Lehrstuhl für Informatik 5
(Mustererkennung)
Universität Erlangen-Nürnberg

Prof. Dr.-Ing. H. Niemann
Dr.-Ing. E. Nöth

Martensstr. 3
D–8520 Erlangen
(09131) 85–7774

Verbundprojekt ASL
- Südverbund -

Architekturen von Systemen
zur integrierten Analyse von
Sprachlauten und Sprachstrukturen

Institut für Deutsche Philologie
Ludwig-Maximilians-Universität

Prof. Dr. H. Altmann

Schellingstr. 3
D–8000 München 40
(089) 2180–2916

Implementation eines Intonationsmodells des Deutschen

Barbara Raithel
Andreas Kießling
Anton Batliner
Ralf Kompe
Elmar Nöth

ASL-Süd—TR—11—92/FAU

Juli 1992

Gehört zum Antragsabschnitt: 4.6 Prosodie

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter den Förderkennzeichen 01IV102H0 und 01IV102F4 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Inhaltsverzeichnis

Übersicht: Prosodische Information gewinnt für sprachverstehende Systeme zunehmend an Bedeutung. Zum Beispiel können der Satzmodus oder die Hervorhebung wichtiger Informationen (Fokussierung) ausschließlich mit prosodischen Parametern wie Tonhöhe, Intensität und Dauer markiert werden. Um einen sinnvollen Einsatz prosodischer Information in einem sprachverstehenden System zu erreichen, ist zunächst Wissen über das charakteristische Intonationsmuster bei unterschiedlichen Satzmodi oder Fokuspositionen notwendig. Die intonatorischen Markierungen des Satzmodus und des Fokus können jedoch nicht separat untersucht werden, da sie sich gegenseitig beeinflussen.
Ein syntaktisch eingeschränktes Modell zur intonatorischen Markierung von Modus und Fokus im Deutschen, das Wissen über die wechselseitige intonatorische Beeinflussung beinhaltet, wurde im Rahmen des von der Deutschen Forschungsgemeinschaft geförderten Projekts "Modus–Fokus–Intonation" entwickelt. In der vorliegenden Arbeit wird dieses eingeschränkte Intonationsmodell implementiert, wobei als Repräsentationsform ein semantisches Netz gewählt wird. Anschließend wird die Gültigkeit des implementierten Modells überprüft.

1 Einleitung

1.1 Begriffserläuterungen

Zunächst sollen die Begriffe Prosodie, Intonation und Fokus erläutert werden. Die hier verwendeten Definitionen wurden im wesentlichen aus [Nöth91] übernommen.

Unter **Prosodie** wird die Untersuchung sprachlicher Eigenschaften verstanden, die sich nicht nur auf einen Laut beziehen, sondern übergreifend mehrere aufeinanderfolgende Laute betreffen. Als kleinste Untersuchungseinheit der Prosodie wird im allgemeinen die Silbe angesehen. Zu den prosodischen Eigenschaften gehören Tonhöhe, zeitliche Strukturierung, Lautheit, Klangfarbe, Sprechgeschwindigkeit u.a. Diesen perzeptiven Parametern stehen akustische Korrelate gegenüber. So wird die Tonhöhe im Sprachsignal durch die Grundfrequenz, die zeitliche Strukturierung durch die Dauer und die Lautheit durch die Intensität realisiert.

Für den Begriff **Intonation** existieren in der Literatur verschiedene Definitionen [Bußm83, Kohl77]. In [Nöth91] wird Intonation als die distinktive Verwendung *aller* prosodischen Eigenschaften zur Bedeutungsdifferenzierung einer Äußerung verstanden, wobei als die wichtigsten prosodischen Parameter Tonhöhe, zeitliche Strukturierung und Lautheit angesehen werden. Die Differenzierung kann in der emotionalen oder intellektuellen Bedeutung einer Äußerung erfolgen. Emotionale Differenzierung bedeutet, daß der Sprecher mit Hilfe der prosodischen Parameter unterschiedliche Reaktionen wie Wut, Ärger oder andere Emotionen ausdrückt. Im folgenden interessiert jedoch nur die intellektuelle Bedeutungsdifferenzierung, die in drei verschiedene Bereiche unterteilt werden kann:

- **Markierung des Akzents**
  Im Deutschen ist für jedes isoliert gesprochene Wort festgelegt, auf welcher Silbe der Wortakzent liegt, das heißt, welche Silbe eines Wortes gegenüber allen anderen hervorgehoben wird. Das Wort *August* beispielsweise nimmt je nach Betonung verschiedene Bedeutungen an. Liegt der Wortakzent auf der ersten Silbe, so ist der männliche Vorname gemeint, bei Betonung der zweiten Silbe der Monatsname. Mit Hilfe der Intonation werden in diesem Fall also die verschiedenen Bedeutungen indiziert.
  In einem Satz werden nur einige Wortakzente auch tatsächlich realisiert. Die am stärksten ausgeprägte Betonung wird als Satz- oder Primärakzent bezeichnet, andere Betonungen in weiteren Phrasen als Phrasen- oder Sekundärakzente. Dadurch können bestimmte Informationen im Satz besonders hervorgehoben werden, wie durch das folgende Beispiel illustriert wird.

  Der Leo <u>säuft</u>.
  vs.
  Der <u>Leo</u> säuft.

  Im ersten Fall wird die Feststellung, daß der Leo Alkoholprobleme hat, intonatorisch gekennzeichnet, im zweiten Fall die Feststellung, daß der Leo säuft und nicht irgendjemand anderes. Als das wichtigste prosodische Mittel zur Akzentuierung wird die Tonhöhe angesehen.

- **Markierung des Satzmodus**
  In [Altm84, Altm88, Altm89] werden unter Satzmodi "...komplexe syntaktische Strukturen (Beispiele: Aussagesatz, Entscheidungsfragesatz, Wunschsatz), denen regelhaft bestimmte abstrakte Funktionstypen (Beispiele: Aussage, Entscheidungsfrage, Wunsch) zugeordnet sind..." ([Altm89, S. 1]) verstanden. Neben anderen Merkmalen wie kategoriale Füllung, Stellungseigenschaften und morphologische Markierungen, kann der Satzmodus auch mit Hilfe der Intonation markiert werden. So ist der folgende Beispielsatz je nach Einsatz prosodischer Parameter entweder dem Funktionstyp Aussage oder Frage zuzuordnen.

  Da fährt noch einer.
  vs.
  Da fährt noch einer?

  Als die wichtigste prosodische Eigenschaft im Zusammenhang mit der intonatorischen Markierung des Satzmodus wird im allgemeinen die Tonhöhe genannt. So kann ein Fragesatz durch das Anheben der Stimme am Äußerungsende charakterisiert werden. Vor allem bei grammatikalisch unvollständigen Sätzen oder bei einzelnen Wörtern kann die Differenzierung der unterschiedlichen Funktionstypen häufig nur auf intonatorische Art und Weise erfolgen (siehe Beispiel in Kapitel 1.3).

- **Gliederung der Äußerung**
  Bei längeren Äußerungen sind oft nur aufgrund der zeitlichen Strukturierung verschiedene Bedeutungen zu unterscheiden. Wie der nachfolgende Beispielsatz (aus [Kohl77]) zeigt, wird in der Schriftsprache die Gliederung durch die korrekte Interpunktion realisiert, in der gesprochenen Sprache müssen jedoch prosodische Parameter, wie vor allem Dauer und Tonhöhe, die Bedeutung differenzieren.

  Der gute Mann denkt an sich, selbst zuletzt.
  vs.
  Der gute Mann denkt an sich selbst zuletzt.

Die drei unterschiedlichen Funktionen der Intonation — Markierung des Akzents, Markierung des Satzmodus und Gliederung der Äußerung — werden teilweise mit den gleichen prosodischen Mitteln realisiert, wobei die Tonhöhe bei allen dreien den bedeutendsten Parameter darstellt. Bei Untersuchungen zu einer dieser Funktionen müssen immer die anderen beiden mitberücksichtigt werden, da sich beispielsweise die intonatorische Markierung des Satzmodus und die Markierung des Satzakzents gegenseitig beeinflussen.

Der semantische Begriff **Fokus** beschreibt die neue oder wichtige Information einer Äußerung. Nach [Jaco88] kann jeder Satz in einen Fokusteil und einen dazu komplementären Hintergrundteil gegliedert werden. Der Hintergrundteil enthält Informationen, die bereits aufgrund der Kontextsituation bekannt sind. Phonetisch kann der Fokus durch den Satzakzent realisiert werden, es werden also prosodische Mittel eingesetzt, um die wichtige Information hervorzuheben. Auch durch die Wortstellung kann der Fokus markiert werden, wie folgendes Beispiel zeigt.

Es ist der Leo, der säuft.

1.2 Das System EVAR

Am Lehrstuhl für Informatik 5 (Mustererkennung) der Universität Erlangen–Nürnberg entsteht ein sprachverstehendes System, das in der Lage sein soll, Dialoge zu führen. Dieses System, genannt EVAR (**E**rkennen, **V**erstehen, **A**ntworten, **R**ückfragen), sieht als einen Bestandteil auch ein Prosodiemodul vor. Eine ausführliche Beschreibung des Systems findet sich in [Niem85, Niem88].

Der Problemkreis, über den der Dialog geführt wird, ist auf die "Intercity–Fahrplanauskunft" beschränkt. Kontinuierlich gesprochene und per Telefon übertragene Sprache soll erkannt und im Laufe eines Dialogs die vom Kunden erfragte Information ausgegeben werden.

Das System ist in mehrere Module unterteilt, die verschiedenen linguistischen Ebenen entsprechen. Das **Akustik–Phonetik–Modul** extrahiert aus dem Sprachsignal einzelne Laute, die mit Hilfe des **Worterkennungsmoduls** zu Worthypothesen zusammengesetzt werden. Über **Syntax-, Semantik-** und **Pragmatikmodul** werden aus den einzelnen Wörtern Sätze erzeugt und innerhalb des Anwendungsbereichs "Intercity–Fahrplanauskunft" interpretiert. Das **Dialogmodul** steuert den Ablauf des Dialogs und somit die Reaktion des Systems.
Der Einsatz des **Prosodiemoduls** kann innerhalb der verschiedenen linguistischen Ebenen von Nutzen sein und so zu einer Erhöhung der Leistung des Gesamtsystems führen. Im folgenden Kapitel werden verschiedene Einsatzmöglichkeiten erläutert.

1.3 Einsatzmöglichkeiten eines Prosodiemoduls

In einem sprachverstehenden System wie EVAR, das kontinuierliche Sprache verstehen soll, kann der Einsatz eines Prosodiemoduls vor allem die folgenden Ziele haben, die sich im wesentlichen auf die unterschiedlichen funktionalen Rollen der Intonation (siehe Kapitel 1.1) abbilden lassen:

- **Unterscheidung des Satzmodus**
  Für ein automatisches Dialogsystem ist vor allem die Unterscheidung zwischen Frage und Nicht–Frage von Bedeutung. Eine weitere Differenzierung der Nicht–Fragen in Aussage, Aufforderung, Ausruf und Wunsch spielt eine sekundäre Rolle. In einem Dialogsystem wie EVAR, das telefonisch Fahrplanauskünfte erteilen soll, kann es zu folgender Situation kommen:

  System: *Nürnberg ab fünfzehn Uhr dreiundzwanzig.*

  Kunde : *dreiundzwanzig !*
  vs.
  Kunde : *dreiundzwanzig ?*

  Nur aufgrund der intonatorischen Markierung von *dreiundzwanzig* kann entschieden werden, ob der Kunde nachfragt oder ob er die Minutenangabe bestätigt. Dies ist eine für das Dialogmodul wichtige Information, da das System dementsprechend reagieren muß.

- **Identifikation betonter Stellen**
  In [Nöth91] wurde die Erstellung einer datengetriebenen Betonungsbeschreibung erläutert. Diese kann direkt aus dem Sprachsignal extrahiert werden und benötigt kein zusätzliches Wissen über das Gesprochene. Die Betonungsbeschreibung kann dazu genutzt werden, die am stärksten betonte Stelle eines Sprachsignals zu finden und so den Satzfokus, der auf Dialogebene eine wichtige Rolle spielt, zu bestimmen. Aber auch auf lexikalischer Ebene können bestimmte Wortklassen an den als betont erkannten Stellen ausgeschlossen werden.
  Liegen bereits Analyseergebnisse aus anderen Modulen vor, so spricht man von einer erwartungsgesteuerten Analyse. Diese können dazu verwendet werden, Wort- oder Satzhypothesen zu verifizieren. Das System liefert beispielsweise folgende segmental ähnlichen Sätze als konkurrierende Satzhypothesen.

  Da <u>fährt</u> noch einer ?
  vs.
  Der fährt um <u>ein</u> Uhr ?

  Kann aufgrund der Vorgeschichte des Dialogs festgestellt werden, daß im ersten Fall der Satzakzent auf *fährt* liegen muß und im zweiten auf *ein*, so bestimmt das Prosodiemodul die Position der am stärksten betonten Stelle der aktuellen Äußerung, vergleicht sie mit der Satzakzentposition in den zwei Hypothesen und kann eine Entscheidung für eine der beiden Hypothesen herbeiführen.

- **Erkennung der gesprochenen Laute**
  Mikroprosodische Eigenschaften beziehen sich auf einzelne Laute und besitzen für bestimmte Lautfolgen charakteristische Ausprägungen. Die Betrachtung solcher Eigenschaften kann innerhalb des Akustik–Phonetik–Moduls zur Lauterkennung genutzt werden. So folgt einem stimmlosen Plosivlaut in der Regel ein Grundfrequenzabfall.

- **Identifikation von Gliederungsgrenzen**
  Erfolgt die Gliederung einer Äußerung nur mittels intonatorischer Mittel, so muß das Prosodiemodul auf syntaktisch–semantischer Ebene eingesetzt werden. Satzgrenzen und inhaltliche Unterschiede, die für den weiteren Verstehensprozeß von Bedeutung sind, können so erkannt werden.

1.4 Aufgabenstellung

Im Mittelpunkt der vorliegenden Arbeit stehen die beiden folgenden Punkte, die der Einsatz prosodischer Information in einem sprachverstehenden Dialogsystem zum Ziel hat:

- Unterscheidung des Satzmodus
- Identifikation betonter Stellen zur Bestimmung des Fokus

Da sowohl die intonatorische Markierung des Satzmodus als auch die intonatorische Markierung des Satzakzents zum Zwecke der Fokussierung mit denselben prosodischen Merkmalen realisiert werden, ergeben sich Überlagerungen. Daher sollte bei der Bestimmung akzentuierter Stellen immer die Annahme über den realisierten Satzmodus mit einbezogen werden. So ist in dem Beispiel (siehe Kapitel 1.3)

Da fährt noch einer ?
vs.
Der fährt um ein Uhr ?

bei der prosodischen Verifikation von Satzhypothesen die intonatorische Fragemarkierung durchaus von Bedeutung.
Für einen sinnvollen Einsatz eines Prosodiemoduls ist zunächst Wissen darüber nötig, wie Modus und Fokus im Deutschen markiert werden, das heißt, wie die verschiedenen prosodischen Eigenschaften bzw. deren akustische Korrelate vom Sprecher eingesetzt werden, um verschiedene Betonungen und Satzmodi zu realisieren, und wie sich die Modus- und Fokusmarkierungen überschneiden.
Erste Untersuchungen dazu wurden in München im Rahmen des DFG-Projekts "Modus–Fokus–Intonation" unternommen. Es wurde ein Intonationsmodell entwickelt, das Wissen über die gegenseitige Beeinflussung der Modus- und Fokusmarkierung beinhaltet. Dieses Modell besitzt aufgrund des eingeschränkten Untersuchungskorpus nur für einen partiellen Bereich der deutschen Sprache Gültigkeit. So wird als Satzmodus nur zwischen Frage und Nicht–Frage differenziert, und als möglicher Träger des Fokus ist nur die vorletzte oder letzte Phrase erlaubt. Das Modell basiert auf der Annahme, daß es für jede Kombination von Modus und Fokus einen oder mehrere charakteristische Vertreter gibt.

Die Aufgabe der vorliegenden Arbeit besteht in der Implementierung dieses Intonationsmodells als semantisches Netz. Außerdem soll die Gültigkeit des implementierten Modells am zugrundeliegenden Korpus belegt werden, wobei die Merkmale zur Repräsentation der unterschiedlichen Intonationsmarkierungen automatisch aus dem Sprachsignal bestimmt wurden.
Kapitel 2 beschreibt die Vorgehensweise bei der Entwicklung des Modells; in Kapitel 3 wird die im Verlauf dieser Arbeit vorgenommene Implementierung erläutert. Auf die Überprüfung des implementierten Intonationsmodells wird in Kapitel 4 eingegangen.

2 Erstellung eines Intonationsmodells

In [Altm89] wird über Untersuchungen zur intonatorischen Markierung von Modus und Fokus im Deutschen berichtet. Eine getrennte Betrachtung der durch akzentuelle Hervorhebung erzielten Fokusmarkierung und der intonatorischen Satzmoduskennzeichnung ist nicht möglich, da sie durch dieselben prosodischen Merkmale, wie beispielsweise die Tonhöhe, realisiert werden, und die Markierungen sich gegenseitig beeinflussen [Batl89c]. Im Laufe dieser Untersuchungen sollte getestet werden, welche prosodischen Eigenschaften sich wie und mit welcher Relevanz auf die intonatorische Fokusmarkierung bei Fragen und Nicht–Fragen auswirken. Dazu wurden vier Modus–Fokus–Korpora gewonnen, die

ausführlich in [Batl89b] beschrieben sind und die Grundlage für die weiteren Untersuchungen darstellten. Eines dieser Korpora, das Fokus–Korpus, diente als Basis zur Erstellung eines Intonationsmodells, welches jedoch nur für einen, durch das Korpus vorgegebenen, syntaktisch eingeschränkten Bereich der deutschen Sprache gültig ist. Dieses Korpus wird im folgenden genauer dargestellt; eine ausführliche Beschreibung befindet sich in [Batl89a].

2.1 Das Untersuchungsmaterial

Basis des Fokus–Korpus sind drei verschiedene Sätze mit ähnlicher syntaktischer Struktur. Die einzelnen Sätze bilden sogenannte intonatorische Minimalpaare, das heißt, nur mit Hilfe intonatorischer Mittel können verschiedene Satzmodi realisiert werden. Ebenso wird die im Fokus stehende Phrase nur durch intonatorische Merkmale angezeigt. Weitere Erläuterungen zu dem Prinzip intonatorischer Minimalpaare befinden sich in [Altm84]. Tabelle 1 zeigt die Testsätze mit ihren möglichen Satzmodi.

SATZ				**SATZMODUS**	
Matrixsatz	1.Phrase	2.Phrase	3.Phrase	Frage	Nicht–Frage
Sie läßt	die Nina	das Leinen	weben	assertive Frage	Aussage
Lassen Sie	den Manni	die Bohnen	schneiden	Verb-Erst-Frage	Imperativ
Lassen wir	den Manni	die Blumen	düngen	Verb-Erst-Frage	Adhortativ

Tabelle 1: Die drei Testsätze des Fokus–Korpus mit möglichen Satzmodi

Das Fokus–Korpus wurde mit sechs 'naiven' Sprechern der süddeutsch/bairisch gefärbten Standardsprache (drei weiblich, drei männlich) aufgenommen. Jeder der drei Sätze wurde den Sprechern in verschiedenen Kontexten vorgelegt, wodurch sowohl Satzmodus als auch Position und Art des Fokus impliziert wurde. Abb. 1 zeigt eine solche Kontextvorgabe, durch die festgelegt wird, daß der Testsatz als Nicht–Frage realisiert wird und der Fokus auf der zweiten Phrase liegt.

Situation	In einem Textilbetrieb; eine Mutter erkundigt sich bei einer Angestellten nach den handwerklichen Fortschritten ihrer Tochter.
Kontextsatz	*Was läßt die Meisterin meine Nina gerade weben ?*
Testsatz	*Sie läßt die Nina das Leinen weben.*

Abb. 1: Beispielkontext mit Testsatz

Die Kontextsätze wurden so gewählt, daß der Fokus nur auf der letzten oder vorletzten Phrase liegen konnte, um so den Aufwand bei den darauffolgenden Untersuchungen einzuschränken. Außerdem wurde nur die Grobunterteilung des Satzmodus in Frage und Nicht–Frage beachtet, die weitere Unterteilung in Aussagesatz, Imperativsatz oder Adhortativsatz interessierten nicht. Insgesamt ergaben sich somit für die weiteren Untersuchungen folgende vier Modus–Fokus–Konstellationen:

Nicht–Frage, Fokus auf der vorletzten Phrase
Nicht–Frage, Fokus auf der letzten Phrase
Frage, Fokus auf der vorletzten Phrase
Frage, Fokus auf der letzten Phrase

Das Fokus–Korpus besteht aus 360 verschiedenen Realisierungen der drei Testsätze, wobei in 48 Prozent der Fälle der Satzmodus eine Frage war und in 76 Prozent der Fälle der Fokus auf der vorletzten Phrase lag.

2.2 Merkmalsberechnung und Klassifikation

Um Aussagen über den unterschiedlichen Einfluß der Grundfrequenz (F_0), der Dauer und der Intensität auf die Fokusmarkierung von Fragen und Nicht–Fragen machen zu können, wurden zunächst für jede Äußerung Merkmale bestimmt. Diese wurden zum Teil per Hand aus dem Zeitsignal und dem mit dem F_0-Meter gemessenen Grundfrequenzverlauf extrahiert. Die Intensitätswerte sowie das Merkmal Steigung wurden am bereits digitalisierten Signal gemessen. Tabelle 2 gibt einen Überblick über diese Merkmale. Die zur Berechnung benötigten Phrasengrenzen wurden ebenfalls per Hand ermittelt. Die

Merkmale	**Abk.**	**M/F**	**Berechnung/Transformationen**
Offset	Off	M	Grundfrequenzwert am Äußerungsende in Halbtönen minus sprecherspezifischen Basiswert.
Steigung	Steig	M	Steigung der Ausgleichsgeraden für die Grundfrequenzwerte in Halbtönen.
F_0-Maximum, 2. und 3.Phrase	Max2, Max3	M/F	Maximaler und minimaler Grundfrequenzwert der 2. und 3.Phrase jeweils
F_0-Minimum, 2. und 3.Phrase	Min2, Min3	M/F	in Halbtönen minus sprechespezifischen Basiswert.
Rel. Position von F_0-Max./Min. (2. und 3.Phrase)	Pos2, Pos3	M/F	Differenz der Positionen von F_0-Minimum und F_0-Maximum auf der Zeitachse; positiver Wert, wenn Maximum vor Minimum, sonst negativer Wert.
Dauer der 2. und 3.Phrase	Dau2, Dau3	F	$[(aktuelleDauer/mittl.\ Phrasendauer) * (Dauer/(Äußerungsdauer/Silbenzahl))]$
Intensität der 2. und 3.Phrase	Int2, Int3	F	Maximale Energie der Phrase im Frequenzbereich 0–5000 Hz, gemessen in relativen Millibelwerten.

Tabelle 2: (aus [Nöth91]) Berechnete Merkmale, die als Variablen zur Klassifikation dienen. In der Spalte M/F wird angegeben, ob sie zur Modusentscheidung (M) oder zur Fokusentscheidung (F) herangezogen wurden.

Grundfrequenzwerte Off, Max2, Max3, Min2 und Min3 wurden zunächst in Halbtöne transformiert[1]; um den Wert sprecherunabhängig zu machen, wurde ein sprecherspezifischer Basiswert (der tiefste vom Sprecher erreichte Wert) subtrahiert.

Der Einfluß der einzelnen Merkmale und damit der entsprechenden intonatorischen Eigenschaften auf die verschiedenen Modus–Fokus–Konstellationen wurde mit einem statistischen Klassifikationsverfahren (Diskriminanzanalyse) getestet. Folgende vier Klassifikationsexperimente, in denen immer nur zwei verschiedene Klassen zu unterscheiden waren, wurden vorgenommen:

Modus	Alle Äußerungen wurden entweder der Klasse Frage oder Nicht-Frage zugeordnet.
Fokus	Für alle Äußerungen wurde entschieden, ob der Fokus auf der zweiten oder auf der dritten Phrase liegt.
Fokus_F	Nur Fragen werden bezüglich der Position des Fokus klassifiziert.
Fokus_N	Nur Nicht–Fragen werden bezüglich der Position des Fokus klassifiziert.

Für nähere Erläuterungen zur Merkmalsextraktion und Klassifikation sei auf [Batl89a] verwiesen, hier sollen nur die erzielten Erkennungsraten angegeben werden.

Tabelle 3 zeigt die Erkennungsraten der einzelnen Klassifikatoren, wenn nur ein Merkmal verwendet wird (un) bzw. wenn zwei Merkmale genutzt werden (bi). In diesen Fällen bestand die Lernstichprobe ebenso wie die Teststichprobe aus den Äußerungen aller sechs Sprecher. Die Erkennungsraten spiegeln

[1] Formel zur Transformation $Ht = 17,31 \ln(Hz)$

die unterschiedliche Bedeutung der prosodischen Parameter für die Fokusmarkierung von Fragen und Nicht–Fragen wider. So besitzten die Dauer und die Intensität bei Fragen einen geringeren Einfluß auf die Markierung des Fokusakzents als bei Nicht–Fragen. Die wichtigste Einflußgröße scheint jedoch die Grundfrequenz und somit die Tonhöhe zu sein.

Merkmale	**Modus**			**Fokus**			**Fokus_F**			**Fokus_N**		
	un		bi	un		bi	un		bi	un		bi
Off	93			-			-			-		
Steig	85			-			-			-		
Max2	73	>	94	60	>	78	80	>	81	54	>	92
Max3	94			60			63			88		
Min2	65	>	84	59	>	62	53	>	71	62	>	70
Min3	84			47			70			47		
Pos2	76	>	85	51	>	69	78	>	82	69	>	52
Pos3	78			51			62			48		
Dau2		-		66	>	74	60	>	71	70	>	82
Dau3		-		72			70			82		
Int2		-		58	>	66	52	>	56	62	>	70
Int3		-		55			51			53		

Tabelle 3: Erkennungsraten der Klassifikatoren Modus, Fokus, Fokus_F und Fokus_N bei Verwendung eines (un) oder zweier (bi) Merkmale (mit Lernstichprobe =Teststichprobe).

Tabelle 4 gibt Aufschluß darüber, wie beim Einsatz aller Merkmale die Erkennungsraten aufgrund verschiedener Lern- und Prüfstichproben schwanken, wobei folgendes gilt:

- Die Zeile multiv-l=t zeigt die Erkennungsraten, wenn die Lernstichprobe der Teststichprobe entspricht. Im Vergleich zu den Erkennungsraten in Tabelle 3 läßt sich feststellen, daß beim Einsatz aller Merkmale bessere Ergebnisse erzielt werden. Die hier gelieferten Erkennungsraten stellen die oberen Grenzen für weitere Experimente dar.
- multiv-l5t1 steht für die Verwendung von fünf Sprechern als Lernstichprobe und einem als Teststichprobe. Dieser Fall simuliert eine sprecherunabhängige Spracherkennung und stellt somit den realistischen Fall dar.
- Bei multiv-l1t5 ist ein Sprecher Lernstichprobe und fünf Sprecher sind Teststichprobe. Damit kann getestet werden, inwiefern ein Sprecher alle anderen repräsentieren kann oder ob er eine sprecherspezifische Intonationsstrategie verfolgt. Diese Werte stellen untere Grenzen dar.

Lern/Test	**Modus**	**Fokus**	**Fokus_F**	**Fokus_N**
multiv-l=t	97	93	95	96
multiv-l5t1	92	84	86	94
multiv-l1t5	92	78	76	82

Tabelle 4: Erkennungsraten bei verschiedenen Lern- und Teststichproben beim Einsatz aller Merkmale.

Die Ergebnisse zeigen, daß sich die berechneten Merkmale für die Beschreibung der unterschiedlichen intonatorischen Modus- und Fokusmarkierung eignen. Außerdem läßt sich feststellen, daß eine Trennung in Fragen und Nicht–Fragen eine Verbesserung der Erkennungsraten der Fokuszuweisung bewirkt. Dies ist auf die Überlagerungen der Intonationsmarkierungen zurückzuführen.

2.3 Hörerurteile

Im Laufe der DFG-Untersuchungen wurden für die 360 Realisierungen Perzeptionstests durchgeführt. Durchschnittlich zwölf Hörer beurteilten die Realisierungen der Testsätze in drei verschiedenen Tests:

Natürlichkeitstest: Die Hörer mußten auf einer Rangskala von 1 bis 5 angeben, wie gut die Realisierung zu dem entsprechenden Kontext paßt. 1 steht dabei für 'paßt sehr gut' und 5 für 'paßt überhaupt nicht'. Als Natürlichkeitsmaß NAT für eine Realisierung wird der Mittelwert über die Urteile aller Hörer verwendet.

Kategorisierungstest: Ohne Wissen des Kontextes sollten die Hörer angeben, ob die Realisierung dem Äußerungstyp Aufforderung, Frage, Aussage, Ausruf, Exklamation oder Wunsch zuzuordnen ist. Im weiteren soll jedoch nur die Unterscheidung von Frage und Nicht–Frage interessieren. Als Maß MOD wird der Prozentsatz an Fragezuweisungen verwendet. Stuften alle Hörer eine Äußerung als Frage ein, so erhält MOD den Wert 100, wurde die Realisierung von keinem als Frage erkannt, so wird MOD der Wert 0 zugewiesen.

Akzenttest: Die Hörer sollten entscheiden, welche Phrase Träger des Satzakzents ist, wobei wiederum keine Kenntnis des Kontextes vorlag. Als Maß FOK wird $\frac{SA_2 - SA_3}{SA_1 + SA_2 + SA_3}$ angegeben, wobei SA_i die Anzahl der Hörer ist, die den Fokus auf der i–ten Phrase wahrnahmen. FOK ist $1,0$ bzw. $-1,0$, wenn alle Hörer sich für Phrase zwei bzw. Phrase drei entschieden.

Abb. 2 zeigt die Verteilung des Natürlichkeitsmaßes NAT für alle Äußerungen. Nahezu der Hälfte aller Äußerungen wurde eine Bewertung bis zu 2 zugeordnet und somit als recht natürlich beurteilt.

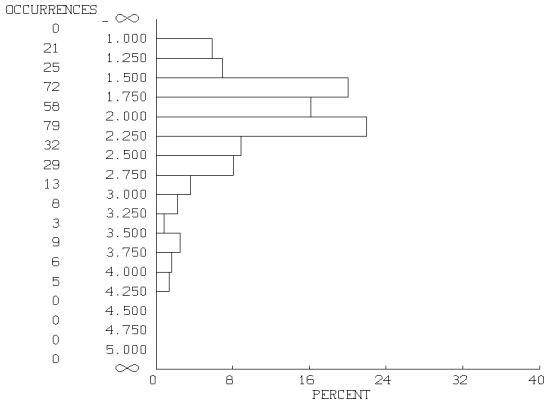
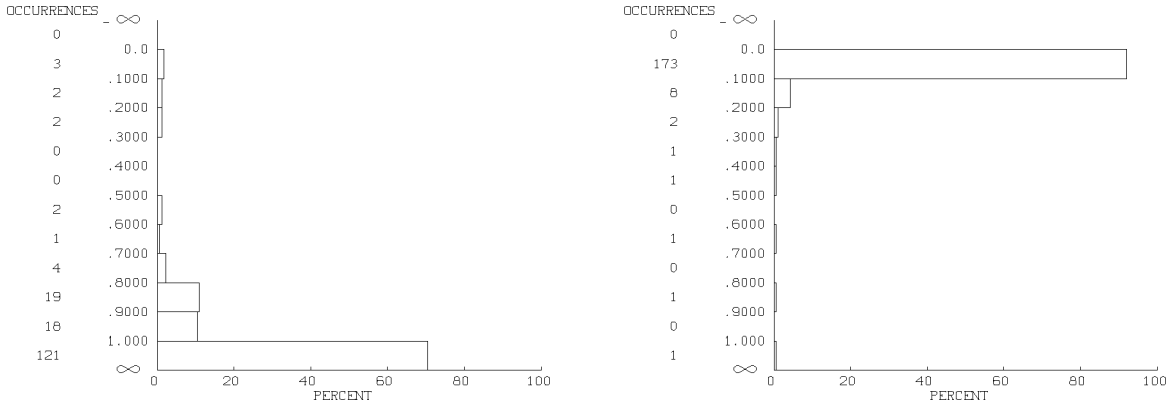


Abb. 2: Verteilung des Hörerurteils NAT für alle Äußerungen

Abb. 3(a) zeigt die Verteilung des Hörerurteils MOD nur für Fragen. 70% aller Fragen wurden eindeutig, das heißt mit der Bewertung $1,0$ als Fragen erkannt. Es gibt jedoch sieben Fälle, in denen die Mehrheit der Hörer auf Nicht-Frage entschied, und deren Maß somit geringer als $0,5$ ist. Diese Fälle werden als sogenannte Fehlproduktionen bezeichnet, da sie vom Sprecher falsch realisiert wurden. Auch für Nicht–Fragen (siehe Abb. 3(b)) sind drei solcher Fehlproduktionen zu beobachten. Es fällt jedoch auf, daß bei Nicht–Fragen die eindeutige Erkennung des Satzmodus höher liegt als bei Fragen.

Die Verteilung des aus dem Akzenttest vorliegenden Hörerurteils FOK für Fragen (siehe Abb. 4(a)) und Nicht–Fragen (siehe Abb. 4(b)) fallen im wesentlichen gleich aus, bei Nicht-Fragen scheint jedoch die Fokuszuordnung etwas eindeutiger zu sein.



(a) (b)

Abb. 3: Verteilung des Hörerurteils MOD für Fragen (a) und für Nicht–Fragen (b). (Die Prozentwerte sind für MOD/100 angetragen.)

Ein Vergleich zwischen den Verteilungen von FOK und MOD zeigt, daß sich die Hörer bei der Zuweisung des Satzmodus wesentlich sicherer waren als bei der Fokuszuordnung.

2.4 Gewinnung von Prototypen

Das Intonationsmodell soll prototypisch strukturiert sein, das heißt, typische Vertreter, welche für eine bestimmte Kategorie charakteristische Merkmale aufweisen, dienen als Prototypen. Für das eingeschränkte Fokus–Korpus heißt das, daß für jede der vier Modus–Fokus–Konstellationen ein oder mehrere typische Vertreter gesucht werden müssen, die sowohl den für diese Konstellation typischen Grundfrequenzverlauf als auch die charakteristischen Kennwerte prosodischer Parameter besitzen.

Die Gewinnung der Prototypen wurde in zwei Schritten vorgenommen (vgl. im einzelnen [Batl89a]). Im ersten Schritt wurden alle Äußerungen einer Modus-Fokus-Konstellation herangezogen, um die Mittelwerte von zehn Merkmalen (siehe Tabelle 2) zu berechnen, die während der Klassifikation ihre Relevanz bei der Beschreibung der Modus- und Fokusmarkierungen bewiesen haben. *Off* und *Steig* wurden dabei nicht betrachtet. Man erhält somit ein erstes quantitatives Modell, das für jede Konstellation einen typischen Vertreter vorgibt, der die durchschnittlichen und charakteristischen Merkmalswerte besitzt. In Abb. 5 sind für jede der Modus–Fokus–Konstellationen die durchschnittlichen Werte von Grundfrequenzmaximum und -minimum der zweiten und dritten Phrase dargestellt sowie ihre Positionen innerhalb der Phrasen. Die Grundfrequenzwerte werden in Halbtönen minus dem tiefsten vom Sprecher realisierten Wert angetragen. An der Zeitachse wird in Centisekunden die Zeit ab Äußerungsbeginn angetragen.
Fragen sind durch einen steigenden Grundfrequenzverlauf in der dritten Phrase gekennzeichnet, während bei Nicht–Fragen ein fallender Tonhöhenverlauf charakteristisch ist. Nicht–Fragen weisen auch auf der zweiten Phrase eine fallende Kontur auf. Die akzentuierte Phrase bei Nicht–Fragen besitzt gegenüber der unakzentuierten einen größeren Tonumfang. Bei Fragen mit Fokus auf der dritten Phrase ist in der zweiten Phrase ein fallender Verlauf zu beobachten.
Die Mittelwertbildung der relevanten Merkmale wurde auch getrennt für jeden Sprecher vorgenommen, und bis auf drei Ausnahmen ergaben sich dieselben charakteristischen Verhältnisse der Merkmalsmittelwerte untereinander, wie sie in Abb.5 dargestellt sind. Die Ausnahmen bildeten die Konstellationen *Frage, Fokus auf der zweiten Phrase* und *Nicht–Frage, Fokus auf der dritten Phrase* von Sprecher 6 und die Konstellation *Frage, Fokus auf der zweiten Phrase* von Sprecher 1.

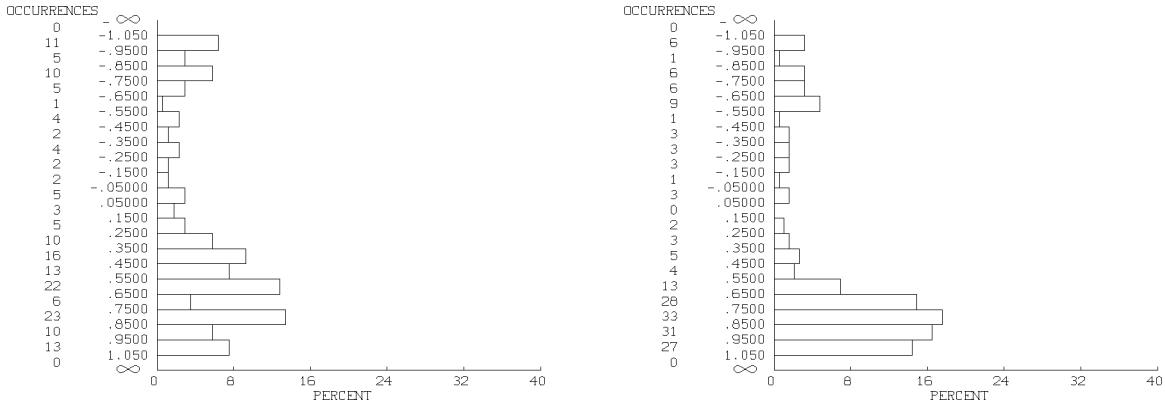
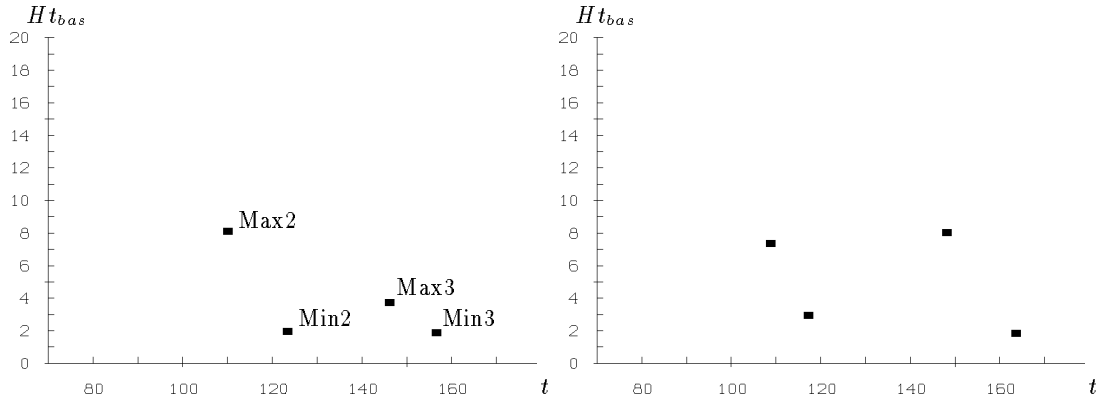


Abb. 4: Verteilung des Hörerurteils FOK für Fragen (a) und für Nicht–Fragen (b)

Der zweite Schritt zur Gewinnung von Prototypen für die jeweiligen Modus–Fokus–Konstellationen erfolgte mit Hilfe der im vorigen Kapitel beschriebenen Hörerurteile NAT, MOD und FOK, für die hohe Schwellwerte festgelegt wurden:

- NAT ≤ 2
- MOD ≥ 80 oder MOD ≤ 20
- $|\text{FOK}| = 1$

24 der insgesamt 360 Realisierungen erfüllen diese Bedingungen. Bei der Fokuszuweisung waren sich alle Hörer einig, bei der Moduszuweisung fast alle und die Natürlichkeit wurde als gut eingestuft. Verglichen mit den charakteristischen Vertretern, die im ersten Schritt gewonnen wurden, stellte man fest, daß 19 der 24 Äußerungen den typischen Verlauf der Mittelwertdarstellungen besaßen. Aus diesen 19 Realisierungen wurde für jede Konstellation eine ausgewählt, die dann als sogenannter Kernprototyp die entsprechende Konstellation repräsentiert. Die fünf Ausnahmen entsprachen den Verläufen, die aus der getrennten Betrachtung der einzelnen Sprecher resultierten und die andere typische Merkmalswerte besaßen. Diese Fälle stellen somit nicht normale, aber dennoch für den Hörer gültige Fälle dar. Dies wird dadurch berücksichtigt, daß für diese Konstellationen zusätzlich zu den Kernprototypen noch sogenannte Randprototypen vorgesehen werden. Für die Konstellation *Frage, Fokus auf der zweiten Phrase* konnten auf diese Weise zwei Randprototypen, für die Konstellation *Nicht–Frage, Fokus auf der dritten Phrase* ein Randprototyp gefunden werden. Abb. 6 zeigt die Kernprototypen für diejenigen Konstellationen, für die nur ein Prototyp gefunden werden konnte. In Abb. 7 sind die zwei Prototypen für die Konstellation *Nicht–Frage, Fokus auf der dritten Phrase* und in Abb. 8 die drei Prototypen für *Frage, Fokus auf der zweiten Phrase* dargestellt.

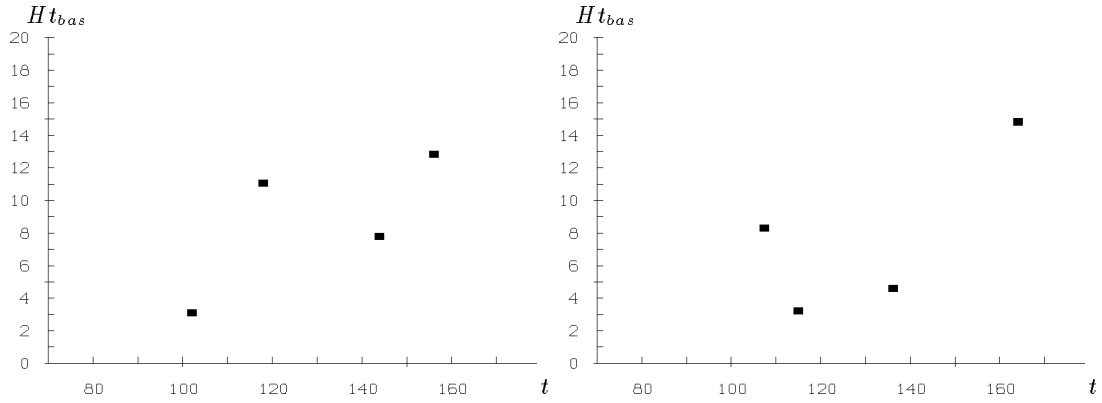


Nicht–Frage, Fokus auf der zweiten Phrase

(a)

Nicht–Frage, Fokus auf der dritten Phrase

(b)



Frage, Fokus auf der zweiten Phrase

(c)

Frage, Fokus auf der dritten Phrase

(d)

Abb. 5: Mittelwerte der Merkmale Max2, Min2, Max3 und Min3. Die Differenz aus den Positionen von Maximum und Minimum einer Phrase spiegelt die Mittelwerte für Pos2 und Pos3 wider.

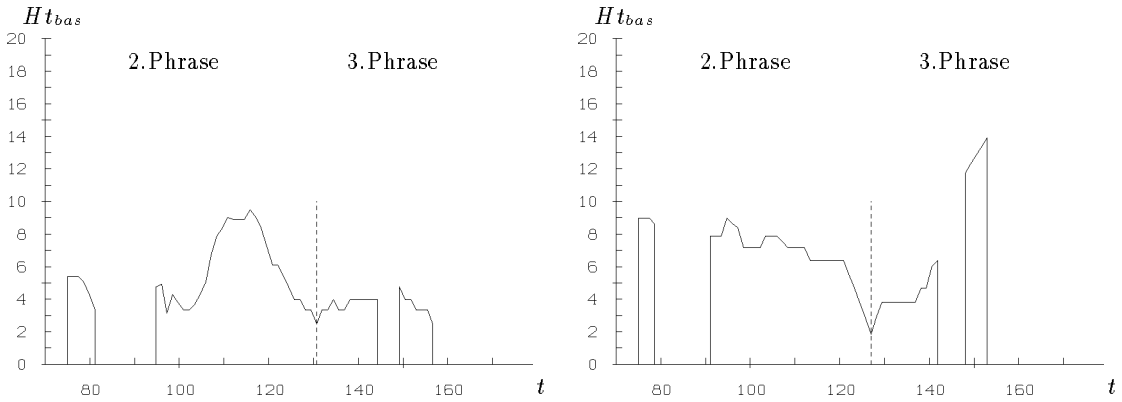


Abb. 6: Konstellationen, die nur einen Kernprototyp besitzen

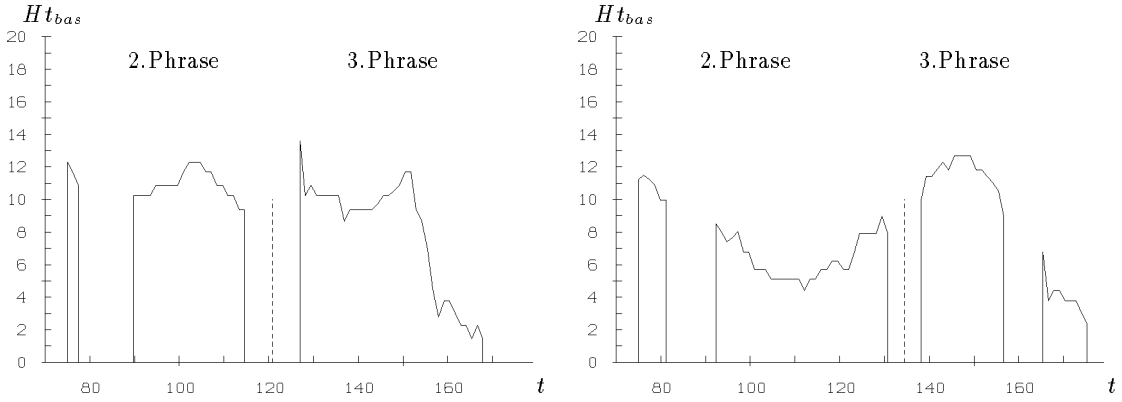


Abb. 7: Nicht–Frage, Fokus auf der dritten Phrase

2.5 Das Modell

Ergebnis der in den vorangegangenen Kapiteln beschriebenen Untersuchungen ist ein partielles Intonationsmodell der deutschen Sprache. In diesem Modell wird die gegenseitige Beeinflussung der intonatorischen Modus- und Fokusmarkierung berücksichtigt, indem für jede mögliche Kombination von Modus und Fokus einer oder mehrere prototypische Vertreter existieren.
Betont werden muß, daß das Modell sehr eingeschränkt ist. Beim Satzmodus wird nur zwischen Frage und Nicht–Frage unterschieden, weitere Differenzierungen werden vernachlässigt. Außerdem ist als mögliche Fokusposition nur die vorletzte oder die letzte Phrase zugelassen. Im weiteren beinhaltet das Modell nur Sätze, die die durch die drei Testsätze (siehe Abb. 1) vorgegebene Struktur besitzen. Bei der Konstruktion der Testsätze wurden die fokussierbaren Phrasen aus Wörtern gebildet, die nur lange Vokale enthalten. Dies stellt wiederum eine Einschränkung des Modells dar.
Abb. 9 gibt einen Überblick über das erstellte Modell.
Jede Modus–Fokus–Konstellation (NFF2 steht für *Nicht–Frage, Fokus auf der zweiten Phrase*) ist

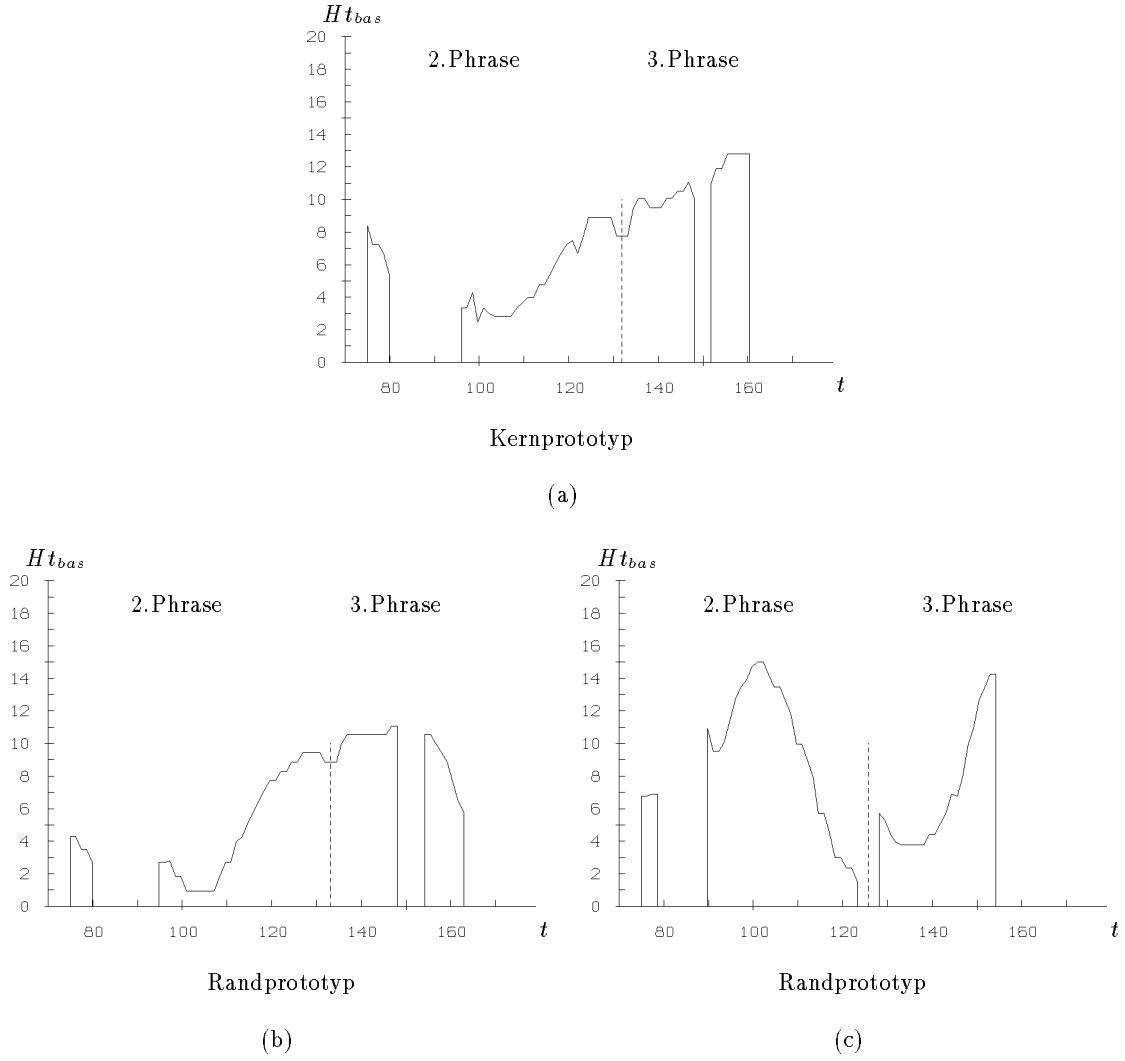


Abb. 8: Frage, Fokus auf der zweiten Phrase

durch eine der 360 Realisierungen gegeben, die den Kernprototypen und damit den charakteristischen und normalen Fall darstellt. Außerdem gibt es noch drei Randprototypen, die 'nichtnormale', aber dennoch gültige Fälle repräsentieren. Die hier verwendete Namensgebung für die einzelnen Prototypen wurde von [Nöth91] übernommen. Die ersten Buchstaben geben die entsprechende Modus-Fokus-Konstellation an, gefolgt von der Charakterisierung, ob es sich um einen Kerntyp (KT) oder einen Randtyp (RT) handelt. Die Symbole *a, A, b, B* beschreiben den typischen Grundfrequenzverlauf. *a* und *A* stehen für eine fallende Bewegung, *b* und *B* für eine steigende Bewegung. Ist die Ausprägung der Bewegung einer Phrase deutlich geringer als bei der benachbarten Phrase, so wird der Kleinbuchstabe verwendet, ansonsten der Großbuchstabe.

3 Implementierung des Intonationsmodells

Das gewonnene und im Kapitel 2 beschriebene Wissen über die intonatorische Markierung von Modus und Fokus soll nun in geeigneter Form maschinell dargestellt werden. Zur Wissensrepräsentation lassen sich verschiedene Formalismen, wie regelbasierte Systeme, Prädikatenlogik erster Ordnung, erweiterte Übergangsnetze, formale Grammatiken, Frames oder semantische Netze verwenden. Einen Überblick über verschiedene Darstellungsmöglichkeiten von Wissen gibt [Sage90].
In [Nöth91] wurde als geeignete Repräsentationsform für das Intonationsmodell das semantische Netz,

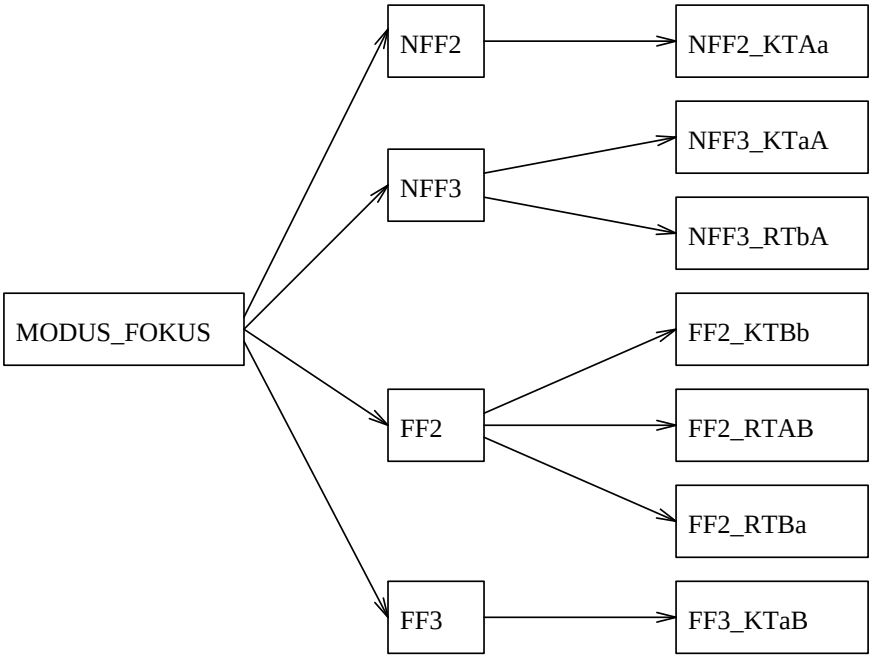


Abb. 9: Eingeschränktes Intonationsmodell

ein von Quillian [Quil68] eingeführter Formalismus, vorgeschlagen. Semantische Netze sind Graphen mit gerichteten markierten Kanten. In den Knoten kann Wissen über beliebige Begriffe (Objekte, Ereignisse, Sachverhalte) abgelegt werden. Die Kanten beschreiben bestimmte Beziehungen zwischen den Knoten. Der Formalismus des semantischen Netzes wurde zum einen deshalb gewählt, da sich die Struktur des Intonationsmodells (siehe Abb. 9) in das Netz übertragen läßt und zusätzliches Wissen, das zu einer Erweiterung des Modells führt, leicht in Form von weiteren Knoten in das Netz übernommen werden kann. Außerdem existiert am Lehrstuhl für Informatik 5 (Mustererkennung) bereits eine Systemumgebung zur Wissensrepräsentation und -nutzung auf der Basis von semantischen Netzen, genannt ERNEST, die auch zur Darstellung der syntakischen, semantischen und pragmatischen Wissensbasis von EVAR genutzt wird.
Zunächst folgt eine kurze Einführung in ERNEST, eine ausführliche Beschreibung befindet sich in [Kumm90, Sage90, Erne90]. Anschließend wird das in ERNEST implementierte *Intonationsnetz* erläutert.

3.1 Die Systemumgebung ERNEST

Die Systemumgebung ERNEST (**ER**langener semantisches **NE**tzwerk **S**ys**T**em) wurde am Lehrstuhl für Informatik 5 (Mustererkennung) zur wissensbasierten Musteranalyse entwickelt. Das System wird sowohl in der Sprachanalyse als auch in der Bildverarbeitung genutzt.
ERNEST stellt Datenstrukturen für ein semantisches Netz zur Verfügung. Abb. 10 zeigt die Struktur für den Knoten, in dem Begriffe modelliert werden können und der in ERNEST **Konzept** genannt wird. Jedes Konzept besitzt Einträge, die der näheren Beschreibung des im Knoten modellierten Begriffes dienen. So können quantitative Eigenschaften durch Attribute und Beziehungen zwischen den Attributen durch Relationen beschrieben werden. Auch für Kanten, die die Verbindungen mit anderen Konzepten herstellen, sind solche Einträge vorgesehen. Eine genaue Beschreibung der Datenstrukturen in ERNEST befindet sich in [Sage90, Erne90].

Zusätzlich zu diesem in Konzepten repräsentierten deklarativen Wissen besitzt ERNEST auch eine prozedurale Komponente, welche die Nutzung des im Netz abgelegten Wissens erlaubt. Die an Konzepte gebundenen Prozeduren zur Berechnung von Attributen, Relationen und Bewertungen stellen dabei das *problemabhängige* prozedurale Wissen dar.
Ziel der Wissensnutzung in ERNEST ist es, bestimmte in Knoten modellierte Begriffe während eines

Name	→	Text
Information	→	Text
Grade	→	4 Integer (KON,SPEZ,BST,MOD)
Prioritäten	→	5 Integer
Modell-von	→	Kante zu einem **Konzept**
Modell	→	Liste von Kanten zu **Konzepten**
Spezialisierung-von	→	Liste von Kanten zu **Konzepten**
Spezialisierung	→	Liste von Kanten zu **Konzepten**
Kontext-von	→	Liste von **Konzepten**
Bestandteil-von	→	Liste von Kanten zu **Konzepten**
Bestandteil	→	Liste von **Kantenbeschreibungen**
Konkretisierung-von	→	Liste von Kanten zu **Konzepten**
Konkretisierung	→	Liste von **Kantenbeschreibungen**
Modalität	→	Liste von **Modalitätsbeschreibungen**
Attribut	→	Liste von **Attributbeschreibungen**
lokales Attribut	→	Liste von **Attributbeschreibungen**
Analyse-Parameter	→	Liste von **Attributbeschreibungen**
Struktur-Relationen	→	Liste von **Relationsbeschreibungen**
Analyse-Relationen	→	Liste von **Relationsbeschreibungen**
Identifikation	→	Liste von **Identifikationen**
Bewertung	→	**Funktionsbeschreibung**
Instanz	→	Liste von **Instanzen**
Grafik	→	**Funktionsbeschreibung**
Spezialisierungs-Auswahl	→	Funktionsname
Akquisitionsregel	→	Funktionsname
Häufigkeit	→	1 Integer

Abb. 10: (aus [Erne90]) Die Struktur eines **Konzeptes** im semantischen Netzwerksystem ERNEST

Analyselaufs in konkreten Aufnahmen eines Sprachsignals oder eines Bildes wiederzufinden. Dies führt dazu, daß in ERNEST zwei weitere Knotentypen existieren, die während eines Analyselaufs erzeugt werden. Diese Knoten heißen **Modifiziertes Konzept** und **Instanz**. Abb. 11 gibt einen Überblick über die in ERNEST vorgesehenen Knotentypen. Abb. 12 erläutert die wichtigsten Kantentypen in ERNEST.

Konzept	In einem Konzept können allgemeine Begriffe oder Sachverhalte dargestellt werden. Jedes Konzept kann u.a. durch Attribut- und Relationsbeschreibungen näher charakterisiert werden.
Modifiziertes Konzept	Während der Analyse gewonnene Informationen und Zwischenergebnisse führen zu einer Modifikation von Konzepten, z.B. zu einer Einschränkung der möglichen Attributwerte eines Konzepts.
Instanz	Instanzen stellen eine Verbindung zwischen Signalausschnitt und Konzept her, das heißt, sie repräsentieren das im Konzept abgelegte Wissen mit konkreten Werten bezüglich des Signalausschnitts. Mit der im entsprechenden Konzept angegebenen Bewertungsfunktion wird für die Instanz eine Bewertung berechnet, die ein Maß dafür ist, wie gut der durch die Instanz gegebene Signalausschnitt zu dem Konzept paßt.

Abb. 11: Die Knotentypen in ERNEST

Das Ziel der Analyse in ERNEST besteht darin, Konzepte zu instantiieren, das heißt, die in der Wissensbasis modellierten Begriffe sollen über im Ergebnisspeicher abgelegte Instanzen einem Signalausschnitt zugeordnet werden. Beim Start der Analyse werden die zu instantiierenden Zielkonzepte angegeben. Konnte eines der Zielkonzepte erfolgreich instantiiert werden, so bricht die Analyse in der Regel ab.
Ein auf dem A^*-Algorithmus basierender *problemunabhängiger* Kontrollalgorithmus [Kumm90] sorgt dafür, daß die bestbewertete Instanz zu einem Konzept gefunden wird. Dazu wird ein sogenannter

spez	Über Spezialisierungskanten kann eine Verbindung zwischen einem allgemeinen Konzept und spezielleren hergestellt werden. Alle Attribute, Relationen und Kanten des allgemeinen Konzepts werden in der Regel an die spezielleren vererbt.
bst	Mit Hilfe von Bestandteilskanten kann ein Konzept in seine einzelnen Bestandteile zerlegt werden, die wiederum als Konzepte modelliert sind.
kon	Die Konkretisierungskante verbindet Konzepte unterschiedlicher Abstraktionsebenen.
inst	Die Instanzkante führt vom Konzept zu den zugehörigen, während der Analyse erzeugten Instanzen.

Abb. 12: Die wichtigsten Kantentypen in ERNEST

Suchbaum aufgebaut, der zu Beginn der Analyse für jedes Zielkonzept einen Knoten besitzt. Werden zum Beispiel zu einem Konzept mehrere konkurrierende Instanzen im Signalausschnitt gefunden, so wird der Suchbaum aufgespalten. In jedem Suchbaumknoten ist ein Ausschnitt des ganzen semantischen Netzes enthalten, der die bis jetzt modifizierten oder instantiierten Knoten beinhaltet. Die Suchbaumknoten werden bewertet, und der bestbewertete wird zur weiteren Analyse ausgewählt. In der Regel entspricht die Bewertung des Suchbaumknotens der Bewertung der aktuell erzeugten Instanz oder des modifizierten Konzepts. Diese Bewertung erfolgt mit Hilfe von Funktionen, die an das Konzept gebunden sind.
Die Grundlage für den Kontrollalgorithmus bilden sechs Inferenzregeln, mit deren Hilfe man modifizierte Konzepte und Instanzen erzeugen kann. Diese Regeln sind *anwendungsunabhängig*, das heißt, sie sind nur vom verwendeten Netzwerkformalismus, nicht aber vom im Netz abgelegten Wissen abhängig. Abb. 13 gibt einen Überblick über die Anwendungsziele der einzelnen Regeln.

Regel 1	Generierung partieller Instanzen
Regel 2	Generierung von Instanzen
Regel 3	Erweiterung von Instanzen
Regel 4	Datengetriebene Generierung modifizierter Konzepte
Regel 5	Modellgetriebene Generierung modifizierter Konzepte
Regel 6	Konzeptschätzung

Abb. 13: Die Ziele der sechs Inferenzregeln in ERNEST

Die Regeln werden während einer sich abwechselnden Expansions- und Instantiierungsphase angewendet. Kann kein Konzept instantiiert werden, so muß expandiert werden, das heißt, im weiteren wird versucht, die Bestandteils- und Konkretisierungskonzepte zu instantiieren. Wurden dazu erfolgreich Instanzen generiert, so kann auch das darüberliegende Konzept instantiiert werden. Fehlt jedoch noch ein Bestandteil, so erfolgt wieder eine Expansionsphase. Dies geschieht solange, bis eines der beim Start der Analyse angegebenen Zielkonzepte instantiiert werden konnte.

Ein weiterer Bestandteil von ERNEST ist die **Erklärungskomponente** [Erne90]. Mit ihrer Hilfe kann das semantische Netz und das darin abgelegte Wissen auf graphische Weise dargestellt werden. Ebenso kann das während eines Analyselaufs gewonnene Wissen und der Ablauf der Analyse dargestellt werden. Alle folgenden Abbildungen, die eine Übersicht des Netzes geben oder den während der Analyse erzeugten Suchbaum zeigen, wurden von dieser Erklärungskomponente erzeugt.

3.2 Einsatzmöglichkeit eines Intonationsmodells

In Kapitel 2 wurde erläutert, wie das zu implementierende Intonationsmodell entwickelt wurde. Bevor nun auf die Implementierung als semantisches Netz in der Systemumgebung ERNEST eingegangen wird, soll zunächst ein möglicher Verwendungszweck des Modells in einem sprachverstehenden System beschrieben werden.

Der Einsatz einer datengetriebenen, das heißt unter alleiniger Verwendung des Sprachsignals erstellten, automatischen Betonungsbeschreibung führt, wie in [Nöth91] erläutert, zu einer Verbesserung der Analyseergebnisse eines sprachverstehenden Systems. Eine Berücksichtigung von Wissen, das während der Verstehensphase bereits gewonnen wurde, läßt eine weitere Verbesserung der Analyseergebnisse erwarten. Dieses Wissen kann die segmentelle und syntaktische Struktur der Äußerung betreffen oder auch in Form von Phrasen- und Satzhypothesen vorliegen. In [Nöth91] wurde aufgezeigt, wie das Intonationsmodell erwartungsgesteuert zur prosodischen Verifikation von Satzhypothesen eingesetzt werden kann. Als Motivation dazu diente das in Kapitel 1.3 bereits aufgeführte Beispiel:

Da fährt noch einer ?
vs.
Der fährt um ein Uhr ?

Diese akustisch sehr ähnlichen Sätze stellen konkurrierende Satzhypothesen für eine vom Benutzer gesprochene Äußerung dar. Als zusätzliches Wissen steht beispielsweise noch zur Verfügung, daß es sich um Fragen handeln muß, wobei die Frage intonatorisch markiert wurde, und daß in der ersten Hypothese der Satzakzent auf dem Wort *fährt* und in der zweiten auf dem Wort *ein* liegen muß. Besitzt man nun Wissen darüber, wie sich in beiden Fällen die intonatorische Frage- und Satzakzentmarkierung auswirken und gegenseitig beeinflussen, kann man durch einen Vergleich mit der tatsächlich realisierten Intonationsmarkierung eine Entscheidung für eine der beiden Satzhypothesen treffen. Dieses Wissen ist im Intonationsmodell für einen eingeschränkten Bereich der deutschen Sprache in Form von prototypischen Realisierungen repräsentiert. Eine maschinelle Darstellung des Intonationsmodells in einem semantischen Netz und seine Nutzung zur prosodischen Verifikation von Satzhypothesen wird in [Nöth91] folgendermaßen vorgeschlagen:
In den Konzepten der Prototypen wird die Information über die charakteristische intonatorische Markierung abgelegt. Dies erfolgt in Form von Steueranweisungen, aufgrund derer die prosodische Komponente eines Synthesegeräts für eine Satzhypothese die entsprechende Grundfrequenzkontur oder andere charakteristische Kennwerte erzeugen kann. Die generierte Referenzkontur kann mit der aktuellen Testkontur verglichen werden, und ein sich daraus ergebendes Abstandsmaß spiegelt wider, wie gut die Äußerung zu der durch den Prototyp repräsentierten Modus–Fokus–Konstellation paßt. Liefert das System mehrere konkurrierende Satzhypothesen, so wird die aktuelle Kontur mit den generierten Referenzkonturen derjenigen Prototypenkonzepte verglichen, die aufgrund des bereits gewonnenen Wissens überhaupt sinnvoll sind und die durch die entsprechende Satzhypothese vorgegeben sind. So interessieren im obigen Beispiel nur Prototypen, die eine Fragemarkierung besitzen und, je nach Satzhypothese, den Satzakzent an der entsprechenden Position haben. Das sprachverstehende System entscheidet sich für diejenige Satzhypothese, für die eine Kontur erzeugt wurde, die der aktuellen am ähnlichsten ist. Kernprototypen können durch unterschiedliche Gewichtung gegenüber den Randprototypen bevorzugt werden.

Da die Möglichkeit der Erzeugung einer charakteristischen Kontur mit Hilfe des Synthesegeräts aufgrund fehlender Steueranweisungen entfällt, wurden bei der Implementierung des Intonationsmodells die Grundfrequenzkonturen der prototypischen Äußerungen, welche in [Batl89a] (siehe Kapitel 2) gefunden wurden, als Referenzkonturen herangezogen. Dementsprechend erfolgte die Wahl von Attributen und Relationen in den Konzepten der Prototypen.

3.3 Überblick über das semantische Netz *intonet*

Das Intonationsmodell wurde im Rahmen dieser Arbeit als semantisches Netz mit dem Namen *intonet* in ERNEST implementiert. Es existiert ein Konzept MODUS_FOKUS, das über Spezialisierungskanten mit den vier Konzepten verbunden ist, welche die unterschiedlichen Modus–Fokus–Konstellationen modellieren. Die Namen für diese Konzepte lauten NFF2 für *Nicht–Frage, Fokus auf der zweiten Phrase* usw. Die sieben prototypischen Äußerungen werden wiederum in spezielleren Konzepten dargestellt. Die Namensgebung der Prototypenkonzepte wurde bereits in Kapitel 2.5 erläutert. Abb. 14 zeigt die im Netz existierende Spezialisierungshierarchie. Es ist deutlich sichtbar, wie die in Abb. 9 vorgegebene Struktur des Intonationsmodells übernommen werden konnte.

Außerdem beinhaltet das Netz das Konzept AEUSSERUNG, das die Schnittstelle zur übrigen Sprachanalyse darstellt. Dieses Konzept modelliert die intonatorische Markierung einer konkreten Äußerung

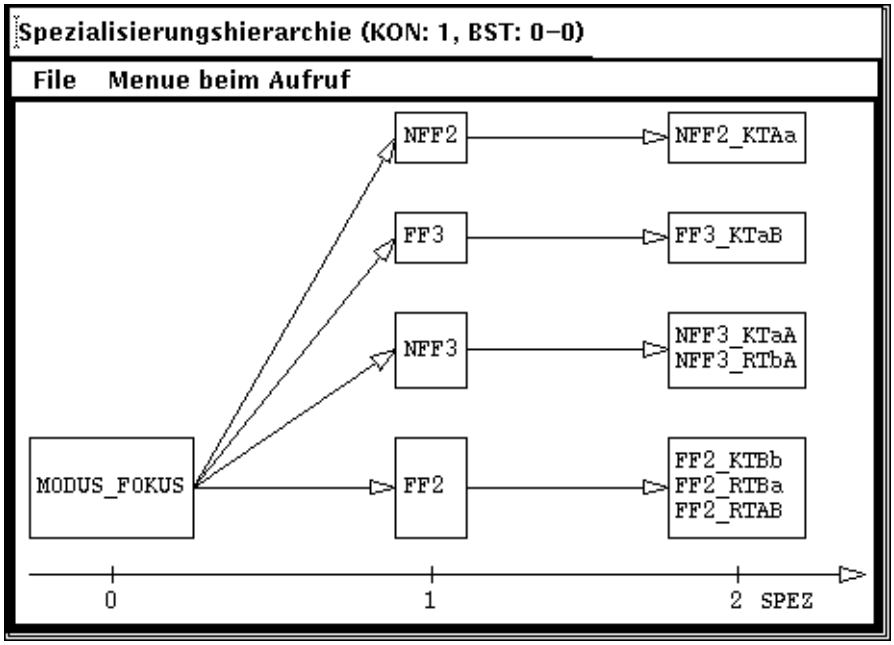


Abb. 14: Übersicht über die Spezialisierungshierarchie im Netz *intonet*

und beinhaltet die tatsächliche Grundfrequenzkontur des Sprachsignals, für das verschiedene Satzhypothesen erzeugt wurden (siehe Kapitel 3.2). Über eine Konkretisierungskante ist das Konzept AEUSSERUNG mit dem Konzept MODUS_FOKUS verbunden. Die Konkretisierungskante wurde gewählt, da die Konzepte MODUS_FOKUS, NFF2, NFF2_KTAa etc. Wissen über mögliche Modus–Fokus–Konstellationen modellieren, das Konzept AEUSSERUNG jedoch eine konkrete Ausprägung einer solchen Konstellation ist und sich somit auf einer niedrigeren Abstraktionsstufe befindet.
Abb. 15 gibt einen Überblick über alle 13 im Netz vorhandenen Knoten sowie über die Spezialisierungs- und Konkretisierungsstufen des Netzes. Die Konkretisierungskanten von Konzepten höherer Spezialisierungsstufen sind auf den Vererbungsmechanismus in ERNEST zurückzuführen.

Da in ERNEST die Möglichkeit der Nutzung des abgelegten Wissens gegeben ist, konnte für das Netz *intonet* automatisch das Analyseprogramm *intonet_cont* erzeugt werden.
Wie bereits erwähnt, enthält jedes Prototypenkonzept die Information über die prototypische Äußerung aus dem Fokus–Korpus. Jede andere Äußerung des Fokus–Korpus kann beim Aufruf des Analyseprogramms herangezogen werden, mit dem Ziel, dafür die entsprechende Modus–Fokus–Konstellation zu bestimmen. Im wesentlichen geschieht dies dadurch, daß die Referenzkontur der prototypischen Äußerung mit der Grundfrequenzkontur der aktuellen Testäußerung, basierend auf dem Prinzip der dynamischen Zeitverzerrung, verglichen wird.
Dasjenige Prototypenkonzept, welches nach Ablauf der Analyse die bestbewertete Instanz besitzt, gibt die Modus–Fokus–Konstellation vor, die der Testäußerung zugewiesen wird.
Beim Aufruf des Analyseprogramms werden

- die zu instantiierenden Zielkonzepte und
- der Name der aktuell zu testenden Äußerung

angegeben. Für jede Äußerung aus dem Fokus–Korpus müssen zwei Dateien mit folgendem Inhalt zur Verfügung stehen:

- Grundfrequenzwerte
- Phrasengrenzen

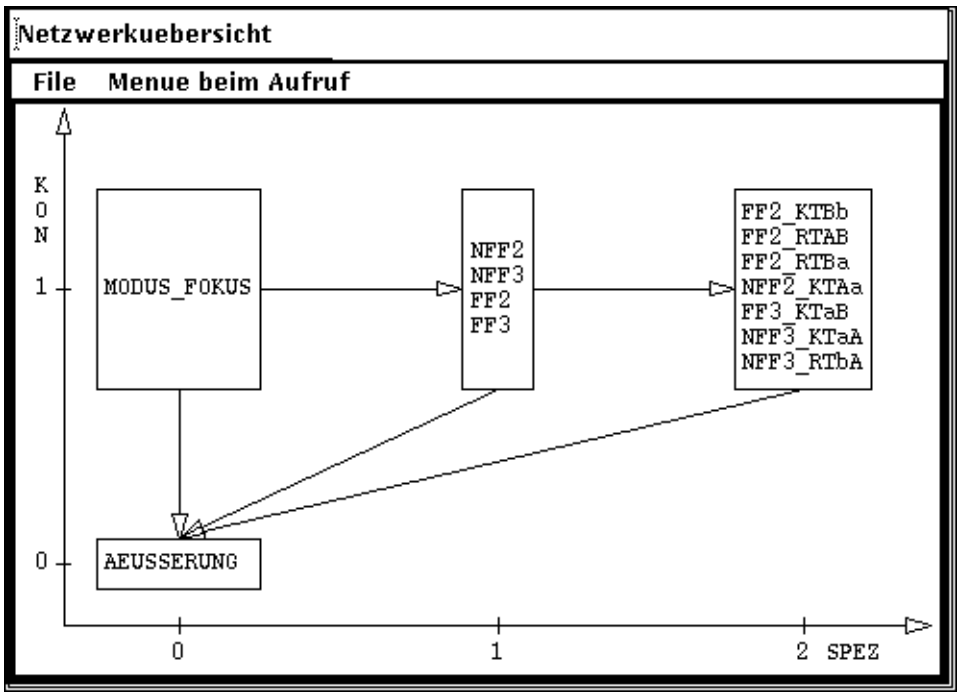


Abb. 15: Übersicht über das Netz *intonet*

3.4 Die Konzepte

Im folgenden werden die einzelnen Konzepte des Netzes vorgestellt. Dabei wird nur auf die Attribute und Relationen sowie die dazugehörigen Prozeduren eingegangen. Eine Erläuterung zu den einzelnen Bewertungsfunktionen erfolgt erst im Rahmen des Kapitels 3.5, da die gewählten Bewertungen den Analyselauf beeinflussen.

3.4.1 Das Konzept AEUSSERUNG

Das Konzept AEUSSERUNG stellt die intonatorische Markierung einer konkreten Äußerung dar. Abb. 16 gibt einen Überblick über alle Einträge des Konzepts.
Es besitzt die drei Attribute

- *gf*
- *phr*
- *max*

und den Analyseparameter

- *a_name*.

Beim Aufruf des Analyseprogramms wird der Name der zu testenden Äußerung mitangegeben. Eine problemabhängige Routine, die vor Beginn der eigentlichen Kontrolle aufgerufen wird, trägt den Namen zeichenweise in den Analyseparameter *a_name* ein.

Das Attribut *gf* beschreibt die intonatorische Markierung der aktuellen Äußerung in Form der Grundfrequenzkontur. Die Grundfrequenzwerte stehen in einer verketteten Liste. Das Struktogramm in Abb. 17 zeigt den Ablauf der Funktion *ber_gf*, die während der Analyse zur Berechnung des Attributs *gf* aufgerufen wird. Als Argument besitzt die Funktion den Analyseparameter *a_name*, der den Äußerungsnamen vorgibt. Damit kann der Name der Datei gebildet werden, aus der die Grundfrequenzwerte eingelesen werden. Diese Grundfrequenzwerte wurden automatisch aus dem Sprachsignal

```
BEGINNE_KONZEPTDEFINITION( AEUSSERUNG , 0000 ,  )

   PRIORITAET( 0 0 0 0 0 )

   START_ATTRIBUTE( 3 )
      ATTRIBUT( gf )
         DEFINITIONSBEREICH( RECORD, gf_list )
         DIMENSION( 1 , 1 , 1 , 1 )
         WERTEBERECHNUNG( ber_gf )
            ARGUMENTE( a_name )
      ATTRIBUT( phr )
         DEFINITIONSBEREICH( INTEGER )
         DIMENSION( 2 , 20 , 1 , 1 )
         WERTEBERECHNUNG( ber_phr )
            ARGUMENTE( a_name )
      ATTRIBUT( max )
         DEFINITIONSBEREICH( INTEGER )
         DIMENSION( 1 , 10 , 1 , 1 )
         WERTEBERECHNUNG( ber_max )
            ARGUMENTE( gf, phr )
   ENDE_ATTRIBUTE

   KONKRETISIERUNG_VON( MODUS_FOKUS )

   START_ANALYSEPARAMETER( 1 )
      ATTRIBUT( a_name )
         DEFINITIONSBEREICH( CHARACTER )
         DIMENSION( 6 , 6 , 1 , 1 )
         WERTEBERECHNUNG( ber_name )
   ENDE_ANALYSEPARAMETER

   BEWERTUNG( bewertung )

BEENDE_KONZEPTDEFINITION( AEUSSERUNG )
```

Abb. 16: Das Konzept AEUSSERUNG

extrahiert. An stimmhaften Bereichen wurde jeweils für einen äquidistanten Zeitbereich (Frame) ein Grundfrequenzwert bestimmt, an stimmlosen Bereichen wurde der Wert mit null vorbesetzt. Im Falle der prosodischen Verifikation von Satzhypothesen könnte anstelle des Einlesens eine Funktion aufgerufen werden, die die Grundfrequenzkontur der aktuellen Äußerung bestimmt.

Dateinamen erzeugen
Grundfrequenzwerte von Datei einlesen
Mittelwertfrei machen
Linear interpolieren an stimmlosen Bereichen
Liste von Werten als Attributwert eintragen

Abb. 17: Ablauf der Attributberechnungsfunktion *ber_gf*

Um die Äußerung von der Tonhöhe her sprecherunabhängig zu machen, wird zunächst der Mittelwert der Grundfrequenzwerte gebildet und dieser Wert subtrahiert. Da eine Äußerung aus stimmlosen und stimmhaften Bereichen besteht, an den stimmlosen Bereichen jedoch keine Grundfrequenzwerte bestimmt werden können, muß an diesen Stellen linear interpoliert werden (siehe Abb. 18).

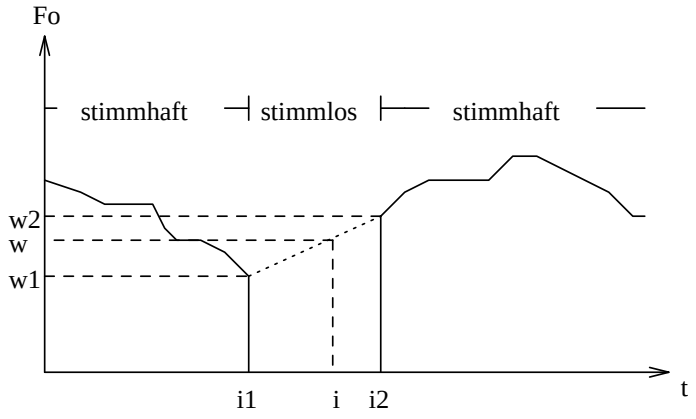


Abb. 18: Lineare Interpolation an stimmlosen Bereichen

Innerhalb eines stimmlosen Bereichs berechnet sich der neue Grundfrequenzwert w an der Position i folgendermaßen:

$$w = \frac{i - i1}{i2 - i1}(w2 - w1) + w1,$$

wobei $i1$ die Position des letzten Grundfrequenzwertes ($w1$) des vorangegangenen stimmhaften Bereichs und $i2$ die Position des ersten Grundfrequenzwertes ($w2$) des nächsten stimmhaften Bereichs ist.

Das Attribut *phr* beschreibt die Phrasengrenzen der aktuellen Äußerung. Es besteht aus einem Integer–Feld mit einer geradzahligen Anzahl von Elementen. Zwei benachbarte Zahlen beschreiben immer Anfangs- und Endposition einer Phrase, wobei die Positionsangabe in Form von Frames vorliegt. Es gilt also:

$$\text{An der Postion } n \text{ steht der } \begin{cases} \text{Anfangsframe der } \frac{n+1}{2} \text{ Phrase,} & \text{wenn } n \text{ ungerade.} \\ \text{Endframe der } \frac{n}{2} \text{ Phrase,} & \text{wenn } n \text{ gerade.} \end{cases}$$

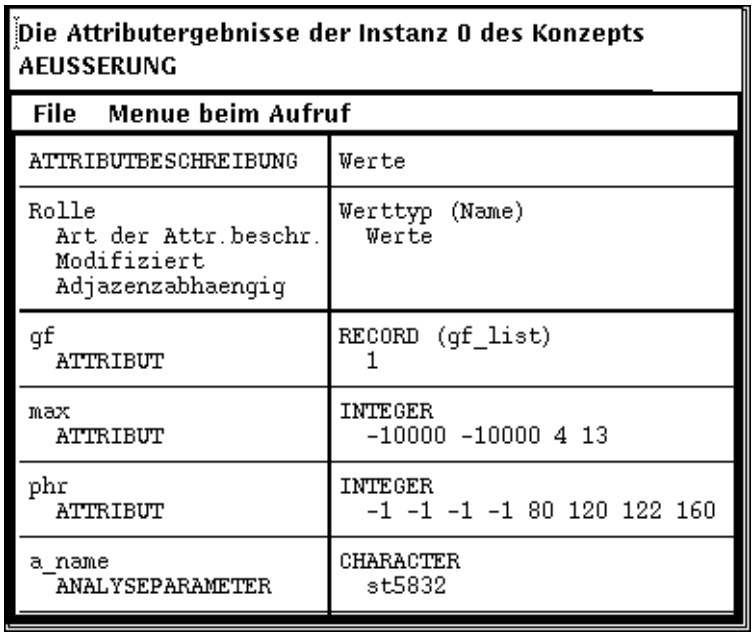
Eine Äußerung, die nur aus einer Phrase besteht, besitzt als Attribut somit nur zwei Einträge, nämlich Anfangs- und Endframe der gesamten Äußerung. Als maximale Anzahl möglicher Phrasen wurde zehn gewählt. Die Phrasengrenzen werden von einer Datei eingelesen, deren Dateiname wie beim Attribut *gf* aus dem Äußerungsnamen *a_name* gebildet wird. Diese Datei enthält entweder per Hand extrahierte Phrasengrenzen [Batl89a] oder automatisch berechnete Grenzen, zum Beispiel mit dem in [Metz91] vorgestellten Verfahren. Für die im Rahmen dieser Arbeit durchgeführten Untersuchungen interessierten nur die vorletzte und letzte Phrase (siehe Kapitel 2). Die Grenzen der übrigen Phrasen wurden mit -1 vorbesetzt.

Das Attribut *max* steht für den maximalen Grundfrequenzwert jeder Phrase, der bei Aufruf der Funktion *ber_max* berechnet wird. Als Parameter dienen die Attribute *gf* und *phr*. Die Gründe für die Einführung des Attributs *max* werden in Kapitel 4.3.4 erläutert. Ebenso wie beim Attribut *phr* wird für diejenigen Phrasen, die bei der weiteren Untersuchung nicht interessieren, ein Dummy-Wert eingetragen.

Abb. 19 zeigt die während der Analyse berechneten Werte der Attribute und des Analyseparameters, wenn der Kontrollalgorithmus für die Realisierung mit dem Namen *st5832* gestartet wurde.

3.4.2 Die Konzepte MODUS_FOKUS, NFF2, NFF3, FF2, FF3

Das Konzept MODUS_FOKUS (siehe Abb. 20) modelliert allgemein jede mögliche Modus–Fokus–Konstellation. Zum gegenwärtigen Zeitpunkt besitzt das Konzept vier Spezialisierungskanten, die zu den vier die verschiedenen Konstellationen beschreibenden Konzepte führen. Eine Erweiterung des intonatorischen Modells um weitere Modus–Fokus–Konstellationen kann durch die Einführung neuer



Die Attributergebnisse der Instanz 0 des Konzepts AEUSSERUNG

File Menue beim Aufruf

ATTRIBUTBESCHREIBUNG	Werte
Rolle Art der Attr.beschr. Modifiziert Adjazenzabhaengig	Werttyp (Name) Werte
gf ATTRIBUT	RECORD (gf_list) 1
max ATTRIBUT	INTEGER -10000 -10000 4 13
phr ATTRIBUT	INTEGER -1 -1 -1 -1 80 120 122 160
a_name ANALYSEPARAMETER	CHARACTER st5832

Abb. 19: Attributwerte der Instanz, die zum Konzept AEUSSERUNG erzeugt wurde.

Spezialisierungskanten berücksichtigt werden. Außerdem enthält das Konzept MODUS_FOKUS eine Konkretisierungskante (*satzhyp*) zum Konzept AEUSSERUNG. Die Konkretisierungskante wird vom Konzept MODUS_FOKUS an alle spezielleren Konzepte, das heißt auch an die Prototypenkonzepte, weitervererbt.

```
BEGINNE_KONZEPTDEFINITION( MODUS_FOKUS , 1000 ,  )

   PRIORITAET( 1 1 1 2 0 )

   SPEZIALISIERUNG( NFF2, NFF3, FF2, FF3 )
   START_KONKRETISIERUNGEN( 1 )
      KANTE( satzhyp )
         DEFINITIONSBEREICH( AEUSSERUNG )
         DIMENSION( 1 , 1 )
   ENDE_KONKRETISIERUNGEN

   BEWERTUNG( bewertung )

BEENDE_KONZEPTDEFINITION( MODUS_FOKUS )
```

Abb. 20: Das Konzept MODUS_FOKUS

Die strukturell identisch aufgebauten Konzepte FF2, FF3, NFF2 und NFF3 besitzen zwei verschiedene Kanteneinträge. Zum einen die Spezialisierungskanten, die zu den entsprechenden Prototypenkonzepten führen, zum anderen den Eintrag, daß sie über eine Spezialisierungskante von MODUS_FOKUS aus erreichbar sind. In Abb. 21 ist das Konzept NFF2 beispielhaft für alle vier Konzepte zu sehen.

3.4.3 Die Konzepte der Prototypen

Alle Prototypenkonzepte besitzen die Attribute

- *ref_gf*

```
BEGINNE_KONZEPTDEFINITION( NFF2 , 1100 ,  )

   PRIORITAET( 1 1 1 1 0 )

   SPEZIALISIERUNG_VON( MODUS_FOKUS )

   SPEZIALISIERUNG( NFF2_KTAa )
   BEWERTUNG( bewertung )

BEENDE_KONZEPTDEFINITION( NFF2 )
```

Abb. 21: Das Konzept NFF2

- *ref_phr*

den Analyseparameter

- *p_name*

sowie die Relation

- *dp_phr*

Die beiden Attribute entsprechen den Attributen *gf* und *phr* im Konzept AEUSSERUNG. *ref_gf* beschreibt die Grundfrequenzkontur der entsprechenden prototypischen Äußerung, *ref_phr* die Phrasengrenzen. Ursprünglich war zur Gewinnung dieser Attribute der Einsatz eines Synthesegeräts vorgesehen. Da dies, wie bereits erwähnt, noch nicht möglich ist, werden stattdessen die Grundfrequenzwerte und Phrasengrenzen derjenigen Aufnahme aus dem Fokus–Korpus von Datei eingelesen, die den entsprechenden Prototyp repräsentiert. Der Name der Realisierung wurde bei Erstellung des Netzes im Analyseparameter *p_name* abgelegt. Zur Berechnung der Attribute und des Analyseparameters können die gleichen Funktionen verwendet werden, die auch an das Konzept AEUSSERUNG gebunden sind.

Die Relation *dp_phr* beschreibt die Ähnlichkeit der Referenzkontur mit der Testkontur. Als Parameter für die Relationsbewertungsfunktion *bew_dp_phr* dienen die Attribute *ref_gf* und *ref_phr* im Konzept des Prototyps sowie die Attribute *gf* und *phr* im Konzept AEUSSERUNG, auf die über den Kantennamen der von MODUS_FOKUS geerbten Konkretisierungskante zugegriffen werden kann. Die Konturen jeder Phrase, die von Interesse ist — also deren Grenzen nicht mit -1 vorbesetzt wurden — werden mit Hilfe dynamischer Zeitverzerrung verglichen. Für jede dieser Phrasen ergibt sich ein Abstandsmaß d_abst, das aufsummiert wird. Der negative Gesamtabstand stellt die Bewertung der Relation *dp_phr* dar. Die Bewertung wird also umso besser, das heißt umso größer, je kleiner der berechnete Gesamtabstand ist. Der Ablauf der Funktion *bew_dp_phr* ist in Abb. 22 dargestellt. Auf die Vorgehensweise bei der dynamischen Zeitverzerrung wird erst in Kapitel 4.2 näher eingegangen, da bei der Überprüfung des Modells auch dazu unterschiedliche Ansätze getestet wurden.

g_abst = 0		
FOR jede Phrase		
	IF	Phrasengrenzen ungleich -1
	THEN	Bestimme Abstandsmaß d_abst mit Hilfe dynamischer Programmierung
		g_abst = d_abst + g_abst
dp_bew = -g_abst		

Abb. 22: Ablauf der Relationsbewertungsfunktion *bew_dp_phr*

Abb. 23 zeigt das Konzpet FF2_KTBb beispielhaft für alle anderen Prototypenkonzepte. Die Konzepte NFF2_KTAa und NFF3_KTaA besitzen noch die zusätzliche Relation *maxphr*. Diese Relation war, ebenso wie das Attribut *max* im Konzept AEUSSERUNG, nicht ursprünglich geplant, sondern wurde aufgrund der Ergebnisse der in Kapitel 4 beschriebenen Testläufe eingeführt. Eine genaue Beschreibung des Attributs *max* und der Relation *maxphr* befindet sich deshalb erst in Kapitel 4.3.4.

```
BEGINNE_KONZEPTDEFINITION( FF2_KTBb , 1200 ,  )

   PRIORITAET( 1 1 1 0 0 )

   SPEZIALISIERUNG_VON( FF2 )

   START_ATTRIBUTE( 2 )
      ATTRIBUT( ref_gf )
         DEFINITIONSBEREICH( RECORD, gf_list )
         DIMENSION( 1 , 1 , 1 , 1 )
         WERTEBERECHNUNG( ber_gf )
            ARGUMENTE( p_name )
      ATTRIBUT( ref_phr )
         DEFINITIONSBEREICH( INTEGER )
         DIMENSION( 2 , 20 , 1 , 1 )
         WERTEBERECHNUNG( ber_phr )
            ARGUMENTE( p_name )
   ENDE_ATTRIBUTE

   START_STRUKTURRELATIONEN( 1 )
      RELATION( dp_phr )
         BEWERTUNG( bew_dp_phr )
            ARGUMENTE( ref_gf, ref_phr, satzhyp.gf, satzhyp.phr )
   ENDE_STRUKTURRELATIONEN

   START_ANALYSEPARAMETER( 1 )
      ATTRIBUT( p_name )
         DEFINITIONSBEREICH( CHARACTER )
         DIMENSION( 6 , 6 , 1 , 1 )
         WERTEBERECHNUNG( ber_name )
         PRAEFERENZ( NIL )
            WERT( p, a, 5, 4, 1, 8 )
   ENDE_ANALYSEPARAMETER

   BEWERTUNG( bew_FK )
      ARGUMENTE( dp_phr )

BEENDE_KONZEPTDEFINITION( FF2_KTBb )
```

Abb. 23: Das Konzept FF2_KTBb

3.5 Anmerkungen zum Analyseablauf

Das Ziel der Analyse im Netz *intonet* besteht darin, bei Angabe einer Testäußerung **alle** Prototypenkonzepte zu instantiieren, die als Zielkonzepte angegeben wurden bzw. die über **Spezialisierungskanten** erreichbar sind. Danach wird diejenige Modus–Fokus–Konstellation, die durch den bestbewerteten Prototyp vorgegeben ist, der aktuellen Testäußerung zugeordnet. Im vorliegenden Kapitel wird nun auf den für das Intonationsnetz typischen Ablauf der Analyse eingegangen, indem folgende zwei Punkte näher erläutert werden:

- Instantiierung von **spezielleren** Konzepten
- Instantiierung **aller** Zielkonzepte und deren Spezialisierungen

Diese müssen bei der Erstellung des Analyseprogramms berücksichtigt werden und rufen im ersten Fall eine Modifikation der Abbruchbedingung der Analyse und im zweiten Fall die passenden Konzeptbewertungsfunktionen hervor.

Instantiierung von spezielleren Konzepten

Beim Start der Analyse müssen ein oder mehrere Zielkonzepte angegeben werden. Abb. 24 zeigt beispielhaft den während der Analyse erzeugten Suchbaum, wenn als Zielkonzept NFF2_KTAa vorgegeben

wurde. Der Suchbaum besteht in diesem Falle nur aus einem einzigen Ast.

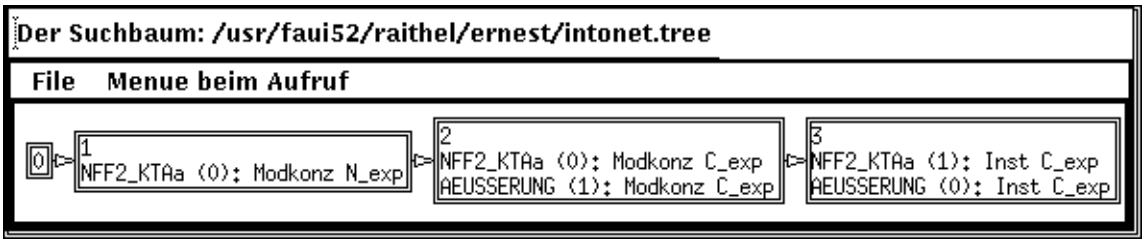


Abb. 24: Während der Analyse erzeugter Suchbaum, wenn als Zielkonzept NFF2_KTAa angegeben wurde.

Zunächst wird ein modifiziertes Konzept von NFF2_KTAa erzeugt. Es wird expandiert, das heißt, da NFF2_KTAa noch nicht instantiiert werden kann, muß das Konzept AEUSSERUNG weiter untersucht werden. Dieses wird modifiziert. Danach folgt die Instantiierungsphase, denn AEUSSERUNG kann sofort instantiiert werden, das heißt, die Attribute *gf* und *phr* werden berechnet. Damit ist auch die Instantiierung von NFF2_KTAa möglich. Die Analyse bricht ab, da zum Zielkonzept eine Instanz generiert werden konnte.

In dem hier vorliegenden Intonationsmodell scheint es jedoch sinnvoller, anstelle der einzelnen Prototypen nur die verschiedenen Modus-Fokus-Konstellationen zu unterscheiden und diese als Zielkonzepte anzugeben. Durch welchen Prototyp die entsprechende Konstellation repräsentiert wird, interessiert weniger. Abb. 25 zeigt den Suchbaum beim Aufruf mit Zielkonzept NFF2.

Abb. 25: Während der Analyse erzeugter Suchbaum, wenn als Zielkonzepte NFF2 angegeben wurde.

Man erkennt, daß keine Instantiierung des spezielleren Prototypenkonzepts NFF2_KTAa erfolgt, was jedoch nötig wäre, um eine sinnvolle Entscheidung zu treffen. In ERNEST ist eine Expansion über Spezialisierungskanten nicht vorgesehen, kann jedoch durch eine Modifikation der Abbruchbedingung des Kontrollalgorithmus erreicht werden (siehe [Kumm90] Kapitel 4.7). Dazu muß eine anwenderdefinierbare Funktion abgeändert werden. Abb. 26 zeigt den Ablauf der modifizierten Abbruchfunktion. Dies hat zur Folge, daß wenn ein Konzept instantiiert wurde, und das Konzept noch weitere Spezialisierungen besitzt, auch diese instantiiert werden. Nach der Instantiierung von NFF2 stellt der Analysealgorithmus also fest, daß zu diesem Konzept noch speziellere existieren und versucht daraufhin, diese zu instantiieren. Abb. 27 zeigt den Suchbaum nach der Modifikation.

Aufgrund der Modifikation der Abbruchbedingung genügt es, als Zielkonzept MODUS_FOKUS anzugeben, um eine vollständige Instantiierung aller Prototypenkonzepte zu erreichen.

Instantiierung aller Zielkonzepte und deren Spezialisierungen

Während der Analyse sollen **alle** Prototypenkonzepte instantiiert werden, die als Zielkonzepte angegeben wurden oder über Spezialisierungskanten von den Zielkonzepten erreichbar sind. Nur so kann die Testäußerung nach Ablauf der Analyse dem besten Prototyp zugeordnet werden. Da als Bewertung des Suchbaumknotens die Bewertung seines Zieleintrags übernommen wird und die Analyse abbricht, sobald der bestbewertete Suchbaumknoten als Zieleintrag die Instanz eines Prototypenkonzepts besitzt, müssen die Bewertungen so vorgenommen werden, daß modifizierte Prototypenkonzepte immer eine bessere Bewertung besitzen als Instanzen von Prototypenkonzepten. Durch diese optimistische

IF	Zieleintrag des bestbewerteten Suchbaumknotens ist eine Instanz eines Zielkonzeptes	
THEN	IF	Zieleintrag besitzt keine Spezialisierungskanten
	THEN	Abbruchbedingung ist erfüllt
	ELSE	Abbruchbedingung ist noch nicht erfüllt
ELSE	Abbruchbedingung ist noch nicht erfüllt	

Abb. 26: Ablauf der anwenderabhängigen Abbruchfunktion der Analyse

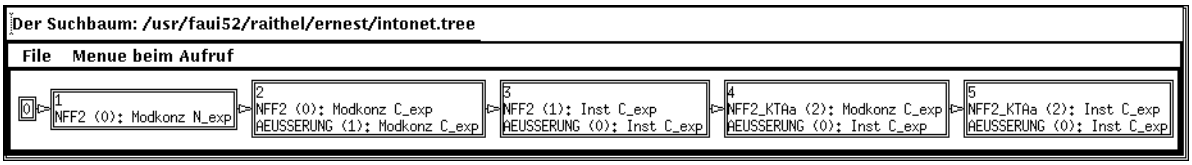


Abb. 27: Suchbaum mit Zielkonzept NFF2 nach Modifikation

Bewertungsweise wird eine Instantiierung aller Prototypenkonzepte erreicht. Der bestbewertete Suchbaumknoten, der zur weiteren Analyse ausgewählt wird, besitzt nur dann als Zieleintrag die Instanz eines Prototyps, wenn bereits alle Prototypenkonzepte instantiiert wurden.

Die Bewertung in ERNEST ist durch einen Vektor $(f_1, ..., f_5, i_1, ...i_5)$, bestehend aus fünf reellen und fünf ganzen Zahlen, vorgegeben. Für das semantische Netz, das Wissen über Syntax, Semantik und Pragmatik in EVAR modelliert, existiert ein festgelegtes Bewertungskalkül [Sage90]. Eine Bewertungskomponente des Vektors stellt zum Beispiel das binäre Maß *Zulässigkeit* dar. Besitzt diese Komponente den Wert eins, so erfüllt eine Hypothese alle linguistischen Restriktionen und widerspricht nicht bestimmten Gesetzmäßigkeiten. Eine weitere Komponente stellt die *Qualität* dar, die ein Maß dafür gibt, wie gut eine Hypothese zum Signal paßt.

Für die Bewertung der Prototypenkonzepte im Netz *intonet* sind zwei Komponenten von Interesse, deren Bedeutungen sich an das erwähnte Bewertungskalkül anlehnen:

- Die erste Komponente ist ein binäres Maß und sagt aus, ob die Grundfrequenzkontur der Testäußerung bestimmten Gesetzmäßigkeiten, die vom entsprechenden Prototyp gefordert werden, erfüllt. Bisher wurden für das Intonationsnetz noch keine Restriktionen festgelegt, die von einer Testäußerung verletzt werden könnten. Die Komponente besitzt somit für modifizierte Konzepte und Instanzen den Wert eins. Erst im Kapitel 4.3.4 wird eine Restriktion festgelegt, in deren Zusammenhang auch das Attribut *max* im Konzept AEUSSERUNG und die Relation *maxphr* in den Konzepten NFF2_KTAa und NFF3_KTaA erläutert werden.

- Die zweite Komponente des Bewertungsvektors erhält für die modifizierten Konzepte der Prototypen den Wert eins, das heißt, noch passen Test- und Referenzkontur optimal zusammen. Bei der Instantiierung wird die Bewertung der Relation *dp_phr* eingetragen, die ein Ähnlichkeitsmaß zwischen Test- und Referenzkontur im Hinblick auf die zweite und dritte Phrase darstellt. Sie spiegelt somit die *Qualität* wider, wie gut die Testäußerung zu der durch den Prototyp vorgegebenen Konstellation paßt. Bei Konzepten, die Randprototypen repräsentieren, ist eine Gewichtung dieses Eintrags möglich. Dadurch kann eine Bevorzugung der Kernprototypen festgelegt und somit die Tatsache berücksichtigt werden, daß die Kernprototypen die 'normalen' Fälle einer Konstellation darstellen (siehe Kapitel 2.4). Bei der Überprüfung des Modells (siehe Kapitel 4) konnte basierend auf dem Koordinatenabstieg der Faktor 1,5 als geeigneter Gewichtungsfaktor bestimmt werden. Dieser Faktor kann jedoch beim Start des Analyseprogramms variiert werden. Da die Relationsbewertung einen negativen Wert besitzt, wird die Instanzbewertung gegenüber der Bewertung der modifizierten Konzepte schlechter. Dadurch wird die Forderung erfüllt, die die Instantiierung aller Prototypenkonzepte bewirkt.

Für alle anderen Konzepte außer den Prototypenkonzepten ist die Bewertung von modifiziertem Konzept und Instanz gleich. Den beiden Bewertungskomponenten wird jeweils der Wert eins zugewiesen.

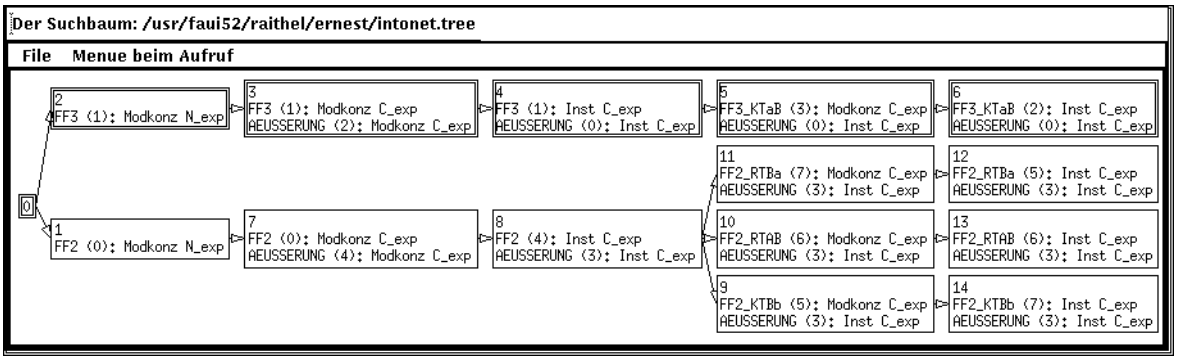


Abb. 28: Während der Analyse erzeugter Suchbaum, wenn als Zielkonzepte FF2 und FF3 angegeben wurden.

Abb. 28 zeigt einen Suchbaum, der während der Analyse erstellt wurde. Die angegebenen Zielkonzepte waren FF2 und FF3. Die doppelt eingerahmten Kästen kennzeichnen den optimalen Pfad durch den Suchbaum. FF3_KTaB stellt somit denjenigen Prototyp dar, der der Testäußerung am ähnlichsten ist. Man erkennt, daß **alle spezielleren** Konzepte der Zielangaben instantiiert wurden. In der Prozedur, die direkt nach Ablauf der Analyse aufgerufen wird, wird diejenige Instanz mit der besten Bewertung ausgewählt.

Für jede Äußerung aus dem Fokus–Korpus, die beim Start des Analyse angegeben wird, findet der gleiche Analyseablauf statt. Es werden die gleiche Anzahl von Suchbaumknoten, modifizierten Konzepten sowie Instanzen erzeugt.
Tabelle 5 gibt einen Überblick über die Anzahl der erzeugten Knoten in Abhängigkeit von den gewählten Zielkonzepten. #N steht für die Anzahl erzeugter Suchbaumknoten, #Q für die Anzahl modifizierter Konzepte und #I für die Anzahl erzeugter Instanzen. Die in der Tabelle angegebenen CPU–Zeiten stellen durchschnittliche Werte dar, die aus den Laufzeiten für 50 Äußerungen ermittelt wurden. Die Ermittlung der CPU–Zeiten erfolgte auf einer VAX–Station 3500 der Firma *Digital Equipment* unter normaler Auslastung.
Die geringere Laufzeit, die bei Angabe der Zielkonzepte NFF2 und NFF3 erzielt wurde, ist auf die geringere Anzahl von Prototypen zurückzuführen, die diese im Vergleich zu FF2 und FF3 besitzen.
Im Hinblick auf die verschiedenen Äußerungen im Fokus–Korpus konnten jedoch kaum Laufzeitdifferenzen festgestellt werden. Dies war auch nicht zu erwarten, da alle Realisierungen auf den gleichen drei segmental ähnlichen Sätzen basieren und nur geringe Differenzen in der Länge der ganzen Äußerungen bzw. der einzelnen Phrasen entstanden.
Die Laufzeiten wurden für das in Kapitel 4.2 ausgewählte Verfahren zur dynamischen Zeitverzerrung ermittelt.

Zielkonzepte	#N	#Q	#I	CPU–Zeit
MODUS_FOKUS	25	12	12	2.30 s
FF2 FF3	14	7	7	1.31 s
NFF2 NFF3	12	6	6	1.16 s

Tabelle 5: Anzahl während der Analyse erzeugter Suchbaumknoten (#N), modifizierter Konzepte (#Q) und Instanzen (#I) sowie benötigte CPU–Zeit in Abhängigkeit unterschiedlicher Zielkonzepte.

4 Überprüfung des Intonationsmodells

Die Gültigkeit des implementierten Intonationsmodells soll am zugrundeliegenden Fokus–Korpus belegt werden. Dazu wird für jede Realisierung aus dem Korpus das in ERNEST erzeugte Analyseprogramm gestartet, wobei als Zielkonzept MODUS_FOKUS angegeben wird. Die Konstellation des

bestbewerteten Prototypen wird mit der tatsächlichen Konstellation der Testäußerung verglichen und die daraus für das ganze Korpus resultierende Erkennungsrate beurteilt. Zum Vergleich werden die Grundfrequenzkonturen der Äußerungen herangezogen und mittels dynamischer Zeitverzerrung ein Ähnlichkeitsmaß berechnet. Im folgenden wird zunächst auf die zur Verfügung stehenden Testdaten eingegangen. Anschließend werden verschiedene Ansätze zur dynamischen Zeitverzerrung beschrieben, die im Rahmen der Arbeit getestet wurden. Danach folgt eine Erläuterung der Ergebnisse.

4.1 Grundfrequenzwerte und Phrasengrenzen

Um das Analyseprogramm starten zu können, müssen für jede Äußerung aus dem Fokus–Korpus zwei Dateien vorhanden sein, die folgenden Inhalt haben:

- Grundfrequenzwerte
- Phrasengrenzen

Die hier zur Verfügung stehenden Grundfrequenzwerte wurden automatisch bestimmt. Die prinzipielle Vorgehensweise zur Berechnung von Grundfrequenzwerten besteht zunächst aus der Extraktion stimmhaft artikulierter Signalbereiche, da nur an diesen Stellen eine Grundfrequenzbestimmung sinnvoll ist. Für die stimmhaft/stimmlos (SH/SL)–Entscheidung existieren verschiedene Ansätze, wie das sogenannte Sonorantenverfahren [Nöth91] oder die in [Kies89] beschriebene Vorgehensweise. Danach wird für jeden Frame innerhalb dieser stimmhaften Bereiche ein Grundfrequenzwert bestimmt [Komp89, Nöth91].

Name	Verfahren
$TD_{K,m}$	SH/SL-Entscheidung nach [Kies89] Grundfrequenzverfahren nach [Komp89] Framelänge: 12,5 ms Phrasengrenzen: per Hand
$TD_{Kg,m}$	SH/SL-Entscheidung nach [Kies89] Grundfrequenzverfahren nach [Komp89] Glättung mit Medianfilter Framelänge: 12,5 ms Phrasengrenzen: per Hand
$TD_{S,m}$	SH/SL-Entscheidung: Sonorantenverfahren Grundfrequenzverfahren nach [Komp89] Framelänge: 12,5 ms Phrasengrenzen: per Hand
$TD_{K,a}$	SH/SL-Entscheidung nach [Kies89] Grundfrequenzverfahren nach [Komp89] Framelänge: 10 ms Phrasengrenzen: automatisch

Abb. 29: Überblick über die Variationen in bezug auf die Grundfrequenzwerte und Phrasengrenzen

Insgesamt stehen zum Start des Analyseprogramms vier verschiedene Typen von Testdaten zur Verfügung, die alle auf dem in [Komp89] vorgestellten Grundfrequenzverfahren basieren. Abb. 29 gibt einen Überblick über die unterschiedlichen Variationen im Hinblick auf die Grundfrequenzwerte und Phrasengrenzen. Der Index i im Namen der unterschiedlichen Testdaten $TD_{i,j}$ bezieht sich auf das verwendete Verfahren zur SH/SL–Entscheidung. K steht dabei für das SH/SL–Verfahren nach [Kies89], S für das Sonorantenverfahren. Der Index j kennzeichnet die Art der Gewinnung der Phrasengrenzen, wobei m manuell und a automatisch bedeutet.
Die Untersuchungsdaten $TD_{Kg,m}$ unterscheiden sich von $TD_{K,m}$ nur durch eine zusätzliche Glättung der Grundfrequenzkontur mit Hilfe eines Medianfilters. Die Grundfrequenzwerte für die Typen $TD_{K,m}$, $TD_{Kg,m}$ und $TD_{S,m}$ wurden jeweils für Frames der Länge 12,5 ms bestimmt, während für $TD_{K,a}$ als Framelänge 10 ms gewählt wurde.

Die Phrasengrenzen, basierend auf der Framelänge 12,5 ms, wurden im Laufe der DFG–Untersuchungen per Hand extrahiert (siehe Kapitel 2.2). Für die Testdaten mit verwendeter Framelänge von 10 ms konnten automatisch berechnete Phrasengrenzen herangezogen werden. Die automatische Extraktion erfolgte mit einem in [Metz91] erläuterten Verfahren und basiert auf der Lautklassifikation von [Wöhr90].

Außerdem werden anstelle der Grundfrequenzwerte für jeden der vier Varianten von Testdaten auch die entsprechenden Halbtonwerte als Untersuchungsgrundlage verwendet. Vor dem Start des Analyseprogramms muß dann statt einer Grundfrequenzdatei die Datei mit den transformierten Halbtonwerten zur Verfügung stehen.

4.2 Dynamische Zeitverzerrung

Um eine Testäußerung einem der Prototypen zuordnen zu können, muß erst ein Maß bestimmt werden, das aufzeigt, wie gut diese Äußerung zu den Prototypen paßt. Zur Berechnung dieses "Ähnlichkeitsmaßes" werden im folgenden nur Tonhöhenmerkmale miteinbezogen, die vielfach als die bedeutendsten prosodischen Eigenschaften angesehen werden und die ihr akustisches Korrelat in der Grundfrequenz finden. Eine zu testende Ansatzmöglichkeit stellt die dynamische Zeitverzerrung (Dynamic Time Warping) dar, die als Maß einen Abstand zwischen nichtlinear verzerrtem Testmuster und Referenzkontur verwendet. Eine nichtlineare Verzerrung ist sinnvoll, da beispielsweise Schwankungen in der Sprechgeschwindigkeit berücksichtigt werden.
Die Verzerrung und Abstandsberechnung erfolgt effizient mit Hilfe dynamischer Programmierung.

Im folgenden wird auf das Prinzip der dynamischen Zeitverzerrung eingegangen:
Eine Verzerrungsfunktion w ordnet jedem Punkt der Referenzkontur $(a_1, ..., a_I)$ einen oder mehrere Punkte des Testmusters $(b_1, ..., b_J)$ zu, wodurch ein Pfad durch die (i, j)–Ebene definiert wird (siehe Abb. 30).

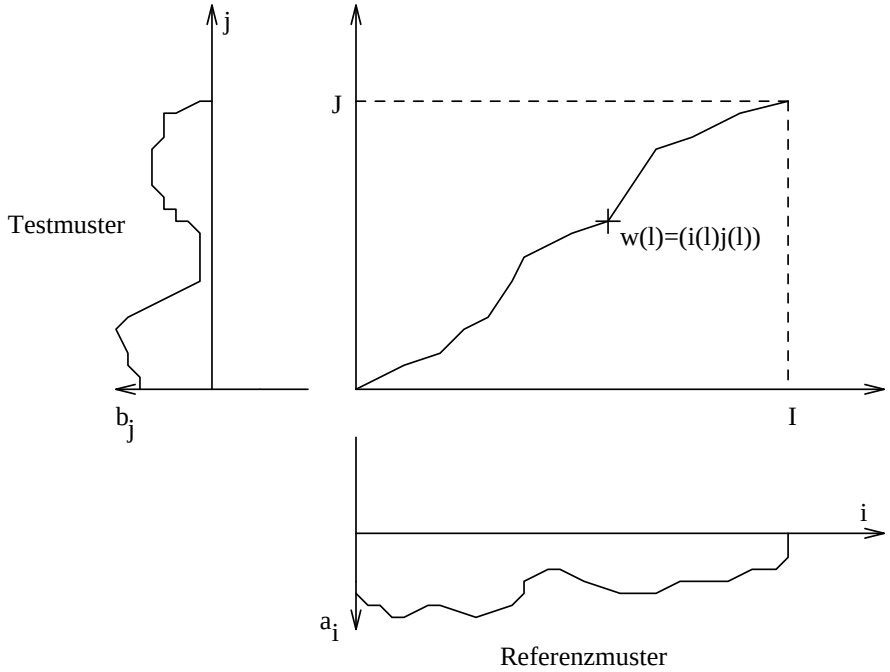


Abb. 30: Verzerrungspfad durch die (i, j)–Ebene

Die Verzerrungsfunktion w ist also definiert durch eine Folge von Punkten

$$\begin{aligned} w &= w(1), \ldots, w(l), \ldots, w(L) \\ &\text{mit} \qquad w(l) = (i(l), j(l)) \end{aligned}$$

Der Abstand zweier Muster bzgl. eines Pfades ist gegeben durch

$$\begin{aligned} D(w) &= \sum_{l=1}^{L} d(w(l)) \\ &= \sum_{l=1}^{L} d(i(l), j(l)), \end{aligned}$$

wobei das lokale Abstandsmaß d verschieden gewählt werden kann, zum Beispiel

$$\begin{aligned} d(i,j) &= |a_i - b_j| \\ \text{oder} & \\ d(i,j) &= (a_i - b_j)^2. \end{aligned}$$

Bei der dynamischen Zeitverzerrung wird die Verzerrungsfunktion Gegenstand einer Optimierung, das heißt, gesucht ist diejenige Funktion, die den minimalen Abstand von Referenzmuster zu verzerrtem Testmuster liefert.

$$D^* = \min_w D$$

Die Bestimmung der optimalen Verzerrungsfunktion kann mit Hilfe dynamischer Programmierung erfolgen, sofern die Kosten des Pfades monoton sind. Dies ist durch die oben angegebenen lokalen Abstandsmaße gewährleistet.
Die Zeitverzerrungsfunktion kann bestimmten Bedingungen unterworfen werden, welche "unsinnige" Pfade ausschließen und den Aufwand reduzieren:

- **Randbedingung**

  $$\begin{aligned} w(1) &= (1,1) \\ w(L) &= (I,J) \end{aligned}$$

  Dadurch wird sichergestellt, daß die Anfangs- und Endpunkte der zu vergleichenden Muster aufeinander abgebildet und nicht nur Teilsequenzen der Kontur zur Abstandsberechnung verwendet werden.

- **Monotoniebedingung**

  $$\begin{aligned} i(l-1) &\leq i(l) \\ j(l-1) &\leq j(l) \end{aligned}$$

  Hiermit erreicht man, daß der Pfad durch die (i,j)–Ebene monoton steigend erfolgt.

- **Lokale Steigungs- und Stetigkeitsbedingung**
  Durch diese Bedingung kann festgelegt werden, welche Vorgänger der Knoten $w(l)$ im Pfad besitzen darf. Mögliche Einschränkungen zeigt Abb. 31.

- **Globale Steigungs- und Stetigkeitsbedingung**
  Pfade, die zu weit von der idealen Diagonalen entfernt liegen, können ausgeschlossen werden. Abb. 32 zeigt zwei Beispiele, wobei das erlaubte Gebiet in 32(a) durch Vorgabe minimaler und maximaler Pfadsteigung entsteht. Durch die lokalen Einschränkungen in Abb. 31 (a) und (b) wird so ein rautenförmiges Gebiet bestimmt, wobei die maximale Steigung zwei beträgt.

Eine Modifikation des bis jetzt dargestellten Abstandsmaßes stellt eine Gewichts- oder eine Längennormierung dar:

- **Gewichtsnormierung**
  Bestimmte Übergänge können verschieden gewichtet werden. Der Gesamtabstand $D^*(I,J)$ wird dann mit Hilfe eines Faktors N normiert:

  $$D_N^*(I,J) = \frac{D^*(I,J)}{N}$$

  In Abb. 33 werden die diagonalen Übergänge stärker gewichtet und es gilt $N = I + J$.

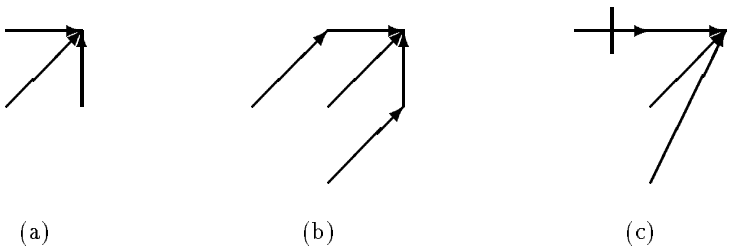


Abb. 31: Lokale Beschränkungen der Verzerrungsfunktion

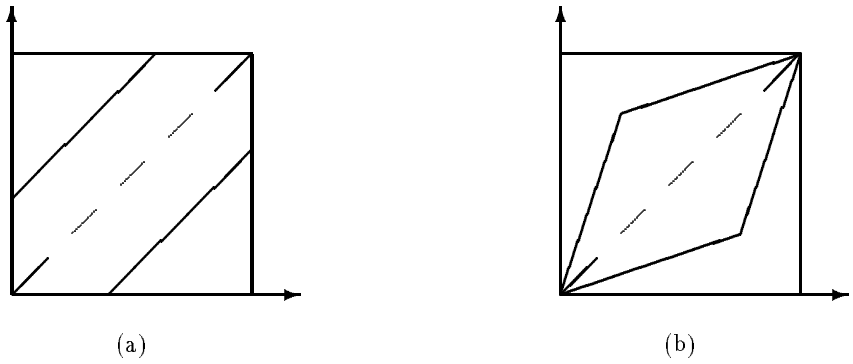


Abb. 32: Globale Pfadeinschränkungen der Verzerrungsfunktion

- **Längennormierung**
  Da längere Pfade immer benachteiligt sind, kann statt über D auch über die längennormierte Distanz D_L minimiert werden:

$$D_L(i,j) \quad = \quad \frac{D(i,j)}{L(i,j)}$$

  $L(i,j)$ ist die Anzahl der Summanden, aus denen $D(i,j)$ gebildet wurde.

Im Laufe der vorliegenden Arbeit wurden verschiedene aus der Literatur [Niem83, Myer80, Sako78, Itak75] bekannte Ansätze zur dynamischen Zeitverzerrung implementiert und an den Testdaten $TD_{K,m}$ getestet, um so für den hier notwendigen Vergleich zweier Grundfrequenzkonturen das beste Verfahren auswählen zu können. Dabei wurde sowohl eine Längen- als auch eine Gewichtsnormierung vorgenommen. Ferner wurden unterschiedliche lokale (siehe Abb. 31) sowie globale (siehe Abb. 32) Pfadeinschränkungen verwendet. Außerdem wurden Testläufe mit den zwei vorgeschlagenen lokalen Abstandsmaßen — Absolutabstand oder quadratischer Abstand — durchgeführt.
Im folgenden wird derjenige Ansatz vorgestellt, mit dem die besten Ergebnisse erzielt wurden und der in der Bewertungsfunktion der Relation *dp_phr* im semantischen Netz *intonet* (siehe Abschnitt 3.4) verwendet wird:

Die Randbedingung ($w(1) = (1,1)$ und $w(L) = (I,J)$) muß erfüllt sein, und als lokale Vorgängerbeschränkung wurde der Ansatz aus Abb. 31(a) gewählt, wodurch die Monotoniebedingung sichergestellt wird.

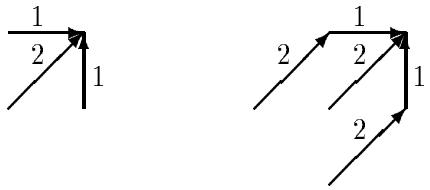


Abb. 33: Stärkere Gewichtung der diagonalen Übergänge

Die rekursive Berechnungsformel zur Berechnung des Abstands lautet somit

$$D^*(i,j) = \min \begin{cases} D^*(i,j-1) & + d(i,j) \\ D^*(i-1,j-1) & + d(i,j) \\ D^*(i-1,j) & + d(i,j) \end{cases}$$

wobei als lokales Abstandsmaß

$$d(i,j) = (a_i - b_j)^2$$

verwendet wird. Als globale Pfadeinschränkung erwies sich eine zur Diagonalen parallele Bereichseinschränkung wie in Abb. 32(a) als günstig, wobei der Bereich zu Beginn 30% des Testmusters enthält.

In der Bewertungsfunktion der Relation *dp_phr* wird für die zweite und dritte Phrase jeweils getrennt mit dem beschriebenen Implementierungsansatz ein Abstand berechnet. Der negative Wert der Summe aus beiden resultierenden Abständen wird als Bewertung der Relation übernommen. Dieser Wert wird auch zur Instanzbewertung herangezogen, wobei für Randprototypen eine stärkere Gewichtung vorgesehen ist.

4.3 Ergebnisse

In diesem Abschnitt werden die Ergebnisse erläutert, die unter Verwendung des im vorangegangenen Kapitel beschriebenen Ansatzes zur dynamischen Zeitverzerrung erzielt wurden. Außerdem wurde bei der Instanzbewertung von Randprototypen das berechnete Abstandsmaß um den Faktor 1,5 verschlechtert. Nachdem mehrere Testläufe — basierend auf dem Koordinatenabstieg — mit unterschiedlichen Faktorwerten durchgeführt wurden, erwies sich dieser als günstig.

4.3.1 Beurteilungskriterien

Das Analyseprogramm des implementierten Intonationsnetzes wurde für alle Realisierungen — außer denjenigen, welche die Protoypen darstellen — gestartet, wobei als Zielkonzept MODUS_FOKUS angegeben wurde. Diese Zielkonzeptangabe führt, wie in Kapitel 3.5 beschrieben, zur Instantiierung aller Prototypenkonzepte. Im Anschluß an jeden Analyselauf erfolgt für die aktuelle Äußerung ein Eintrag in eine Ergebnistabelle. Der Eintrag gibt an, ob die Modus–Fokus–Konstellation des bestbewerteten Prototypen mit der Konstellation der Äußerung oder ob zumindest der Satzmodus bzw. die Position des Fokus übereinstimmen.
Bei der Beurteilung der Instantiierungsergebnisse dienen der vom Kontext (siehe Abb. 1) vorgegebene Satzmodus sowie das Hörerurteil FOK als Vergleichskriterien. FOK beschreibt das Resultat aus dem Akzenttest (siehe Kapitel 2.4), bei dem die Hörer ohne Wissen des Kontextes angeben sollten, welche Phrase sie als fokussiert wahrnahmen. Dieses perzeptive Maß wurde herangezogen, da einige Kontexte die Fokuspositionen nicht eindeutig auf der zweiten oder dritten Phrase indizierten, sondern beide Positionen als mögliche Träger der Fokusinformation vorsahen. Im Gegensatz dazu ist der Satzmodus immer eindeutig festgelegt. Daher konnte der indizierte Modus als Beurteilungskriterium für die richtige Entscheidung herangezogen werden.

Besitzt der bestbewertete Prototyp also denselben Satzmodus, der auch durch die Situationsbeschreibung vorgegeben ist, so wird der Modus als richtig erkannt eingestuft. Bei der Beurteilung der Fokusentscheidung orientiert man sich an FOK, das heißt, wird die zweite Phrase erkannt und ist FOK $>$ 0, so ist die Entscheidung richtig ausgefallen. Ist FOK $\leq$ 0 und wird eine Konstellation mit Fokus auf der dritten Phrase als die beste erkannt, so ist die Entscheidung ebenfalls richtig. Es muß aber beachtet werden, daß Äußerungen, für die FOK den Wert 0 besitzt, eigentlich keine falsche Fokusposition zugewiesen werden kann.
Nachdem die Analyse für alle Äußerungen durchgelaufen ist, können somit die Erkennungsraten für folgende drei Entscheidungen berechnet werden:

Modus/Fokus: Die richtige Modus–Fokus–Konstellation wurde erkannt.

Modus: Der Satzmodus stimmte überein.

Fokus: Die richtige Fokusposition wurde zugewiesen.

Die Erkennungsraten für die Entscheidung **Modus** und **Fokus** dienen als Vergleichszahlen zu den Ergebnissen der gleichnamigen Klassifikationsexperimente, die bei der Erstellung des Intonationsmodells durchgeführt wurden (siehe Kapitel 2.2).
Um auch für die Experimente **Fokus_F** und **Fokus_N** Vergleichszahlen zu erhalten, wird das Analyseprogramm getrennt für alle Fragen und alle Nicht–Fragen gestartet. Dabei werden für Fragen als Zielkonzepte FF2 und FF3 (vgl. den Suchbaum aus Abb. 28) und für Nicht–Fragen NFF2 und NFF3 vorgegeben. Es ergeben sich also noch zwei zusätzliche Erkenungsraten für **Fokus_F** und **Fokus_N**.

Bei den Untersuchungen zur Ursache der Fehlentscheidungen werden die folgenden drei Aspekte berücksichtigt:

- **Fehler bei der Grundfrequenzberechnung**
- **Berücksichtigung der Hörerurteile NAT, FOK und MOD**
- **Grundfrequenzverlauf**

Fehler bei der Grundfrequenzberechnung
Die Fehlentscheidungen bei der Modus–Fokus–Zuweisung sind zum Teil auf Fehler zurückzuführen, die während der Grundfrequenzberechnung entstanden.
Die in Abb. 34 gezeigte Grundfrequenzkontur der zweiten und dritten Phrase einer Äußerung ist am Ende durch hohe Frequenzwerte gekennzeichnet. In diesem Fall handelt es sich um einen sogenannten Oktavfehler. Anstelle des korrekten Wertes wird das Doppelte als Grundfrequenzwert bestimmt. Dies entspricht einer Erhöhung um zwölf Halbtöne, das heißt, einer Erhöhung um genau eine Oktave.
Wird nur die Hälfte des tatsächlichen Grundfrequenzwertes bestimmt, so spricht man von subharmonischen Grundfrequenzwerten (siehe [Nöth91]).
Ebenso finden einige Fehlentscheidungen bei der Modus–Fokus–Zuweisung ihre Ursache bereits in der SH/SL–Entscheidung, wenn nämlich stimmhafte Bereiche nicht als stimmhaft erkannt wurden und deshalb für diese Bereiche keine Grundfrequenzwerte bestimmt werden konnten.

Berücksichtigung der Hörerurteile NAT, FOK und MOD
In die Untersuchung der Fehlentscheidungen werden auch immer die im Rahmen der DFG–Untersuchungen unternommenen Perzeptionstests einbezogen. So sind Fehlurteile über Äußerungen, die Hörer mit einer geringen Natürlichkeit bewerteten oder bei denen sie sich in der Modus- und Fokuszuweisung nicht einig waren, in ihrer Gewichtigkeit geringer einzuschätzen.
Zur Untersuchung der Fehler in der Modusentscheidung wird auch das Hörerurteil MOD herangezogen. Es treten Realisierungen im Fokus–Korpus auf, bei denen der vom Kontext vorgegebene Modus nicht mit dem übereinstimmte, was die Mehrzahl der Hörer wahrnahm (vgl. auch Abb. 3). Bei diesen Äußerungen lag also schon beim Sprechen eine sogenannte Fehlproduktion vor. Abb. 35 zeigt die Grundfrequenzkontur der zweiten und dritten Phrase einer solchen Äußerung. Diese Äußerung sollte vom Kontext her eigentlich als Frage realisiert werden, es ist jedoch kein für Fragen charakteristischer Grundfrequenzanstieg in der dritten Phrase zu beobachten.
Bei Fehlentscheidungen in der Fokusbestimmung wird nochmals das Hörerurteil FOK herangezogen, um zu beurteilen, inwieweit sich die Hörer bei der Akzentzuweisung einig waren.

Grundfrequenzverlauf
Als letztes Kriterium zur Beurteilung der Fehlentscheidungen wird der Grundfrequenzverlauf selbst

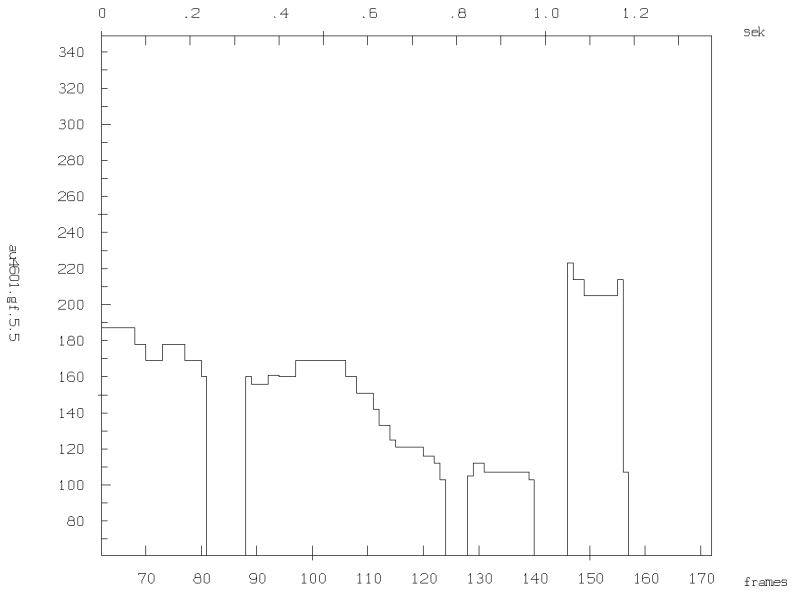


Abb. 34: Grundfrequenzkontur mit Oktavfehler am Äußerungsende

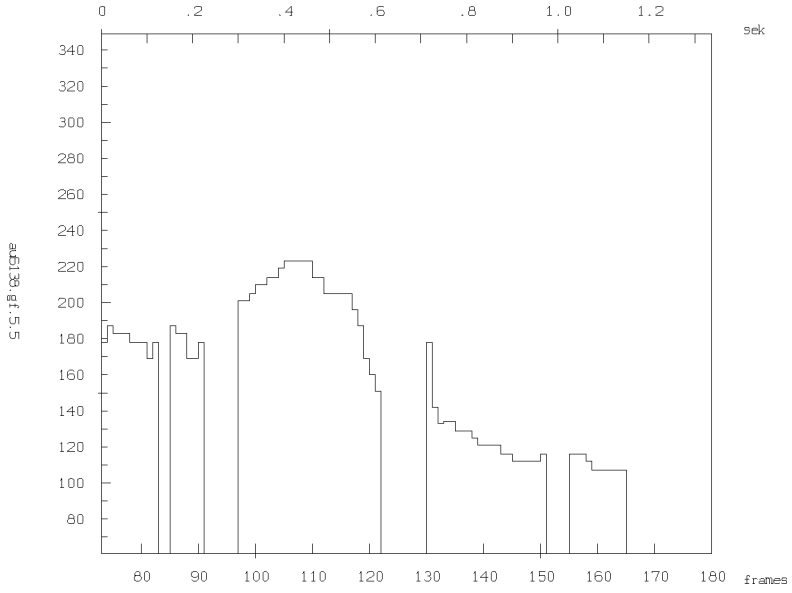


Abb. 35: Fehlproduktion: Die Äußerung wurde als Nicht–Frage realisiert, obwohl durch den Kontext eine Frage intendiert wurde.

	NFF2	NFF3	FF2	FF3	
NFF2	97	42	1	5	145
NFF3	6	30	1	0	37
FF2	7	0	111	3	121
FF3	4	2	13	26	45
	114	74	126	34	348

Tabelle 6: Zuweisungstabelle, $n(X,Y)$: Anzahl, wie oft Äußerungen mit der Konstellation Y die Konstellation X zugewiesen wurde.

verwendet. Es konnten einige Äußerungen beobachtet werden, die zwar den charakteristischen Grundfrequenzverlauf besitzen, aber trotzdem vom Kontrollalgorithmus dem falschen Prototypenkonzept und damit der falschen Modus–Fokus–Konstellation zugewiesen wurden.

Da sich für die unterschiedlichen Variationen von Grundfrequenz- und Phrasengrenzen (siehe Abschnitt 4.1) in etwa die gleichen Aussagen über Erkennungsraten und Ursachen der Fehlentscheidungen machen lassen, werden im folgenden Kapitel beispielhaft die Fehlentscheidungen für einen Testdatentyp näher untersucht. Die Daten $TD_{K,a}$, bei denen sowohl Grundfrequenzwerte als auch Phrasengrenzen automatisch bestimmt wurden und die daher im Hinblick auf einen Einsatz in einem Spracherkennungssystem eigentlich von Interesse wären, standen erst zu einem späteren Zeitpunkt für die Testläufe zu Verfügung. Deshalb wurde die Fehleranalyse am Typ $TD_{K,m}$ vorgenommen. Nach der Fehleranalyse für diesen Typ wird anschließend noch vergleichend auf die Erkennungsraten eingegangen, die mit den übrigen Testdaten erzielt wurden.

4.3.2 Beurteilung von Fehlentscheidungen für $TD_{K,m}$

Für die Testdaten $TD_{K,m}$ stehen insgesamt 357 Realisierungen in Form von Grundfrequenzwerten zur Verfügung, von denen sieben die prototypischen Äußerungen darstellen. Bei zwei Realisierungen wurden in der dritten Phrase keine stimmhaften Bereiche gefunden und somit konnten keine Grundfrequenzwerte bestimmt werden. Diese Äußerungen müssen daher von weiteren Tests ausgeschlossen werden. Es bleiben also noch 348 Äußerungen zur Überprüfung des Modells übrig.

Insgesamt ergibt sich eine Erkennungsrate von 76%, das heißt, 264 der 348 Testäußerungen wurde nach den im vorangegangenen Abschnitt genannten Beurteilungskriterien die richtige Modus–Fokus–Konstellation zugewiesen. 328mal (94%) wurde zumindest der richtige Satzmodus bestimmt und 274mal (79%) wurde diejenige Phrase als Träger des Fokus erkannt, die auch die Mehrzahl der Hörer als fokussiert wahrnahmen. Tabelle 6 gibt einen Überblick über Art und Anzahl der aufgetretenen Entscheidungen.

Im folgenden werden zunächst diejenigen Fälle genauer untersucht, bei denen ein Fehler in der Satzmodusentscheidung auftritt. Anschließend erfolgt eine Analyse der Fälle, bei denen zwar der richtige Satzmodus festgestellt werden konnte, jedoch die falsche Phrase als fokussiert erkannt wurde. Die Untersuchung der Fehlurteile erfolgt relativ detailliert, läßt sich aber in ihren Grundaussagen auf die übrigen Testdaten übertragen.

Modusfehler Bei der Modusentscheidung wurde eine Erkennungsrate von 94% erzielt. 20 Äußerungen wurde somit der falsche Satzmodus zugeordnet. Siebenmal war der Satzmodus des bestbewerteten Prototypen eine Frage, während die Testäußerung eine Nicht–Frage war. Im weiteren wird auf jede Fehlerart einzeln eingegangen.

NFF2 nach FF2 (1/145): Einer von 145 Nicht–Fragen (siehe Tabelle 6), die nach dem Hörerurteil FOK die zweite Phrase als Träger der Fokusinformation besitzen, wurde der Satzmodus Frage zugewiesen, wobei jedoch die Fokusposition übereinstimmt. Dieses Fehlurteil ist auf einen Oktavfehler in der Grundfrequenzberechnung, wie er in Kapitel 4.3.1 beschrieben wurde, zurückzuführen.

NFF2 nach FF3 (5/145): Auch diese Art von Fehlentscheidung ist in allen fünf Fällen auf einen Oktavfehler in der Grundfrequenzberechnung zurückzuführen. Abb. 34 zeigt den Grundfrequenzverlauf der zweiten und dritten Phrase für eine der fünf Realisierungen.

NFF3 nach FF2 (1/37): Die hier fehlkategorisierte Äußerung besitzt einen steigenden Grundfrequenzverlauf am Äußerungsende und wurde von der Kontur her durchaus dem richtigen Satzmodus zugewiesen. Die Hörer waren sich im Kategorisierungstest jedoch völlig einig, daß es sich um eine Nicht–Frage handelt. Eine genauere Untersuchung dieser Äußerung ergab, daß eine Fehlentscheidung bei der SH/SL–Bestimmung vorliegen muß: Die letzte Silbe der dritten Phrase wurde als stimmloser Bereich erkannt. Dadurch fehlt der für die Konstellation charakteristische Grundfrequenzabfall am Ende.

NFF3 nach FF3 (0/37): Dieser Fehler ist nicht aufgetreten.

Äußerungen mit Satzmodus Frage wurde 13mal der Modus Nicht–Frage zugeordnet:

FF2 nach NFF2 (7/121): Bei fünf der sieben Fehlurteile handelt es sich um Fehlproduktionen. Das Analyseprogramm entscheidet sich für denjenigen Satzmodus, für den sich auch die Mehrzahl der Hörer im Kategorisierungstest entschied. Die Fehlurteile der beiden anderen Äußerungen sind auf sogenannte subharmonische Grundfrequenzwerte zurückzuführen. In der dritten Phrase wurde als Grundfrequenzwert durchgehend die Hälfte des tatsächlichen bestimmt.

FF2 nach NFF3 (0/121): Dieser Fehler ist nicht aufgetreten.

FF3 nach NFF2 (4/45): Die Äußerungen, die von dieser Art Fehler betroffen sind, besitzen alle den für die Konstellation *Frage, Fokus auf der dritten Phrase* typischen Grundfrequenzverlauf, jedoch mit zu geringen Ausprägungen. Abb. 36 zeigt die Kontur einer solchen Äußerung, beginnend ab der zweiten Phrase, zusammen mit der Kontur des bestbewerteten Prototypen und des Kernprototypen, der die Konstellation eigentlich repräsentiert. Man erkennt deutlich, daß der Grundfrequenzanstieg am Äußerungsende wesentlich geringer ausgeprägt ist als bei der prototypischen Äußerung dieser Konstellation. Im Kategorisierungstest waren sich alle Hörer einig, daß es sich bei den Äußerungen um Fragen handelt.

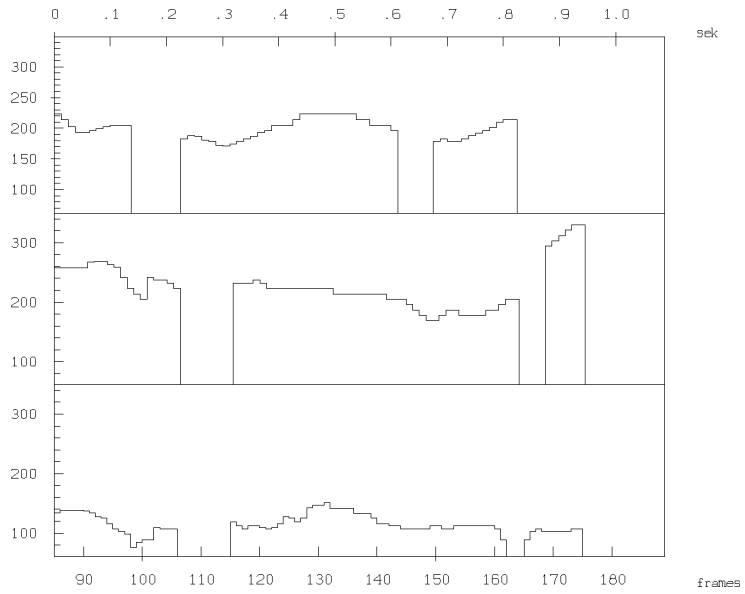


Abb. 36: Äußerung der Konstellation *Frage, Fokus auf der dritten Phrase* (obere Kontur) zusammen mit prototypischer Äußerung dieser Konstellation (mittlere Kontur) und prototypischer Äußerung der Konstellation, die ihr zugewiesen wurde (untere Kontur).

FF3 nach NFF3 (2/45): Für die beiden Fehlentscheidungen konnten dieselben Gründe wie im vorangegangenen Fall festgestellt werden.

Zusammenfassend läßt sich für die Fehler in der Modusentscheidung folgendes feststellen: Von den insgesamt 20 Fehlurteilen sind neun auf Fehler in der Grundfrequenzbestimmung zurückzuführen. Bei fünf Äußerungen handelt es sich um Fehlproduktionen, im Hinblick auf die Hörerurteile trat somit keine Fehlentscheidung auf. In sechs Fällen lag zwar der charakteristische Grundfrequenzverlauf vor, jedoch mit zu geringen Ausprägungen.
Werden die Äußerungen mit Grundfrequenzfehlern bei der Berechnung der Erkennungsrate ausgeschlossen, sowie fehlproduzierte Äußerungen als richtig erkannt eingestuft, so ergibt sich mit dem implementierten Verfahren für die Entscheidung **Modus** sogar eine Erkennungsrate von 98%.

Fokusfehler Nun werden die 64 Fälle näher betrachtet, bei denen die Moduszuweisung korrekt ausfiel, jedoch die Position des Fokus falsch bestimmt wurde (siehe Tabelle 6). Hier soll nochmals erwähnt werden, daß sich die Beurteilung, ob die Fokusposition richtig oder falsch bestimmt wurde, nach dem von den Hörern gelieferten Akzentmaß FOK richtet. Wenn FOK für eine Äußerung größer als 0 ist, so wird die zweite Phrase als die richtige Fokusposition betrachtet, ansonsten die dritte Phrase.

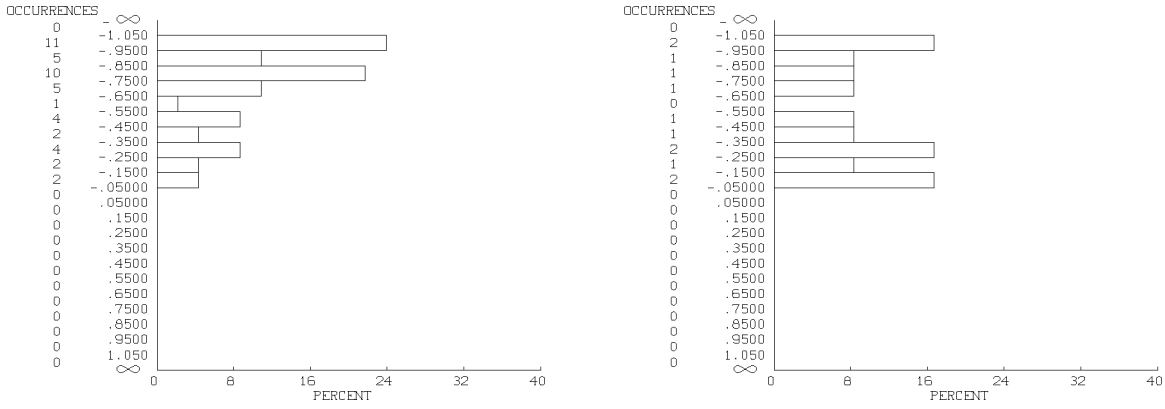
FF2 nach FF3 (3/121): Bei zwei der drei Äußerungen kann man nicht von einem Fehlurteil reden, da das Maß FOK genau den Wert 0 besitzt, das heißt, die gleiche Anzahl von Hörern entschied sich für die zweite Phrase wie für die dritte Phrase als Träger der Fokusinformation.
Bei genauerer Untersuchung der dritten Äußerung stellte sich heraus, daß nach der vorgegebenen Kontextsituation der Fokus auf der dritten Phrase liegen müßte und das Analyseprogramm infolgedessen die richtige Entscheidung liefert. Beim Anhören der betreffenden Äußerung konnte zusammen mit anderen Zuhörern festgestellt werden, daß die zweite Phrase keinesfalls als Träger des Fokus zu erkennen ist. Vermutlich trat bei der Übertragung der Hörerurteile in Tabellen etc. ein Fehler auf, der an dieser Stelle nicht mehr nachvollziehbar ist.

FF3 nach FF2 (13/45): Eines der 13 Fehlurteile ist auf einen Fehler in der Grundfrequenzberechnung zurückzuführen.
Die Anzahl von zwölf weiteren Fehlurteilen stellt ein Resultat dar, welches durchaus noch verbessert werden sollte. Bei einer Beurteilung des Hörerurteils FOK konnte jedoch festgestellt werden, daß dieses Maß bei den zwölf fehlkategorisierten Äußerungen einen Durchschnittswert von -0,52 besitzt und die Hörer sich somit weniger einig in der Akzentzuweisung waren als bei allen Fragen mit Fokus auf der dritten Phrase, für die ein Durchschnittswert von -0,69 ermittelt wurde. Abb. 37(a) zeigt die Verteilung des Hörerurteils für alle Fragen aus dem Korpus mit Fokus auf der dritten Phrase, Abb. 37(b) die Verteilung für die zwölf Äußerungen, die von dem Fehler betroffen sind. Man erkennt eine Verlagerung zu Werten, die eine größere Uneinigkeit der Hörer widerspiegeln.

NFF2 nach NFF3 (42/145): Auf diese Art von Fehler sind 50% aller Fehler zurückzuführen. Bei zwei Fällen kann man wiederum nicht von einer Fehlentscheidung sprechen, da das Maß FOK den Wert 0 besitzt.
40mal wurde somit die dritte Phrase als Träger des Fokus bestimmt, obwohl die Mehrzahl der Hörer die zweite Phrase als fokussiert wahrnahm. In den meisten Fällen besitzt das Maximum der zweiten Phrase eine zu geringe Ausprägung, was jedoch sprecherspezifisch zu sein scheint, da 29 der 40 Äußerungen von nur zwei der sechs Sprecher realisiert wurden. 62% aller Nicht–Fragen mit Fokus auf der zweiten Phrase, die von diesen beiden Sprechern realisiert wurden, sind von der Fehlentscheidung betroffen. Weitere 24% der Nicht–Fragen eines Sprechers wurden aufgrund eines Oktavfehlers einer Konstellation mit Satzmodus Frage zugewiesen. Vermutlich wäre, wenn nicht der Grundfrequenzfehler aufgetreten wäre, auch diese Äußerungen der Konstellation mit der falschen Fokusposition zugewiesen worden. Das Maß FOK der einzelnen Äußerungen ließ **nicht** auf eine größere Uneinigkeit der Hörer bei der Akzentzuweisung schließen.

NFF3 nach NFF2 (6/37): Bei der Untersuchung des Hörermaßes FOK für die sechs fehlkategorisierten Äußerungen konnte auch hier festgestellt werden, daß sich die Hörer im Durchschnitt weniger einig waren als bei allen anderen Nicht–Fragen mit Fokus auf der dritten Phrase (Durchschnittswerte: -0,63 vs. -0,48).

Wie die oben erläuterten Fehlentscheidungen bereits andeuten, und wie die Betrachtung der Erkennungsraten für **Fokus_F** und **Fokus_N** (siehe Tabelle 7) erkennen läßt, ist die geringere Erkennungsrate von **Modus/Fokus** im Vergleich zu **Modus** auf die schlechte Fokuszuweisung bei Nicht–Fragen



(a) (b)

Abb. 37: Verteilung des Hörerurteils FOK für alle Fragen mit Fokus auf der dritten Phrase (a) und für die zwölf Äußerungen, denen die zweite Phrase als Fokusposition zugewiesen wurde (b).

NFF3	NFF3_KTaA	NFF3_RTbA
30	21	9

(a)

FF2	FF2_KTBb	FF2_RTBa	FF2_RTAB
111	62	34	15

(b)

Abb. 38: Zahlenmäßige Verteilung auf Kern- und Randprototypen bei der Modus–Fokus–Zuweisung

zurückzuführen. Es muß jedoch nochmals betont werden, daß auch die Fokusentscheidung bei Fragen mit Fokus auf der dritten Phrase wenig zufriedenstellend verläuft, jedoch aufgrund der geringeren Anzahl von Äußerungen mit dieser Konstellation nicht zu sehr ins Gewicht fällt.

Von den insgesamt 84 Äußerungen, die von einem Fehler betroffen sind, lassen sich zehn Fehlurteile (circa 12%) auf Fehler in der Grundfrequenzbestimmung zurückführen. Werden die Äußerungen, die von Grundfrequenzfehlern betroffen sind, bei der Berechnung der **MODUS/FOKUS**–Erkennungsrate ausgeschlossen, so erhöht sich die Erkennungsrate von 76% auf 78%. Werden weiterhin derjenige Fall (siehe **FF2 nach FF3**), bei dem offensichtlich ein Übertragungsfehler auftrat, und die fünf fehlproduzierten Äußerungen (siehe **FF2 nach NFF2**) sowie die vier Äußerungen mit dem Wert FOK = 0 (siehe **FF2 nach FF3** und **NFF2 nach NFF3**) als richtig erkannt eingestuft, so ergibt sich eine Erkennungsrate von 81% für **MODUS/FOKUS**.

Die Abb. 38 soll zum Abschluß dieses Abschnitts noch einen Eindruck vermitteln, wie sich die richtigen Zuweisungen zu Modus–Fokus–Konstellationen, die sowohl Kern- als auch Randprototypen besitzen, anzahlenmäßig auf die einzelnen Prototypenkonzepte verteilen. Von 30 Äußerungen der Konstellation *Nicht–Frage, Fokus auf der dritten Phrase*, die auch dieser Konstellation zugewiesen wurden (siehe Tabelle 6), besaß in 21 Fällen der Kernprototyp die beste Bewertung, in neun Fällen der Randprototyp.

4.3.3 Vergleich zu den übrigen Testdaten

Nun soll in bezug auf die Fehleranalyse im vorangegangenen Kapitel noch auf die Ergebnisse eingegangen werden, die durch Verwendung anderer Testdaten (Grundfrequenzwerte, Phrasengrenzen) erzielt wurden. Ferner werden die Ergebnisse angesprochen, die aufgrund der Transformation der Grundfrequenzwerte in Halbtonwerte erzielt werden.

	Modus/Fokus	**Modus**	**Fokus**	**Fokus_F**	**Fokus_N**
$TD_{K,m}$	76%	94%	79%	87%	71%
$TD_{Kg,m}$	73%	94%	76%	89%	64%
$TD_{S,m}$	77%	93%	80%	90%	71%
$TD_{K,a}$	77%	94%	80%	90%	72%

Tabelle 7: Erkennungsraten für die unterschiedlichen Testdaten

$TD_{K,m}$	NFF2	NFF3	FF2	FF3	
NFF2	97	42	1	5	145
NFF3	6	30	1	0	37
FF2	7	0	111	3	121
FF3	4	2	13	26	45
	114	74	126	34	348

(a)

$TD_{Kg,m}$	NFF2	NFF3	FF2	FF3	
NFF2	83	56	1	5	145
NFF3	5	31	1	0	37
FF2	6	1	111	3	121
FF3	4	2	10	29	45
	98	90	123	37	348

(b)

$TD_{S,m}$	NFF2	NFF3	FF2	FF3	
NFF2	102	33	1	9	145
NFF3	9	26	2	0	37
FF2	6	1	111	2	120
FF3	3	3	11	27	44
	120	63	125	38	346

(c)

$TD_{K,a}$	NFF2	NFF3	FF2	FF3	
NFF2	98	39	3	2	142
NFF3	8	28	1	0	37
FF2	8	0	109	3	120
FF3	6	1	8	29	44
	120	68	121	34	343

(d)

Abb. 39: Zuweisungstabellen für die vier Typen von Testdaten, $n(X,Y)$: Anzahl, wie oft Äußerungen mit der Konstellation Y die Konstellation X zugewiesen wurde.

Tabelle 7 zeigt die Erkennungsraten für die unterschiedlichen Testdaten im Überblick. Abb. 39 zeigt die einzelnen Zuweisungstabellen. Die niedrigere Gesamtanzahl von Testäußerungen bei $TD_{S,m}$ und $TD_{K,a}$ ist auf Äußerungen zurückzuführen, für die mit den unterschiedlichen Verfahren in der dritten Phrase keine stimmhaften Bereiche gefunden werden konnten und die somit von den Tests ausgeschlossen werden mußten. Bei genauerer Analyse der Tabelle 7 und der Abbildung 39 lassen sich deutlich dieselben Tendenzen in den Fehlurteilen feststellen. Auch die im vorangegangenen Kapitel beschriebenen Feststellungen in bezug auf die Verteilung des Hörerurteils FOK können auf das übrige Material übertragen werden. Im folgenden werden vergleichend die einzelnen noch nicht näher untersuchten Testdatentypen angesprochen.

$TD_{Kg,m}$:

Die Testdaten $TD_{Kg,m}$ unterscheiden sich von $TD_{K,m}$ lediglich durch eine zusätzliche Glättung der

Grundfrequenzkontur. Bei der Modusentscheidung bewirkt dies keinerlei Veränderung in der Erkennungsrate. Ebenso viele und genau die gleichen Äußerungen einer Konstellation werden dem falschen Satzmodus zugewiesen. Bei der Fokusentscheidung für Fragen führte die Glättung zu einer geringfügigen Verbesserung der Erkennungsrate. Nur die Fehlerrate bei der Fokus–Zuweisung von Nicht–Fragen nimmt zu und macht 60% aller Fehlurteile bei diesem Typ aus.

$TD_{S,m}$:
Für diesen Typ lassen sich im Vergleich zu $TD_{K,m}$ geringfügig bessere Ergebnisse feststellen. Die etwas schlechtere Erkennungsrate bei der Modusentscheidung ist auf eine größere Anzahl von Äußerungen zurückzuführen, bei denen ein Oktavfehler in der Grundfrequenzberechnung aufgetreten ist.

$TD_{K,a}$:
Auch bei diesem Typ wurden bessere Ergebnisse als bei $TD_{K,m}$ erzielt. Dieses Ergebnis ist insofern überraschend, da sowohl die Grundfrequenzwerte als auch die Phrasengrenzen automatisch extrahiert wurden. Zum einem spricht dies für das automatische Verfahren zur Extraktion von Phrasengrenzen und zum anderen für das zugrundeliegende Verfahren zur Lautklassifikation. Vermutlich trägt auch die kleinere Framelänge von 10 ms zu den guten Ergebnissen bei.

Vergleich zwischen Grundfrequenzwerten und Halbtonwerten

Tabelle 8 zeigt die erzielten Erkennungsraten, wenn anstelle der Grundfrequenzwerte Halbtonwerte verwendet werden. Es läßt sich im Vergleich zu Tabelle 7 eine deutliche Verschlechterung der Ergebnisse feststellen. Für die Modusentscheidung erhält man zwar in etwa dieselbe Rate, jedoch werden in der Fokuszuweisung die Ergebnisse vor allem bei $TD_{K,m}$ und $TD_{Kg,m}$ wesentlich schlechter. Hier ist wieder die Tatsache festzustellen, daß die schlechten Ergebnisse in der Fokusentscheidung auf die Fokusentscheidung bei Nicht–Fragen zurückzuführen sind. Für die Fokusentscheidung bei Fragen bewirkt die Transformation sogar eine gewisse Verbesserung. Abb. 40 zeigt die unterschiedlichen Erkennungsraten für $TD_{K,a}$ nochmals graphisch.

	Modus/Fokus	**Modus**	**Fokus**	**Fokus_F**	**Fokus_N**
$TD_{K,m}$	61% (–15)	93% (–1)	65% (–14)	89% (+2)	42% (–29)
$TD_{Kg,m}$	63% (–10)	94% (+0)	66% (–10)	91% (+2)	42% (–22)
$TD_{S,m}$	73% (–4)	93% (+0)	76% (–4)	91% (+1)	63% (–8)
$TD_{K,a}$	73% (–4)	94% (+0)	77% (–3)	91% (+1)	66% (–6)

Tabelle 8: Erkennungsraten bei vorangegangener Transformation in Halbtonwerte. Die Veränderungen gegenüber den Erkennungsraten bei Verwendung von Grundfrequenzwerten ist in Prozentpunkten angegeben.

4.3.4 Fokusentscheidung bei Nicht-Fragen

Für die Fokusentscheidung von Nicht–Fragen werden bei jedem der vier unterschiedlichen Typen von Testdaten deutlich geringere Erkennungsraten erzielt als für die Modusentscheidung oder die Fokusentscheidung bei Fragen. Obwohl auch für Fragen mit Fokus auf der dritten Phrase die Fokuszuweisung wenig zufriedenstellend ausfällt, kann jedoch bei den Fehläußerungen eine größere Uneinigkeit der Hörer im Hinblick auf die Fokusposition festgestellt werden. Bei den Nicht–Fragen, für die anstelle der zweiten Phrase die dritte Phrase als Fokusposition bestimmt wird, konnte dies jedoch nicht bemerkt werden.
Die Fehlentscheidungen sind im wesentlichen auf den zu ähnlichen Verlauf der Grundfrequenzkonturen der beiden Kernprototypen zurückzuführen. Eine zusätzliche Glättung der Kontur ($TD_{Kg,m}$) sowie eine Transformation in Halbtonwerte verschlechterten die Ergebnisse noch zusätzlich. Viele Realisierungen mit der Konstellation *Nicht–Frage, Fokus auf der zweiten Phrase* besitzen ein relativ kleines Maximum in der zweiten Phrase und werden deshalb dem Kernprototypen der Konstellation *Nicht–Frage, Fokus auf der dritten Phrase* zugewiesen.

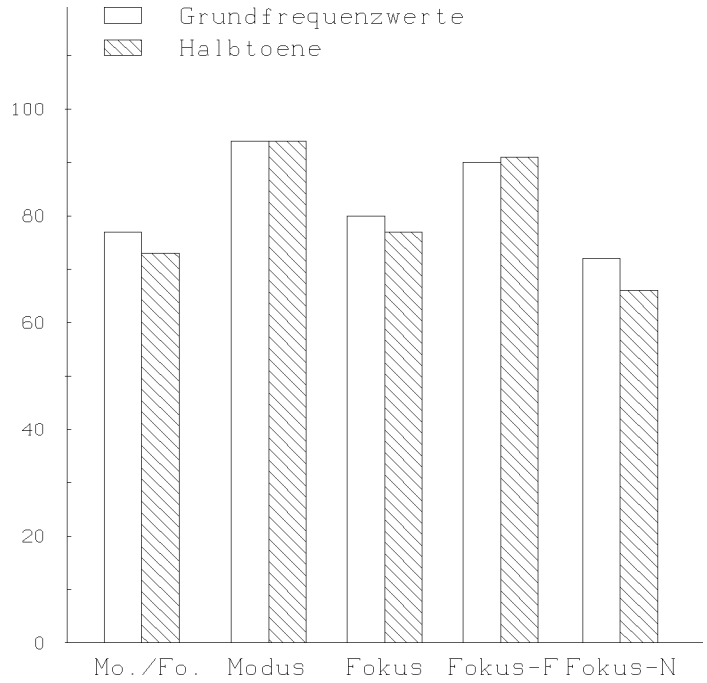


Abb. 40: Vergleich der Erkennungsraten bei Verwendung von Grundfrequenzwerten und Halbtonwerten für $TD_{K,a}$

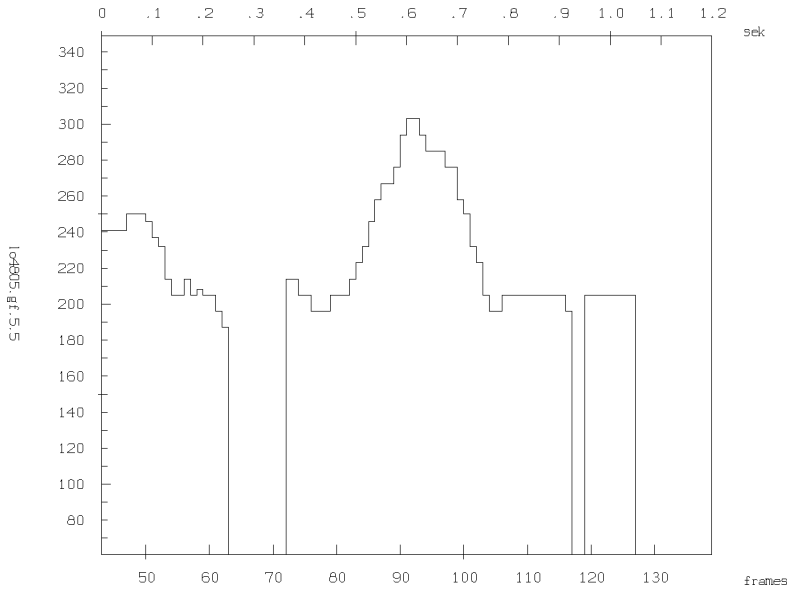


Abb. 41: Grundfrequenzkontur mit ausgeprägtem Maximum in der zweiten Phrase

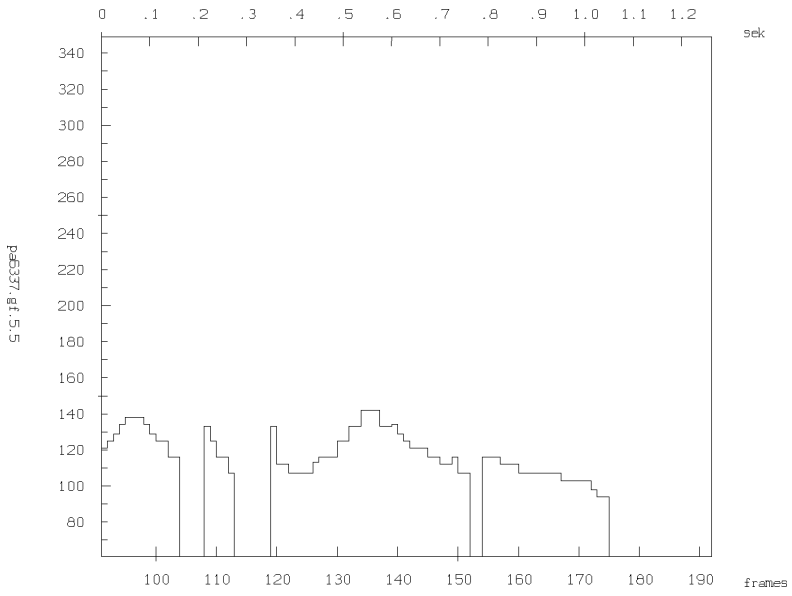


Abb. 42: Grundfrequenzkontur mit schwach ausgeprägtem Maximum in der zweiten Phrase

Abb. 41 und Abb. 42 zeigen die Grundfrequenzkonturen zweier Nicht–Fragen, bei denen die jeweiligen Maxima der zweiten Phrase in ihrer Ausprägung stark differieren, die aber in etwa die gleichen Bewertungen von den Hörern erhielten. (Dieser Unterschied ist nach einer Halbton-Transformation zwar weniger ausgeprägt, aber immer noch vorhanden.) Da die meisten der falsch beurteilten Nicht–Fragen von nur zwei Sprechern realisiert wurden, scheint die geringer ausgeprägte Grundfrequenzbewegung in der zweiten Phrase ein sprecherabhängiges Merkmal zu sein. Es stellt sich die Frage, ob diese Sprecher zusätzlich zur Tonhöhe noch andere prosodische Parameter, wie zeitliche Strukturierung oder Lautheit, zur Markierung der Fokusposition einsetzen. Untersuchungen, ob die Sprecher die Fokusposition durch eine längere Dauer der fokussierten Phrase markieren, brachten keine positiven Ergebnisse.

Die dynamische Zeitverzerrung stellt bei diesen Realisierungen kein geeignetes Verfahren zur Unterscheidung dar. Deshalb wurde versucht, außer auf das durch die dynamische Zeitverzerrung gelieferte Abstandsmaß noch auf ein weiteres Unterscheidungskriterium zurückzugreifen. Aufgrund der in Kapitel 2.2 beschriebenen Ergebnisse, aus denen sich die Grundfrequenzmaxima der zweiten und dritten Phrase als adäquate Merkmale zur Unterscheidung der Fokusposition bei Nicht–Fragen herausstellten, wurden diese Werte in die Entscheidung einbezogen. Dies führte zu der in Kapitel 3.4 erwähnten Einführung des Attributs *max* im Konzept AEUSSERUNG und der Relation *maxphr* in den Konzepten NFF2_KTAa und NFF3_KTaA, welche die Kernprototypen für die Nicht–Fragen repräsentieren.

Damit die Testäußerung diesen Kernprototypen zugeordnet werden kann, sollte die Größe der Maxima von zweiter und dritter Phrase bestimmte Bedingungen erfüllen. Es wurde daher festgelegt, daß die Differenz aus den Maxima bei Nicht–Fragen mit Fokus auf der dritten Phrase — sofern sie dem prototypischen Verlauf des Kerntypen besitzen — einen bestimmten Schwellwert nicht überschreiten darf. Wird dieser Schwellwert bei einer aktuellen Testäußerung überschritten, so scheidet der Kernprototyp der Konstellation *Nicht–Frage, Fokus auf der dritten Phrase* als bestbewertete Instanz von vorneherein aus. Ebenso wird für die Konstellation *Nicht–Frage, Fokus auf der zweiten Phrase* vorausgesetzt, daß das Maximum der zweiten Phrase nicht unter dem der dritten Phrase liegen darf. Diese Bedingungen werden in den Bewertungsfunktionen für die Relation *maxphr* der Konzepte NFF2_KTAa und NFF3_KTaA geprüft. Für die Bewertung der Relation *maxphr* im Konzept NFF2_KTAa gilt somit:

$$max\_bew = \begin{cases} 0, \text{wenn} & max2 < max3 \\ 1, \text{wenn} & max2 \geq max3 \end{cases}$$

Für das Konzept NFF3_KTaA gilt:

$$max\_bew = \begin{cases} 0, \text{wenn} & max2 > max3 + schwelle \\ 1, \text{wenn} & max2 \leq max3 + schwelle \end{cases}$$

Durch diese Bedingungen werden Restriktionen festgelegt, die von einer Testäußerung verletzt werden können. In Kapitel 3.5 wurde beschrieben, daß zur Bewertung von Prototypenkonzepten die ersten beiden Komponenten des Bewertungsvektors herangezogen werden. Die erste Komponente ist ein binäres Maß und gibt an, ob die Testäußerung bestimmte Restriktionen verletzt. Da davon bisher noch keine festgelegt wurden, besaß die Komponente immer den Wert eins. Durch die oben genannten Bedingungen werden nun solche Restriktionen aufgestellt und können somit auch verletzt werden. In diesem Falle muß die erste Bewertungskomponente auf null gesetzt werden. Dies erfolgt durch die Übernahme der Bewertung der Relation *maxphr*.

Tabelle 9 zeigt die Erkennungsraten nach Einführung dieses zusätzlichen Unterscheidungskriteriums bei Nicht–Fragen, wobei als *schwelle* der Wert zehn gewählt wurde. Beim Start des Analyseprogramms kann dieser Wert beliebig festgelegt werden. Man erkennt, daß bereits durch diesen einfachen Ansatz eine wesentliche Erhöhung der Erkennungsraten erzielt werden kann. In Abb. 43 sind für $TD_{K,a}$ die unterschiedlichen Erkennungsraten dargestellt.

	Modus/Fokus	**Modus**	**Fokus**	**Fokus_F**	**Fokus_N**
$TD_{K,m}$	82% (+6)	94%	84% (+5)	87%	82% (+11)
$TD_{Kg,m}$	84% (+11)	94%	86% (+10)	89%	84% (+20)
$TD_{S,m}$	82% (+5)	93%	85% (+5)	90%	84% (+13)
$TD_{K,a}$	84% (+7)	94%	87% (+7)	90%	84% (+12)

Tabelle 9: Erkennungsraten bei Berücksichtigung der Grundfrequenzmaxima. Die Veränderungen gegenüber den Erkennungsraten unter alleiniger Verwendung dynamischen Zeitverzerrung sind in Prozentpunkten angegeben.

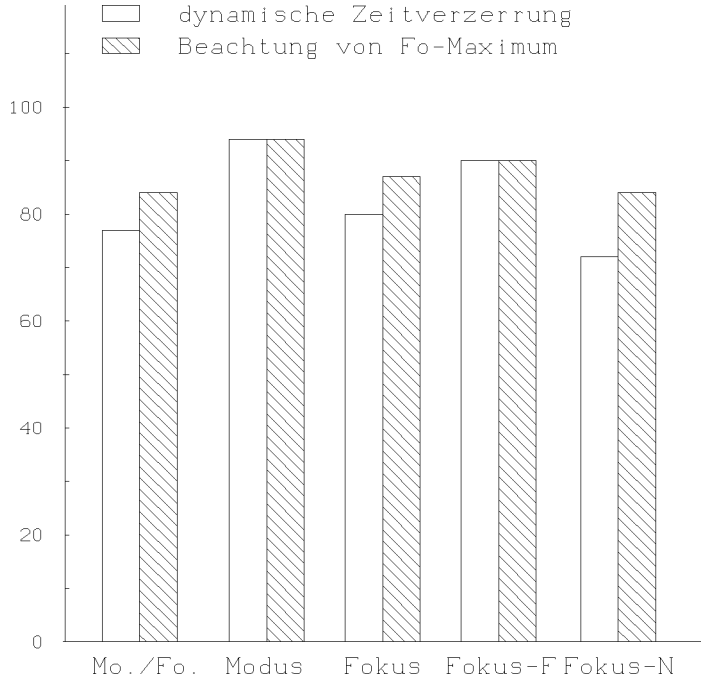


Abb. 43: Vergleich der Erkennungsraten für $TD_{K,a}$

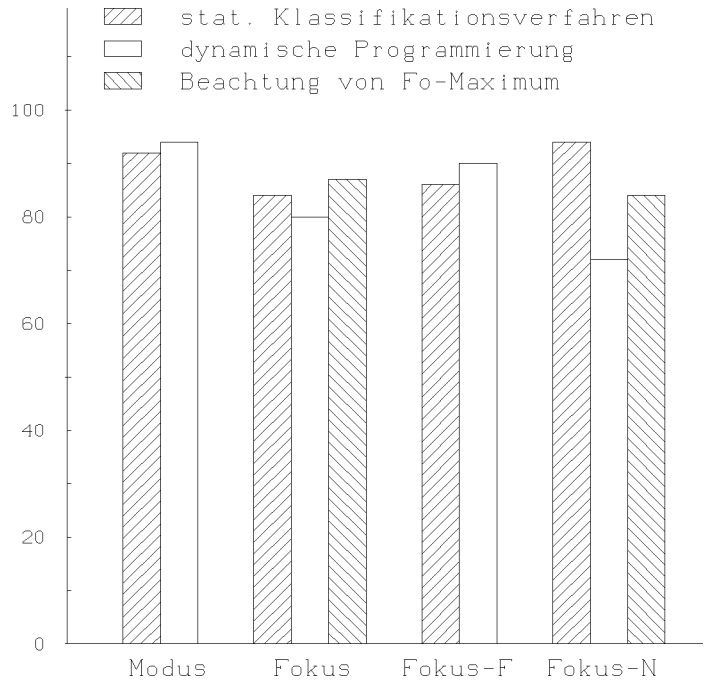


Abb. 44: Vergleich der Erkennungsraten

4.3.5 Abschließende Bemerkungen

Im Rahmen dieser Arbeit wurde einer Testäußerung eine Modus–Fokus–Konstellation zugewiesen, nachdem die Grundfrequenzkontur der prototypischen Äußerungen mit der Kontur der Testäußerung verglichen wurde. Der Vergleich erfolgt nicht über einzelne Merkmale, die den Grundfrequenzverlauf repräsentieren, sondern die komplette Kontur wird als Vergleichskriterium herangezogen. Mittels dynamischer Zeitverzerrung wird ein Abstandsmaß zwischen Test– und Referenzkontur berechnet, wobei das Abstandsmaß der Randprototypen durch einen gewissen Faktor verschlechtert wird. Aufgrund der in den vorangegangenen Abschnitten erläuterten Erkennungsraten und unter der Berücksichtigung, daß die Grundfrequenzkontur automatisch bestimmt wurde, läßt sich die Gültigkeit des implementierten Intonationsmodells durchaus bestätigen.

Abb. 44 zeigt einen Vergleich der hier erzielten Erkennungsraten für den Testtyp $TD_{K,a}$ mit jenen aus den Klassifikationsexperimenten, die zur Erstellung des Intonationsmodells durchgeführt wurden (siehe Tabelle 4 in Kapitel 2.2). Als Vergleichszahlen dienen diejenigen Erkennungsraten, die sich bei einer Lernstichprobe von fünf Sprechern und einer Teststichprobe von einem Sprecher ergaben. Die Testdaten $TD_{K,a}$ stellen im Hinblick auf den Einsatz in einem Spracherkennungssystem den interessantesten der vier Testtypen dar, da bei ihnen sowohl Grundfrequenzwerte als auch Phrasengrenzen automatisch extrahiert wurden.
Bei der Satzmodusentscheidung konnte bei dem in dieser Arbeit implementierten Verfahren ein sehr gutes Ergebnis erzielt werden, das sogar über der Vergleichsrate liegt. Es soll an dieser Stelle nochmals betont werden, daß die hier verwendete Grundfrequenzkontur automatisch berechnet wurde, während die Merkmale für den statistischen Klassifikator zum Teil per Hand extrahiert wurden. Für die Satzmodusentscheidung eignet sich somit der Vergleich zweier Grundfrequenzkonturen mittels dynamischer Zeitverzerrung besonders gut.
Bei der Fokusentscheidung liefert die dynamische Zeitverzerrung mit zusätzlicher Gewichtung der Randprototypen zufriedenstellende Erkennungsraten, die jedoch durchaus noch verbesserungswürdig erscheinen. Ein Vergleich mit den Erkennungsraten des Klassifikators **Fokus_F** zeigt, daß die hier erzielten Ergebnisse besser sind. Die Erkennungsrate der Fokusentscheidung bei Nicht–Fragen fällt dagegen wesentlich schlechter aus. Dies ist darauf zurückzuführen, daß sich die Grundfrequenzverläufe derjenigen Äußerungen, die die Kernprototypen von Nicht–Fragen repräsentieren, zu ähnlich sind. Realisierungen der Konstellation *Nicht–Frage, Fokus auf der zweiten Phrase*, die ein schwach ausgeprägtes Maximum in der zweiten Phrase besitzen, werden meist der Konstellation *Nicht–Frage, Fokus auf der*

dritten Phrase zugewiesen. Vermutlich ist diese geringe Ausprägung der Grundfrequenzbewegung in der zweiten Phrase sprecherabhängig, da ein Großteil der Fehlurteile auf Äußerungen zurückzuführen sind, die von zwei Sprechern realisiert wurden. In diesem Fall, läßt sich bereits durch relativ einfache Ansätze, wie durch den im vorigen Kapitel realisierten Vergleich der Grundfrequenzmaxima, eine Verbesserung erzielen. In Abb. 44 sind für **Fokus** und **Fokus_N** noch zusätzlich die Erkennungsraten eingetragen, die sich infolge der Modifikation ergeben.

Hinsichtlich der unterschiedlichen Testdaten konnten keine großen Unterschiede festgestellt werden, außer daß eine Glättung der Kontur offensichtlich zu einer Verschlechterung der Ergebnisse in der Fokuszuweisung bei Nicht–Fragen führt. Dies ist allerdings nur der Fall, wenn als alleiniges Unterscheidungskriterium der aus der dynamischen Zeitverzerrung gelieferte Abstand verwendet wird. Ebenso führt eine Transformation in Halbtonwerte in diesem Fall zu schlechteren Erkennungsraten. Überraschend war, daß für Testdaten $TD_{K,a}$ insgesamt die besten Erkennungsraten erzielt wurden. Überraschend besonders deshalb, da hier sowohl Phrasengrenzen als auch Grundfrequenzwerte automatisch erzeugt wurden.

Zusammenfassend kann festgestellt werden, daß das Modell durchaus seine Gültigkeit bewiesen hat. Das Verfahren der dynamischen Programmierung liefert bei der Modusentscheidung sehr gute Ergebnisse, erscheint jedoch vor allem bei der Fokusentscheidung unter Hinzunahme einzelner aus der Grundfrequenzkontur extrahierter Merkmale noch verbesserbar.

Literatur

[Altm84] H. Altmann: *Linguistische Aspekte der Intonation am Beispiel Satzmodus.* Forschungsbericht des Instituts für Phonetik und Sprachliche Kommunikation der Universität München 19, S.132-152, 1984.

[Altm88] H. Altmann (Hg.): *Intonationsforschungen.* Niemeyer Verlag, Tübingen, 1988.

[Altm89] H. Altmann, A. Batliner, W. Oppenrieder (Hgg) : *Zur Intonation von Modus und Fokus im Deutschen.* Niemeyer Verlag, Tübingen, 1989.

[Batl89a] A. Batliner: *Fokus,Modus und die große Zahl. Zur intonatorischen Indizierung des Fokus im Deutschen.* In [Altm89], S.21-70, 1989.

[Batl89b] A. Batliner, W. Oppenrieder: *Korpora und Auswertung.* In [Altm89], S.281-320, 1989.

[Batl89c] A. Batliner, E. Nöth: *The Prediction of Focus.* Proceedings European Conference on Speech Technology, Vol.1, S.210-213, Paris, 1989.

[Bußm83] H. Bußmann: *Lexikon der Sprachwissenschaft.* Alfred Körner Verlag, Stuttgart, 1983.

[Erne90] ERNEST–Manual: Version 1.5, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen–Nürnberg, 1990.

[Itak75] F. Itakura: *Minimum Prediction Residual Principle Applied to Speech Recognition.* IEEE Transaction ASSP–23, S.67-72, 1975.

[Jaco88] *Fokus–Hintergrund–Gliederung und Grammatik.* In [Altm88], S.89-132, 1988.

[Kies89] A. Kießling: *Ein interaktives System zur periodenweisen Bestimmung der Grundfrequenz.* Studienarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen–Nürnberg, 1989.

[Kohl77] K. Kohler: *Einführung in die Phonetik des Deutschen.* Erich Schmidt Verlag, Berlin, 1977.

[Komp89] R. Kompe: *Ein Mehrkanalverfahren zur Berechnung der Grundfrequenzkontur unter Einsatz der Dynamischen Programmierung.* Diplomarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen–Nürnberg, 1989.

[Kumm90] F. Kummert: *Flexible Steuerung eines sprachverstehenden Systems mit homogener Wissensbasis.* Dissertation, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen–Nürnberg, 1990.

[Metz91] B. Metzner: *Erstellung einer automatischen Lautzuordnung und deren Einsatz bei der Fokusanalyse.* Diplomarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen–Nürnberg, 1991.

[Myer80] C. Myers, L. R. Rabiner, A. E. Rosenberg: *Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition* IEEE Transaction ASSP–28, S.623-635, 1980.

[Niem83] H. Niemann: *Klassifikation von Mustern.* Springer-Verlag, Berlin, 1983.

[Niem85] H. Niemann, A. Brietzmann, R. Mühlfeld, P. Regel, G. Schukat–Talamazzini: *The Speech Understanding and Dialog System EVAR.* In R. DeMori, C. Y. Suen (Hgg.): *New Systems and Architectures for Automatic Speech Recognition and Synthesis*, Springer-Verlag, Berlin, S.271-302, 1985.

[Niem88] H. Niemann, A. Brietzmann, U. Ehrlich, S. Posch, P. Regel, G. Sagerer, G. Schukat–Talamazzini: *A Knowledge Based Speech Understanding System.* International Journal of Pattern Recognition and Artificial Intelligence, Vol.2, No.2, S.321–350, 1988.

[Nöth91] E. Nöth: *Prosodische Information in der automatischen Spracherkennung - Berechnung und Anwendung*. Niemeyer Verlag, Tübingen, 1991.

[Quil68] M. R. Quillian: *Semantic Memory*. In M. Minsky (Hg.):*Semantic Information Processing*, S.216-270, MIT Press, Cambridge, MA, 1968.

[Rabi80] L. R. Rabiner, C. E. Schmidt: *Application of Dynamic Time Wraping to Connected Digit Recognition*. IEEE Transaction ASSP–28, S.377-388, 1980.

[Sage90] G. Sagerer: *Automatisches Verstehen gesprochener Sprache*. Vol.74 von Reihe Informatik, Bibliographisches Institut, Mannheim, 1990.

[Sako78] H. Sakoe, S. Chiba: *Dynamic Programming Algorithm Optimization for Spoken Word Recognition*. IEEE Transaction ASSP–26, S.43-49, 1978.

[Wöhr90] Th. Wöhrle: *Textüberwachtes Training von Lautkomponenten mit Hilfe eines akustisch-phonetischen Netzwerks*. Diplomarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen–Nürnberg, 1990.