Akzenterkennung gleichermaßen starken Gebrauch von der F_0-Information machen, ist zu erwarten, daß die Module bei fehlerhafter Grundfrequenzbestimmung auch die gleichen Fehler begehen.

Die in dieser Arbeit entwickelte Fokuserkennung sollte in erster Linie auf akustischer Information basieren. Dazu kann auch noch die Worthypotheseninformation gerechnet werden, die in Verbmobil erfolgreich eingesetzt wird (Abschnitt 4.3.2). Weitere lexikalisch-syntaktische Information wie Wortart oder Wortstellung sollten aber von einem linguistischen Modul beurteilt werden. Wenn ein starker Akzent auf einem Funktionswort oder in ungewöhnlicher Position gefunden wird, sollte er deswegen nicht automatisch verworfen werden. Insbesondere die Interpretation von engem Fokus kann nur von der Semantik vorgenommen werden, da sie auch über den Dialogkontext und Weltwissen verfügt. Wenn der erkannte Fokusakzent zu keiner sinnvollen Interpretation führt, kann er im Zweifelsfall immer noch ignoriert werden.

In Verbmobil werden zur Verbesserung der Prosodieerkennung noch zusätzlich Informationen aus der Worterkennung genutzt (siehe Abschnitt 4.3.2). Da Phrasengrenzen nicht innerhalb von Wörtern auftreten, können diese mit Hilfe von Wortgrenzen zuverlässiger bestimmt werden. Die Kenntnis von Segmentgrenzen erlaubt eine Einbeziehung von Dauerinformation für die Prosodieerkennung (Nöth et al., 1997). Zur Bewertung der ein-

geschränkten Information in Intarc wurde untersucht (Strom, 1998), inwieweit auch die Erkennungsleistung von menschlichen Hörern begrenzt ist, wenn ihnen nur die prosodischen Informationen (aber nicht die einzelnen Wörter) zur Verfügung stehen.

Zu diesem Zweck wurde mehreren Versuchspersonen delexikalisierte Sprache dargeboten, um die Informationen, die die Prosodieerkenner in Intarc zur Verfügung haben, zu simulieren. Es stellte sich heraus, daß es für die Hörer relativ schwierig war, ohne Wortinformation prosodische Ereignisse wie Phrasengrenzen oder Akzente zu erkennen. Der Prosodieerkenner schnitt bei der Erkennung von Phrasengrenzen etwas schlechter ab als die Hörer, bei der Akzenterkennung war er sogar besser (Strom, 1998). Die Erkennungsleistung der Hörer variierte in bezug auf das verwendete Delexikalisierungsverfahren; auch die Form der Markierung von erkannten Ereignissen war für die Hörer noch nicht optimal gelöst, da es schwierig war, die erkannten Zeitpunkte nur im Signal zu markieren.

Für die Fokusakzenterkennung ist die Wortinformation zur Zeit noch nicht nutzbar. Zu diesem Zweck müßten weitere Experimente zur Dauerinformation gemacht werden. In Verbmobil hat es sich gezeigt, daß die Wortinformation zu einer Steigerung in der Akzenterkennung beiträgt (Kießling, 1997). Auch für die Prosodieerkennung in Intarc (Strom, 1998) wird vermutet, daß eine wesentliche Verbesserung nur noch mit Hilfe von Wortinformation erreicht werden kann.

9.2 Beitrag der Fokusinformation zur Unterstützung anderer Module

Charakteristisch für Verbmobil und Intarc ist eine starke Vernetzung, d. h. die einzelnen Verarbeitungsmodule arbeiten nicht isoliert, sondern geben schnell Daten und Ergebnisse weiter. Jedes Modul kann so zur Verbesserung von anderen Modulen und damit des Gesamtsystems beitragen. In den folgenden Abschnitten wird noch einmal zusammengefaßt, welchen Beitrag die Fokuserkennung für die anderen Module leisten kann.

9.2.1 Worterkennung

Von verschiedenen Autoren (z. B. Lea, 1980) wurde vermutet, daß akzentuierte Bereiche auf der akustischen Ebene leichter zu erkennen sind. Diese Bereiche sind meist besonders deutlich, laut oder sorgfältig gesprochen, sie bilden sog. *'islands of reliability'*. Wenn solche zuverlässigeren Stellen erfolgreich analysiert sind, könnte auch der Kontext sicherer erschlossen werden.

Eine Einschränkung findet sich allerdings bei Waibel (1988). Seine Untersuchungen konnten die Verbesserung für die phonetische Erkennung an akzentuierten Stellen nicht bestätigen. Er räumt aber die Möglichkeit ein, daß die Übereinstimmung zwischen der phonetischen Realisierung und idealisierter lexikalischer Repräsentation an akzentuierten Stellen eine Erleichterung für den Lexikonzugriff darstellen kann.

Im Teilprojekt Intarc wurde eine informelle Untersuchung darüber durchgeführt, inwie-

weit fokussierte Wörter von der Worterkennung (Hübener et al., 1996) besser erkannt werden. Dies wurde für 330 Äußerungen der CD 4 untersucht. Die Erkennungsrate lag bei 73 % für willkürlich fokussierte Wörter, und bei 75.6 % für die von der Fokuserkennung als fokussiert bestimmten Wörter (H. Heine, 1996, persönliche Mitteilung). Die Unterschiede scheinen nicht sehr groß, zum Testzeitpunkt hatten allerdings sowohl die Worterkennung als auch die Fokuserkennung noch einige Performanzmängel. Es ist außerdem zu vermuten, daß bei angestrebter Erkennung von Sprache in schlechter Signalqualität fokussierte Wörter noch wichtiger für die Erkennung werden können.

9.2.2 Semantik

In den Semantikmodulen werden Merkmalsstrukturen aufgebaut, die den Inhalt der Äußerung und die Beziehungen zum Dialogverlauf möglichst eindeutig beschreiben. Akustisch bestimmte Fokusakzente können einen wichtigen Beitrag dazu leisten, diese Merkmalsstrukturen korrekt und eindeutig zu gestalten, außerdem kann die Analyse beschleunigt werden, weil möglicherweise eine langwierige Kontextsuche entfällt.

Zur endgültigen Bestimmung des semantischen Fokus ist eine linguistische Analyse unerläßlich. Die akustisch bestimmten Fokusakzente können erst durch die linguistischen Module eindeutig zugeordnet und genutzt werden. Das dafür benötigte Zusatzwissen kann die Wortart oder die Wortstellung sein, auch der Dialogkontext und ein bestimmtes Situationswissen sind wichtig.

In vielen Fällen läßt sich der semantische Fokus einfach anhand von syntaktischen Regeln oder aufgrund von Erwartungen im Dialogverlauf bestimmen. Die Kenntnis des prosodischen Fokus kann besonders effektiv genutzt werden, wenn vor allem die Fokusakzente an ungewöhnlichen Positionen zur Verfügung stehen, die nicht automatisch vorhersagbar sind (also meist enger Fokus). Die Fokusakzenterkennung ist dann entsprechend 'verbraucherfreundlich' zu trainieren: Es sollte nicht angestrebt werden, auch schwache Default-Akzente oder Nebenakzente zu erkennen - sonst steigt die Anzahl der Einfügungsfehler beträchtlich an, und der Nutzen für die Semantik wäre dahin.

Für das hier vorgestellte Verfahren zur Fokuserkennung bedeutet das eine Konzentration auf die Kategorien Ff, Fk und Fq (siehe auch Abschnitt 6.8). Wenn ein Default-Akzent Fd erkannt wird, ist dies aber nicht als Einfügungsfehler zu bewerten, denn die semantische Analyse wird dadurch nicht abgelenkt, sondern bestätigt.

9.2.3 Transfer

Der Transfer in Intarc nutzt die Fokusinformation auf zwei verschiedenen Ebenen: In der tiefen Analyse wird die Information indirekt genutzt, sie ist bereits in der Merkmalstruktur und im Dialogakt enthalten, die von der Semantik-Auswertung geliefert werden. In einem zusätzlichen Pfad wird mit der besten Wortkette (aus der Worterkennung) und der Fokusinformation ein flacher Transfer durchgeführt. Die Fokusakzente werden dabei den Wörtern aus der besten Wortkette zugeordnet.

Die Erfolgsrate der flachen Übersetzung ist erstaunlich hoch, fast die Hälfte der Verbmobil-Äußerungen kann damit inhaltlich adäquat, wenn auch nicht in vollem Wortlaut, übersetzt werden. Da die spontanen Dialoge vielfach grammatisch unkorrekte Abschnitte, Abbrüche und Reparaturen enthalten, ist die syntaktisch-semantische Analyse oft langwierig und führt nicht zum gewünschtem Erfolg.

Die Steuerung der Übersetzung mit fokussierten Wörtern birgt natürlich auch einige Fallstricke. Bei fehlerhafter Worterkennung aufgrund von schlechter Signalqualität kann auch eine korrekte Fokuserkennung nicht weiterhelfen. Es besteht aber eine gewisse Wahrscheinlichkeit, daß die Wörter an den fokussierten Stellen deutlicher gesprochen wurden und damit besser zu erkennen sind.

9.2.4 Synthese

In der Forschung hat eine gute Modellierung der Prosodie in der Synthese mehr und mehr Beachtung gefunden. Auch beim Menschen hängt die Schnelligkeit des Verständnisses davon ab, wie gut die prosodische Struktur eines Satzes erkennbar ist (Birch und Clifton, 1995). Für ein Synthesemodul gibt es verschiedene Möglichkeiten, die prosodische Struktur eines Textes zu bestimmen. Liegt nur ein geschriebener Text vor, kann zunächst mit einigen einfachen Regeln eine Grobstrukturierung vorgenommen werden (Satzzeichen liefern Informationen für Phrasengrenzen und Satzmodus, unterschiedliche Akzentuierungswahrscheinlichkeiten für bestimmte Wortarten, Position im Satz, etc.). Für eine weitere Verfeinerung der prosodischen Gestaltung muß der Text linguistisch analysiert werden (siehe auch Abschnitt 3.2).

In Verbmobil wird ein "Concept-to-speech"-Ansatz angestrebt. Da ja zunächst eine Erkennung stattfindet, existiert bereits eine linguistische Analyse des gesprochenen Textes in Form von semantischen Merkmalsstrukturen und Dialogakten. Diese werden als abstraktes Konzept an das Generierungsmodul weitergegeben. Mit Hilfe dieses Konzepts kann die Generierung den zu synthetisierenden Text mit den entsprechenden linguistischen Strukturen annotieren. Die Prosodie leistet hier auch wieder einen indirekten Beitrag, indem sie den korrekten Aufbau der semantischen Merkmalsstrukturen beschleunigt bzw. erst ermöglicht. Idealerweise sollten die realisierten Akzente direkt von der Prosodie-Erkennung über die linguistischen Module zur Synthese weitergereicht werden (siehe Abschnitt 4.1); bisher konnte dies jedoch von der Generierung noch nicht technisch umgesetzt werden (Lieske et al., 1997).

Auch wenn die linguistisch-prosodische Struktur vorliegt, hat die Synthese noch einige 'technische' Probleme zu bewältigen. Wie verschiedene Untersuchungen gezeigt haben (siehe auch Abschnitt 3.4.3), ist die Markierung von Fokusakzenten stark abhängig von der Stellung eines zu akzentuierenden Wortes im Satz. In (Wolters und Wagner, 1998) gelang es beispielsweise nicht, einen Initialfokus für alle Hörer perzeptiv wahrnehmbar zu erzeugen. Eine lokale Manipulation des zu akzentuierenden Wortes reicht offensichtlich nicht aus, wichtig scheint vor allem die Deakzentuierung der folgenden Wörter. Darüber hinaus muß noch weiter erforscht werden, wie sich die akustischen Parameter lokal und global auf die Wahrnehmung von Fokusakzenten auswirken.

9.3 Zusammenfassung und Ausblick

Vor dem Hintergrund der eingangs gesteckten Ziele (siehe Abschnitt 1.2) bleiben noch einige Fragen offen. Nach wie vor ist es ein schwieriges Problem, die Intention eines Sprechers im Dialog herauszufinden. Es kann aber davon ausgegangen werden, daß Sprecher in der akustischen Realisierung einer Äußerung wichtige Information markieren und unwichtige in den Hintergrund schieben. Umgekehrt können auch bestimmte syntaktische und semantisch-pragmatische Regeln genutzt werden, um eine Akzentrealisierung vorherzusagen (siehe Abschnitt 3.2); letztendlich gibt aber immer die Sprecherentscheidung den Ausschlag.

Eine Akzentrealisierung kann also Hinweise zur linguistischen Interpretation einer Äußerung geben. Daher ist es wünschenswert, in einem sprachverstehenden System zusätzlich zur linguistischen Analyse auch eine akustisch-prosodische Auswertung vorzunehmen. In dieser Arbeit konnte gezeigt werden, daß es möglich ist, mit relativ geringen Mitteln eine automatische Fokusakzenterkennung durchzuführen. Im Gegensatz zu stark von der Datenmenge abhängenden Erkennungsverfahren (HMM, Neuronale Netze) konnte mit einem einfachen Regelsystem bereits eine gute Erkennungsleistung erzielt werden.

Das Verfahren läßt sich möglicherweise auch auf andere Sprachen anwenden, da die Idee des Verfahrens (siehe Abschnitt 6.1) auf einem physiologischen Korrelat basiert: Der *'physical effort'*, mit dem eine Äußerung produziert wird, scheint ungleich verteilt zu sein: Der Aufwand ist hoch bis zu dem Zeitpunkt, wo der Fokus erreicht ist - danach sinkt der Sprachproduktionsaufwand auf eine deutlich niedrigere Ebene. Diese unmittelbare Abbildung auf linguistische Kategorien scheint zunächst nur für die germanische Sprachgruppe anwendbar. Für romanische Sprachen muß die Nutzung von Fokusakzentinformation in der automatischen Sprachverarbeitung noch weiter untersucht werden (Delin und Zacharski, 1997).

Es hat sich gezeigt, daß verschiedene Sprecher stark unterschiedlichen Gebrauch von prosodischen Mitteln machen. Wenn nur sehr monoton gesprochen wird, äußert sich dies in einer recht schwachen Erkennung. Bei einer Hyperartikulation gibt es andererseits Probleme für die Worterkennung, wenn sie darauf nicht trainiert ist (Oviatt et al., 1998). Die Fokuserkennung muß versuchen, einen Mittelweg zur Erkennung von möglichst vielen unterschiedlichen Sprechern zu finden.

Die Zusammenarbeit der Module in Verbmobil und Intarc hat sich als ein sehr nützliches Konzept erwiesen. Die Prosodie stellt dabei eine wichtige Verknüpfung von akustischer und semantischer Information dar. Die Experimente in Intarc haben gezeigt, daß es auch lohnenswert ist, den Weg der 'klassischen' tiefen Analyse zu verlassen und neue Kooperationen, wie z. B. zwischen Prosodie und Transfer zu untersuchen.

Literaturverzeichnis

C. Adams und R. Munro. In search of acoustic correlates of stress: fundamental frequency, amplitude and duration in the connected utterance of some native and non-native speakers of English. *Phonetica*, 35:125 – 156, 1978.

K. Alter, H. Pirker und W. Finkler (Hrsg.). *Concept to Speech Generation Systems*, Madrid, 1997.

H. Altmann, A. Batliner und W. Oppenrieder. Das Projekt Modus-Fokus-Intonation. Ausgangspunkt, Konzeption und Resultate im Überblick. In: H. Altmann, A. Batliner und W. Oppenrieder (Hrsg.), *Zur Intonation von Modus und Fokus im Deutschen*, S. 1 – 19. Niemeyer, Tübingen, 1989.

J.L. Austin. *How to do things with words*. Oxford, 1962.

P. C. Bagshaw. An investigation of acoustic events related to sentential stress and pitch accents, in English. *Speech Communication*, 13:333 – 342, 1993.

R. Bannert. Fokus, Kontrast und Phrasenintonation im Deutschen. *Zeitschrift für Dialektologie und Linguistik*, 52:289 – 305, 1985a.

R. Bannert. Towards a Model for German Prosody. *Folia Linguistica*, XIX:321 – 341, 1985b.

R. Bannert. Automatic Recognition of Focus Accents in German. *Journal of Semantics*, 8:191 – 218, 1991.

W. J. Barry. Prosodic functions revisited again! *Phonetica*, 38:320–340, 1981.

A. Batliner. Eine Frage ist eine Frage ist keine Frage. Perzeptionsexperimente zum Fragemodus im Deutschen. In: H. Altmann, A. Batliner und W. Oppenrieder (Hrsg.), *Zur Intonation von Modus und Fokus im Deutschen*, S. 87 – 110. Niemeyer, Tübingen, 1989a.

A. Batliner. Fokus, Modus und die große Zahl. Zur intonatorischen Indizierung des Fokus im Deutschen. In: H. Altmann, A. Batliner und W. Oppenrieder (Hrsg.), *Zur Intonation von Modus und Fokus im Deutschen*, S. 21 – 70. Niemeyer, Tübingen, 1989b.

A. Batliner. Deciding upon the relevancy of intonational features for the marking of focus: a statistical approach. *Journal of Semantics*, 8:171 – 189, 1991.

A. Batliner. Protokoll des Prosodie-Workshops. Erlangen, 1999.

A. Batliner, A. Feldhaus, S. Geissler, A. Kießling, T. Kiss, R. Kompe und E. Nöth. Integrating syntactic and prosodic information for the efficient detection of empty categories. In: *Proceedings of the Int. Conf. on Computational Linguistics*, Vol. 1, S. 71 – 76. Kopenhagen, 1996.

A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann und E. Nöth. M = Syntax and Prosody: A syntactic-prosodic labelling scheme for large spontaneous speech databases. *Speech Communication*, 25:193 – 222, 1998a.

A. Batliner, W. Oppenrieder, E. Nöth und G. Stallwitz. The intonational marking of focal structure: Wishful thinking or hard fact? In: *Proceedings of XIIth ICPhS*, Vol. 1, S. 278 – 281. Aix-en-Provence, 1991.

A. Batliner und M. Reyelt. Ein Inventar prosodischer Etiketten für VERBMOBIL. Verbmobil Memo 33, Ludwig-Maximilian-Universität München und TU Braunschweig, 1994.

A. Batliner, V. Warnke, E. Nöth, J. Buckow, R. Huber und M. Nutt. How to label accent position in spontaneous speech automatically with the help of syntactic-prosodic boundary labels. Verbmobil-Report 228, Universität Erlangen-Nürnberg, 1998b.

M. Beckman. *Stress and Non-Stress Accent*. Foris, Dordrecht, 1986.

S. Birch und C. Clifton. Focus, Accent, and Argument Structure: Effects on Language Comprehension. *Language and Speech*, 38:365–391, 1995.

H. U. Block. The language components in Verbmobil. In: *Proceedings of ICASSP 97*, Vol. 1, S. 79 – 82. München, 1997.

D. Bolinger. A theory of pitch accent in English. *Word*, 14:109 – 149, 1958.

D. Bolinger. Contrastive accent and contrastive stress. *Language*, 37:83 – 96, 1961.

D. Bolinger. Accent is predictable (if you're a mind reader). *Language*, 48:633–644, 1972.

D. Bolinger. *Intonation and its uses*. Edward Arnold, London, 1989.

J. Bos, A. Batliner und R. Kompe. On the use of Prosody for Semantic Disambiguation in VERBMOBIL. Verbmobil-Memo 82, Universität München, 1995.

J. Bos und M. Schiehlen. Klassifikation der deutschen Partikeln in Verbmobil. Verbmobil-Memo 141, Universität des Saarlandes, Universität Stuttgart, 1999.

A. Bradlow, G. Torretta und D. Pisoni. Intelligibility of normal speech I: Global and fine-grained acoustic talker characteristics. *Speech Communication*, 20:255 – 272, 1996.

G. Brown, K. L. Currie und J. Kenworthy. *Questions of intonation*. Croom Helm, London, 1980.

G. Bruce und P. Touati. On the Analysis of Prosody in Spontaneous Dialogues. In: D. House und P. Touati (Hrsg.), *Working Papers*, Vol. 36, S. 37–55. Lund University, Department of Linguistics, Lund, 1990.

G. Bruce und P. Touati. On the analysis of prosody in spontaneous speech with exemplification from Swedish and French. *Speech Communication*, 11:453 – 458, 1992.

T. Bub, W. Wahlster und A. Waibel. Verbmobil: The combination of deep and shallow processing for spontaneous speech translation. In: *Proceedings of ICASSP 97*, Vol. 1, S. 71 – 74. München, 1997.

K. Bühler. *Sprachtheorie*. Gustav Fischer, Jena, 1934.

H. Bußmann. *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart, 2. Auflage, 1990.

N. Campbell. Prosodic influence on segmental quality. In: *Proceedings of EUROSPEECH '95*, Vol. 2, S. 1011 – 1014. Madrid, 1995.

J. Catford. *A practical introduction to phonetics*. Oxford Press, 1988.

W. Chafe. Language and consciousness. *Language*, 50:111 – 133, 1974.

W. Chafe. Givenness, contrastiveness, definiteness, subjects, topics and point of view. In: C. N. Li (Hrsg.), *Subject and Topic*, S. 25 – 55. Academic Press, New York, 1976.

N. Chomsky und M. Halle. *The Sound pattern of English*. Harper & Row, New York, 1968.

A. Cohen, R. Collier und J. t' Hart. Declination: construct or intrinsic feature of speech pitch. *Phonetica*, 39:254 – 273, 1982.

R. Collier. Physiological correlates of intonation patterns. *J. Acoust. Soc. Am.*, 58:249–255, 1975.

B. Collins. Convergence of fundamental frequencies in conversation: If it happens, does it matter? In: *Proceedings of ICSLP'98*. Sydney, 1998.

W. Cooper, S. Eady und P. Mueller. Acoustic aspects of contrastive stress in question-answer contexts. *J. Acoust. Soc. Am.*, 77:2142 – 2156, 1985.

W. Cooper und J. M. Sorensen. Fundamental frequency contours at syntactic boundaries. *J. Acoust. Soc. Am.*, 62:683 – 692, 1977.

W. Cooper und J. M. Sorensen. *Fundamental frequency in Sentence production*. Springer-Verlag, New York, 1981.

E. Couper-Kuhlen. *An Introduction to English Prosody*. Niemeyer, Tübingen, 1986.

A. Cruttenden. Falls and rises: meanings and universals. *Journal of Linguistics*, 17:77 – 91, 1981.

A. Cruttenden. *Intonation*. Cambridge: Cambridge University Press, 1986.

D. Crystal. *Prosodic systems and intonation in English*. Cambridge University Press, Cambridge, 1969.

D. Crystal. *A Dictionary of Linguistics and Phonetics*. Blackwell, Oxford, 4.. Auflage, 1997.

D. Crystal und R. Quirk. *Systems of prosodic and paralinguistic features in English*. Mouton, The Hague, 1964.

P. Culicover und M. Rochemont. Stress and focus in English. *Language*, 59:123 – 165, 1983.

A. Cutler und D. R. Ladd (Hrsg.). *Prosody: Models and Measurements*. Springer, Berlin, 1983.

F. Daneš. Sentence intonation from a functional point of view. *Word*, 16:34 – 54, 1960.

F. Daneš. Order of elements and sentence intonation. In: *To Honor Roman Jakobson*, S. 499 – 512. Mouton, The Hague, 1967.

F. Daneš. Zur linguistischen Analyse der Textstruktur. *Folia Linguistica*, 4:72 – 78, 1970.

F. Daneš. Functional sentence perspective and the organization of the text. In: F. Daneš (Hrsg.), *Papers on Functional Sentence Perspective*, S. 106 – 128. Academia, Prag, 1974.

R. Daniloff, G. Shuckers und L. Feth. *The Physiology of Speech and Hearing - an Introduction*. Englewood Cliffs, New Jersey, 1980.

K. J. de Jong. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *J. Acoust. Soc. Am.*, 97:491 – 504, 1995.

J. R. de Pijper. *Modelling British English Intonation*. Foris, Dordrecht, 1983.

J. R. de Pijper und A. A. Sanderman. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *J. Acoust. Soc. Am.*, 96:2037–2047, 1994.

J. Delin und R. Zacharski. Pragmatic Determinants of Intonation Contours for Dialogue Systems. *International Journal of Speech Technology*, 1:109 – 120, 1997.

A. di Cristo. *De la microprosodie à l'intonosyntaxe*. Publications Université de Provence, 1985.

A. di Cristo und D. Hirst. Modelling French micromelody: Analysis and synthesis. *Phonetica*, 43:11 – 30, 1986.

S. Eady und W. Cooper. Speech intonation and focus location in matched statements and questions. *J. Acoust. Soc. Am.*, 80:402 – 415, 1986.

S. Eady, W. Cooper, G. Klouda, P. Mueller und D. Lotts. Acoustical characteristics of sentential focus: narrow vs. broad and single vs. dual focus environments. *Language and Speech*, 29:233 – 255, 1986.

A. Elsner. Distinction between 'normal' and 'contrastive/emphatic' Focus. In: *Proceedings ICSLP'96*. Philadelphia, 1996.

A. Elsner. Focus detection with additional information of phrase boundaries and sentence mode. In: *Proceedings EUROSPEECH'97*, Vol. 1, S. 227 – 230, Rhodes, Greece, 1997a.

A. Elsner. Realisierung des Fokus in bezug auf Satzmodus, Satzposition und Sprecherabhängigkeit. In: *Fortschritte der Akustik - DAGA '97*, S. 384 – 385, 1997b.

A. Elsner. Prediction and perception of focal accents. In: *Proceedings of XIVth ICPhS*. San Francisco, 1999.

A. Elsner und A. Klein. Erkennung des prosodischen Fokus und die Anwendung im dialogaktbasierten Transfer. Verbmobil-Memo 107, Universität Bonn, Universität Hamburg, 1996.

D. Erickson. Effects of Contrastive Emphasis on Jaw Opening. *Phonetica*, 55:147 – 169, 1998.

G. Fant. *Acoustic Theory of Speech Production*. Mouton and Co., 's-Gravenhage, The Netherlands, 1960.

G. Fant und A. Kruckenberg. Preliminaries to the study of Swedish prose reading and reading style. *KTH Stockholm, Speech Transmission Laboratory - Quarterly Progress and Status Report*, 2/1989:1–83, 1989.

C. Féry. *German Intonation Patterns*. Niemeyer, Tübingen, 1993.

D. Fry. Duration and intensity as physical correlates of stress. *J. Acoust. Soc. Am.*, 27: 765 – 768, 1955.

A. Fuchs. Normaler und kontrastiver Akzent. *Lingua*, 38:293 – 312, 1976.

H. Fujisaki (Hrsg.). *Proceedings International Symposium on Prosody*. Yokohama, 1994.

R. Gartenberg. Artikulatorische Faktoren in der Ausprägung von Intonationsmustern. Magisterarbeit, Universität Kiel, 1987.

E. Grabe. Pitch accent realization in English and German. *Journal of Phonetics*, 26: 129–143, 1998.

N. Grønnum. Prosodic parameters in a variety of regional Danish standard languages, with a view towards Swedish and German. *Phonetica*, 47:182 – 214, 1990.

N. (Thorsen) Grønnum. Stress group patterns, sentence accents and sentence intonation in southern Jutland (Sønderborg and Tønder) - with a view to German. *Annual Report of the Institute of Phonetics, University of Copenhagen*, 23:1 – 86, 1989.

B. Grosz und J. Hirschberg. Some intonational characteristics of discourse structure,. In: *Proceedings of ICSLP*, S. 429 – 432. Banff, 1992.

J. Gundel, N. Hedberg und R. Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274 – 307, 1993.

C. Gussenhoven. Focus, mode and the nucleus. *Journal of Linguistics*, 19:377 – 417, 1983a.

C. Gussenhoven. Testing the reality of focus domains. *Language and Speech*, 26:61 – 80, 1983b.

C. Gussenhoven, B. H. Repp, A. Rietveld, H. H. Rump und J. Terken. The perceptual prominence of fundamental frequency peaks. *J. Acoust. Soc. Am.*, 102:3009 – 3022, 1997.

C. Gussenhoven und T. Rietveld. Intonation contours and the prominence of F0 peaks. In: *Proceedings of the ICSLP*, Vol. I, S. 339 – 342. Yokohama, 1994.

M. Halle und K. Stevens. A note on laryngal features. *M.I.T Progress Report*, 101:198 – 213, 1971.

M. Halliday. Notes on transitivity and theme in english, part 2. *Journal of Linguistics*, 3:199 – 244, 1967.

C. Hauenschild, S. Heizmann, S. Petzolt und B. Prahl. Übersetzungsstrategien, Bewertung und Kontrolle für VERBMOBIL. Verbmobil-Report 203, Universität Hildesheim, 1997.

R. Hayashi und S. Kirirtani. A Study on Production and Perception of Focus in German by Japanese Learners. In: *Annual Bulletin of Research Institute of Logopedics and Phoniatrics (RILP)*, S. 65 – 68. University of Tokyo, 1994.

P. Hedelin und D. Huber. Pitch period determination of aperiodic signals. In: *Proceedings of the ICASSP*, Vol. 1, S. 361 – 364, 1990.

W. Hess. *Pitch Determination of Speech Signals*. Springer Series of Information Sciences. Springer Verlag, Berlin, 1983.

W. Hess, A. Batliner, A. Kießling, R. Kompe, E. Nöth, A. Petzold, M. Reyelt und V. Strom. Prosodic modules for speech recognition and understanding in VERBMOBIL. In: Y. Sagisaka und N. Campbell und N. Higuchi (Hrsg.), *Computing Prosody*, S. 363 – 383. Springer-Verlag, New York, 1995.

B. Heuft. *Eine prominenzbasierte Methode zur Prosodieanalyse und -synthese*. Peter Lang, Frankfurt am Main, 1999.

J. Hirschberg. Studies of Intonation and Discourse. In: D. House und P. Touati (Hrsg.), *Proceedings of an ESCA Workshop on Prosody*, S. 90 – 95. Lund University, Department of Linguistics, Lund, 1993.

J. Hirschberg, C. Nakatani und B. Grosz. Conveying discourse structure through intonation variation. In: *Proceedings of ESCA Workshop on Spoken Dialogue Systems*, S. 189 – 192. Vigsø, Denmark, 1995.

D. Hirst und A. di Cristo. A survey of intonation systems. In: D. Hirst und A. di Cristo (Hrsg.), *Intonation Systems: a Survey of Twenty Languages*, S. 1–44. Cambridge University Press, Cambridge, 1998.

L. Hiyakumoto, S. Prevost und J. Cassell. Semantic and discourse information for Text-to-Speech Intonation. In: K. Alter, H. Pirker und W. Finkler (Hrsg.), *Concept to Speech Generation Systems*, S. 47 – 56, Madrid, 1997.

J. Hoepelman und J. Machate (Hrsg.). *Modellbildung für die Auswertung der Fokusintonation im gesprochenen Dialog (MAFID)*. Niemeyer Verlag, Tübingen, 1994. Beiträge zur Dialogforschung.

H. Hollien. On vocal registers. *Journal of Phonetics*, 2:125 – 143, 1974.

S. Hoskins. A phonetic study of focus in intransitive verb sentences. In: *Proceedings ICSLP'96*. Philadelphia, 1996.

S. Hoskins. The prosody of broad and narrow focus in English: two experiments. In: *Proceedings of EUROSPEECH'97*, Vol. 2, S. 791 – 794, Rhodes, Greece, 1997.

A. S. House und G. Fairbanks. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J. Acoust. Soc. Am.*, 25:105–113, 1953.

K. Hübener, U. Jost und H. Heine. Speech recognition for spontaneously spoken German dialogues. In: *Proceedings ICSLP'96*, Vol. 1. Philadelphia, 1996.

D. Huber. *Aspects of the Communicative Function of Voice in Text Intonation*. Dissertation, Universität Göteborg/Lund, 1988.

A. Ichikawa und S. Sato. Some prosodical characteristics in spontaneous spoken dialogue. In: *Proceedings of ICSLP*, S. 147 – 150. Yokohama, 1994.

J. Jacobs. Fokus-Hintergrund-Gliederung und Grammatik. In: H. Altmann (Hrsg.), *Intonationsforschungen*, S. 89 – 134. Niemeyer, Tübingen, 1988.

R. Jakobson. Linguistics and poetics. In: T. A. Sebeok (Hrsg.), *Style in language*. M.I.T. Press, Cambridge, Mass., 1960.

S.J. Jekat. Automatic Interpretation of Dialogue Acts. Natural Language Processing Series. Amsterdam/Berlin: Mouton de Gruyter, 1996.

S.J. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast und J.J. Quantz. Dialogue Acts in VERBMOBIL. Verbmobil-Report 65, Universität Hamburg, 1995.

W. Kasper und H.-U. Krieger. Integration of prosodic and grammatical information in the analysis of dialogs. Verbmobil-Report 141, DFKI Saarbrücken, 1996.

A. Kießling, R. Kompe, A. Batliner, H. Niemann und E. Nöth. Automatic labeling of phrase accents in German. In: *Proceedings of ICSLP*, Vol. 1, S. 115 – 118. Yokohama, 1994.

A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Spracherkennung*. Shaker, Aachen, 1997.

K. Kohler. Macro and micro F_0 in the Synthesis of Intonation. In: J. Kingston und M. Beckman (Hrsg.), *Papers in laboratory phonology I*, S. 115 – 138. Cambridge University Press, 1989.

K. Kohler. A model of German Intonation. In: K. Kohler (Hrsg.), *Studies in German intonation*, Vol. 25, S. 295 – 360. Arbeitsberichte des Instituts für Phonetik der Universität Kiel, 1991.

K. Kohler. *Einführung in die Phonetik des Deutschen*. Erich Schmidt Verlag, Berlin, 2. Auflage, 1995. (1. Auflage 1977).

K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson und W. Thon. Handbuch zur Datenerhebung und Transliteration in TP14 von Verbmobil - 3-0. Verbmobil Technisches Dokument Nr. 11, Universität Kiel, 1994.

R. Kompe. *Prosody in Speech Understanding Systems*, Vol. 1307 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin, New York, 1997.

F. J. Koopmans-Van Beinum. The role of focus word in natural and in synthetic continuous speech: acoustic aspects. *Speech Communication*, 11:439 – 452, 1992.

J. C. Kowtko. *The Function of Intonation in Task-Oriented Dialogue*. Dissertation, University of Edinburgh, 1996.

E. Krahmer und M. Swerts. Reconciling two competing views on contrastiveness. In: *Proceedings of ICSLP'98*. Sydney, 1998.

D. R. Ladd. *The Structure of Intonational Meaning*. Indiana University Press, 1980.

D. R. Ladd. *Intonational Phonology*. Cambridge University Press, Cambridge, 1996.

P. Ladefoged. The features of the larynx. *Journal of Phonetics*, 1:73 – 83, 1973.

J. Laver. *Principles of phonetics*. Cambridge University Press, Cambridge, 1994.

W. Lea. Prosodic aids to speech recognition. In: W. Lea (Hrsg.), *Trends in speech recognition*, S. 166 – 205. 1980.

I. Lehiste. *Suprasegmentals*. MIT Press, Cambridge, MA, 1970.

I. Lehiste und G. Peterson. Vowel amplitude and phonemic stress in American English. *J. Acoust. Soc. Am.*, 31:428 – 435, 1959.

I. Lehiste und G. Peterson. Some basic considerations in the analysis of intonation. *J. Acoust. Soc. Am.*, 33:419 – 425, 1961.

P. Lieberman. Some acoustic correlates of word stress in American English. *J. Acoust. Soc. Am.*, 32:451 – 453, 1960.

P. Lieberman. *Intonation, perception and language*. M.I.T Press, Cambridge, 1967.

C. Lieske, J. Bos, M. Emele, B. Gambäck und C. J. Rupp. Giving prosody a meaning. In: *Proceedings of EUROSPEECH'97*, Vol. 3, S. 1431 – 1434, Rhodes, Greece, 1997.

J. Lyons. *Semantics. Volume I*. Cambridge University Press, 1977.

S. Maeda. *A characterization of American English Intonation*. Dissertation, M.I.T, 1976.

K. Maekawa. Effects of Focus on Vowel Formant Frequencies in Japanese. In: *International Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing*, S. 2–2 – 2–11. ATR, Japan, 1995.

M. Mast. Schlüsselwörter zur Detektion von Diskontinuitäten und Sprechhandlungen. Verbmobil-Memo 57, Universität Erlangen-Nürnberg, 1995.

M. Mast, R. Kompe, S. Harbeck, A. Kießling, H. Niemann und E. Nöth. Dialog Act Classification with the help of Prosody. Verbmobil-Report 130, Universität Erlangen-Nürnberg, 1996.

V. Mathesius. Zur Satzperspektive im modernen Englisch. *Archiv für das Studium der neueren Sprachen und Literaturen*, 84:202 – 210, 1929.

J. Mayer. *Intonation und Bedeutung*. Dissertation, Fakultät für Philosophie, Universität Stuttgart, 1997.

W. Menzerath und A. de Lacerda. *Koartikulation, Steuerung und Lautabgrenzung*. Dümmler, Berlin, 1933.

B. Möbius. *Ein quantitatives Modell der deutschen Intonation — Analyse und Synthese von Grundfrequenzkonturen*. Niemeyer, Tübingen, 1993.

R. Most und E. Saltz. Information structure in sentences: New information. *Language and Speech*, 22:89 – 95, 1979.

S. Nakajima und J. Allen. Prosody as a cue for discourse structure. In: *Proceedings of ICSLP*, S. 425 – 428. Banff, 1992.

S. Nakajima und J. Allen. A Study on Prosody and Discourse Structure in Cooperative Dialogues. *Phonetica*, 50:197–210, 1993.

J. Neppert und M. Pétursson. *Elemente einer akustischen Phonetik*. Buske, Hamburg, 1986.

H. Niemann, E. Nöth, A. Batliner, J. Buckow, F. Gallwitz, R. Huber, A. Kießling, R. Kompe und V. Warnke. Using prosodic cues in spoken dialog systems. In: Y. Kosarev (Hrsg.), *Proceedings of SPECOM'98 Workshop*, S. 17 – 28. St. Petersburg, 1998.

S. Nooteboom und J. Kruyt. Accents, focus distribution, and the perceived distribution of given and new information: An experiment. *J. Acoust. Soc. Am.*, 82:1512 – 1524, 1987.

S. G. Nooteboom. The prosody of speech: Melody and rhythm. In: W. J. Hardcastle und J. Laver (Hrsg.), *The Handbook of Phonetic Sciences*, S. 641–673. Blackwell, Oxford, 1997.

E. Nöth. *Prosodische Information in der automatischen Spracherkennung*. Niemeyer, Tübingen, 1991.

E. Nöth, A. Batliner, A. Kießling, R. Kompe und H. Niemann. Prosodische Information: Begriffsbestimmung und Nutzen für das Sprachverstehen. Verbmobil-Report 219, Universität Erlangen-Nürnberg, 1997.

J. Ohala und W. Ewan. The speed of pitch change. *J. Acoust. Soc. Am.*, 53:345, 1973.

S. Oviatt, G.-A. Levow, E. Moreton und M. MacEachern. Modeling global and focal hyperarticulation during human-computer error resolution. *J. Acoust. Soc. Am.*, 104: 3080 – 3098, 1998.

R. Petersen. Variation in inherent F_0 level differences between vowels as a function of position in utterance and the stress group. *Annual Report of the Institute of Phonetics, University of Copenhagen*, 13:27 – 57, 1979.

R. Petersen. Perceptual compensation for segmentally conditioned fundamental frequency perturbation. *Phonetica*, 43:31 – 42, 1986.

A. Petzold. Nachverarbeitung bei der Grundfrequenzbestimmung von Sprachsignalen zur Erfassung von Intonationskonturen. Magisterarbeit, Universität Bonn, 1993.

A. Petzold. Nachverarbeitung bei der Grundfrequenzbestimmung von Sprachsignalen zur Erfassung von Intonationskonturen. In: *Fortschritte der Akustik - DAGA '94*, S. 1345–1348, 1994.

A. Petzold. Strategies for Focal Accent Detection in Spontaneous Speech. In: *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Vol. III, S. 672–675. Stockholm, 1995.

A. Petzold. Energieverteilung in fokussierten und nichtfokussierten Bereichen. In: *Fortschritte der Akustik - DAGA '96*, S. 484 – 485, 1996.

J. Pierrehumbert. The perception of fundamental frequency declination. *J. Acoust. Soc. Am.*, 66:363 – 369, 1979.

J. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. Dissertation, M.I.T, Cambridge, 1980.

T. Portele. Perceived prominence and acoustic parameters in American English. In: *Proceedings of ICSLP'98*. Sydney, 1998.

T. Portele. A perceptually motivated intonation model for German. In: *Proceedings of XIVth ICPhS*. San Francisco, 1999.

T. Portele und B. Heuft. Towards a prominence-based speech synthesis system. *Speech Communication*, 21:61–72, 1997.

S. Prevost und M. Steedman. Specifying intonation from context for speech synthesis. *Speech Communication*, 15:139 – 153, 1994.

P. J. Price, M. Ostendorf, S. Shattuck-Hufnagel und C. Fong. The use of prosody in syntactic disambiguation. *J. Acoust. Soc. Am.*, 90:2956–2970, 1991.

M. Reyelt. Consistency of prosodic transcriptions. Labelling experiments with trained and untrained transcribers. In: *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Vol. 4, S. 212 – 215, Stockholm, 1995.

B. Ripplinger und J. Alexandersson. Disambiguation and translation of German Particles in VERBMOBIL. Verbmobil-Memo 70, IAI / DFKI Saarbrücken, 1996.

M. Rooth. A theory of focus interpretation. *Natural Language Semantics*, 1:75–117, 1992.

M. Rossi. L'intensité spécifique des voyelles. *Phonetica*, 24:129 – 161, 1971.

A. A. Sanderman. *Prosodic phrasing*. Dissertation, Technische Universiteit Eindhoven, 1996.

D. Schaffer. The role of intonation as a cue to topic management in conversation. *Journal of Phonetics*, 12:327 – 344, 1984.

S. Schmerling. A re-examination of normal stress. *Language*, 50:66 – 73, 1974.

B. Schmitz und K. Fischer. Pragmatisches Beschreibungsinventar für Diskurspartikeln und Routineformeln anhand der Demonstratorwortliste. Verbmobil-Memo 75, Technische Universität Berlin, Universität Bielefeld, 1995.

B. Schmitz und J. J. Quantz. Dialogue Acts in Automatic Dialogue Interpreting. Verbmobil-Report 173, TU Berlin, 1996.

B. Schmitz, J. J. Quantz, N. Ruge, D. Kochanowska und J. Lagunov. Übersetzung von Dialogen ins Englische - Interpretationshypothesen am Beispiel von Verben und Präpositionen. Verbmobil-Memo 10, TU Berlin, 1994.

J.R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, 1969.

K. Silverman. F_0 segmental cues depend on intonation: The case of the rise after voiced stops. *Phonetica*, 43:76 – 91, 1986.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert und J. Hirschberg. ToBI: A standard for labeling English prosody. In: *Proceedings of the International Conference on Spoken Language Processing, Banff*, S. 867–870, 1992.

A. Sluijter. *Phonetic Correlates of Stress and Accent*. Holland Academic Graphics, The Hague, 1995.

A. Sluijter und V. van Heuven. Perceptual cues of linguistic stress: intensity revisited. In: *Proceedings of ESCA Workshop on Prosody*, Vol. 41, S. 246 – 249. Workings Papers, Lund, 1993.

A. Sluijter und V.J. van Heuven. Effects of Focus Distribution, Pitch Accent and Lexical Stress on the Temporal Organization of Syllables in Dutch. *Phonetica*, 52:71–89, 1995.

C. Sorin. Functions, roles and treatments of intensity in speech. *Journal of Phonetics*, 9: 359–374, 1981.

D. Sperber und D. Wilson. *Relevance: Communication and Cognition*. Blackwell, Oxford, 2. Auflage, 1995. 1. Auflage 1986.

M. Stede und B. Schmitz. Partikeln: Diskursfunktionen für die Übersetzung. Verbmobil-Memo 139, TU Berlin, 1998.

B. Streefkerk, L. Pols und L. ten Bosch. Automatic detection of prominence (as defined by listeners) in read aloud sentences. In: *Proceedings of ICSLP'98*. Sydney, 1998.

V. Strom. Detection of Accents, Phrase Boundaries, and Sentence Modality in German with Prosodic Features. In: *Proceedings EUROSPEECH'95*, S. 2039–2041. Madrid, Spain, 1995.

V. Strom. *Automatische Erkennung von Satzmodus, Akzentuierung und Phrasengrenzen in einem sprachverstehenden System*. Dissertation, Universität Bonn, 1998.

V. Strom, A. Elsner, W. Hess, W. Kasper, A. Klein, U. Krieger, J. Spilker, H. Weber und G. Görz. On the use of prosody in a speech-to-speech translator. In: *Proceedings EUROSPEECH'97*, Vol. 3, S. 1479 – 1481, Rhodes, Greece, 1997.

M. Swerts. *Prosodic features of discourse units*. Dissertation, University of Antwerpen, 1993.

M. Swerts. Prosodic features at discourse boundaries of different strength. *J. Acoust. Soc. Am.*, 101:514–521, 1997.

M. Swerts, C. Avesani und E. Krahmer. Reaccentuation or deaccentuation: A comparative study of Italian and Dutch. In: *Proceedings of XIVth ICPhS*, S. 1541 – 1544. San Francisco, 1999.

M. Swerts, D. Bouwhuis und R. Collier. Melodic cues to the perceived finality of utterances. *J. Acoust. Soc. Am.*, 96:2064 – 2075, 1994.

P. Taylor. *A phonetic model of English intonation*. Dissertation, University of Edinburgh, 1992.

P. Taylor. The rise/fall/connection model of intonation. *Speech Communication*, 15:169 – 186, 1994.

J. Terken. Fundamental frequency and perceived prominence of accented syllables. *J. Acoust. Soc. Am.*, 89:1768 – 1776, 1991.

J. Terken. Fundamental frequency and perceived prominence of accented syllables II. Nonfinal accents. *J. Acoust. Soc. Am.*, 95:3662 – 3665, 1994.

J. Terken, E. Lathouwers, M. Theune und G. Veldhuijzen van Zanten. Prosodic structuring of system utterances in man-machine dialogues. In: *Proceedings of SPECOM'97 Workshop*, S. 71 – 76. Cluj-Napoca, 1997.

J. t'Hart, R. Collier und A. Cohen. *A perceptual study of intonation*. Cambridge University Press, Cambridge, 1990.

N. Trubetzkoy. *Grundzüge der Phonologie*. Göttingen, 1939.

S. Uhmann. *Fokusphonologie*. Niemeyer, Tübingen, 1991.

N. Umeda. F_0 declination" is situation dependent. *Journal of Phonetics*, 10:279 – 290, 1982.

G. Ungeheuer. *Elemente einer akustischen Theorie der Vokalartikulation*. Springer, Berlin, 1962.

J. Vaissière. Language-independent prosodic features. In: A. Cutler und D. R. Ladd (Hrsg.), *Prosody: Models and Measurements*, S. 53 – 66. Springer, Berlin, 1983.

J. Vaissière. The use of prosodic parameters in Automatic Speech Recognition. In: H. Niemann, M. Lang und G. Sagerer (Hrsg.), *Recent Advances in Speech Understanding and Dialog Systems*, S. 71 – 99. Springer, Berlin, 1988.

K van Deemter. A blackboard model of accenting. *Computer Speech and Language*, 12: 143 – 164, 1998.

J. van den Berg. Myoelastic-aerodynamic theory of voice production. *Journal of Speech and Hearing Research*, 1:227 – 244, 1958.

O. von Essen. *Allgemeine und angewandte Phonetik*. Akademieverlag, Berlin, 1953.

P. Wagner. The synthesis of German contrastive focus. In: *Proceedings of XIVth ICPhS*. San Francisco, 1999.

W. Wahlster. Verbmobil – Translation of Face-to-Face dialogs. In: *Proceedings of EUROSPEECH '93*, Vol. 1, S. 29–38, Berlin, 1993. Opening and plenary sessions.

W. Wahlster. Verbmobil: Erkennung, Analyse, Transfer, Generierung und Synthese von Spontansprache. Verbmobil-Report 198, DFKI Saarbrücken, 1997.

A. Waibel. *Prosody and Speech Recognition.* Morgan Kaufmann Publishers Inc., San Mateo, California, 1988.

H. Weber, J. W. Amtrup und J. Spilker. Innovative Systemarchitekturen zur inkrementellen interaktiven Verarbeitung. *KI*, 4:26 – 30, 1997.

M. Wolters und P. Wagner. Focus perception and prominence. In: *Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache - Konvens'98*, S. 227 – 236, Bonn, 1998.