Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

Prosodische Information: Begriffsbestimmung und Nutzen für das Sprachverstehen.

E. Nöth, A. Batliner, A. Kießling,
R. Kompe, H. Niemann

F.–A.–Universität Erlangen–Nürnberg

Vm
Report 219
September 1997

September 1997

E. Nöth, A. Batliner, A. Kießling,
R. Kompe, H. Niemann

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Tel.: (09131) 85 - 7883
e-mail: `noeth@informatik.uni-erlangen.de`

Gehört zum Antragsabschnitt: 1.7.2

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Bildung, Wissenschaft, Forschung und Technologie unter dem Förderkennzeichen 01 IV 701 K/5 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.
ISSN 0949-6084

From Paulus, Wahl (ed.), Mustererkennung 1997, Informatik aktuell, Springer, Heidelberg, 1997, S.37–52.

Prosodische Information: Begriffsbestimmung und Nutzen für das Sprachverstehen

Elmar Nöth, Anton Batliner, Andreas Kießling*, Ralf Kompe + und Heinrich Niemann

Lehrstuhl für Mustererkennung (Informatik 5)
Universität Erlangen–Nürnberg, Martensstr. 3, 91058 Erlangen, Germany
Tel.: +49 (9131) 857888 Fax.: +49 (9131) 303811
email: noeth@informatik.uni-erlangen.de
* jetzt bei Ericsson Eurolab, Nürnberg
+ jetzt bei Sony Stuttgart Technology Center, Fellbach

Zusammenfassung Prosodische Information spielt in der Mensch–Mensch–Kommunikation eine große Rolle, in der automatischen Sprachverarbeitung (ASV) wurde diese Informationsquelle bisher jedoch nicht benutzt. Erst seitdem sich die automatische Sprachverarbeitung der Spontansprache und weniger restringierten Aufgabenstellungen zugewandt hat, ist der Einsatz der Prosodie wirklich wesentlich geworden. Wir beschreiben im einzelnen die Gründe dafür und zeigen an der Integration der Prosodie in das automatische Übersetzungssystem VERBMOBIL, daß dieser Einsatz auch erfolgreich ist. VERBMOBIL ist weltweit das erste ASV–Gesamtsystem, welches prosodische Information während der linguistischen Analyse einsetzt. Die zur Zeit wirkungsvollste prosodische Information wird von den Wahrscheinlichkeiten für Satzgrenzen geliefert. Diese werden zu 94% richtig erkannt. Während des syntaktischen Parsens von Worthypothesengraphen führt die Benutzung der Satzgrenzen–Information zu einer Beschleunigung der syntaktischen Analyse um 92% und zu einer Reduktion der syntaktischen Lesarten um 96%.

1 Was ist Prosodie

Die Prosodie beschäftigt sich mit suprasegmentalen (lautübergreifenden) sprachlichen Ereignissen. Diese Ereignisse überlagern sprachliche Einheiten, die mehr als einen Laut umfassen, also *Silben, Wörter, Phrasen, Sätze*, usw. Zur spektralen Dimension zählen *Klangfarbe, Tonhöhe, Stimmlage* und *Stimmqualität*, zur

[1] Die diesem Bericht zugrundeliegenden Untersuchungen wurden mit Mitteln des Bundesministers für Bildung, Wissenschaft, Forschung und Technologie (*BMBF*) unter den Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 F/4 im Rahmen des Verbundprojektes VERBMOBIL gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren. Wir danken allen VERBMOBIL–Partnern, die die prosodische Information etikettiert bzw. in ihre Analysemodule integriert haben, für die außerordentlich gute Kooperation.

Intensität *Lautheit*, und die zeitliche Dimension umfaßt *Pausensetzung, Dauerverhältnisse, Rhythmus, Sprechgeschwindigkeit* und *Tempo*. Diesen perzeptiven Einheiten entsprechen akustische Parameter. So ist z.B. die *Grundfrequenz* des Sprachsignals (F0) das akustische Korrelat der *Tonhöhe*.

Ein anderer Begriff, der prosodische Ereignisse und ihre Verwendung in der Mensch–Mensch–Kommunikation bezeichnet, ist die *Intonation*; häufig wird allerdings nur die distinktive Verwendung der Tonhöhe als Forschungsgegenstand der Intonation bezeichnet (siehe hierzu z.B. die Diskussion in [21, 14]).

Der Zuhörer extrahiert Information aus den wahrgenommenen prosodischen Ereignissen, d.h. wir können den Ereignissen funktionale Rollen zuordnen. Als wichtigste Funktionen werden allgemein die prosodische Markierung von *Grenzen, Betonung* und *Satzmodus* angesehen.

Bereits Lea [18] hat den Einsatz dieser prosodischen Information für ASV–Systeme vorgeschlagen. Trotz einer wachsenden Zahl von Forschungsprojekten zur Prosodie und ASV dauerte es 17 Jahre, bis mit VERBMOBIL das erste komplette ASV–System implementiert wurde, in dem prosodische Information eingesetzt wird. Wir sehen die folgenden Gründe für diesen langen Zeitraum:

- Das zentrale Problem im Zusammenhang mit der rechnergestützten Analyse prosodischer Information ist der hohe Komplexitätsgrad, der im wesentlichen durch folgende Faktoren verursacht wird: Beeinflussung durch segmentale Information, Interferenzen in den akustischen Dimensionen der unterschiedlichen prosodischen Funktionen, Interaktion der verschiedenen prosodischen Parameter, Fakultativität der prosodischen Mittel sowie sprecher- bzw. sprachenspezifische Faktoren.
- Die wichtigste Rolle der Prosodie, die Segmentierung und Disambiguierung von Äußerungen, kam in den bisherigen Anwendungen oft nicht zum Tragen, da entweder diese Analyseaufgaben nicht notwendig waren (z.B. in Diktiersystemen) oder da die Äußerungen der Benutzer zu kurz waren. So ist zum Beispiel die durchschnittliche Länge einer Äußerung in einem Feldexperiment mit einem Zugauskunftssystem 3.5 Wörter, vgl. [10].

VERBMOBIL ist dagegen fast das einzige System, in dem "real life"–Sprache in einer Mensch–Mensch–Kommunikation verarbeitet werden muß. 70% der Äußerungen enhalten mehr als einen Satz und im Durchschnitt ist eine Äußerung 20 Wörter lang [28]. Auch die in Spontansprache häufigen elliptischen Konstruktionen sowie Abbrüche und Neuansätze tragen zur Komplexität und Ambiguität bei. Mit deutlichem Erfolg kann Prosodie bisher allerdings noch nicht in der Erkennungsphase (Umsetzung in die am besten passende Wortkette bzw. in einen Worthypothesengraphen), sondern erst in der Verstehensphase (Interpretation der Äußerung) eingesetzt werden. Deshalb zeigt sich die Bedeutung der Prosodie erst in einem System wie VERBMOBIL, das eines der ersten Systeme ist, in dem eine "end-to-end"–Evaluation das Optimierungskriterium darstellt und in dem auch eine tiefe linguistische Analyse durchgeführt wird.

In diesem Beitrag wollen wir die Berechnung und den Nutzen prosodischer Information beschreiben. Da die Autoren das Prosodiemodul des VERBMOBIL–Systems entwickelt haben und da die Benutzung prosodischer Information auf allen Ebenen der linguistischen Analyse in VERBMOBIL möglich ist, werden wir die Anwendungsbeispiele aus diesem System nehmen. Die vorgestellten Algorithmen und ihre Integration in die linguistischen Analysemodule lassen sich aber u.E. in praktisch jedes System einbringen.

Nach einer kurzen Beschreibung des VERBMOBIL–Systems (Kap. 2) werden wir die Berechnung der prosodischen Information beschreiben (Kap. 3). In Kap. 4 zeigen wir den Nutzen der berechneten Information für die linguistischen Analyse–Ebenen. In Kap. 5 präsentieren wir erste Ergebnisse und enden mit einem Ausblick auf zukünftige Arbeiten (Kap. 6).

2 Das VERBMOBIL–System

VERBMOBIL ist ein Projekt zur Übersetzung spontan gesprochener Verhandlungsdialoge mit eingeschränkter Domäne (Terminabsprachen). Ein Überblick über das System findet sich z.B. in [30, 4]. Übersetzungsrichtung ist im Augenblick hauptsächlich Deutsch nach Englisch. Die Prosodie wird z.Zt. nur für das Deutsche eingesetzt. Bild 1 zeigt einen Überblick über die Architektur des Prototypen. Nach der Aufnahme des Sprachsignals wird mit einem von zwei alternativen Hidden–Markov–Erkennern (HMM) ein Worthypothesen–Graph (WHG) berechnet. Dieser wird mit prosodischer Information angereichert (Kap. 3). Der WHG wird von einem von zwei alternativen Syntaxmodulen analysiert, das heißt, die syntaktisch beste Wortkette zusammen mit möglichen Ableitungsbäumen (Lesarten) wird an die semantische Analyse weitergegeben. Die semantische Repräsentation der Äußerung wird unter Verwendung des Dialogmoduls vom Transfermodul in die semantische Repräsentation einer englischen Äußerung übertragen, welche im Generierungsmodul in eine textuelle Repräsentation überführt und schließlich synthetisiert wird. Parallel zu dieser tiefen linguistischen Analyse führt das Dialogmodul eine flache linguistische Analyse durch. Diese wird als "Rückfall"–Systemreaktion genutzt, falls die tiefe Analyse nach Ablauf einer Zeitschranke noch keine Lösung gefunden hat. Hierbei wird die Äußerung nach Dialogakten klassifiziert und mit dialogaktabhängigen Schablonen übersetzt. Alternativ zu den Schablonen können die Dialogakte auch durch ein zweites flaches Verfahren (Übersetzungsgedächtnis, *example based translation*) übersetzt werden [30]. Eine detaillierte Beschreibung der VERBMOBIL–Architektur findet sich z.B. in [8].

Bild 1 zeigt die Interaktion des Prosodiemoduls mit den anderen Modulen des VERBMOBIL–Systems. Die durchgezogenen Linien zeigen Datenschnittstellen, die gestrichelten Linien den Informationsfluß; hier werden die Daten des Prosodiemoduls nicht direkt, sondern von zwischengeschalteten Modulen zusammen mit ihren eigenen Analyseergebnissen weitergereicht. Zur Zeit benutzen die folgenden Module des VERBMOBIL–Systems prosodische Information: Syntax, Semantische Konstruktion, Dialog, Transfer und Sprachsynthese. Bevor wir beschreiben, wie die Module prosodische Information einsetzen, wollen wir zunächst eine Übersicht darüber geben, wie die prosodische Information berechnet wird.

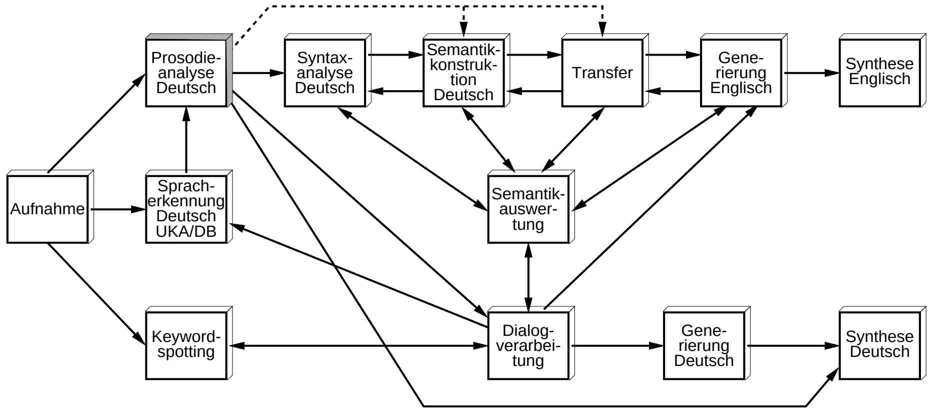

Bild 1. Architektur des VERBMOBIL–Forschungsprototypen.

3 Berechnung der prosodischen Information

Unter Berechnung der prosodischen Information wollen wir die zwei folgenden Schritte verstehen:

- Extraktion von prosodischen Merkmalvektor für suprasegmentale Einheiten; momentan berechnen wir Merkmalvektoren für Wörter (bzw. Worthypothesen) und für ganze Redebeiträge (Turn) eines Sprechers
- Klassifikation des Vektors nach funktionalen prosodischen Klassen wie Grenze/Nicht–Grenze; die Wahrscheinlichkeit für die Zugehörigkeit zu diesen Klassen stellt die prosodische Information dar, die von den linguistischen Modulen benutzt wird.

Eingabe des Prosodiemoduls ist ein Sprachsignal und ein dazugehöriger WHG, Ausgabe ein mit prosodischer Information angereicherter WHG. Dazu werden jeder Worthypothese, d.h. jeder Kante im WHG, Wahrscheinlichkeiten zugeordnet dafür, daß

- das Wort Träger eines Phrasenakzents ist
- nach dem Wort eine Satzgrenze folgt
- es sich (im Falle einer Satzgrenze) bei dem Bereich links von der Grenze um einen Satz mit Aussage–, Frage– oder progredienter (weiterweisender) Intonation handelt (Satzmodus).

Die Berechnung der prosodischen Information wird detailliert in [14, 13] beschrieben und die Anwendung der Information auf WHG in [17, 16].

Aus dem Sprachsignal wird frameweise die F0 und die Lautheit berechnet. Diese Merkmale werden als prosodische Basismerkmale bezeichnet. Zu jeder Worthypothese wird die Zeitzuordnung zwischen dem akustischen Signal und den Phonemen seiner Standardaussprache mit einem HMM erzeugt. Somit liegt auch eine Zeitzuordnung zu den Silben der Worthypothese vor. Mit Hilfe der Viterbi–Suche wird für jede Worthypothese der optimale Pfad im WHG durch dieses Wort bestimmt. Für jede Worthypothese wird aufgrund dieses Pfades unter Berücksichtigung der Vorgänger und Nachfolger ein Merkmalvektor berechnet. Die Merkmale werden auf den Basismerkmalen berechnet, und zwar in dem Zeitbereich, welcher von der Worthypothese und den optimalen Vorgängern/Nachfolgern überdeckt wird. Es werden für Silben– und für Wortsegmente Merkmale berechnet. Auf der einen Seite sind diese Merkmale zum Teil miteinander hoch korreliert und enthalten dadurch redundante Information. Auf der anderen Seite konnten wir in [2] zeigen, daß jede der Merkmalgruppen zu einer Verbesserung der Erkennungsleistung beiträgt. Die Größe des Kontexts für die Experimente in Kap. 5 betrug ± 2 Wörter sowie ± 2 Silben und Silbenkerne, bezogen auf die aktuelle Worthypothese bzw. seine wortfinale Silbe. Für jeden dieser Kontexte werden folgende Merkmale berechnet:

- absolute und normalisierte Dauer jeder Silbe, jedes Silbenkerns und jedes Wortes; die Normalisierung geschieht wie in [32];
- Merkmale zur Beschreibung des F0– und Energieverlaufs; hierzu gehören
  - Regressionskoeffizienten
  - erster Wert (*Onset*), Minimum, Maximum und letzter Wert (*Offset*)
  - die relative Position dieser Werte auf der Zeitachse;
- die Länge einer Pause vor und nach der betrachteten Worthypothese (falls vorhanden);
- die Sprechgeschwindigkeit.

Diese Merkmale bezeichnen wir als prosodische Strukturmerkmale, da sie in ihrer Gesamtheit die Struktur der akustisch–prosodischen Parameter widerspiegeln. So schlägt sich ein fallender Intonationsverlauf vor einer potentiellen Grenze in der Steigung der Regressionsgeraden ebenso nieder wie in der Tatsache, daß die zeitliche Position des Maximums weiter entfernt ist als die des Minimums.

Zusätzlich können wir jedem Kontext sogenannte linguistisch–prosodische Merkmale zuweisen, wie etwa, ob eine betrachtete Silbe Träger des lexikalischen Wortakzents ist (lexikalische Merkmale), oder welche syntaktisch/semantischen Eigenschaften die aktuelle Worthypothese hat. In VERBMOBIL werden z.Zt. nur lexikalisch–prosodische Merkmale verwendet, da die Prosodie ja gerade zur syntaktisch/semantischen Analyse eingesetzt werden soll. In Zukunft ist aber der Einsatz solcher Merkmale durchaus auch vor der linguistischen Analyse denkbar, z.B. durch einen stochastischen *Part–of–Speech Tagger*. Bild 2 gibt einen Überblick über die Extraktion der prosodischen Merkmale.

Die unterschiedlichen Merkmale werden in [14] ausführlich beschrieben und evaluiert. Dort werden die besten Ergebnisse erzielt, wenn **alle** (276) Merkmale für die Klassifikation herangezogen werden [2]. Aus Aufwandsgründen haben wir unter Verwendung klassischer Merkmalsauswahlverfahren [20] die Merkmale auf 121 für die Satzgrenzen–Klassifikation und auf 113 für die Akzent–Klassifikation reduziert. Für die Satzmodus–Klassifikation werden momentan 14 aus der F0–Kontur abgeleitete Merkmale verwendet.

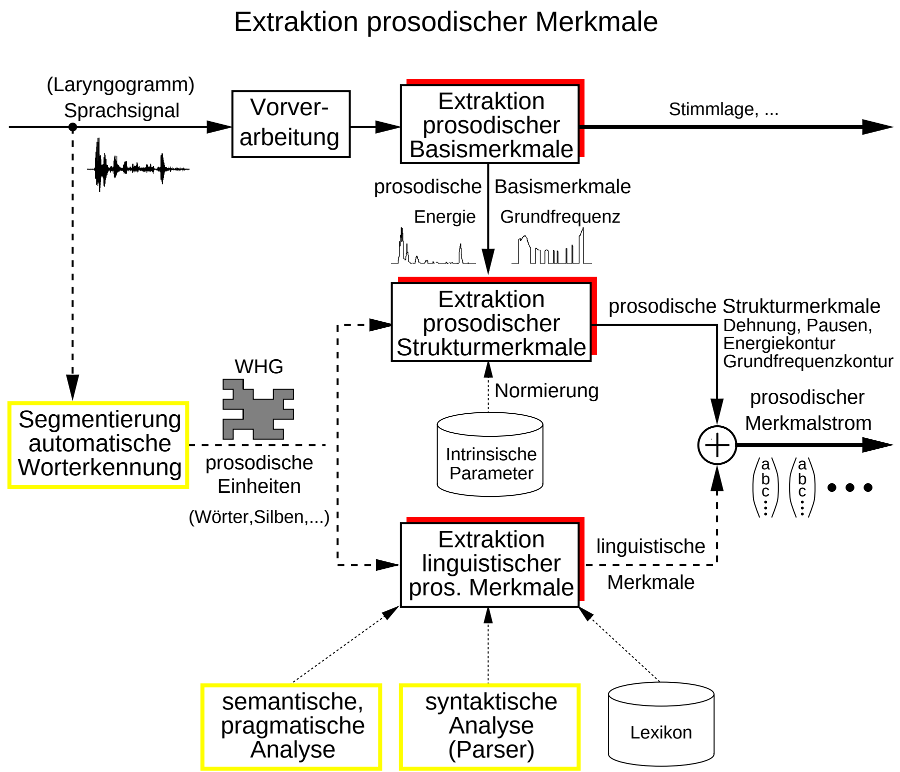

Bild 2. Schematischer Überblick über die Extraktion prosodischer Merkmale.

Für jede dieser Klassifikationsaufgaben wird ein eigenes Mehrschichten–Perzeptron (MLP) benutzt. In Voruntersuchungen erreichten diese NN–Klassifikatoren signifikant bessere Erkennungsraten als multivariate Normalverteilungsklassifikatoren und als Polynomklassifikatoren [15]. Als Referenzdaten dienten dabei perzeptiv gelabelte Grenzen, die von unserem VERBMOBIL–Partner Universität Braunschweig erstellt wurden [25].

Bei der Klassifikation der Satzgrenzen wird das Ergebnis des MLP mit einem kategoriebasierten n-Gramm–Sprachmodell verknüpft, das die Wahrscheinlichkeit für eine Satzgrenze unter Berücksichtigung der Kontextwörter berechnet [16, 17]. Zur Zeit ist der Kontext auf ± 2 Wörter beschränkt. Durch die Einführung eines neuen syntaktisch–prosodischen Labelsystems für Satzgrenzen konnte die Erkennungsrate des Sprachmodells im Vergleich zum Training mit perzeptiv gelabelten Grenzen drastisch erhöht werden [3]. Dies lag in erster Linie daran, daß die syntaktisch–prosodischen Label auf der Transliteration wesentlich schneller erzeugt werden können als die perzeptiven durch Anhören des Sprachsignals; somit lag ungefähr die zehnfache Menge an Trainingsdaten vor.

In [31] werden Klassifikationsbäume für die Klassifikation von Satzgrenzen verwendet; bei unseren Experimenten waren jedoch n-Gramme deutlich besser als Klassifikationsbäume [22].

4 Anwendung der prosodischen Information

Im folgenden wollen wir den Nutzen der prosodischen Information für die ASV beschreiben. Hierzu werden wir für die einzelnen linguistischen Analyseebenen des VERBMOBIL–Systems beschreiben, welche prosodische Information genutzt wird. Da die Interaktion mit der Syntax–Analyse z.Zt. den Hauptbeitrag zur Verbesserung des Gesamtsystems leistet, werden wir diesen Aspekt detaillierter darstellen.

Syntaktische Analyse:
Es gibt zwei Gründe, warum die Syntax–Analyse sehr stark von prosodischer Information abhängt.

- Um sicherzustellen, daß die meisten der gesprochenen Wörter erkannt werden, muß bei Spontansprache ein großer WHG erzeugt werden. Bei dem momentanen VERBMOBIL–System werden WHG mit ca. 10 Hypothesen pro gesprochenem Wort erzeugt. Das Finden des korrekten (oder eines approximativ korrekten) Pfades durch den Graphen stellt ein schwieriges Suchproblem dar.
- Selbst wenn die gesprochene Wortfolge korrekt vorliegt, sind normalerweise immer noch viele verschiedene Lesarten möglich, da in spontaner Sprache durch elliptische Konstrukte viele Ambiguitäten vorhanden sind, und da die Äußerungen in den VERBMOBIL–Dialogen vergleichsweise lang sind.
  In dem folgenden Beispiel liegen zwei unterschiedliche syntaktische Lesarten einer Wortfolge aus der VERBMOBIL–Domäne vor, bei denen nur die Satzgrenzen zur Unterscheidung zwischen den unterschiedlichen syntaktischen Strukturen, ihrer semantischen Bedeutung, ihrer pragmatischen Bedeutung und schließlich ihrer adäquaten Übersetzung in die Zielsprache führen.

  (1) "*Vielleicht. Am Montag bei mir. Paßt das?*"
  "*Maybe. On Monday, at my place. Is that OK?*"

  (2) "*Vielleicht am Montag. Bei mir paßt das.*"
  "*Maybe on Monday. That's possible for me.*"

Beide VERBMOBIL–Syntaxmodule benutzen die Bewertung der Satzgrenzen des Prosodiemoduls zusammen mit den akustischen Bewertungen der Worthypothesen und der n-Gramm–Sprachmodelle, um eine Präselektion zwischen den kombinatorisch möglichen Pfaden im WHG durchzuführen. Diese ausgewählten Wortfolgen, die Information über Satzgrenzen enthalten, werden syntaktisch analysiert. Hierbei wird im Siemens–Modul eine "Trace Unification Grammar" (TUG) benutzt [5, 4], im IBM–Modul eine "Head–driven Phrase Structure Grammar" (HPSG) [23].

rule 1: S	→ PHRASE		S
rule 2: PHRASE	→ SATZ		PSSG
rule 3: PHRASE	→ TOPIK_ELLIPSE		PSSG
rule 4: PHRASE	→ ELLIPTISCHE_PHRASE		PSSG
rule 5: PHRASE	→ EXKLAMATIV		PSSG
rule 6: PHRASE	→ EXKLAMATIV		

Tabelle 1. Teil der Grammatik, die prosodisch-syntaktische Satzgrenzen (PSSG) berücksichtigt. Die erste Regel bedeutet beispielsweise, daß das Startsymbol S durch die nicht–terminalen Symbole PHRASE und S ersetzt werden kann. PHRASE steht dabei für eine beliebige satzäquivalente Phrase.

Im Siemens–Parser werden Hypothesen über prosodisch–syntaktische Satzgrenzen (PSSG) direkt zur Steuerung der Suche des Parsers verwendet. Ausgangspunkt für die Verwendung von PSSG beim Parsen ist die vom Prosodiemodul mittels NN/n–Gramm–Klassifikator für jede Worthypothese im WHG berechnete Wahrscheinlichkeit für eine syntaktisch–prosodische Wortgrenze. Aus diesen Wahrscheinlichkeiten ergeben sich für jede Worthypothese Hypothesen für PSSG bzw. ¬PSSG. Um diese beim Parsen verwenden zu können, mußten die TUG Grammatik und der WHG–Parser [27] von Siemens erweitert werden.

Eine kontextfreie Grammatik für die Analyse spontaner Sprache muß die verschiedensten Abfolgen von Phrasen innerhalb einer einzigen Äußerung erlauben; der dafür relevante Teil der Grammatik ist in Tabelle 1 dargestellt. Die erste Regel definiert, daß eine Äußerung aus mehreren Phrasen bestehen kann. Solche Phrasen sind normale Sätze (Regel 2), Sätze mit Topik–Ellipse (Regel 3), elliptische Phrasen wie Präpositional– oder Nominal–Phrasen (Regel 4) oder Diskurs– bzw. "Exklamativ"–Partikeln (Regeln 5 und 6). Bei solchen "Exklamativen" werden zwei Regeln benötigt, um die Optionalität von PSSG nach Wörtern wie *oh* zu definieren. Nach allen anderen Phrasentypen ist ein PSSG Symbol obligatorisch.

Die Analyse von Wortgraphen basiert auf einer solchen Grammatik. Der Parser ist dabei in eine A*-Suche integriert (siehe z.B. [20]). Die Analyse verläuft im Wortgraphen von links nach rechts und startet beim ersten logischen Knoten des Graphen (l_1). Ein Suchraumknoten n_i entspricht einem zyklenfreien Pfad in einem Wortgraphen, der am Knoten l_1 beginnt, und der sich zusätzlich durch genau eine Segmentierung der Wortkette durch PSSG–Symbole auszeichnet. Mit anderen Worten: zwei unterschiedliche Segmentierungen desselben Pfades im Wortgraphen bilden zwei verschiedene Suchbaumknoten. In jedem Punkt der Suche werden grundsätzlich beide Alternativen (PSSG und ¬PSSG) verfolgt. Die Prosodie steuert dabei über die im Wortgraphen eingetragenen Satzgrenzenbewertungen die Suche. Keine Alternative wird frühzeitig durch die Prosodie ausgeschlossen. Allerdings werden immer nur grammatikalisch (im Sinne der deutschen spontanen Sprache) zulässige Wort-PSSG-Folgen weiterverfolgt.

Die einzelnen Suchbaumknoten repräsentieren bewertete konkurrierende Teilanalysen. In jedem Schritt der Suche wird der am besten bewertete Eintrag (der mit minimalem Schätzwert für die Gesamtkosten) aus der Agenda der noch zu bearbeitenden Teilanalysen entfernt und vom Parser analysiert. Wenn die Analyse fehlschlägt, d.h. wenn die dem Knoten n_i zugrundeliegende Symbolfolge grammatikalisch inkorrekt ist, wird die Hypothese verworfen. Wenn die Analyse erfolgreich war, d.h. wenn die Symbolfolge den Anfang einer grammatikalisch korrekten Symbolfolge bildet, wird der Knoten expandiert und die Nachfolgeknoten werden bewertet und in die Suchagenda einsortiert.

Die Gesamtkosten eines Knotens setzen sich zusammen aus

- den vom Worterkenner berechneten akustischen Bewertungen der Worthypothesen,
- der n-Gramm–Bewertung für die Wortkette und
- der vom Prosodiemodul berechneten Wahrscheinlichkeiten für PSSG bzw. ¬PSSG nach den einzelnen Wörtern der dem Knoten zugrundeliegenden Wortkette.
- geeignet abgeschätzten Restkosten.

Im IBM–Modul werden Präselektion und tiefe Analyse sequentiell ausgeführt: zuerst werden die n besten Wortfolgen aus dem WHG extrahiert. Je zwei solcher Wortfolgen unterscheiden sich in Bezug auf die enthaltenen Wörter und/oder die Position eines PSSG Symbols und/oder in der Position des *leeren Elements*. In einem deutschen Hauptsatz ist das Verb normalerweise an zweiter Stelle, wohingegen es in einem Nebensatz an letzter Stelle ist; allerdings muß diese "letzte" Stelle nicht notwendigerweise das Ende des Satzes sein. Das *leere Element* in Verb–Zweit–Sätzen ist an der Position, an der das Verb wäre, wenn es sich um einen Verb–Letzt–Satz handeln würde. Die Bestimmung dieser Position ist sehr aufwendig und der Suchraum wird durch prosodische Grenzinformation drastisch eingeschränkt. Die Benutzung prosodischer Information durch das IBM Syntaxmodul ist in [1] ausführlich beschrieben.

Semantische Konstruktion:
Das VERBMOBIL–Semantikmodul erhält einen Ableitungsbaum, die dazugehörige Wortfolge und die prosodischen Betonungs–Bewertungen vom Syntaxmodul. Aufgrund dieser Eingabe werden unterspezifizierte *Diskurs–Repräsentationsstrukturen* (DRS) [12, 7] erzeugt.

Wenn aufgrund von Ambiguitäten mehrere DRS plausibel sind, wird Betonungs–Information dazu benutzt, die falschen DRS zu verwerfen. Zur Disambiguierung der Interpretation könnte auch Dialog–Kontextinformation benutzt werden; allerdings ist diese Disambiguierung wesentlich aufwendiger als die mit prosodischer Information [6]. Die Benutzung der Prosodie kann mit dem folgenden Beispiel aus dem VERBMOBIL–Korpus verdeutlicht werden. Hierbei ist zwar die Bedeutung der beiden Sätze gleich, aber die Position der Satzbetonung ändert den Skopus und somit die Präsupposition der Äußerung. Dies führt zu unterschiedlichen Übersetzungen des Partikels *noch* (*still, another*).

(3) "*Dann müssen wir noch einen <u>Termin</u> ausmachen.*"
"*Then we still have to fix a <u>date</u>.*"

(4) "*Dann müssen wir <u>noch</u> einen Termin ausmachen.*"
"*Then we have to fix <u>another</u> date.*"

Dialogverarbeitung:
Eine der Aufgaben des Dialogmoduls [24] ist es, den Verhandlungsdialog grob zu verfolgen. Hierzu werden die Äußerungen in Dialogakte segmentiert. Dialogakte sind Kommunikationsschritte wie *Begrüßung*, *Ablehnung* oder *Vorschlag*. Zur Zeit werden 18 Dialogakte unterschieden (siehe [11] für eine detaillierte Beschreibung). Die Erkennung der Dialogakte geschieht mit statistischen Klassifikatoren. Da ein Turn im Durchschnitt 2 Dialogakte enthält, wird der Turn zunächst in Dialogaktsegmente zerlegt, welche anschließend klassifiziert werden. Hierzu wird zunächst aus dem WHG diejenige Wortsequenz extrahiert, die vom akustischen und vom Sprachmodell–Klassifikator am besten bewertet wurde. Mit Hilfe der prosodischen Satzgrenzen–Wahrscheinlichkeiten wird die Wortsequenz in Teilsequenzen zerlegt. Diese Teilsequenzen werden in die 18 Dialogaktklassen klassifiziert. Über Schwellwerte, ab welchen eine Satzgrenzen–Wahrscheinlichkeit zu einer Dialoaktgrenze führt, wird die Über–/Untersegmentierung geregelt. Da im Durchschnitt nur jede zweite Satzgrenze auch eine Dialogaktgrenze ist, aber andererseits praktisch jede Dialogaktgrenze eine Satzgrenze, wird somit die gleiche Information verwendet wie bei der syntaktischen Analyse, allerdings nach anderen Richtlinien ausgewertet. Sobald genügend Datenmaterial vorhanden ist, sollen allerdings Klassifikatoren verwendet werden, welche nur auf Dialogaktgrenzen trainiert wurden. Diese so gewonnenen Dialogakte sind Basis für beide flachen Übersetzungsverfahren. Weitere Details zur Dialogaktsegmentierung finden sich z.B. in [17, 19].

Transfer:
Das Transfermodul des VERBMOBIL–Systems übersetzt DRS, welche die semantische Information der Äußerungen repräsentieren, in DRS für die entsprechenden englischen Sätze [9]. Hierfür muß u.U. eine pragmatische Analyse und Disambiguierung durchgeführt werden. In den folgenden Fällen greift der Transfer auf Betonungs– und Satzmodus–Information zu:

– Die Satzmodus–Information wird benutzt, um zwischen Fragen und Nicht–Fragen zu unterscheiden, wenn andere grammatische Indikatoren fehlen. Betrachten wir das folgende Beispiel, wo das Verb *treffen* in Verb–Erst–Stellung ist, und der Satzmodus nur von der prosodischen Information abhängt [6]:

  (5) "*Treffen wir uns dann beim Informationsbüro der IAA!*"
  "*So, let us meet at the IAA information office.*"

  (6) "*Treffen wir uns dann beim Informationsbüro der IAA?*"
  "*Do we meet then at the IAA information office?*"

– Die Betonungs–Information wird hauptsächlich zur Interpretation von Partikeln herangezogen. Im folgenden Beispiel hängt die Bedeutung der Wortfolge davon ab, ob der Satzakzent auf *schon* oder auf *finde* liegt. Weitere Beispiele zum Nutzen prosodischer Information im VERBMOBIL–Transfermodul finden sich in [26].

(7) "*Finde ich <u>schon</u>.*" "*I really believe that.*"
(8) "*<u>Finde</u> ich schon.*" "*I'll find it certainly.*"

Sprachsynthese:
Um eine höhere Benutzerakzeptanz zu erreichen, sollte sich die synthetisierte Sprachausgabe eines Übersetzungssystems an die Stimme des Sprechers anpassen (vor allem in einem Multiparty–Szenario). In Bezug auf Prosodie bedeutet dies, daß Parameter wie Stimmlage und Sprechgeschwindigkeit angepaßt werden sollten. Bisher wird im VERBMOBIL–System lediglich zwischen einer männlichen und weiblichen Stimme umgeschaltet. Die Entscheidung wird aufgrund der F0–Kontur der Benutzeräußerung gefällt.

5 Experimente und Ergebnisse

Tabelle 2 zeigt die wichtigsten Ergebnisse des momentanen VERBMOBIL–Prosodiemoduls bei isolierter Auswertung. Um das Modul besser beurteilen zu können, wurde auf der gesprochenen Wortkette gearbeitet, d.h. es wurde eine 100%-ig korrekte Worterkennung angenommen. Wenn man das Modul mit WHG der Dichte 10 Hypothesen/gesprochenes Wort auswertet, und sich auf WHG beschränkt, in denen die gesprochenen Wörter im WHG enthalten sind, sinkt die Erkennungsrate um ca. 2 Prozentpunkte (im WHG sind alle gesprochenen Wörter enthalten, in der besten Worthypothesen–Kette des WHG dagegen nur ca. 75% der gesprochenen Wörter).

Klassifikationsaufgabe	Gesamt-erkennungsrate	klassenweise gemittelte Erkennungsrate
Satzgrenze vs. Nicht–Grenze	94%	90%
Dialogaktgrenze vs. Nicht–Grenze	86%	89%
Betont vs. Nicht–Betont	83%	82%

Tabelle 2. Erkennungsraten des VERBMOBIL–Prosodiemoduls.

Die Erkennungsraten für die Grenzen beziehen sich auf alle Wortgrenzen eines Turns außer dem Turnende (hier ist die Klassifikation trivial). Bei den Satzgrenzen wurde ein Klassifikator verwendet, der die mit dem MLP gewonnene akustisch–prosodische Information mit einem n-Gramm–Sprachmodell verknüpft. Das MLP alleine erreichte eine Erkennungsrate von 86%. Bei der Betonungsbewertung wurde auf den gesprochenen Wörtern, nicht den Silben ausgewertet.

Der im Oktober 1996 präsentierte Forschungsprototyp von VERBMOBIL [29, 30] übersetzt mit den einzelnen Übersetzungsansätzen jeweils ca. 50% der Äußerungen approximativ korrekt. Werden die flachen Analysestrategien als Rückfall beim Scheitern der tiefen Analyse verwendet, so können 74% der Äußerungen approximativ korrekt übersetzt werden.

Bis jetzt liegen systematische Auswertungen der Wirkung der prosodischen Information nur für die Interaktion mit den beiden Syntaxmodulen vor, bei denen der Einfluß der prosodischen Information durch An–/Abschalten des Moduls ausgewertet wurde.

	ohne Prosodie	mit Prosodie	Verbesserung
# erfolgreiche Analysen	368	359	-2%
# Lesarten	137.7	5.6	96%
Parse Zeit (Sekunden)	38.6	3.1	92%

Tabelle 3. Ergebnisse des Siemens WHG Parsers.

Tabelle 3 zeigt die Verbesserung des Siemens–WHG–Parsers, wenn die Wahrscheinlichkeiten über prosodische Satzgrenzen eingesetzt wird. Wie man sieht, wird sowohl die Zahl der Lesarten als auch die Parse–Zeit drastisch reduziert. Diese Ergebnisse wurden auf 594 echt spontanen Turns erzielt. Diese Turns wurden weder zum Training des Prosodiemoduls noch zur Entwicklung der Grammatik oder des Parsers verwendet.

Die Tatsache, daß 9 WHG beim Einsatz prosodischer Information nicht analysiert werden konnten, ist darauf zurückzuführen, daß der Suchraum durch das Einreihen der partiellen Parses in die Suchagenda anders durchsucht wird, und daß die Analyse nach einer Zeitschranke abgebrochen wird. Wir halten diese geringe Verschlechterung allerdings für vernachlässigbar, wenn man bedenkt, daß ohne Prosodie der Echtzeitfaktor für die Syntaxanalyse durchschnittlich 6.1 beträgt und durch die Prosodie auf durchschnittlich 0.5 absinkt. Da die Berechnung der prosodischen Information bei einer WHG–Dichte von 10 Hypothesen/gesprochenes Wort einen Echtzeitfaktor von 1.0 hat, liegt die Beschleunigung bei ca. 75%.

Für den IBM–Parser liegen nur Resultate für transliterierte echte VERBMOBIL–Dialoge vor, welche von geübten Sprecher nachgesprochen wurden. Mit diesem Sprachmaterial konnte eine Beschleunigung der Parsezeit um 46% erzielt werden, wenn die prosodische Satzgrenzen–Information benutzt wurde.

Eine systematische Auswertung des Einflusses der Prosodie ist nicht ohne weiteres möglich, da beide flache Übersetzungsverfahren ohne die prosodische Segmentierung eines Turns in Dialogaktsegmente nicht funktionieren. Betrachtet man diese Tatsache sowie die inakzeptable Bearbeitungszeit in der tiefen Analyse für die Syntax ohne prosodische Information, so kann man sagen, daß die prosodische Information für die Analyse in VERBMOBIL unverzichtbar ist (Hinzu kommt ja auch noch die Bearbeitungszeit der Semantik für 137.7 vs. 5.6 Lesarten!).

Die Wichtigkeit der prosodischen Information für die verschiedenen Module, die an der linguistischen Analyse von VERBMOBIL beteiligt sind, konnte mit all diesen Ergebnissen eindrucksvoll demonstriert werden.

6 Zusammenfassung und Ausblick

Die drei wichtigsten Funktionen der Prosodie sind die Markierung von Grenzen, Akzenten und Satzmodus. Obwohl die Bedeutung der prosodischen Information in der Mensch–Mensch–Kommunikation allgemein anerkannt wird, wird diese Informationsquelle in der ASV bisher nur spärlich benutzt. Häufig liegt dies am Szenario, in dem an der Entwicklung der ASV–Technologie gearbeitet wird. Für die Verarbeitung von Spontansprache, von relativ langen Redebeiträgen sowie bei der Anforderung, diese Beiträge nicht nur zu erkennen, sondern auch inhaltlich zu erschließen, konnte am Beispiel von VERBMOBIL gezeigt werden, daß prosodische Information nicht nur sehr wichtig, sondern sogar unabdingbar wird. Von den drei wichtigsten Funktionen der Prosodie liefert die Grenzmarkierung in VERBMOBIL den größten Beitrag zur Gesamtleistung des Systems, insbesondere bei der syntaktischen Analyse und bei der flachen Inhaltserschließung. Hier liegen auch bereits systematische Auswertungen zum Beitrag der Prosodie im Gesamtsystem vor: Die Zeit für die Syntaxanalyse konnte um 92% reduziert werden, die Zahl der Lesarten um 96%.

Abgesehen davon, daß erst aufgrund der Art des in VERBMOBIL untersuchten Datenmaterials prosodische Information wesentlicher Bestandteil der linguistischen Analyse wurde, sehen wir folgende Gründe für den großen Erfolg unserer Arbeiten:

- Sinnvoller Einsatz linguistischen und prosodischen Wissens bei der Annotation großer Trainingskorpora.
- Benutzung des statistischen Ansatzes: Das Prosodiemodul fällt keine harten Entscheidungen, sondern bewertet die unterschiedlichen möglichen prosodischen Ereignisse.
- Die prosodischen Bewertungen steuern lediglich den Ablauf der Suche in der linguistischen Analyse. Eine fehlerhafte Bewertung führt nicht dazu, daß richtige Teilanalysen verworfen werden.
- Einbezug des Ergebnisses der Worterkennung: Somit können bekannte intrinsische Einflüsse der realisierten Phoneme in die Bewertung mit einbezogen werden [21, Kapitel 2].
- Berechnung der Wahrscheinlichkeiten für prosodische Ereignisse für jede Worthypothese: Werden prosodische Ereignisse für einen Zeitpunkt bewertet, erhalten alle Worthypothesen, welche in einem Knoten enden, die gleiche Bewertung; unser Ansatz erlaubt daher einen sehr differenzierten Einsatz in der linguistischen Analyse.
- Bereitstellung der prosodischen Information vor der linguistischen Analyse: Der Einsatz der prosodischen Information in der Suche ist effizienter, als wenn Alternativen erst vollständig linguistisch analysiert werden, und die prosodische Information danach zur Disambiguierung eingesetzt wird.

– Entwicklung neuer, bisher noch nicht untersuchter Algorithmen.

In Zukunft planen wir, neben einer für das VERBMOBIL–System notwendigen Anpassung der Verfahren an das Englische und Japanische, systematische Auswertungen des Einflusses der Akzent– und Satzmodus–Information bei der semantischen Auswertung, dem Transfer und bei der Extraktion des zentralen Gehalts eines Dialogaktes für die flachen Übersetzungsverfahren.

Weiterhin arbeiten wir daran, weg von der strikt sequentiellen Architektur (Worterkennung $\rightarrow$ Prosodie $\rightarrow$ Syntax) zu einer integrierten Verarbeitung zu gelangen. Ziel ist es, daß das Worterkennungsmodul die prosodische Information zusammen mit stochastischen Sprachmodellen bereits während der Erkennungsphase benutzt, um Hypothesen zu generieren, welche Einheiten oberhalb der Wortebene überdecken, also Phrasen und Teilsätze. Über die Integration der prosodischen Information in die stochastischen Sprachmodelle erwarten wir auch einen Einfluß der Prosodie auf die Erkennungsphase, was durch unseren bisherigen sequentiellen Ansatz nicht möglich war.

Schließlich wollen wir den Bereich der prosodischen Charakterisierung von Sprechertypen und die Erkennung von Emotionen untersuchen, eine Informationsquelle, die sowohl für die schnelle Adaption des Erkennungssystem als auch für die Reaktion des Gesamtsystems von Bedeutung ist.

Literaturverzeichnis

1. A. Batliner, A. Feldhaus, S. Geissler, A. Kießling, T. Kiss, R. Kompe, and E. Nöth. Integrating Syntactic and Prosodic Information for the Efficient Detection of Empty Categories. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 71–76, Kopenhagen, 1996.
2. A. Batliner, A. Kießling, R. Kompe, H. Niemann, and E. Nöth. Can We Tell apart Intonation from Prosody (if we Look at Accents and Boundaries)? In G. Kouroupetroglou, editor, *Proc. of an ESCA Workshop on Intonation*, pages 39–42, Athens, 1997. University of Athens, Department of Informatics.
3. A. Batliner, R. Kompe, A. Kießling, H. Niemann, and E. Nöth. Syntactic–prosodic Labelling of Large Spontaneous Speech Data–bases. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1720–1723, Philadelphia, 1996.
4. H.U. Block. The Language Components in Verbmobil. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 79–82, page 1, München, 1997.
5. H.U. Block and S. Schachtl. Trace & Unification Grammar. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 87–93, Nantes, 1992.
6. J. Bos. Personal communication, July 1996.
7. J. Bos, B. Gambäck, Ch. Lieske, Y. Mori, M. Pinkal, and K. Worm. Compositional Semantics in Verbmobil. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 131–136, Kopenhagen, 1996.
8. T. Bub and J. Schwinn. Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. In *Int. Conf. on Spoken Language Processing*, volume 4, pages 1026–1029, Philadelphia, 1996.
9. K. Eberle. Disambiguation by Information Structure in DRT. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 334–339, Kopenhagen, 1996.

10. W. Eckert, E. Nöth, H. Niemann, and E. Schukat-Talamazzini. Real Users Behave Weird — Experiences made collecting large Human–Machine–Dialog Corpora. In P. Dalsgaard, L.B. Larsen, L. Boves, and I. Thomsen, editors, *Proc. of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, pages 193–196, Vigsø, Denmark, 1995. ESCA.
11. S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. Quantz. Dialogue Acts in Verbmobil. Verbmobil Report 65, 1995.
12. H. Kamp and U. Reyle. *From Discourse to Logic and DRT; An Intorduction to Modeltheoretic Semantics of Natural Language.* Kluwer Academic Publishers, Dordrecht, 1993.
13. A. Kießling, R. Kompe, A. Batliner, H. Niemann, and E. Nöth. Classification of Boundaries and Accents in Spontaneous Speech. In R. Kuhn, editor, *Proc. of the 3rd CRIM / FORWISS Workshop*, pages 104–113, Montreal, 1996.
14. Andreas Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.
15. R. Kompe, A. Batliner, A. Kießling, U. Kilian, H. Niemann, E. Nöth, and P. Regel-Brietzmann. Automatic Classification of Prosodically Marked Phrase Boundaries in German. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 173–176, Adelaide, 1994.
16. R. Kompe, A. Kießling, H. Niemann, E. Nöth, E.G. Schukat-Talamazzini, A. Zottmann, and A. Batliner. Prosodic Scoring of Word Hypotheses Graphs. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 1333–1336, Madrid, 1995.
17. Ralf Kompe. *Prosody in Speech Understanding Systems.* Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
18. W. Lea. Prosodic Aids to Speech Recognition. In W. Lea, editor, *Trends in Speech Recognition*, pages 166–205. Prentice–Hall Inc., Englewood Cliffs, New Jersey, 1980.
19. M. Mast, R. Kompe, S. Harbeck, A. Kießling, H. Niemann, E. Nöth, and V. Warnke. Dialog Act Classification with the Help of Prosody. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1728–1731, Philadelphia, 1996.
20. H. Niemann. *Pattern Analysis and Understanding*, volume 4 of *Series in Information Sciences.* Springer–Verlag, Heidelberg, 1990.
21. E. Nöth. *Prosodische Information in der automatischen Spracherkennung — Berechnung und Anwendung.* Niemeyer, Tübingen, 1991.
22. E. Nöth, R. De Mori, J. Fischer, A. Gebhard, S. Harbeck, R. Kompe, R. Kuhn, H. Niemann, and M. Mast. An Integrated Model of Acoustics and Language Using Semantic Classification Trees. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 419–422, Atlanta, 1996.
23. C. Pollard and I. Sag. *Information–based Syntax and Semantics, Vol. 1*, volume 13 of *CSLI Lecture Notes.* CSLI, Stanford, CA, 1987.
24. N. Reithinger, E. Maier, and J. Alexandersson. Treatment of Incomplete Dialogues in a Speech–to–speech Translation System. In P. Dalsgaard, L.B. Larsen, L. Boves, and I. Thomsen, editors, *Proc. of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, pages 33–36. ESCA, Vigsø, Denmark, 1995.
25. M. Reyelt. Ein System zur prosodischen Etikettierung von Spontansprache. In R. Hoffmann and R. Ose, editors, *Elektronische Sprachsignalverarbeitung*, volume 12 of *Studientexte zur Sprachkommunikation*, pages 167–174. TU Dresden, Wolfenbüttel, 1995.

26. B. Ripplinger and J. Alexandersson. Disambiguation and Translation of German Particles in Verbmobil, Verbmobil Memo 70, 1996.
27. L.A. Schmid. Parsing Word Graphs Using a Linguistic Grammar and a Statistical Language Model. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 41–44, Adelaide, 1994.
28. H. Tropf. Spontansprachliche syntaktische Phänomene: Analyse eines Korpus aus der Domäne "Terminabsprache". Technical report, Siemens AG, ZFE ST SN 54, München, 1994.
29. W. Wahlster. Presseerklärung zum Verbmobil-Forschungsprototypen am 25.10.1996 in München, 1996. http://www.dfki.uni-sb.de/verbmobil.
30. W. Wahlster, T. Bub, and A. Waibel. Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 71–74, München, 1997.
31. M.Q. Wang and J. Hirschberg. Automatic Classification of Intonational Phrase Boundaries. *Computer Speech & Language*, 6(2):175–196, 1992.
32. C.W. Wightman and M. Ostendorf. Automatic Labeling of Prosodic Patterns. *IEEE Trans. on Speech and Audio Processing*, 2(3):469–481, 1994.