Automatische Erkennung von Akzentuierungen und Phrasierungen in Sprachsynthesekorpora

Diplomarbeit

Timo Baumann

Automatische Erkennung von Akzentuierungen und Phrasierungen in Sprachsynthesekorpora: Diplomarbeit

Timo Baumann

Abgabetermin 29. Mai 2007

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Danksagung

Diese Diplomarbeit ist durch die IBM Deutschland Entwicklung GmbH in Böblingen unterstützt worden, die mir Know-How, Arbeitsmittel und Infrastruktur zur Verfügung stellte und dadurch erst diese Arbeit ermöglichte.

Meinem Betreuer bei IBM, Gregor Möhler, danke ich besonders, dass er mich nach Kräften unterstützt und mich im schwierigen Terrain der Prosodie geleitet hat.

Auch den anderen Mitgliedern der IBM Voice Group, ihnen voran Volker Fischer, Ruth Fuchss und Carsten Günther möchte ich danken, dass sie immer ein offenes Ohr für mich hatten und viele Fragen beantworten konnten.

Wolfgang Menzel, meinem Betreuer am Fachbereich Informatik der Universität Hamburg möchte ich für die Bereitschaft danken, diese Arbeit zu betreuen. Obwohl er mir alle Freiheiten bei ihrer Durchführung ließ, hat er mir bei Schwierigkeiten stets geholfen. Nicht zuletzt ermöglichte er die Anschaffung des Kiel-Korpus aus Fachbereichsmitteln.

Auch von außerhalb habe ich vielfache Hilfe erfahren: Caren Brinckmann danke ich für die Übersendung der Prosodielabel-Dateien des Kiel-Korpus, die auf den Original-CDs des IPDS leider nicht mitgeliefert wurden. Hannes Pirker vom OFAI hat mir das Wien-Korpus zur Verfügung gestellt, wofür ich sehr dankbar bin. Antje Schweitzer vom IMS suchte für mich das Manuskript des IMS-Korpus. Außerdem danke ich allen, die an meinem Perzeptionstest teilgenommen haben.

Ganz persönlich möchte ich mich bei meinen Kommilitonen Michael Dotzer, Christof Kubosch, Simon Strecker und Janne Zeller bedanken. Sie haben mich nicht nur jahrelang ertragen, sondern sich schließlich auch noch als fleißige Korrekturleser erwiesen.

Die übrigen Praktikanten und Diplomanden bei IBM möchte ich grüßen. Stella Müller bin ich darüber hinaus besonders dankbar dafür, dass sie nicht nur jederzeit Lust auf eine Kaffeepause hatte, sondern mir auch tatkräftig beim Labelling zur Seite stand.

Vor allen anderen danke ich Verena Willkomm. Sie richtet mich auf, wenn ich am Boden liege.

Versicherung an Eides statt

Ich versichere, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe angefertigt habe. Ich habe mich anderer als der im beigefügten Verzeichnis angegebenen Hilfsmittel nicht bedient.

Alle Zitate und sinngemäßen Entlehnungen sind als solche gekennzeichnet. Die zugehörigen Quellen sind jeweils angegeben.

Ich wünsche mir die Aufnahme in den Bestand der Bibliothek des Fachbereichs.

Hamburg, den 29. Mai 2007

Timo Baumann

Kapitel 1. Einleitung

Diese Arbeit beschäftigt sich mit der Prosodieannotierung in Sprachsynthesekorpora. Bei der Prosodieannotierung werden Markierungen benutzt, um zu beschreiben, wie eine Äußerung gesprochen wurde. Dies kann Hervorhebungen durch den Sprecher, seine Unterbrechungen im Redefluss und so weiter betreffen.

Für die Prosodieannotierung hat sich durchgesetzt, Akzentuierungen und Phrasierungen zu kennzeichnen. Mehrere prosodieannotierte Korpora sind verfügbar und konnten für diese Arbeit genutzt werden. Die Unterschiede zwischen den Korpora sind teilweise groß, und konnten nicht vollständig überwunden werden. Der Vergleich unterschiedlicher Korpora erlaubt aber auch Schlüsse darüber, was allgemeingültig und was korpusspezifisch ist.

Ziel der Arbeit ist eine möglichst hohe Leistung der Prosodieannotierung. Gleichzeitig wird auf die Nutzung zu komplexer Prosodiemodelle verzichtet. Auf diese Weise werden Zirkelschlüsse vermieden. Die Ergebnisse der Prosodieannotierung sind deshalb für die unabhängige Untersuchung von generativen Prosodiemodellen, wie sie in der Sprachsynthese verwendet werden, geeignet.

Außerdem soll untersucht werden, welche Merkmale für die Erkennung von Akzentuierungen und Phrasierungen besonders geeignet sind. Es stellt sich heraus, dass für die unterschiedlichen Korpora teilweise verschiedene Merkmale wichtig waren. Dies legt den Schluss nahe, dass Prosodie teilweise sprecherabhängig ist.

Die Erkennung von Akzentuierungen und Phrasierungen erfolgt in dieser Arbeit über Klassifizierer, die mit Methoden des maschinellen Lernens trainiert werden. Der Wert der Merkmale für die Klassifizierung wird mithilfe automatischer Merkmalsauswahl bestimmt.

1.1. Aufbau der Arbeit

In Kapitel 2 definiere ich zunächst, was in dieser Arbeit unter Prosodie verstanden werden soll und gehe auf Systeme der Prosodie und ihre Anwendbarkeit im Rahmen dieser Arbeit ein. Die Teilsysteme der Prosodie werden vorgestellt und die Silbe und das Wort als Konstituenten von Akzentuierung beziehungsweise Phrasierung motiviert.

Kapitel 3 behandelt Maschinelles Lernen. Zunächst werden die Methoden zur Evaluierung maschinellen Lernens beschrieben und wichtige Leistungsmaße vorgestellt. Dann werden die in dieser Arbeit benutzten Methoden maschinellen Lernens, Naïve-Bayes-Klassifizierer und Entscheidungsbäume, vorgestellt und der zusätzliche Schritt einer automatischen Merkmalsauswahl erläutert.

Kapitel 4 beschäftigt sich mit den Korpora, die für diese Arbeit zur Verfügung stehen. Ein Abschnitt beschäftigt sich mit dem Aufbau eines minimal prosodieannotierten Korpus für die Sprachsynthese und den dabei auftretenden Problemen. Die untersuchten Korpora werden jeweils statistisch untersucht, um eine Grundlage für die Merkmalsauswahl des folgenden Kapitels zu legen.

Kapitel 5 stellt das entwickelte System zur Prosodieerkennung vor. Dazu werden die für die Akzentuierungs- und Phrasierungsklassifizierung benötigten Merkmale und ihre Extraktion aus den Korpora beschrieben.

In Kapitel 6 wird die Leistung des in Kapitel 5 vorgestellten Systems auf den einzelnen Korpora evaluiert. Die Ergebnisse der automatischen Merkmalsauswahl werden diskutiert. Außerdem wird versucht, den Nutzen der einzelnen Merkmale zu bewerten.

In Kapitel 7 soll die Sprecherunabhängigkeit des Verfahrens evaluiert werden. Dazu werden die Korpora wechselweise für Training und Test benutzt. Eine besondere Rolle kommt hier dem Kiel-Korpus zu, da es einheitlich annotierte Daten verschiedener Sprecher enthält.

Eine mögliche Anwendung für die automatische Prosodieannotierung wird in Kapitel 8 durchgeführt. Ein bisher nicht prosodieannotiertes Korpus wird automatisch annotiert und die Ergebnisse werden

für das Training der symbolischen und akustischen Prosodiemodule eines TTS-Systems benutzt. Die Ausgabe des modifizierten TTS-Systems wird auf Basis des ursprünglichen Systems evaluiert und es werden Schlüsse für eine Verbesserung der erzielten Ergebnisse gezogen.

Kapitel 2. Prosodie

Dieses Kapitel stellt die Systeme, Phänomene und Merkmale vor, die zusammengefasst die Prosodie bilden.

2.1. Der Begriff der Prosodie

Der Begriff *Prosodie* (griech. „pros-odé“: Hinzusingen, Hinzutönen; Neppert 1999) beschreibt das System aller *suprasegmentalen Merkmale* der Sprache. Das sind Merkmale, die nicht direkt einem bestimmten Segment zugehörig sind und sich meist über einen längeren Bereich erstrecken (Neppert 1999, S. 155).

Das Vorhandensein von Prosodie ist eine der Universalien menschlicher Sprache (Hocket 1963, nach Hirst und Di Cristo 1998). Beim Spracherwerb wird die Prosodie als erstes erlernt und bleibt bei Aphasikern als letzte erhalten (Neppert 1999, S. 155). Bei den prosodischen Merkmalen handelt es sich „um Merkmale der Lautbildung, die auf den am wenigsten spezialisierten Aktivitäten des menschlichen Stimmapparates beruhen“ (Günther 1999).

Prosodie ist also ein elementarer Bestandteil menschlicher Sprache. Vielleicht ist gerade ihre Allgegenwärtigkeit hinderlich bei ihrer linguistischen Beschreibung. Eine detaillierte und allgemeingültige Definition der Prosodie steht jedenfalls aus:

> Auch wenn der Gebrauch dieses Begriffs bis in die griechische Antike zurückverfolgt werden kann, muß noch heute am Anfang einer Arbeit zum Phänomenbereich der Prosodie definiert werden, was im folgenden unter dem Begriff der Prosodie im einzelnen zu verstehen sein soll.
>
> —(Günther 1999, S. 15)

Ähnlich kontrovers wird der Begriff der Intonation benutzt. Das Spektrum reicht vom synonymen Gebrauch mit Prosodie (vgl. Hirst und Di Cristo 1998, S. 3; Bußmann 1990, S. 352) bis zur Gleichsetzung von Intonation mit dem Grundfrequenzverlauf in Literatur zu TTS-Systemen. Auch Intonation bedarf also einer eingehenden Bedeutungsbeschreibung.

Beide Begriffe müssen jeweils auf phonologischer und auf phonetischer Ebene betrachtet werden. Insbesondere die komplizierte Abbildung der phonologischen Teilsysteme der Prosodie auf phonetische Systeme und Merkmale erschwert die Situation: „the correspondence between abstract prosodic characteristics and acoustic features is far from simple.“ (Hirst und Di Cristo 1998, S. 5)

Die folgenden Abschnitte stellen Prosodie als Menge von phonologischen und phonetischen Systemen vor, erläutern ihre Verhältnisse untereinander, zu den realisierten suprasegmentalen Merkmalen und ihrer jeweiligen akustischen Repräsentation.

2.2. Phonologische Teilsysteme

Hirst und Di Cristo (1998, S. 5) ordnen die Prosodie in lexikalische und nicht-lexikalische Teilsysteme.

Als lexikalische Teilsysteme betrachten sie Akzent, Ton und Quantität. Weder Ton noch Quantität müssen im Deutschen als eigene Systeme angesehen werden: Töne kommen (anders als in *Tonsprachen* wie im Chinesischen oder eingeschränkt im Schwedischen) nicht vor.

Die *Quantität* beschränkt sich im Deutschen auf die Unterscheidung von langen und kurzen (bzw. *gespannten* und *ungespannten*) Vokalen. In dieser Arbeit werden die gespannten und ungespannten Vokale bereits auf segmentaler Ebene unterschieden, sodass die Quantität hier kein suprasegmentales Merkmal darstellt.

Die nicht-lexikalischen Teilsysteme werden von Hirst und Di Cristo als Intonation zusammengefasst. In Anlehnung an Günther (1999) werden hier die Teilsysteme Junktur und Intonation weiter unterschieden.

Betrachten wir also die drei Teilsysteme Akzent, Junktur und Intonation.

2.2.1. Akzent

Akzent (auch *Wortakzent*, engl. stress) bezeichnet die Stellen von Wörtern, die akzentuiert, also hervorgehoben werden *können*: „Word stress […] is the position in a word to which a phonetic accent may be assigned." (Gibbon 1998, S. 80)

Es stellen sich zwei Fragen: Welches sind die Stellen, die hervorgehoben werden können? Und: Welcher Art sind diese Stellen? Letztere Frage, die Frage nach den *Konstituenten*, den akzenttragenden Elementen wird zuerst beantwortet.

Chomsky und Halle (1968, nach Hirst und Di Cristo 1998, S. 8) gingen noch davon aus, dass der Akzent ein segmentales Merkmal der Vokale in einem Wort ist. Die Versprecherforschung widerlegt diese Annahme, da das Vertauschen von Vokalen in einem Wort die Akzentposition meist nicht mitvertauscht:

Wohnmo*BIL* – Wohnmi*BOL*[1]

Akzent wird deswegen besser als abstraktes Merkmal auf Silbenebene repräsentiert. Welche möglicherweise akzenttragenden Silben sind nun akzentuiert? Akzent ist, wie bereits erwähnt, ein lexikalisches System. Teilweise dient die Akzentposition der Unterscheidung von segmental homophonen Wörtern:

um*FAHR*en – *UM*fahren

*TE*nor – Te*NOR*

Insbesondere die Ungleichverteilung der Akzente spricht aber dagegen, dass die Akzentposition systematisch zwischen Wörtern unterscheidet. Mengel (2000, S. 161) berichtet, dass bei zweisilbigen Wörtern das Verhältnis Erstakzentuierung zu Zweitakzentuierung bei 6:1 liegt. Dies spricht für eine überwiegend regelgeleitete Zuweisung des Wortakzents.

Nach Gibbon (1998, S. 80) tragen native Wortstämme des Deutschen den Akzent auf der ersten Silbe, während er bei fremden Wortstämmen auf der letzten oder vorletzten Silbe liegt. Es liegt also nahe, dass der Akzent fremder Wörter lexikalisch repräsentiert wird, während er für native Stämme fest auf der ersten Silbe liegt. Für manche Fremdwörter kann sich zudem der Wortakzent verschieben wie in „Marzipan" (ursprünglich Marzi*PAN*, häufig auch *MAR*zipan; Mengel 2000, S. 177).

Affixe, die Wortstämme zu Wörtern ableiten, können (1) den Akzent innerhalb des Stammes verändern, (2) den Akzent tragen oder (3) sich neutral verhalten (Gibbon 1998, S. 81). Durch die Ableitung ergibt sich insgesamt ein freier Akzent: Die akzentuierte Silbe eines Wortes kann prinzipiell an jeder Position stehen.[2]

Die Kompositabildung im Deutschen führt dazu, dass Wörter mehr als einen Akzent tragen können. Die Akzente werden dann in einer *Akzenthierarchie* in *Haupt-* und *Nebenakzent* unterteilt (Neppert 1999, S. 167).

Die Rangfolge der Akzente innerhalb der Akzenthierarchie bestimmt sich aus der Semantik des Kompositums: Im Normalfall beschreibt die erste Komponente die zweite und erhält deswegen den Hauptakzent. Verstärkt hingegen die erste Komponente die zweite, beschreibt sie aber nicht näher, erhält sie den Nebenakzent (Mengel 2000, S. 11). Einige Komposita sind gerade dadurch Minimalpaare:

*VOLL*zug (im Eisenbahnwesen) – Voll*ZUG* (in der Justiz)[3]

Die *metrische Phonologie* spezifiziert die Akzenthierarchie weiter. Nicht nur die Rangfolge von Akzenten untereinander, sondern alle Silben werden hinsichtlich ihres Akzents strukturiert. Die genaue Dar-

[1] Beispiel aus (Günther 1999, S. 172).
[2] Im Gegensatz dazu hat beispielsweise das Ungarische einen festen Akzent: Er liegt immer auf der ersten Silbe eines Wortes.
[3] Beispiel aus (Zehnpfenning 2005, S. 9).

stellung dieser Theorien würde hier den Rahmen sprengen, insbesondere weil im Praxisteil keine weitere Strukturierung der Silben im Wort vorgenommen wird.

2.2.2. Junktur

Die *Junktur* strukturiert den Sprechstrom durch Grenzsignale. Neppert versteht unter der Junktur die „mehr oder weniger verbindende oder abgrenzende Übergangsart", die „durch ihre Gliederungsfunktion in erheblichem Maße zur Verstehbarkeit des Gesprochenen" beiträgt (Neppert 1999, S. 190).

Als Grenzsignale wirken alle suprasegmentalen Merkmale, Eigenschaften der segmentalen Elemente (Dehnung, Aspiration, Qualität) sowie segmentale Elemente ohne eigenen Phonemstatus. Letzteres betrifft den Glottalverschluss vor wortinitialen Vokalen im Deutschen und Pausen im Allgemeinen. Im Regelfall wird ein Grenzsignal durch eine Kombination von Merkmalen repräsentiert (Neppert 1999, S. 156).

Neppert bezeichnet auch die Signale selbst als *Junkturen* und unterscheidet interne geschlossene und externe offene Junkturen (Neppert 1999, S. 188). *Externe offene* Junkturen liegen immer an Morphem- oder Wortgrenzen. Nur sie können akustisch manifestiert sein. *Interne geschlossene* Junkturen hingegen stehen für die Tatsache, dass eben keine mögliche Grenze vorliegt und deswegen auch keine Grenzsignale manifestiert sein können.

Im folgenden (englischen) Beispiel ist die Silbengrenze von „nitrate" nicht markiert, da es sich um eine interne geschlossene Junktur handelt. „night rate" hingegen hat eine externe offene Junktur, sodass die Silben stärker voneinander abgegrenzt sind.

„nitrate" `/naItrEIt/` – „night rate" `/naIt.rEIt/`[4] [5]

Die Junktur hilft also dabei, die im Sprechstrom enthaltenen Silben, Morpheme und Wörter zu dekodieren. Sie ist nicht nur verständnisunterstützend, sondern häufig auch disambiguierend. Das beliebte Beispiel „Staubecken" ist einerseits in Flusstälern andererseits in Studentenwohnungen anzutreffen.

Die Gliederungsfunktion der Junktur wirkt auf allen prosodischen Ebenen. Was sind also die Konstituenten der Junktur? Untere anderem zeigt Pétursson (1978, nach Neppert 1999, S. 190), dass externe offene Junkturen nur an Morphemgrenzen auftreten können, sodass kleinere Konstituenten für Junkturen ausgeschlossen sind.

Da Junkturen auf mehreren Ebenen wirken, liegt es nahe, dass die Konstituenten mit der jeweiligen Funktion verbunden sind: Signale zur Abgrenzung zwischen Morphemen wirken auf Silbenebene beziehungsweise an den Grenzen zwischen Silben. Signale zur Abgrenzung von Wörtern und größeren Einheiten wirken an den respektiven Grenzen, teilweise schon im Vorfeld der Grenze.

Die Junktur bestimmt insbesondere auch den Zusammenhalt zwischen Wörtern. Vor allem werden durch Grenzsignale Wortgruppen voneinander abgegrenzt. Die Funktion dieser Abgrenzung ist sowohl syntaktisch, als auch semantisch motiviert, jedenfalls werden strukturelle Grenzen im Sprechstrom markiert (Günther 1999, S. 70).

Damit steuert die Junktur die Einteilung des Gesagten in Phrasen. Im Rahmen dieser Arbeit ist besonders die Phrasierungsfunktion der Junktur wichtig.

Zur Konstituentenstruktur von Phrasen gibt es widersprüchliche Ansichten. Als kleinstes mögliches Element der Phrasierung gilt das Wort. Dies verdeutlicht auch folgendes Beispiel, bei dem innerhalb der Wörter jeweils keine phrasierende Junktur vorkommen kann:

Fass voll Bier – Fass Vollbier – Fassvollbier[6]

[4] Beispiel aus (Neppert 1999, S. 188).
[5] In dieser Arbeit werden phonologische und phonetische Transkriptionen durchgehend in SAMPA (Wells o. J.) dargestellt. Siehe auch Abschnitt 4.1.2.
[6] Beispiel aus (Zehnpfenning 2005, S. 9).

Viele Autoren unterscheiden darauf aufbauend unterschiedlich viele verschachtelte Ebenen (Grice und Baumann 2000, S. 28). Der im folgenden Abschnitt vorgestellte Tonsequenzansatz unterscheidet die intermediäre Phrase und die aus intermediären Phrasen aufgebaute Intonationsphrase.

Die *Intonationsphrase* entspricht in etwa einem Satz (Grice und Baumann 2000, S. 29), wobei *Satz* „als eine in sich geschlossene Ausdruckseinheit“ (Neppert 1999, S. 157) zu verstehen ist.

Phonetisch betrachtet sind Äußerungen aus Expirationsgruppen aufgebaut. Eine *Expirationsgruppe* ist „eine nicht durch Atmungsvorgänge unterbrochene Äußerung“ (Neppert 1999, S. 161). Auf der phonologischen Ebene entsprechen Intonationsphrasen den phonetischen Expirationsgruppen.

Nicht nur das eigentliche Atemholen, auch kurze Unterbrechungen des Redeflusses können schon als Atmungsvorgang innerhalb einer Expirationsgruppe gelten. Die innerhalb der Intonationsphrasen liegenden *intermediären Phrasen* werden durch solche kürzeren Unterbrechungen markiert.

Beide Ebenen der Phrasen (intermediäre Phrasen und Intonationsphrasen) entsprechen syntaktischen oder semantischen Einheiten. Sie unterstützen dadurch den Hörer bei der Dekodierung des Gesprochenen.

2.2.3. Intonation

Intonation ist „die distinktive Verwendung prosodischer Eigenschaften zur Bedeutungsdifferenzierung ganzer Äußerungen.“ (Nöth 1990, S. 24 nach Günther 1999, S. 62)

Die Intonation erreicht die Bedeutungsdifferenzierung durch Merkmale, unter denen die Melodiebewegung, mit ihrem akustischen Korrelat Grundfrequenzverlauf, die wichtigste ist (Neppert 1999, S. 158). Die *Melodiebewegung* soll hier durch phonologisch abstrakte *Töne* beschrieben werden, deren Aufbau später erläutert wird. Zunächst muss aber der Funktion und Wirkung von Intonation Aufmerksamkeit geschenkt werden.

Die Junktur gibt zwar die grobe Gliederung der Äußerung durch die Unterteilung in Intonationsphrasen vor. Der *Typ* der Phrasen wird durch Junkturen jedoch nicht weiter spezifiziert. Ebenso können Junkturen zwar die Stärke der Dissoziation zwischen Phrasen, nicht aber die Art beschreiben.

Die Intonation verstärkt und spezifiziert die von der Junktur vorgegebenen Intonationsphrasen durch Intonationsmuster, die die Phrase und ihre Grenzen näher beschreiben (Neppert 1999, S. 162).

Intonationsmuster haben unterschiedliche syntaktische und semantische Funktionen. Die *Frageintonation*, die bei Ja/Nein-Fragen besonders ausgeprägt ist, sowie die weiterführende *Zwischengrenze* äußern sich durch ein Anheben der Tonhöhe vor der Grenze. Eine gleichbleibende Tonhöhe am Phrasenende hat im Diskurs eine weiterführende Funktion (Neppert 1999, S. 162).

Ein Tonabfall am Ende der Intonationsphrase resultiert in der *unmarkierten Exspirationsgruppe*. Diese wird häufig als Universalie angenommen, da sie automatisch durch den über den Verlauf der Äußerung sinkenden subglottalen Druck entsteht (Neppert 1999, S. 161).

In gleicher Weise wie bei der Junktur, gibt der Akzent lediglich die möglichen Hervorhebungen innerhalb eines Wortes vor. Im Folgenden wird mit *Akzentuierung* ein tatsächlich realisierter Akzent, also eine tatsächliche Hervorhebung bezeichnet. Die Selektion der Akzente in der Äußerung zu Akzentuierungen ist Aufgabe der Intonation.

Die Bestimmung der Akzentuierungen verläuft auf Wortebene. Dabei werden solche Wörter hervorgehoben, die den hauptsächlichen semantischen Inhalt der Äußerung repräsentieren (Neppert 1999, S. 164). Die Auswahl der zu akzentuierenden Wörter heißt auch *Satzakzent* (Günther 1999, S. 48). Er disambiguiert zwischen unterschiedlichen Lesarten der Äußerung und hilft bei der Gliederung des Gesagten in Fokus und Hintergrund.

Zusammengesetzte Wörter tragen mehrere, möglicherweise durch eine Akzenthierarchie geordnete Akzente, aus denen die zu realisierenden selektiert werden müssen. Die Auswahl der Akzentuierung beruht dabei auf semantischen, kontrastiven sowie rhythmischen Grundlagen.

*EIN*hundertdreiundzwanzig – einhundert*DREI*undzwanzig –

– einhundertdreiund*ZWAN*zig – ein*HUN*dertdreiundzwanzig[7] –
– *einhun*DERT*dreiundzwanzig – *einhundertdrei*UND*zwanzig

Im Beispiel sind mögliche Akzentuierungen dargestellt, die je nach pragmatischem Ziel geäußert werden können. Beachte: Die letzten beiden Beispiele sind ungrammatisch, weil diese Silben keinen Akzent tragen, also nicht akzentuiert werden können.

Akzentuierungen werden durch die Intonation mit unterschiedlichen *Intonationskonturen* verknüpft. Zum Aufbau dieser Intonationskonturen und dem Zusammenspiel der Phrasenintonation und Akzentuierungsintonation gibt es zwei hauptsächliche, konkurrierende Theorien (Günther 1999, S. 63).

In holistischen Ansätzen werden den Äußerungen und Äußerungsteilen als ganzes komplexe Intonationskonturen zugewiesen. Dabei bleibt die genaue Verteilung der Konturteile auf die Silben teilweise offen (Günther 1999, S. 64). Nach der *Superpositionstheorie* ist die Intonation hierarchisch aufgebaut: Der Intonationsverlauf der Äußerung entsteht aus der Summe der globalen Satzintonation und lokalen Akzentuierungen.

Die *Tonsequenztheorie* (Pierrehumbert 1980) hingegen nimmt keine Hierarchie von Intonationskonturen, sondern eine Sequenz von Akzent-, Phrasen- und Grenztönen an. Auf der Tonsequenztheorie baut ToBI (*tones and break indices*, Pierrehumbert 1980), ein Modell zur Beschreibung intonatorischer Ereignisse auf. ToBI beziehungsweise ToBI-Derivate haben sich zur prosodischen Annotierung durchgesetzt.

Das für das Deutsche überwiegend verwendete Modell ist GToBI (Grice et al 1996, Reyelt et al 1996 nach Grice und Baumann 2000). Deswegen wird im Folgenden auf die Tonsequenztheorie und GToBI weiter eingegangen.

Töne bestehen aus einem oder mehreren atomaren Elementen. Diese liegen entweder auf der oberen oder unteren tonalen Lage und werden daher mit `H` (*hoch*, engl. high) beziehungsweise `L` (*niedrig*, engl. low) bezeichnet.[8].

Akzenttöne können aus mehreren atomaren Elementen bestehen. Die einzelnen Elemente sind dann aufeinander folgenden Silben zugeordnet. Das atomare Element, das der akzentuierten Silbe zugeordnet ist, wird durch einen nachgestellten Stern (`*`) gekennzeichnet. Der Akzentton `H*L` beschreibt also eine Akzentuierung, bei der einem Gipfel auf der akzentuierten Silbe ein deutlich hörbarer Abfall folgt. Dieser dem Ton zugehörige Abfall unterscheidet `H*L` von `H*`.

Phrasenton und Grenzton bestehen nur aus je einem der Elemente `H` und `L`. Der *Phrasenton* markiert das Ende intermediärer Phrasen und wird mit einem Minus (`-`) markiert.

Die Wirkung des Phrasentons erstreckt sich bis zum Ende der intermediären Phrase. Sie beginnt auf dem Akzent, der dem letzten Akzentton in der intermediären Phrase folgt (Grice und Benzmüller 1998, nach Grice und Baumann 2000, S. 29).

Der *Grenzton* steht am Ende der Intonationsphrase und wird mit einem Prozentzeichen (%) markiert. Jede Intonationsphrase enthält mindestens eine intermediäre Phrase. Deswegen fällt der Grenzton immer mit einem Phrasenton zusammen und kann folglich indirekt (wenn sich Phrasenton und Grenzton unterscheiden) aus zwei Elementen bestehen.

Die genaue Realisierung des Übergangs zwischen hoher und tiefer Lage (genaue Position, Steilheit, Höhenunterschied, …) wird in der phonemischen Ebene nicht beschrieben, um eine abstrakte und nur funktional notwendige Repräsentation der Intonationskonturen zu erreichen. Die genaue Realisierung der Töne ist Gegenstand der phonetischen Ebene.

Zusätzlich zu den Tönen wird in GToBI noch ein *break index* angegeben. Im ursprünglichen ToBI war der break index als unabhängige Beschreibungsebene vorgesehen, um die Stärke von Grenzen im Verlauf der Äußerung zu beschreiben (Pierrehumbert 1980).

[7] Zum Beispiel als Korrektur wenn der Hörer unsinnigerweise „ein Hund hat dreiundzwanzig“ verstanden hatte.
[8] Einige Derivate (Féry 1993, nach Grice und Baumann 2000, S. 24) nehmen eine dritte Lage `M` (*mittig*, engl. middle) an, die im aktuellen GToBI (Grice und Baumann 2000) abgelehnt wird

Im aktuellen GToBI ist die Zuordnung von break index zu Phrasierung hingegen eineindeutig: Intermediäre Grenzen werden mit *3*, Grenzen zwischen Intonationsphrasen mit *4* gekennzeichnet.

Zusammenfassung

Auf der Ebene der Phonologie wurde der *Akzent* betrachtet, der mögliche Akzentuierungsstellen vorgibt. Die *Junktur* unterteilt eine Äußerung in Phrasen. Die *Intonation* nimmt auf Grundlage der beiden zuvor genannten Module die prosodische Ausgestaltung der Äußerung durch die Kennzeichnung mit Akzenttönen, Phrasentönen und Grenztönen vor.

2.3. Phonetische Phänomenbereiche und Merkmale

Akzentuierung und Phrasierung bilden die beiden phonetischen Phänomenbereiche der Prosodie. Sie entstehen als Resultat der abstrakten prosodischen Merkmale, die durch die phonologischen Teilsysteme bedingt sind. Akzentuierung und Phrasierung wiederum äußern sich in einer großen Vielfalt von prosodischen Merkmalen unter denen die Melodiebewegung das prominenteste ist.

Die Melodie bewegt sich innerhalb eines gewissen Rahmens, der unten und oben von der *Grundlinie* und der *Dachlinie* begrenzt ist. Die beiden Linien fallen im Verlauf der Intonationsphrase etwas ab, was als *Deklination* bezeichnet wird.

Die *Akzentuierung* entsteht als Resultat der Verbindung von Akzent und intonatorischen Akzenttönen. Sie äußert sich besonders, aber nicht ausschließlich durch das intonatorische Tonmuster. Die Akzentuierung hat außerdem einen Einfluss auf die Sprechintensität und führt zur Längung des Silbenkerns. Die Längung und die höhere Artikulationsintensität führen zu Qualitätsänderungen: Die Formanten des Silbenkerns sind in akzentuierten Silben besonders ausgeprägt.

Die Phrasierung entsteht aus der Verbindung von Junktur und intonatorischen Phrasen- und Grenztönen. Die intermediären Grenzen und die Intonationsphrasengrenzen, die durch Phrasen- beziehungsweise Grenztöne angezeigt werden, unterscheiden sich nicht systematisch voneinander. Vielmehr fallen intermediäre Grenzen (im Folgenden *Zwischengrenzen*) hauptsächlich nur schwächer aus als Intonationsphrasengrenzen (im Folgenden *Vollgrenzen*).

Die Phrasen- und Grenztöne werden ähnlich den Akzenttönen realisiert. Allerdings ist der realisierte Verlauf stärker auf die Grenze ausgerichtet als es bei den Akzenttönen mit den zugrundeliegenden Silben der Fall ist.

Hauptmerkmal der Phrasengrenzen ist die Pause, die an Vollgrenzen fast immer und an Zwischengrenzen häufig auftritt. Weiteres konsistentes Merkmal ist die Dehnung der letzten Silbe vor der Grenze, die sich auf die Dauer der beteiligten Phoneme auswirkt (Neppert 1999, S. 160).

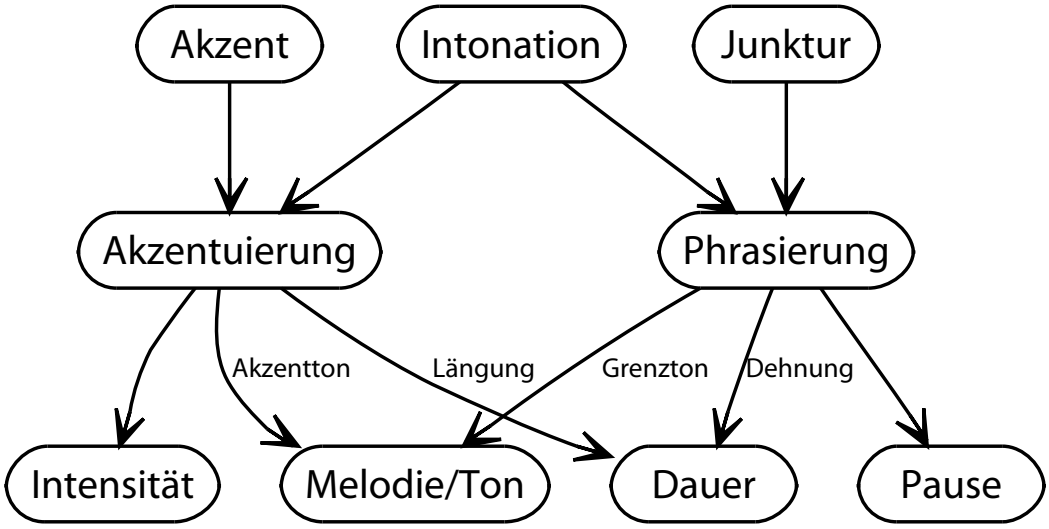
An Vollgrenzen werden außerdem Grund- und Dachlinie für die nächste folgende Frage rückgesetzt, sodass die nächste Intonationsphrase wieder etwas höher beginnt.

Die Phänomene Akzentuierung und Phrasierung werden also durch die prosodischen auditorischen Merkmale Lautheit und Qualität, Tonhöhe und Dauer repräsentiert, denen die akustischen Parameter Schalldruckpegel und Spektralcharakter, Grundfrequenz und Zeit entsprechen.

Abbildung 2.1 zeigt den skizzierten Zusammenhang der phonologischen Teilsysteme Akzent, Junktur und Intonation, ihren gemeinsamen Einfluss auf die phonetischen Phänomenbereiche Akzentuierung und Phrasierung sowie deren Einflüsse auf die prosodischen Merkmale Intensität, Melodieverlauf, Dauer und Pause[9].

[9] Dauer und Pause stehen eng miteinander in Beziehung. In einer nichtlinearen Phonologie in der anstatt von Phonemen Lautbildungskommandos angenommen würden, entspräche die Pause einfach dem längeren Ausbleiben eines solchen Kommandos.

Abbildung 2.1. Abhängigkeiten zwischen den prosodischen Teilsystemen



2.4. Mikroprosodie

Mit *Mikroprosodie* werden „kontextgebundene Variationen“ (Neppert 1999, S. 174) der prosodischen Merkmale bezeichnet. Günther führt aus, „die Realisierung gleicher abstrakter prosodischer Muster führt in unterschiedlichen segmentalen Kontexten zu verschiedenen akustischen Parametern.“ (Günther 1999, S. 88)

Die Mikroprosodie wird durch die jeweilige Stellung der Artikulationsorgane bei der Lautbildung bedingt. So haben hohe Vokale eine höhere Grundfrequenz, geringere Intensität und geringere Dauer als tiefe Vokale (Günther 1999, S. 88).

Für die unterschiedlichen Laute kann jeweils die inhärente Tonhöhe, Intensität und Dauer angegeben werden und die Prosodie als Abweichung der Parameter von diesen inhärenten Werten angesehen werden. Im System der Abweichungen bestehen komplizierte Zusammenhänge, da die lautinhärenten Parameter durch die Artikulation bestimmt sind. Die *makroprosodische* Realisierung wird dadurch eingeschränkt (Günther 1999, S. 89).

Weiteren Einfluss haben Übergänge zwischen Konsonanten und Vokalen. Im Übergangsbereich zwischen Konsonant und Vokal sowie im folgenden Übergang zum nächsten Konsonanten kommt es zu Einstellungsbewegungen der Artikulationsorgane sowie Änderungen im Luftdurchfluss an den Stimmlippen (sogenannte *Koartikulationseffekte*). Dadurch bestimmen auch benachbarte Konsonanten lautinhärente Parameter von Vokalen (Günther 1999, S. 89).

Eine wichtige Einschränkung ergibt sich aus der Stimmhaftigkeit: Nur stimmhafte Abschnitte im Signal können den Melodieverlauf repräsentieren. Die tonale Realisierung von Intonationskonturen beschränkt sich also auf solche Abschnitte oder muss stimmlose Abschnitte überspannen, um hörbar zu sein.

An dieser Stelle sei angemerkt, dass dennoch auch beim Flüstern die Prosodie und Intonation deutlich wahrnehmbar sind, obwohl das wichtigste konstituierende Merkmal wegen der Stimmlosigkeit geflüsterter Äußerungen ausfällt (Neppert 1999, S. 158). Dies weist darauf hin, dass die Intonationskontur neben der Tonhöhe noch weitere wichtige Konstituenten besitzt.

Auch der als Deklination bezeichnete Abfall der Tonhöhe über die Intonationsphrase kann der Mikroprosodie zugeschlagen werden. Die Deklination wird in der Tonsequenztheorie als rein physiologisches Phänomen betrachtet. Sie entsteht durch das Nachlassen des subglottalen Druckes und damit der Stimmlippenschwingung über den Verlauf der Exspirationsgruppe (Neppert 1999, S. 161).

Meiner Meinung nach entsprechen die unter Mikroprosodie zusammengefassten Effekte insgesamt der Koartikulation. Die Koartikulation muss zu diesem Zweck allerdings weiter gefasst werden, um nicht nur lautnachbarschaftliche Effekte, sondern alle Effekte durch die gleichzeitige Realisierung von segmentalen und suprasegmentalen Merkmalen abzudecken.

Insbesondere die Deklination lässt sich dann als koartikulativer Effekt beschreiben, der sich aus der Position innerhalb der Phrase ergibt. Daneben wird auch die Intensität von Äußerungsteilen von der Position innerhalb der Phrase bestimmt: Sie lässt über den Phrasenverlauf nach. Auch dies erachte ich aber als hauptsächlich koartikulativen Effekt. Deswegen wirkt sich in Abbildung 2.1 das phonetische Phänomen Phrasierung nicht auf das prosodische Merkmal Intensität aus.

Kapitel 3. Maschinelles Lernen

Maschinelles Lernen ist ein Teilgebiet der künstlichen Intelligenz. Es verfolgt das Ziel, Algorithmen zu entwickeln, die mithilfe von Trainingsmaterial lernen, bestimmte Aufgaben möglichst gut zu lösen.

John (1997, Kapitel 1.3) unterscheidet folgende *Aufgabentypen*:

- Klassifizierung: Bestimmung der Zugehörigkeit von Instanzen zu Klassen
- Numerische Vorhersage: Zuordnung eines Wertes zu einer Instanz
- Assoziation: Herstellung von Zusammenhängen zwischen Merkmalen der Instanzen
- Clusterung: Abgrenzung von in sich ähnlichen und untereinander verschiedenen Anhäufungen von Instanzen

Das *Trainingsmaterial* enthält Beispielinstanzen des Problems. Diese liegen in Form von Vektoren vor, deren einzelne Komponenten die Merkmale der Instanz beschreiben. Es gibt kategorielle und kontinuierliche Merkmale (Witten und Frank 2000, S. 51): *Kategorielle* Merkmale können nur Werte aus einer bestimmten Menge von Kategorien annehmen, wie zum Beispiel den Wert `/o:/` aus der Menge der Vokale. Im Gegensatz dazu nehmen *kontinuierliche* Merkmale beliebige Werte innerhalb eines Wertebereichs an, wie zum Beispiel Grundfrequenz oder Dauer.

Die Aufgabentypen unterscheiden sich hinsichtlich des benötigten Trainingsmaterials: Für Klassifizierung und numerische Vorhersage wird die Zielklasse beziehungsweise der Zielwert den Beispielinstanzen hinzugefügt. Der Lernalgorithmus ermittelt dann strukturelle Muster in den Merkmalen der Beispielinstanzen, die mit den unterschiedlichen Klassen korrelieren (*überwachtes Lernen*). Die Zielklassen der Beispielinstanzen werden typischerweise aufwendig von Menschen handannotiert. Das Ergebnis ist ein *Klassifizierer* (für numerische Vorhersagen ein Prädiktor), der anhand der Merkmale einer nichtklassifizierten Instanz deren Klassenzugehörigkeit vorhersagt.

Assoziations- und Clusterungsalgorithmen erhalten keine Zielvorgaben sondern finden selbständig geeignete Verknüpfungen innerhalb des Trainingsmaterials (*unüberwachtes Lernen*). Diese Arbeit behandelt ausschließlich Klassifizierungsaufgaben. Deshalb wird nur auf diese im Folgenden näher eingegangen.

3.1. Evaluierung

Was bedeutet Lernen? Wodurch zeichnet sich erfolgreiches Lernen aus? Gerade für maschinelles Lernen steht das operationale Verhalten des lernenden Algorithmus im Vordergrund, da Aussagen über den „Erwerb von [...] Kenntnissen und Fertigkeiten oder Fähigkeiten“[1] eines Automaten anderweitig (beispielsweise durch Introspektion) unmöglich sind. Zur Bewertung seines operationalen Verhaltens wird der Klassifizierer auf *Testmaterial* angewendet, für das ebenfalls eine Handannotierung der Zielklassen (der *gold standard*) vorliegt. Die Klassifizierung durch den Algorithmus wird dann mit dem gold standard verglichen und daraus werden Leistungsmaße über die Klassifizierungsgüte des Algorithmus abgeleitet.

3.1.1. Methodik

Sowohl zum Training als auch zur Evaluierung wird also annotiertes Material benötigt, welches nur in begrenztem Umfang verfügbar ist. Eine strikte Trennung von Trainings- und Testmaterial ist jedoch für eine objektive Bewertung der Lernleistung unerlässlich. Üblich ist eine Teilung von 90 % Trainings- zu 10 % Testmaterial.

Sofern ein Lernverfahren im Zuge seiner Lerntätigkeit (zum Beispiel zur Einstellung von Parametern) eigenes Testmaterial benötigt, so darf dafür unter keinen Umständen das Testmaterial benutzt werden! Sonst würde nicht sauber zwischen Trainings- und Testmaterial unterschieden und das Ergebnis der Evaluierung zu optimistisch ausfallen. In solchen Fällen wird vom Trainingsmaterial ein kleiner Teil

[1] http://de.wikipedia.org/wiki/Lernen

als *Holdout-Menge* abgetrennt. Das endgültige Training nach Einstellung der Parameter kann dann auf dem gesamten Trainingsmaterial inklusive der Holdout-Menge erfolgen.

Die Aufteilung des handannotierten Materials in die unterschiedlichen Teilmengen erfolgt entweder zufällig oder als *geschichtete Stichprobe* (auch stratifizierte Stichprobe, *stratified sample*), die sicherstellt, dass die Stichprobe bezüglich ausgewählter Schichtungsmerkmale der Grundgesamtheit möglichst genau entspricht (Ludwig-Mayerhofer et al o. J.). Geschichtete Stichproben erlauben bessere Vorhersagen als rein zufällige, jedoch hängt dieser Gewinn an der nichttrivialen Auswahl der Schichtungsmerkmale, die möglichst alle relevanten aber keine irrelevanten Merkmale umfasst.

Die notwendige Unterteilung in Trainings- und Testmaterial macht es zunächst unmöglich, das gesamte annotierte Material zum Training zu nutzen. Eine Abhilfe schafft *k-fache Kreuzvalidierung*, bei der das annotierte Material in `k` Teilmengen geteilt wird. Training und Test werden nun `k`-mal ausgeführt, wobei jeweils die `k`-te Teilmenge als Testmaterial und die restlichen Teilmengen gemeinsam als Trainingsmaterial benutzt werden. Als Leistung des Lernverfahrens wird dann die mittlere Leistung der einzelnen Klassifizierer angegeben. Kohavi (1995) ermittelt auf Basis einer breiten Auswahl von Daten aus unterschiedlichen Domänen 10-fach stratifizierte Kreuzvalidierung als beste uninformierte Wahl für Evaluierungen.

3.1.2. Leistungsmaße

Für die Analyse der Lernleistung eines Klassifizierers stehen unter anderem die *quantitativen* Werte Korrektheit, Reinheit, Vollständigkeit sowie F-Maß zur Verfügung, die aus der Vertauschungsmatrix abgeleitet sind (Witten und Frank 2000, Kapitel 5). Leistungsmaße ersetzen nur bedingt eine *manuelle Fehleranalyse*, da nur diese *qualitative* Aussagen über die Leistung eines Klassifizierers erlaubt.

Die *Vertauschungsmatrix* (engl. confusion matrix) enthält detaillierte Angaben über die Verteilung der Klassifizierungsergebnisse bezogen auf die Klassifizierung im gold standard. Für ein Zwei-Klassen-Problem (mit den Klassen *positiv* `p` und *negativ* `n`) ergibt sich eine Matrix wie in Tabelle 3.1. Die einzelnen Zellen zählen Instanzen des Testmaterials, die korrekt als positiv (`tp`, true positive), falsch als positiv (`fp`, false positive), falsch als negativ (`fn`, false negative) und korrekt als negativ (`tn`, true negative) erkannt werden. Für Klassifizierungsaufgaben mit mehr als zwei Klassen ergibt sich eine größere Vertauschungsmatrix aus der zum Beispiel abgelesen werden kann, welche Klassen mit welchen anderen verwechselt und welche Klassen vergleichsweise gut erkannt werden. Die verwendeten Variablen `fp`, `fn` und `tn` werden für mehrklassige Aufgaben jeweils als Summen der entsprechenden Felder errechnet.

Tabelle 3.1. Vertauschungsmatrix eines Zwei-Klassen-Problems

wirkliche Klasse	erkannte Klasse	
	p	**n**
p	`tp`	`fn`
n	`fp`	`tn`

Die *Korrektheit* (engl. correctness, `c`) des Klassifizierers ist der Anteil der korrekten Vorhersagen bezogen auf die Gesamtzahl der Vorhersagen:

$$c = (tp + tn) / (tp + fn + fp + tn)$$

Insbesondere bei ungleichverteilten Klassen ist die Korrektheit wenig aussagekräftig, da sie sich nicht auf eine bestimmte Klasse, sondern auf das Gesamtergebnis bezieht: Die Korrektheit kann sehr hoch sein, obwohl für die relevante Klasse kaum korrekte Entscheidungen gefallen sind.

Reinheit (engl. precision, `p`) und *Vollständigkeit* (engl. recall, `r`) sind jeweils auf eine Klasse bezogene Leistungsmaße, die angeben, wie viele der als positiv erkannten Instanzen tatsächlich positiv sind beziehungsweise wie viele der positiven Instanzen tatsächlich als positiv erkannt wurden:

$$p = tp / (tp + fp)$$

$$r = tp / (tp + fn)$$

Das *F-Maß* (engl. f-measure, `f`; van Rijsbergen 1979) vereinigt Reinheit und Vollständigkeit in genau einem Wert, was den direkten Vergleich zweier Lernverfahren erleichtert.

$$f = (2 \cdot r \cdot p) / (r + p)$$

Bei einem Mehrklassenproblem ergibt sich für die Klassen jeweils ein eigenes F-Maß. Die Leistung zweier Klassifizierer kann so noch nicht auf einer linearen Skala verglichen werden. Zu diesem Zweck wird das *kombinierte F-Maß* ($\mathtt{f}_{\mathtt{komb}}$) definiert: Es ist der Mittelwert der F-Maße für die relevanten Klassen.

$$f_{komb} = avg(f_{1..n})$$

3.1.3. Interpretierbarkeit

Neben hoher Klassifizierungsgüte wird für diese Arbeit außerdem eine gute Interpretierbarkeit der trainierten Klassifizierer angestrebt. Dadurch können qualitative Aussagen über die Adäquatheit des Erlernten gemacht werden und außerdem die Relevanz der einzelnen Merkmale in Hinsicht auf die Ermittlung der Klassenzugehörigkeit bestimmt werden. Die Auswahl der in der Arbeit benutzten Klassifizierungsalgorithmen ist auch im Hinblick darauf erfolgt.

3.2. Klassifizierungsalgorithmen

Für die Klassifizierungsaufgaben in dieser Arbeit werden zwei unterschiedliche Verfahren der Klassifizierung benutzt und im folgenden mit ihren zugehörigen Lernverfahren vorgestellt.

3.2.1. Naïve-Bayes

Ein Naïve-Bayes-Klassifizierer (Duda und Hart 1973; Witten und Frank 2000, Kapitel 4.2) ordnet eine Instanz mit Merkmalsvektor `X` derjenigen Klasse `c` zu, der sie mit größter Wahrscheinlichkeit angehört (Auswahl der maximalen Wahrscheinlichkeit, *maximum likelihood estimation*):

$$c = arg\ max_c\ P(c|X)$$

Diese Wahrscheinlichkeit wird mit dem *Satz von Bayes* ermittelt.

$$P(c|X) = P(c) \cdot P(X|c) / P(X)$$

Die Wahrscheinlichkeiten auf der rechten Seite des Satzes von Bayes werden aus dem Trainingsmaterial ermittelt.

3.2.1.1. Training

Das Training des Naïve-Bayes-Klassifizierers besteht nur aus der Abschätzung der einzelnen Auftretenswahrscheinlichkeiten aus dem Trainingsmaterial.

Die *a priori* Klassenwahrscheinlichkeit *P(c)* ergibt sich durch einfaches Zählen der zu den einzelnen Klassen gehörigen Instanzen im Trainingsmaterial und anschließendes Teilen durch die Zahl der Instanzen insgesamt.

Die *a posteriori* Merkmalswahrscheinlichkeit *P(X|c)* ergibt sich aus dem Produkt der bedingten Wahrscheinlichkeiten für die einzelnen Merkmale x_i des Merkmalsvektors unter der Annahme, dass diese stochastisch unabhängig voneinander sind:

$$P(X|c) = \Pi_i\ P(x_i|c)$$

Für kategorielle Merkmale wird $P(x_i|c)$ analog zur Klassenwahrscheinlichkeit durch Auszählen der Merkmalswerte innerhalb der Instanzen der Klasse `c` bestimmt und eventuell geglättet. Die benutzte Implementierung nutzt Laplace-Glättung (Jurafsky und Martin 2007, Kapitel 4.5.1) um die Null-Wahrscheinlichkeit zu vermeiden.

Für kontinuierliche Merkmale wird aus dem Trainingsmaterial für jede Klasse eine Dichtefunktion f_c ermittelt, die die Verteilung der einzelnen Werte für diese Klasse im Trainingsmaterial widerspiegelt. Bei der Klassifizierung wird anstatt $P(x_i|c)$ der Wert der Dichtefunktion f_c an der Stelle x_i benutzt. Der Wert der Dichtefunktion ist keine Wahrscheinlichkeit: Die Wahrscheinlichkeit ist für jeden einzelnen Wert x_i einer kontinuierlichen Zufallsvariablen null, erst die Wahrscheinlichkeit innerhalb einer gewissen ε-Umgebung ist größer. Die gleiche ε-Umgebung müsste aber für alle Klassen gelten und kürzt sich so wieder. Deshalb wird auf die Berechnung der Wahrscheinlichkeit verzichtet und der Wert der Dichtefunktion direkt benutzt.

Die *a priori* Merkmalswahrscheinlichkeit $P(X)$ braucht nicht berücksichtigt zu werden, da auch sie für alle Klassen gleich ist und deswegen zu keinem Unterschied bei der Argument-Maximierung führt.

Insgesamt ergibt sich für den Naïve-Bayes-Klassifizierer also

$$c = arg\,max_c \prod_i P(x_i|c) \cdot P(c)$$

wobei $P(c)$ und die unterschiedlichen $P(x_i|c)$ aus dem Trainingsmaterial ermittelt werden.

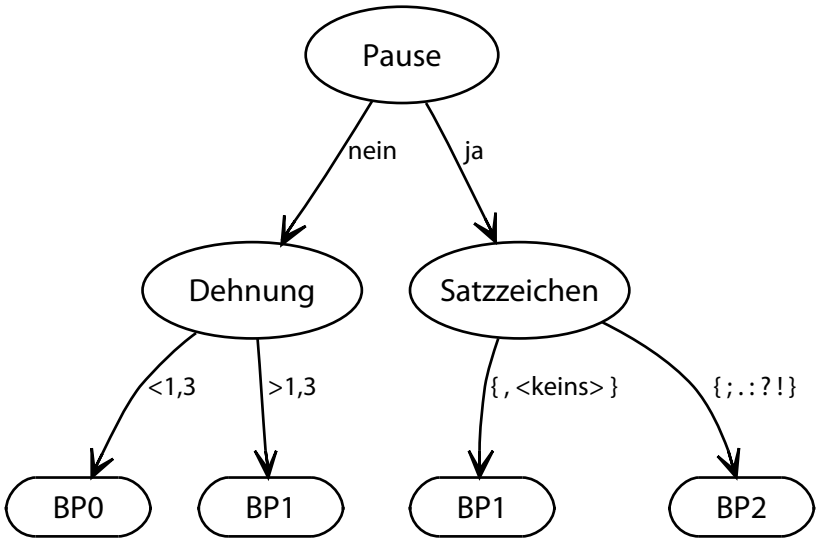
3.2.2. Entscheidungsbäume

Ein Entscheidungsbaum (Breiman et al 1984; Witten und Frank 2000, Kapitel 4.3) klassifiziert, indem er eine hierarchische Auswahl von Merkmalstests ausführt, an deren Ende die Zuordnung einer Instanz zu einer Klasse steht. Der Baum enthält in seinen inneren Knoten jeweils einen Merkmalstest; für jedes mögliche Ergebnis verweist er auf Tochterknoten. In den Blättern des Baumes sind die zuzuordnenden Klassen vermerkt.

Die Merkmalstests unterscheiden sich für kategorielle und kontinuierliche Merkmale: Für kontinuierliche Merkmale wird einer von zwei Tochterknoten ausgewählt, abhängig davon ob das Merkmal einen im Test festgelegten Schwellenwert übersteigt.

Für kategorielle Merkmale gibt es zwei Möglichkeiten: Entweder für jede Kategorie des Merkmals wird auf einen Tochterknoten verwiesen oder aber Kategorien werden zusammengefasst und gemeinsam mehreren Tochterknoten zugeordnet.

Abbildung 3.1. Einfacher Entscheidungsbaum für Phrasengrenzen



Die Klassifizierung mit einem Entscheidungsbaum beginnt an der Wurzel des Baumes. Der Knoten reicht eine Instanz abhängig vom Ausgang des Merkmalstests an den entsprechenden Tochterknoten weiter. Dies setzt sich rekursiv fort. Wenn ein Blatt erreicht ist, zeigt es die ermittelte Klasse für die Instanz an.

Als Beispiel soll die Instanz (*Pause*: `nein`, *Satzzeichen*: `keins`, *Dehnung*[2]: `1,45`) mit dem Entscheidungsbaum in Abbildung 3.1 klassifiziert werden. Das zuerst geprüfte Merkmal *Pause* ist negativ, also ist als nächstes der linke Tochterknoten zu prüfen. Das kontinuierliche Merkmal *Dehnung* ist größer als `1,3`, also wird von hier aus der rechte Tochterknoten gewählt. Dies ist ein Blatt und die Instanz wird als `BP1` (Zwischengrenze) klassifiziert.

3.2.2.1. Konstruktion von Entscheidungsbäumen

Die Konstruktion eines Entscheidungsbaumes im *Top-Down-Verfahren* (Quinlan 1986) verläuft folgendermaßen: Für jeden möglichen Merkmalstest wird errechnet wie groß der Informationsgewinn innerhalb der Trainingsmenge durch eine Unterteilung anhand dieses Tests wäre. Der Test mit dem größten Gewinn wird ausgewählt und für die Ergebnisse der Merkmalstests werden Tochterknoten angelegt. Die Trainingsmenge wird – je nach Ergebnis des Merkmalstests – in lokale Trainingsmengen für die Tochterknoten aufgeteilt und der Prozess in den Tochterknoten rekursiv wiederholt.

Der *Informationsgewinn* ist die Differenz des Informationsgehalts des Klassenattributs vor der Unterteilung und dem gewichteten Mittel der Informationsgehalte des Klassenattributs in den Unterteilungen.

Der *Informationsgehalt* des Klassenattributs in der lokalen Trainingsmenge nach Shannon ist die gewichtete Summe der Auftretenswahrscheinlichkeiten der einzelnen Klassen:

$$I = -\Sigma_c\, p_c \cdot log_2(p_c)$$

Ein Beispiel soll die Berechnung des Informationsgewinns verdeutlichen. Eine Trainingsmenge enthalte elf Instanzen die wie folgt auf drei Klassen verteilt sind: `{6 × BP0, 2 × BP1, 3 × BP2}`. Der Informationsgehalt nach Shannon für das Klassenattribut ist also

$$-(6/11 \cdot log_2(6/11) + 2/11 \cdot log_2(2/11) + 3/11 \cdot log_2(3/11)) = 1{,}435 \text{ bit}$$

Angenommen, das Merkmal *Pause* (mit den Werten `ja` und `nein`) unterteilt die Trainingsmenge in die Teilmengen `{6 × BP0, 1 × BP1}` und `{1 × BP1, 3 × BP2}`. Für diese Teilmengen ergeben sich Informationsgehalte von 0,592 bit beziehungsweise 0,811 bit. Das gewichtete Mittel davon ist

$$7/11 \cdot 0{,}592 \text{ bit} + 4/11 \cdot 0{,}811 \text{ bit} = 0{,}672 \text{ bit}$$

und es ergibt sich der Informationsgewinn durch das Merkmal *Pause* mit

$$1{,}435 \text{ bit} - 0{,}672 \text{ bit} = 0{,}764 \text{ bit}$$

Der Informationsgewinn wird nun für alle möglichen Merkmalstests und die daraus resultierenden Unterscheidungen errechnet und derjenige Merkmalstest ausgewählt, der den größten Informationsgewinn und damit die beste Unterscheidung zwischen den Klassen erzielt.

Das Verfahren terminiert, wenn entweder ein Knoten ausschließlich Instanzen einer Klasse enthält, die lokale Trainingsmenge zu klein wird oder kein Merkmalstest zu einem Informationsgewinn mehr führt. Das entsprechende Blatt wird dann mit der Klasse markiert, die die Mehrzahl innerhalb der lokalen Trainingsmenge stellt.

Die Anzahl möglicher Merkmalstests für kategorielle Merkmale ist gewaltig. Wenn alle möglichen Zusammenfassungen von `n` Kategorien betrachtet werden sollen, ergeben sich schon für eine binäre

[2] siehe auch Kapitel 5

Unterteilung 2^n Möglichkeiten. Praktisch muss die Zahl möglicher Merkmalstests deswegen verringert werden. Entweder jede Kategorie erhält einen Tochterknoten, oder aber es werden nur binäre Unterteilungen betrachtet bei der eine Kategorie allen anderen gegenübergestellt wird.

Im Anschluss an die Konstruktion des Baumes wird dieser beschnitten (engl. *pruning*, hier *postpruning*)[3]. Dies verhindert eine Überanpassung des Baumes an das Trainingsmaterial. Die *Überanpassung* (engl. overfitting) spiegelt keine generelle Struktur des Trainingsmaterials wider, wie sie sich auch im Testmaterial finden wird, sondern rein zufällige Eigenheiten des Trainingsmaterials. Um Überanpassung zu erkennen wird auf dem lokalen Trainingsmaterial eine Fehlerabschätzung durchgeführt und ein Teilbaum beschnitten, falls der geschätzte Fehler größer ist als ein Schwellenwert.

Der in der Arbeit verwendete Algorithmus zum Lernen von Entscheidungsbäumen *C4.5* (Quinlan 1993) in der hier benutzten Revision 8 verwendet zusätzlich Subtree-Raising zur Vermeidung von Überanpassung. Beim *Subtree-Raising* kann der an einem Knoten `k` hängende Tochterknoten `t` mit der größten lokalen Trainingsmenge[4] den usprünglichen Knoten `k` ersetzen, falls sich so eine günstigere Fehlerabschätzung ergibt. Der Schwellenwert für den geschätzten Fehler beträgt 25%. Außerdem wurde die kleinstmögliche lokale Trainingsmenge auf mindestens 10 Instanzen festgelegt und es werden nur binäre Unterteilungen von kategoriellen Merkmalen erlaubt.

Pruning eliminiert Konstruktionsentscheidungen die aufgrund des Rauschens in den Trainingsdaten gefallen sind. Insbesondere werden auch Effekte durch Ausreißer innerhalb der Trainingsdaten vermindert. Das Pruning kommt also einer *Glättung* der Trainingsdaten gleich (John 1995, S. 176). Die Pruning-Entscheidungen werden jedoch lokal in Ästen des Baumes getroffen. Der resultierende Baum ist inkonsistent: Während das Pruning bestimmte Instanzen lokal aus der tatsächlichen Trainingsmenge entfernt hat, waren diese Instanzen an der Bildung höherer Strukturen des Baumes noch beteiligt.

Robust C4.5 (John 1995) entfernt aus der Trainingsmenge die vom C4.5-Algorithmus durch Pruning fehlklassifizierten Trainingsinstanzen und wiederholt das Training solange, bis keine weiteren Instanzen mehr durch Pruning fehlklassifiziert werden. John (1995, S. 178) erreicht in einer Auswahl von Domänen zwar nur eine leichte Verbesserung der Klassifizierung, jedoch mit deutlich kleineren Bäumen.

3.2.3. Diskussion

„Entzwei und gebiete! Tüchtig Wort. – Verein und leite! Besserer Hort."
—J. W. v. Goethe

Beide vorgestellten Lernverfahren zur Klassifizierung haben ihre Tücken. Entscheidungsbäume werden nach dem Prinzip *teile und herrsche* konstruiert. Dem offensichtlichen Vorteil der rekursiven Unterteilung des Merkmalsraums bis in einzelne Klassen steht der damit einhergehende *Datenmangel* (engl. data sparsity) durch die wiederholte Unterteilung in lokale Trainingsmengen entgegen. Die Auswahl der richtigen Unterteilungen benötigt aber eine große Repräsentativität der lokalen Trainingsmenge für die zu erwartende (lokale) Testmenge. Die Repräsentativität schrumpft aber mit der Größe der lokalen Trainingsmenge bei zunehmender Rekursionstiefe. Die enthaltenen Strukturen werden immer stärker von individuellen Ausreißern innerhalb des Trainingsmaterials überlagert.

Gerade bei verrauschten Daten führt der Mangel an Repräsentativität dazu, dass die Auswahl des Algorithmus sich immer weniger auf tatsächlich relevante Merkmale stützt (Witten und Frank 2000, S. 288f.). Stattdessen steigt die Wahrscheinlichkeit, dass ein objektiv betrachtet weniger gutes Merkmal lokal betrachtet gegenüber dem objektiv besseren den größeren Gewinn verspricht. Sogar ein völlig zufälliges Merkmal wird bei genügend großer Rekursionstiefe irgendwann dasjenige Merkmal mit dem zufällig größten Informationsgewinn sein. Allerdings hat dies dann keinerlei Aussagekraft auf unabhängigem Testmaterial, bei dem das zufällige Merkmal einen vom lokalen Trainingsmaterial unabhängigen Wert annimmt, während ein objektiv gutes Merkmal mit dem Trainingsmaterial übereinstimmen würde. Pruning kann dieses Problem nicht lösen, da hier für die Fehlerabschätzung ebenfalls die lokale Trainingsmenge betrachtet wird und kein unabhängiges (Holdout-)Material.

[3] Es ist auch möglich, den Baum schon während der Konstruktion zu beschneiden, sogenanntes *prepruning*.
[4] Diese Einschränkung ist nötig um die Komplexität des Subtree-Raisings zu reduzieren.

Naïve-Bayes-Klassifizierer lernen in jedem Schritt auf dem gesamten Trainingsmaterial, daher leiden sie weniger stark an Datenmangel. Auch *irrelevante* Merkmale stören ihr Ergebnis nicht, da die benutzten Wahrscheinlichkeiten global berechnet und damit für alle Klassen gleich bewertet werden.

Anders sieht es mit *redundanten*, also miteinander korrelierten Merkmalen aus: Oben wurde am Rande erwähnt, dass die Errechnung der a posteriori Merkmalswahrscheinlichkeit aus den Wahrscheinlichkeiten für die einzelnen Merkmale nur für stochastisch unabhängige Merkmale gültig ist. Diese ist allerdings nur selten gegeben: Ein länger ausgesprochener Vokal wird in einer länger dauernden Silbe gesprochen. Wegen der größeren Zeit für die Einstellung der Artikulationsorgane wird er klarer gesprochen, was seine Formanten und die Schalldruckleistung beeinflusst. Entsprechendes gilt für die meisten anderen Klassifizierungsprobleme mit ihren Merkmalen.

Miteinander korrelierte Merkmale bestimmen das Klassifizierungsergebnis überproportional stark, sodass andere, nichtkorrelierte Merkmale weniger stark in die Entscheidung miteinbezogen werden (Witten und Frank 2000, S. 96f.). Eine theoretische Vorraussetzung des Naïve-Bayes-Klassifizierers ist also verletzt. Dennoch ist die Leistung des Naïve-Bayes-Klassifizierers auf echten Daten mit anderen Lernverfahren vergleichbar wie unter anderem Langley et al (1992) zeigen.

Obwohl Lernalgorithmen unwichtige Merkmale robust ignorieren sollten, verschlechtert sich C4.5 durch Hinzufügen eines zufälligen binären Merkmals um 5–10 % (John 1997), durch ein stark korreliertes Merkmal immer noch um 1–5 %. Naïve-Bayes-Klassifizierer ignorieren zwar zufällige Merkmale, werden aber durch korrelierte Merkmale negativ beeinflusst. Die richtige Auswahl der Merkmale ist deshalb besonders wichtig.

3.3. Automatische Merkmalsauswahl

> *"The best way to select relevant attributes is manually, based on a deep understanding of the learning problem and what the attributes actually mean. However, automatic methods can also be useful."*
> *—Witten und Frank 2000, S. 289*

Die Relevanz der manuellen Merkmalsauswahl ist unbestritten. Erst der Spezialist kann aus der beliebig großen Anzahl von Merkmalen vielversprechende Merkmale auswählen. Aber gerade die Korrelation der Merkmale untereinander und der durch das Merkmal mögliche Gewinn ist auch für den Experten schwer abzuschätzen. Nicht zuletzt soll diese Arbeit einen Beitrag dazu liefern, den Wert von Merkmalen zu bestimmen. Dabei hilft eine automatische Merkmalsauswahl.

Neben der höheren Leistung der Klassifizierer durch die Beschränkung auf weniger Merkmale verringert sich gleichzeitig die Komplexität der erlernten Klassifizierer. Auch dies hilft, die Ergebnisse besser zu interpretieren.

Zur Merkmalsauswahl gibt es mehrere Möglichkeiten. Bei der *Filtermethode* ist die Auswahl dem Klassifizierer vorgeschaltet und wählt Merkmale aufgrund der erwarteten Anforderungen des Klassifizierers. Für einen Naïve-Bayes-Klassifizierer könnte beispielsweise eine *Hauptkomponentenanalyse* durchgeführt werden, die die Merkmale dekorreliert. Der resultierende Klassifizierer wäre allerdings sehr schwer interpretierbar, da nicht mehr die ursprünglichen Merkmale verwendet würden, die direkte Entsprechungen in der Domäne haben.

3.3.1. Merkmalsauswahl durch Suche

Die *Umhüllungsmethode* von Kohavi und John (1997; engl. wrapping) umschließt den Klassifizierer und trainiert ihn wiederholt mit unterschiedlichen Teilmengen der Merkmalsmenge. Die jeweilige Leistung wird an einer Holdout-Menge überprüft.

Bei m Merkmalen gibt es 2^m mögliche Teilmengen. Eine vollständige Suche nach der besten Merkmalsauswahl wird also schnell unmöglich. Der Suchraum wird wesentlich schneller durch eine *gierige Hill-Climbing-Suche* abgedeckt. Sie läuft allerdings Gefahr, nur ein lokales Maximum zu erreichen. Die *n-Bestensuche* (Russel und Norvig 2003) verfolgt in einem lokalen Maximum noch die n nächstbesten Möglichkeiten entlang des Pfads um von dort aus ein höheres Niveau als das derzeitige Maximum

zu erreichen. Falls dies gelingt, läuft die Suche weiter und verfährt am nächsten lokalen Maximum entsprechend. Mit steigendem n erhöht sich die Chance, das globale Maximum zu erreichen.

Die Suche im Merkmalsraum kann *vorwärts* nach und nach Merkmale hinzufügen oder *rückwärts* nach und nach Merkmale entfernen. Das Training von Klassifizierern mit weniger Merkmalen ist deutlich schneller. Daher wird die Vorwärtssuche bevorzugt. Sie hat außerdem den Vorteil, dass sie im Unterschied zur Rückwärtssuche eine zu kleine und damit weniger komplexe Auswahl trifft, falls sie schon vor dem globalen Maximum terminiert.

Wenn die Leistung mit den einzelnen Merkmalsmengen an einer Holdout-Menge überprüft werden soll, stellt sich unweigerlich die Frage nach dem richtigen Leistungsmaß. Die verwendete Implementierung nutzt hierfür ausschließlich die Korrektheit. Diese ist – wie oben dargelegt – für ungleichverteilte Klassen ein schlechtes Maß. Sinnvoller wäre das F-Maß beziehungsweise bei einem Mehrklassenproblem das kombinierte F-Maß.

Als Ausweg wird die Holdout-Menge so neu geschichtet, dass die Klassen gleichverteilt sind. Die Korrektheit stimmt auf diese Weise mit dem F-Maß beziehungsweise kombinierten F-Maß überein. Die Neugewichtung innerhalb der Holdout-Menge entspricht also der Nutzung des F-Maßes (bzw. kombinierten F-Maßes) als Leistungsmaß bei der Merkmalsauswahl. Die Trainingsmenge der Klassifizierer wird nicht neu gewichtet. Dies würde den Aufbau der Klassifizierer beeinflussen und die Leistung (F-Maß) auf einer Testmenge verschlechtern, die die Klassen schließlich auch in ihrer ursprünglichen Gewichtung enthält.

3.4. Implementierung und Versuchsaufbau

Für diese Arbeit wird das freie Programmierpaket Weka[5] benutzt, das neben den vorgestellten Algorithmen viele andere Verfahren des maschinellen Lernens implementiert. Wenn im Folgenden nicht anders angegeben, wird die Leistung der Klassifizierer immer mit vorgeschalteter Merkmalsauswahl durch n-Bestensuche mit n=20 angegeben. Für C4.5 wird außerdem der Klassifizierer nur auf den zuvor korrekt klassifizierten Trainingsinstanzen neu trainiert (Robust C4.5).

[5] http://www.cs.waikato.ac.nz/ml/weka/ sowie (Witten und Frank 2000, Teil II)

Kapitel 4. Datengrundlage

Als Datengrundlage für Training und Test der Prosodieerkennung in Sprachsynthesekorpora werden bereits prosodieannotierte Aufnahmen benötigt, die denen für ein Sprachsynthesesystem entsprechen. Einige Sprachsynthesekorpora sind bereits prosodieannotiert und bieten so die Basis für Training und Test.

Frei verfügbare Sprachsynthesekorpora haben oft den Nachteil, dass sie nur Daten eines einzelnen Sprechers enthalten, sodass sprecherübergreifende Prosodieerkennung mit ihrer Hilfe schwer möglich ist. Die *Sprecherunabhängigkeit* von Erkenntnissen zur Prosodieerkennung kann sich bereits auf zwei Ziele erstrecken:

1. Vergleich der Akzentuierungsmuster und der Merkmale von Akzentuierung und Phrasierung zwischen unterschiedlichen Sprechern sowie
2. Generierung sprecherübergreifender Klassifizierer

Das erste Ziel wird bereits durch den Vergleich der sprecherabhängigen Ergebnisse für unterschiedliche Sprecher erreicht. Die Unterschiede zwischen den Daten der Sprecher werden analysiert, bewertet und beim Vergleich miteinbezogen. Die Daten der einzelnen Sprecher können also unterschiedlichen Korpora entstammen, deren direkte Vergleichbarkeit nachrangig ist.

Für das zweite Ziel, das Training von Klassifizierern, die für beliebige weitere Sprecher Akzentuierungs- und Phrasierungsvorhersagen treffen, müssen Test- und Trainingsdaten von unterschiedlichen Sprechern stammen, aber nach den gleichen Kriterien erstellt sein. Hierfür ist ein sprecherübergreifendes Korpus ideal. Nur so sind einheitliche Bedingungen bei den Aufnahmen und gleichartige Texte für die Sprecher im Korpus sicherstellt.

Das *Kiel Corpus of Read Speech* enthält Aufnahmen vieler unterschiedlicher Sprecher und ist so für die Auswahl sprecherunabhängiger Merkmale und den Vergleich zwischen unterschiedlichen Sprechern geeignet. Allerdings handelt es sich nicht um ein Sprachsynthesekorpus. Die Daten sind als Grundlage für die Spracherkennung aufgenommen worden.

Deshalb werden noch zwei weitere öffentlich zugängliche prosodieannotierte Korpora mit jeweils einem Sprecher genutzt und hier beschrieben: Das *IMS-Unit-Selection-Korpus* und das *Vienna Prosodic Speech Corpus*. Außerdem wird die Prosodieannotierung eines nicht-öffentlichen Sprachsynthesekorpus der IBM beschrieben.

4.1. Gemeinsame Grundlagen

Die Erstellung der verwendeten Sprachkorpora verlief nach einem einheitlichen Schema, das hier für alle Korpora gemeinsam vorgestellt wird.

4.1.1. Aufbau und Aufnahmen

Zunächst wurden zu sprechende Äußerungen nach bestimmten Kriterien zusammengestellt. Diese Äußerungen bestehen jeweils aus einem oder mehreren Sätzen, Satzfragmenten, Namen oder ähnlichem. Die Äußerungen werden durch einen Sprecher nacheinander von einer Vorlage (Papier oder Bildschirm) abgelesen und aufgezeichnet. Die Aufteilung des Korpus in einzelne Äußerungen erleichtert die Aufnahme: Bei Versprechern, Häsitationen, Übersteuern der Aufnahme oder Nebengeräuschen muss jeweils nur die Äußerung neu gesprochen werden.

Fast alle Äußerungen in den verwendeten Korpora sind unter 20 Sekunden lang, die längsten unter einer Minute. Die Aufteilung des Korpus in Äußerungen führt dazu, dass die meisten Phrasengrenzen trivial zu bestimmen sind:

Jede Äußerung schließt mit einer starken Phrasengrenze.

Diese Phrasengrenzen durch einen Klassifizierer zu ermitteln ist unnötig. Wichtige Merkmale für die Erkennung von Phrasengrenzen ergeben sich aus einem Vergleich der Wörter oder des Signals vor und nach der möglichen Phrasengrenze. Für die abschließenden Phrasengrenzen fehlt der rechte Kontext, weshalb viele Merkmale nicht berechnet werden können. Deshalb sind sie auch als Trainingsmaterial für die Klassifizierung der übrigen Wortgrenzen ungeeignet.

4.1.2. Segmentale Transkription

Im Anschluss an die Aufnahme eines Korpus wird dieses segmentiert. Die Grundlage für die Segment-Annotierung für alle benutzten Korpora bildet SAMPA (Wells o. J.). *SAMPA* ist ein phonetisches Alphabet bei dessen Entwicklung die Computerlesbarkeit und die Kompatibilität mit dem internationalen phonetischen Alphabet (IPA 1999) im Vordergrund stand.

SAMPA wurde für mehrere Sprachen spezifisch angepasst, fürs Deutsche durch Kohler (1992a), sodass ein phonologisches Alphabet entsteht, das nur durch wenige phonetische Merkmale ergänzt wird.[1]

Die in den Korpora verwendeten SAMPA-Dialekte machen unterschiedliche Erweiterungen gegenüber Kohler (1992a). Insbesondere werden zusätzliche Symbole zur besseren Beschreibung von fremdsprachlichem Material eingeführt. Diese Erweiterungen sind miteinander kompatibel, weshalb hier die Obermenge aller verwendeten Zeichen angegeben wird.

- 21 Vokale: `/i, I, y, y:, Y, e:, e, e~, 2:, E:, E, 9, a, a:, a~, O, O:, o:, o~, u:, u, u~, U/`
- 2 Schwa-Laute: `/@, 6/`
- 5 Diphthonge: `/eI, aI, aU, @U, OY/`
- 26 Konsonanten: `/p, b, t, d, k, g, ?, m, n, N, r, R, f, v, T, D, s, z, S, Z, C, x, h, w, j, l, L/`

Da die verwendeten Korpora SAMPA benutzen, werden Phoneme auch in der vorliegenden Arbeit durch SAMPA-Symbole und nicht die entsprechenden IPA-Symbole dargestellt.

4.1.3. Prosodische Annotierung

In zwei der vier verwendeten Korpora ist die Prosodie GToBI-annotiert, das auf der Ton-Sequenz-Theorie aufbaut. Das Kiel-Korpus ist mit PROLAB (Kohler 1992b, Peters und Kohler 2004) annotiert, das auf dem *Kieler Intonationsmodell* (KIM; Kohler 1991 nach Kohler 1995) aufbaut.

Die beiden Schemata annotierten jeweils Akzentuierungen und Phrasengrenzen, jedoch nach unterschiedlichen Systemen. Gemein ist beiden Systemen, dass sie primär auf dem auditorischen Eindruck der Akzentuierung beziehungsweise Phrasierung beruhen.

Die systematische Einordnung der Akzentuierungen unterscheidet sich dahingehend, dass die Akzenttöne der Ton-Sequenz-Theorie aus atomaren Elementen aufgebaut sind, deren Kombination den Intonationsverlauf beschreibt. In PROLAB hingegen wird die zeitliche Ausrichtung und Stärke der Akzentuierung unterschieden und außerdem der Intonationsverlauf zwischen den Akzentuierungen angegeben.

Aufgrund der theoretischen Unterschiede zwischen GToBI und PROLAB können beide nicht eins-zu-eins aufeinander abgebildet werden (Kohler 1995[2]).

In der Arbeit soll von den verschiedenen Akzenttypen abstrahiert werden. Es soll nur unterschieden werden, ob eine Silbe akzentuiert ist oder nicht. Die Frage der „richtigen“ Theorie stellt sich dadurch

[1] Beispielsweise die Notierung von Glottalverschlüssen obwohl diese im Deutschen phonologisch bedingt sind. Da sie allerdings nicht durchgehend an den zu erwartenden Stellen realisiert werden, wird ein eigenes Symbol notwendig.
[2] (Kohler 1995, S. 11) enthält dennoch eine Tabelle die die Systeme zumindest teilweise aufeinander abbildet.

nicht. Deshalb wird aus der Akzentuierungsannotierung nur die Information genutzt, ob eine Silbe akzentuiert ist, nicht jedoch die Information wie.

Hinsichtlich der Phrasierungen unterscheidet GToBI zwei unterschiedlich starke Phrasengrenzen. Diese werden durch die Ton-Sequenz-Theorie motiviert. PROLAB unterscheidet nur eine Sorte von Phrasengrenzen. Das Phrasierungslabel (derzeit `PGn`) legt jedoch nahe, dass eine nachträgliche Unterscheidung verschieden starker Phrasengrenzen (`PG1`, `PG2`, ...) vorgesehen war.

Nicht alle Phrasengrenzen sind gleich, eine Abstufung ist sinnvoll. Deshalb wird in dieser Arbeit die volle Phrasierungsinformation genutzt und Zwischengrenzen von Vollgrenzen unterschieden.

4.2. Das Kiel Corpus of Read Speech

Das Kiel Corpus of Read Speech (IPDS 1994) basiert auf Aufnahmen, die im Rahmen der Projekte PHONDAT 90 und 92 am Institut für Phonetik und digitale Sprachverarbeitung in Kiel erstellt wurden. Ziel der Aufnahmen war „die Schaffung einer repräsentativen Materialgrundlage auf Diphonem- sowie Silbenbasis“ (Kohler 1992c, S. 7) der deutschen Lesesprache, insbesondere zum Training von Spracherkennern der Industriepartner.

4.2.1. Textmaterial

Das Textmaterial von *PHONDAT 90* besteht aus mehreren Satzlisten mit dem Ziel, alle Phoneme und Phonem-Übergänge des Deutschen zu erfassen. Die *Berlin-* und *Marburg-Sätze* die zuerst Sotscheck (1976a, 1976b, nach Kohler 1992a) veröffentlicht hat, sind kurze Sätze mit Alltagsvokabular, die durch die weiteren Satzlisten (*Kohler*, *Schiefer/Sommer*, *Tillmann/Kohler*) vervollständigt werden. Hinzu kommen zwei Listen der beteiligten Industrie (CNET, SEL1 und SEL2). Die Auswahl der Sätze diskutiert Kohler (1992a) ausführlich. Außerdem sind zwei Kurzgeschichten aufgenommen: Die *Buttergeschichte* mit phonetisch ausgeglichenem Material und *Nordwind und Sonne*[3] (Thon und van Dommelen 1992, S. 47).[4]

Für *PHONDAT 92* wurden Texte der Domäne Bahnauskunft ausgewählt. Die 100 *Siemens-Sätze* sind konstruierte, vollständige Sätze; die teils unvollständigen 100 *Erlangen-Sätze* sind aus mitgeschnittenen und verschrifteten Dialogen entnommen (Thon 1992, S. 115).[5]

4.2.2. Aufnahmen

Das Kiel-Korpus enthält Aufnahmen von 51 norddeutschen Sprechern, die jeweils nur kürzere Abschnitte des Textmaterials sprechen. Eine Frau und ein Mann[6] sprechen das gesamte Textmaterial von 603 Sätzen. Insgesamt ergeben sich 3876 Äußerungen. Die genaue Aufteilung ergibt sich aus Tabelle 4.1.

Tabelle 4.1. Teilkorpora im Kiel-Korpus

Teilkorpus	Anzahl Äußerungen	Anzahl Sprecher
Berlin-Sätze	100	10
Marburg-Sätze	100	10
Buttergeschichte	2	14
Nordwind und Sonne	3	14
Siemens- und Erlangen-Sätze	200	3
weitere Satzlisten zur vollständigen Abdeckung deutscher Phonemübergänge	198	2*

*Diese Beiden sprechen das gesamte Korpus. Die Gesamtzahl der Sprecher in den oberen Zeilen liegt also jeweils um 2 höher.

[3] *Nordwind und Sonne* wird von der IPA bei der Beschreibung von Sprachen als Beispieltext empfohlen. Das IPA-Handbuch (IPA 1999) enthält den Text und die phonetische Transkription in 27 Sprachen.
[4] Das gesamte Textmaterial von PHONDAT 90 findet sich im Anhang von (Thon und van Dommelen 1992).
[5] Das vollständige Textmaterial von PHONDAT 92 enthält der Anhang von (Thon 1992).
[6] offensichtlich Herr Kohler selbst

4.2.3. Transkription

Für die 603 unterschiedlichen Sätze wurde zunächst maschinell eine kanonische Transkription erzeugt und von Hand korrigiert. Eine *kanonische Transkription* gibt die normale Aussprache für eine Äußerung an, wie sie durch einen beliebigen Sprecher des Satzes zu erwarten ist. Mithilfe dieser Transkriptionsvorgabe wurden dann die einzelnen Äußerungen im Korpus manuell segmentiert und die Phonem-Etiketten vergeben (van Dommelen 1992).

Die tatsächliche Aussprache weicht von der kanonischen Transkription durch Einfügung, Tilgung und Ersetzung von Lauten ab. Diese Unterschiede werden für Konsonanten genau notiert: `/n-m/` bedeutet `/n/` wurde als `/m/` realisiert, wie zum Beispiel in „geben“ `/g e b @- n-m/`. Für Vokale wird bei Tilgung (vgl. `/@-/` im Beispiel) und Einfügung ebenso verfahren. Eine Ersetzung wird hingegen nur dann gekennzeichnet, wenn sich dadurch die phonemische Kategorie verändert. Die Reduktion eines gespannten Vokals zum entsprechenden ungespannten Vokal oder Schwa hingegen wird nicht notiert (van Dommelen 1992, S. 203f.). Vokale mit folgendem vokalisierten `/r/` werden als Diphthonge (`/a6/`, `/o6/` und so weiter) notiert.

Das Annotierungsformat des Kiel-Korpus ist kryptisch. Alle Transkriptionsebenen zu einer Äußerung werden in derselben Datei gespeichert. Die Transkription der unterschiedlichen Ebenen ist ineinander verwoben und die einzelnen Markierungen sind aus mehreren Elementen zusammengesetzt. Brinckmann (2004, Kapitel 2.2) beschreibt es in allen seinen Einzelheiten.

4.2.4. Prosodische Annotierung

Das Kiel-Korpus ist nach dem PROLAB-System (siehe oben) prosodieannotiert. Für Akzentuierungen unterscheidet PROLAB vier unterschiedliche Stärken: keine Akzentuierung (`0`), „Deakzentuierung“ (`1`) also sehr schwache Akzentuierung, normale Akzentuierung (`2`) und Emphase (`3`; Kohler 1992b, S. 243). Sowohl Deakzentuierung als auch Emphase sind im Kiel-Korpus selten (6–7 %; Abbildung 2.3 in Brinckmann 2004, S. 32).

Nach informeller Prüfung wurden nur die Stärken `3` und `4` als akzentuiert gewertet, deakzentuierte Silben hingegen wie unakzentuierte Silben behandelt.

PROLAB unterscheidet nur eine Stärke der Phrasierung. Meines Erachtens nach wurden nur Vollgrenzen, nicht aber Zwischengrenzen annotiert. Möglicherweise liegt dies daran, dass das Kieler Intonationsmodell neben Phrasierungen und Akzentuierungen noch weitere Intonationskonturen annimmt, die die Funktion von Zwischengrenzen übernehmen können.

Die Untersuchungen zur Phrasierung im Kiel-Korpus beschränken sich deswegen auf Vollgrenzen. Das Kiel-Korpus enthält nur sehr wenige Vollgrenzen. Dies liegt auch an der generellen Kürze der Äußerungen im Kiel-Korpus.

4.3. Das Vienna Prosodic Speech Corpus

Das Vienna Prosodic Speech Corpus (Neubarth et al 2000) wurde speziell für die Untersuchung der Prosodie in gelesener Sprache entwickelt. Insbesondere sollten auch Fokus-Einflüsse auf Akzentuierung und Phrasierung untersucht werden. Dies zeigt sich bereits im verwendeten Textmaterial.

4.3.1. Textmaterial

Das Korpus besteht aus drei unterschiedlich aufgebauten Teilkorpora:

1. 300 Sätze phonetisch ausgeglichenes Material aus dem PHONDAT-Projekt, nämlich die oben erwähnten Marburg-, CNET-, Kohler-, SEL1-, SEL2-, Schiefer/Sommer- und Tillmann/Kohler-Sätze,
2. Nordwind und Sonne, die Buttergeschichte und 22 Artikel aus österreichischen Tageszeitungen,

3. 250 systematisch variierte Antworten aus Frage/Antwort-Paaren.

Dem Frage/Antwort-Korpus liegen zehn (konstruierte) Antworten zugrunde. Die Antworten liegen jeweils in Varianten vor, die sich in einem Verb unterscheiden, sodass sich eine unterschiedliche syntaktische Struktur und damit auch unterschiedliche Informationsstruktur ergibt:

Peter verspricht dem Freund zu verweilen und den Dieb zu bewachen.
Peter verspricht, den Freund zu entlasten und den Dieb zu bewachen.

Außerdem werden die Antworten weiter systematisch variiert, unter anderem durch die Einfügung von Fokuspartikeln. Für alle Antworten werden dann unterschiedliche Fragen gestellt, die unterschiedliche Foki in der Antwort hervorrufen. Teilweise ist der Fokus weit, teilweise eng auf eine der Informationen im Satz gerichtet.

Was ist passiert? Was geschieht?

Wer verspricht dem Freund zu verweilen und den Dieb zu bewachen?
Wer verspricht, den Freund zu entlasten und den Dieb zu bewachen?

Wem verspricht Peter zu verweilen und den Dieb zu bewachen?
Wen verspricht Peter zu entlasten und den Dieb zu bewachen?[7]

Der Sprecher sollte die Antworten jeweils so sprechen, wie es für ihn als Antwort auf die vorgelegten Fragen natürlich war. Diese feine Kontrolle über die Informationsstruktur der Antworten erlaubt systematische Analysen der Auswirkungen auf die prosodische Realisierung.

Das Korpus lag in MaryXML (Schröder und Breuer 2004) vor. Ob Leerzeichen vor oder nach Satzzeichen stehen sollen, gibt diese Repräsentation nicht an. Insbesondere ist diese Angabe aber bei Anführungszeichen sowie Bindestrichen/Gedankenstrichen wichtig. („Peter – nicht Paul – liebt ‚Lach- und Sachgeschichten'."). Die Zuordnung der Anführungszeichen zum rechten oder linken Wort sowie die Unterscheidung zwischen Binde- und Gedankenstrichen wurde deswegen von Hand vorgenommen.

4.3.2. Aufnahmen und Transkription

Das Korpus umfasst knapp 14.000 Silben. Ein Wiener Sprecher ohne besondere Vorkenntnisse spricht das Korpus in der österreichischen Varietät des Standarddeutschen. Das Sprachmaterial wurde von Hand segmentiert und silbifiziert und hält sich bei der Phonem-Etikettierung eng an die kanonische Aussprache.

Häufig weicht die Aussprache durch Reduktion einiger Segmente am Wortende von der kanonischen Aussprache ab. Diese Segmente sind dennoch im Korpus annotiert, jedoch mit der einheitlichen (und unrealistischen) Dauer von einer Millisekunde.

Außerdem wird noch `/N:/` und `/S:/` für besonders gedehnte `/N/` und `/S/` benutzt. Diese Unterscheidung ist bei der weiteren Verarbeitung unnötig und wird deshalb fallengelassen: Die Dehnung ergibt sich schon aus der Dauer der Segmente.

4.3.3. Prosodieannotierung

Die Akzentuierungsannotierung wurde nach dem GToBI-System durchgeführt. Die Phrasierung hingegen entspricht nicht Standard-GToBI: Der *break index* für Phrasengrenzen unterscheidet drei Klassen (1, 3 und 4). Für die Klassen 1 und 4 ist jeweils der GToBI-Grenzton angegeben, für Klasse 3 fehlt diese Angabe.

Sowohl ToBI-Grenztöne (`%`) als auch -Phrasentöne (`-`) sind den break indices 3 und 4 zugeordnet. In Standard-GToBI sind Grenztöne fest mit dem break index 4 und Phrasentöne fest mit dem break index 3 verbunden.

[7] Die Frage/Antwort-Paare sind vollständig unter http://www.ofai.at/~hannes.pirker/speedurcont/ aufgelistet.

Eine informelle Überprüfung ergab, dass Grenzen der Klasse 1 eher schwache Grenzen bezeichnen und die Klassen 3 und 4 eher starke Grenzen. Zwischen Grenzen der Klassen 3 und 4 bestand kein genereller Unterschied in der Grenzstärke. Sie unterschieden sich nur in der Angabe des Grenztones. Ein enger Zusammenhang von Phrasen- beziehungsweise Grenzton mit der Grenzstärke war nicht ersichtlich. Deswegen wird in der Weiterverarbeitung zwischen zwei unterschiedlich starken Grenzen unterschieden: Grenzen der Klasse 1 als Zwischengrenzen und Grenzen der Klassen 3 und 4 als Vollgrenzen.

4.4. Das IMS-Unit-Selection-Korpus

Das Unit-Selection-Korpus des Instituts für Maschinelle Sprachverarbeitung (IMS) in Stuttgart wurde im Rahmen des SmartKom-Projekts (Wahlster et al 2001) aufgenommen (Schweitzer et al 2003). Ziel war die „Restricted Unlimited Domain Synthesis“ (Schweitzer et al 2003). Es sollte ein Korpus aufgenommen werden, bei dem innerhalb einer bestimmten Domäne unrestringierter Text möglichst gut synthetisiert werden kann.

4.4.1. Textmaterial

Das Korpus besteht aus zwei Teilen, einem domänenunspezifischen aus 1636 Äußerungen phonetisch ausgeglichenem Material und einem domänenspezifischen mit 953 Äußerungen. Das domänenspezifische Teilkorpus führt zu einer gegenüber herkömmlicher Synthese höheren Qualität innerhalb der durch das Korpus zur Verfügung gestellten Bausteine. Gleichzeitig ermöglicht das domänenunspezifischen Teilkorpus die Synthese von beliebigen anderen Äußerungen.

Das phonetisch ausgeglichene Material wurde aus einem großen Textkorpus so ausgewählt, dass die zu erwartende Segment- und Segmentübergangsabdeckung bei einer vorgegebenen Korpusgröße maximiert wurde. Anschließend wurden einige sehr konstruierte Sätze[8] hinzugefügt um die Diphon-Abdeckung zu vervollständigen.

Der domänenspezifische Teil umfasst überwiegend Film- und Schauspielernamen mit einem großen Anteil englischsprachiger Wörter sowie wichtige Sätze für ein Dialog-System zur Kino-Auskunft. Einige Äußerungen wiederholen sich, wohl um bei der Sprachsynthese innerhalb der Domäne zwischen unterschiedlichen Realisierungen derselben Äußerung auswählen zu können und so die Lebendigkeit und Qualität des Systems zu verbessern.

Die englischsprachigen Wörter sind für die Domäne Kino-Auskunft unerlässlich. Hier soll die deutsche Prosodie im Allgemeinen untersucht werden, nicht speziell die Prosodie in der Domäne Kino-Auskunft. Die überproportional vielen englischen Äußerungen stören dabei.

Ein- oder Zweiwort-Äußerungen (meist Titel von Filmen oder Namen) enthalten keine Phrasengrenzen. Außerdem ist für sie die Akzentuierungsvorhersage insofern einfacher, als jedes Wort akzentuiert wird und lediglich die Position des Wortakzents bestimmt werden muss.

Wiederholungen von Äußerungen bergen die Gefahr, dass gleiche Äußerungen sowohl im Trainings- wie im Testmaterial auftauchen und somit die Ergebnisse des Lernverfahrens verfälschen. Wenn Äußerungen mehrfach im Trainingsmaterial auftauchen, dann verschieben sie das Gewicht des Trainings zugunsten dieser Äußerungen, was die Generalität der gelernten Klassifizierer einschränkt.

Aus dem domänenspezifischen Teil werden deshalb nur die Äußerungen benutzt, die (1) ganze Sätze oder Satzteile umfassen, (2) nicht überwiegend englischsprachig sind und (3) keine Wiederholung einer anderen Äußerung sind.

Das Korpus selbst enthält nicht das ursprünglich verwendete orthographische Material sondern nur die normalisierten Wörter ohne Satzzeichen. Um auch textbasierte Merkmale nutzen zu können, ist das ursprüngliche Textmaterial aber wichtig: Zum Beispiel korrelieren Phrasengrenzen häufig mit Satzzeichen. Für die erfolgreiche Suche nach der originalen Satzliste des phonetisch ausgeglichenen

[8] zum Beispiel *Nö öffentlich Anaphorä üben beim Gourmet-Elch.*

Materials bin ich Frau Antje Schweitzer vom IMS zu Dank verpflichtet. Für den domänenspezifischen Teil habe ich selbst nach bestem Wissen dem normalisierten Material Satzzeichen hinzugefügt.

4.4.2. Aufnahmen und Annotierungen

Die Aufnahmen eines Sprechers aus dem Raum Stuttgart sind durch Forced Alignment automatisch segmentiert, mit SAMPA-Etiketten versehen und anschließend von Hand korrigiert worden. Beim *Forced Alignment* wird ein automatischer Spracherkenner benutzt, dem die zu erkennende Wort- oder Segmentfolge vorgegeben ist. Der Erkenner liefert dann die für ihn optimalen Grenzen zwischen den vorgegebenen Segmenten.

Die prosodische Annotierung wurde nach dem GToBI-System durchgeführt. Die Akzentuierungs- und Phrasierungsinformation wurde wie oben beschrieben übernommen.

4.5. Ein minimal prosodieannotiertes Korpus

Im Rahmen der Arbeit wurde ein Teil eines Sprachsynthesekorpus (im weiteren: IBM-Korpus) minimal prosodieannotiert. Es stehen noch weitere Korpora anderer Sprecher zur Verfügung, die im Aufbau identisch sind und für die eine automatische Annotierung der Akzentuierungen und Phrasengrenzen wünschenswert ist. Die Ergebnisse der Untersuchung des IBM-Korpus sind deswegen besonders relevant, weil die Ergebnisse direkt auf andere Korpora übertragen und für die Verbesserung der Sprachsynthese benutzt werden können (vgl. Kapitel 8).

Die Textvorlage besteht aus phonetisch ausgeglichenem Material, das, wie beim IMS-Korpus, automatisch aus einem großen Textkorpus ausgewählt wurde. Es handelt sich überwiegend um Zeitungstexte unterschiedlicher Couleur. Die segmentale Transkription wurde automatisch durch Forced-Alignment auf Basis des normalisierten Textes durchgeführt.

4.5.1. Prosodische Annotierung

Die Annotierung von Phrasengrenzen im Schrifttext (Zwischengrenzen `(BP1)` und Vollgrenzen `(BP2)`) lag bereits aus einer Untersuchung zur automatischen Phrasierung in TTS-Systemen vor (Qin und Fischer 2004) und wurden entsprechend genutzt.

Im Vorfeld der Arbeit wurden, ebenfalls im Schrifttext, akzentuierte Wörter markiert[9]. Vorgabe bei dieser Annotierung war es, dass in jeder Intonationsphrase wenigstens eine Akzentuierung liegen muss.

Für die mit der Arbeit verfolgten Ziele war die Annotierung von Akzentuierungen auf Wortebene nicht ausreichend. Vielmehr wird eine Annotierung auf Silbenebene benötigt, um so Eigenschaften von akzentuierten und nicht akzentuierten Silben voneinander abgrenzen zu können. Zu diesem Zweck wurde das Korpus silbifiziert und eine Zuordnung von Schrifttext zu Sprechtext hergestellt (siehe die beiden folgenden Abschnitte).

Für die Akzentuierungszuweisung auf Silbenebene wurde mithilfe der Textanalyse eines TTS-Systems für den Schrifttext eine Silbifizierung und die Zuweisung des Wortakzents an die Silben automatisch erzeugt. Für jedes manuell als akzentuiert markierte Wort wurden dann diejenigen Silben als akzentuiert markiert, denen vom TTS-System ein Wortakzent oder Wortnebenakzent zugewiesen wurde.

Probleme ergeben sich, wenn (1) die aus dem Text ermittelte Silbenzahl[10] und die aus der akustischen Realisierung ermittelte Silbenzahl nicht übereinstimmen oder (2) das TTS-System keiner Silbe[11] oder einer falschen Silbe[12] den Wortakzent zuweist.

[9] Mein großer Dank hierfür gilt Stella Müller, zu der Zeit Praktikantin bei IBM.
[10] Ein besonders kurioses Beispiel ist *Lü-b-eck*.
[11] Dies betrifft vor allem Funktionswörter, aber auch zum Beispiel *Korrektur*.
[12] Entweder gibt es für das Wort alternative Akzentuierungsmöglichkeiten (*UMfahren*/*umFAHRen*) oder die Zuweisung war schlicht falsch (*ChemNITZ*).

Die Akzentuierungen wurden deswegen einzeln von Hand korrigiert. Dabei wurden auch offensichtliche Fehler der Phrasierungsannotierung behoben. In einigen Fällen fielen einige grobe Fehler im automatischen Segment-Alignment auf und die betroffenen Äußerungen wurden entweder korrigiert oder aus dem Korpus entfernt.

4.6. Silbifizierung

Sind Silbengrenzen für die Prosodieannotierung notwendig? Die Silbe bildet die Konstituente der Akzentuierung (vgl. Kapitel 2). Ihr Aufbau ist jedoch kompliziert und die genaue Bestimmung der Grenzen zwischen zwei Silben nicht immer möglich.

Nicht immer ist die Zugehörigkeit der Segmente zu den Silben eindeutig: Manchmal scheint ein Laut zu zwei Silben zu gehören (*ambisilbischer Konsonant*: `/m/` in „Hammer“), manchmal einer Silbe voranzugehen ohne ihr wirklich zuzugehören (*extrasilbischer Konsonant*: `/S/` in „Streit“).

Die Rolle und schon die Existenz extrasilbischer Konsonanten ist umstritten. So wird das `/S/` im obigen Beispiel durch den folgenden Vokal koartikulativ gefärbt (vgl. „Streit“ und „Strolch“) und gleichzeitig die Koartikulation als nur innerhalb der Silbe wirksam angesehen.

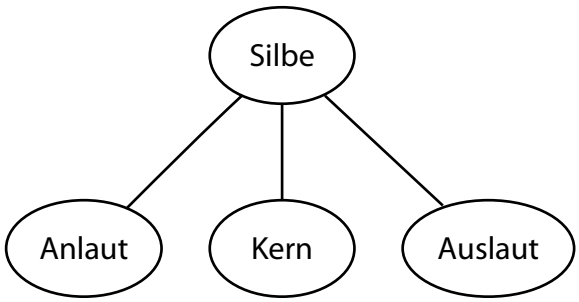
Es erscheint dennoch sinnvoll, bei Silbendauermessungen extrasilbische Konsonanten nicht zu berücksichtigen, da ihre Dauer unabhängig von den Dauern der übrigen Laute der Silbe ist. Ebenso sollten ambisilbische Konsonanten bei Silbendauermessungen anteilig den beteiligten Silben zugeschlagen werden, wobei unklar bleibt, welcher Anteil zu welcher Silbe.

Tamburini (2003) zeigt, dass der eigentliche Träger von Akzentuierungen der Kern der jeweiligen Silbe ist und als Merkmal für die Akzentuierungsentscheidung die Silbenkerndauer ebenso gut geeignet ist, wie die Silbendauer insgesamt (Tamburini 2003, S. 130). Auch andere Merkmale wie Grundfrequenzverlauf, Leistung oder Spektralcharakteristik werden am natürlichsten auf dem Silbenkern berechnet (vgl. Kapitel 5) .

Es würde somit genügen, nur Silbenkerne zu kennzeichnen und die genaue Grenze im Konsonantencluster zwischen zwei Silbenkernen offenzulassen. Leider ist dies nicht möglich. Eines der verwendeten Werkzeuge zur Merkmalsextraktion, PaIntE (vgl. Kapitel 5), benötigt die Angabe von Silbengrenzen und die eindeutige Zuordnung jedes Lautes zu einer Silbe. Ambisilbische und extrasilbische Konsonanten können deshalb nicht berücksichtigt werden.

Nur in zwei der benutzten Korpora sind Silbengrenzen annotiert. Für die anderen mussten sie zunächst ermittelt werden. Dafür wurde ein einfaches Verfahren implementiert. Um den Aufwand dafür möglichst klein zu halten, wurden keine morphologischen Informationen genutzt.

Abbildung 4.1. Vereinfachtes Modell der Silbenstruktur



Stattdessen beruht der Algorithmus auf Regeln zum Silbenaufbau, deutschen phonotaktischen Restriktionen zur Phonemabfolge und verfolgt das *Maximum Onset Principle*, bei dem nur unter bestimmten Voraussetzungen Konsonanten dem Silbenauslaut zugeordnet und alle übrigen dem Silbenanlaut zugeordnet werden.

Nicht zuletzt stand das Ziel, jedes Phonem genau einer Silbe zuzuordnen und jede Silbe genau einem Wort. Eine Silbifizierung über Wortgrenzen hinweg wurde vermieden.

- Zunächst werden die Silbenkerne bestimmt. Jeder Vokal gilt zunächst als Silbenkern. Direkt auf einen Kurz- oder Langvokal folgende reduzierte Vokale (`/@/` und `/6/`) werden dem vorangehenden Vokal zugeschlagen, um der unterschiedlichen Notierung von Schwa-Diphthongen in den Korpora Rechnung zu tragen.[13]

- Stehen die Silbenkerne fest, so wird für Silbenkerne, die aus einem Kurzvokal bestehen, der folgende Konsonant dem Auslaut zugeordnet, da Kurzvokale keine offenen Silben bilden.

- Iterativ werden bis zu drei Konsonanten vor jedem Silbenkern dem Anlaut zugeschlagen, wobei jeweils geprüft wird, ob das entstehende Konsonantencluster den Bedingungen der Anlaut-Klassen aus (Spranger 2001, S. 21f.) genügt.

- Hinter den Kernen im Wort verbliebene Konsonanten werden den Auslauten zugeschlagen. Die Beschränkung des Deutschen auf maximal 5 Konsonanten im Silbenauslaut wird nicht durchgesetzt, um sicherzustellen, dass jedes Segment einer Silbe zugeordnet ist.

- Wortinitial verbliebene Konsonanten werden aus demselben Grund dem Anlaut der ersten Silbe des Wortes zugeschlagen.

Das angegebene Verfahren annotiert die im Kiel-Korpus häufig vorkommenden silbischen Konsonanten (wie in „geben“: `/g e b . =m`[14]`/`) nicht korrekt als Silben. Das zur Merkmalsextraktion verwendete Werkzeug PaIntE erwartet zwingend vokalische Silbenkerne. Eine korrekte Behandlung silbischer Konsonanten ist also nicht möglich. Stattdessen werden silbische Konsonanten der vorstehenden Silbe zugeschlagen.

Die Silbifizierung ist also fehlerhaft. hinsichtlich der Erkennung von Akzentuierungen ergeben sich jedoch keine Probleme: Silbische Konsonanten kommen im Deutschen nur vor, wenn der phonemisch vorhandenene vokalische Silbenkern durch Reduktion verschwindet. Ein silbischer Konsonant steht deswegen niemals in einer akzentuierten, sondern immer in einer reduzierten Silbe. Die eingeschlagene Vorgehensweise führt also lediglich dazu, dass weniger und dafür längere Silben entstehen. Diese Silben können nur an einer Stelle eine Akzentuierung tragen, die dann vom Klassifizierer gelernt beziehungsweise durch ihn vorhergesagt wird.

4.7. Zuordnung von Schrifttext zu Sprechtext

Für zwei der Korpora (IBM und IMS) lag keine direkte Zuordnung von Schrifttext zu Segmentfolge vor. Stattdessen war in den Korpora lediglich eine Zuordnung der Segmente zum normalisierten Sprechtext gegeben.

Jeder Sprecher setzt automatisch *zusammenhängende Ausdrücke* wie Zahlen, Uhrzeiten, Daten, Abkürzungen, URLs, usw. in die gesprochene Form um. „700 000 DM am 3.1.1981 um 6:38 Uhr“ wird zu „sieben hundert tausend deutsche mark am dritten januar neun zehn hundert ein und achtzig um sechs uhr dreißig“. In der Sprachsynthese heißt dies *normalisieren* und ist ein wichtiger Schritt zur korrekten Aussprache beliebiger Texte.

In Sprachsynthesekorpora ist die zugrundeliegende Form unwichtig, solange das Korpus nur zur Auswahl möglichst gut verknüpfbarer Einheiten genutzt wird. Viele für die Prosodieanalyse wertvolle Informationen liegen aber im Schrifttext (Satzzeichen, Worthäufigkeiten der nichtnormalisierten Wörter) oder können nur aus ihm gewonnen werden (Wortarten, Syntaxbäume). Diese Ergebnisse können nur verarbeitet werden, wenn eine Zuordnung von Schrifttext zu Sprechtext vorliegt.

Für diese Zuordnung wird der Schrifttext durch das TTS-System normalisiert. Im Anschluss wird die Normalisierung mit der im Korpus annotierten Wortfolge abgeglichen. Es ergeben sich Unterschiede

[13] Möglicherweise hätten auf einen Kurzvokal folgende Kurz-, Lang- oder Nasalvokale in einen Silbenkern integriert werden sollen um nichtnative Diphthonge zu modellieren.
[14] Mit „=“ vor einem Konsonanten kennzeichnet SAMPA den Konsonanten als silbisch.

zwischen Normalisiererungsergebnis und annotiertem Sprechtext, zum Beispiel die unterschiedliche Behandlung von Zahlen oder Abkürzungen. Der Abgleich der beiden Wortfolgen findet deswegen mit einem Minimum-Edit-Distance-Algorithmus statt, der die wahrscheinlichste Anordnung der unterschiedlichen Wortfolgen ermittelt. Am Ende werden nur jene Grenzen zwischen Wörtern übernommen, die im Korpus und im normalisierten Text übereinstimmen.

Damit steht eine Annotierung zur Verfügung, in der Grenzen zwischen Schrifttextwörtern annotiert sind. Nur an diesen Grenzen können Phrasengrenzen liegen, nur für diese Einheiten lassen sich Wortart oder Entfernung im Parsing-Baum zuverlässig mit externen Werkzeugen erzeugen.

Ebenfalls werden Wortgrenzen innerhalb von zusammenhängenden Ausdrücken entfernt. Eine Analyse am IMS-Korpus ergab, dass dies 423 Wortgrenzen (rund 3 % der Wortgrenzen) betrifft und darunter nur 2 Phrasengrenzen auftreten. Die Tilgung dieser Wortgrenzen aus der Liste der potentiellen Phrasengrenzen ist also zu über 99,5 % korrekt.

4.8. Vergleichende Statistiken

Abschließend werden einige statistische Daten der verwendeten Korpora angegeben und erläutert. Für das Kiel-Korpus sind nur die Daten für die beiden Sprecher angegeben, die jeweils das ganze Korpus gesprochen haben (*kko* und *rtd*). Nur diese Daten werden im sprecherabhängigen Teil dieser Arbeit verwendet. Gleichzeitig repräsentieren sie den Rest des Kiel-Korpus gut.

Tabelle 4.2 gibt neben der absoluten Größe der Korpora in Äußerungen, Wörtern und Silben auch mittlere Längen der Äußerungen und Wörter an. Außerdem sind die absolute Zahl und der jeweilige Anteil akzentuierter Silben und Phrasengrenzen angegeben.

Tabelle 4.2. Übersicht verschiedener Kennzahlen der verwendeten Korpora

	IBM	IMS	VPSC	KCoRS(kko)	KCoRS(rtd)
Äußerungen	1841	1780	733	595	585
Wörter	20188	14966	7489	4799	4579
Silben	38667	29996	13778	7324	6849
Akzentuierungen	8211	6551	3078	2063	1945
starke Phrasengrenzen[*]	971	1311	273	344	376
schwache Phrasengrenzen	3095	1099	588	keine[†]	
durchschnittliche Zahl Wörter pro Äußerung	11,0	8,4	10,2	8,1	7,8
durchschnittliche Zahl Silben pro Wort	1,9	2,0	1,8	1,5	1,5
Prosodieannotierung	minimal	GToBI	GToBI	PROLAB	PROLAB
Anteil akzentuierter Silben	21,2 %	21,8 %	22,3 %	28,2 %	28,4 %
Anteil Wortgrenzen mit starker Phrasengrenze[*]	5,2 %	9,9 %	4,0 %	8,1 %	9,4 %
Anteil Wortgrenzen mit schwacher Phrasengrenze	16,8 %	8,3 %	8,7 %		

[*] Die äußerungsfinalen Phrasengrenzen sind nicht mitgezählt.

[†] Im Kiel-Korpus wird die Stärke der Phrasengrenzen nicht unterschieden.

Die Tabelle zeigt, dass sich die Korpora nicht nur in ihrer Gesamtgröße, sondern auch in der Länge der einzelnen Äußerungen unterscheiden: Das Kiel-Korpus mit seinen manuell ausgewählten Sätzen zeichnet sich durch im Schnitt besonders kurze Äußerungen aus. Die Äußerungen im IMS-Korpus enthalten zwar ähnlich wenige Wörter pro Äußerung, jedoch sind die einzelnen Wörter länger.

Die Länge der Äußerungen in den Korpora ist in den Histogrammen in Abbildung 4.2 und Abbildung 4.3 weiter aufgeschlüsselt.

Abbildung 4.2. Relative Anzahl der Wörter pro Äußerung in den Korpora

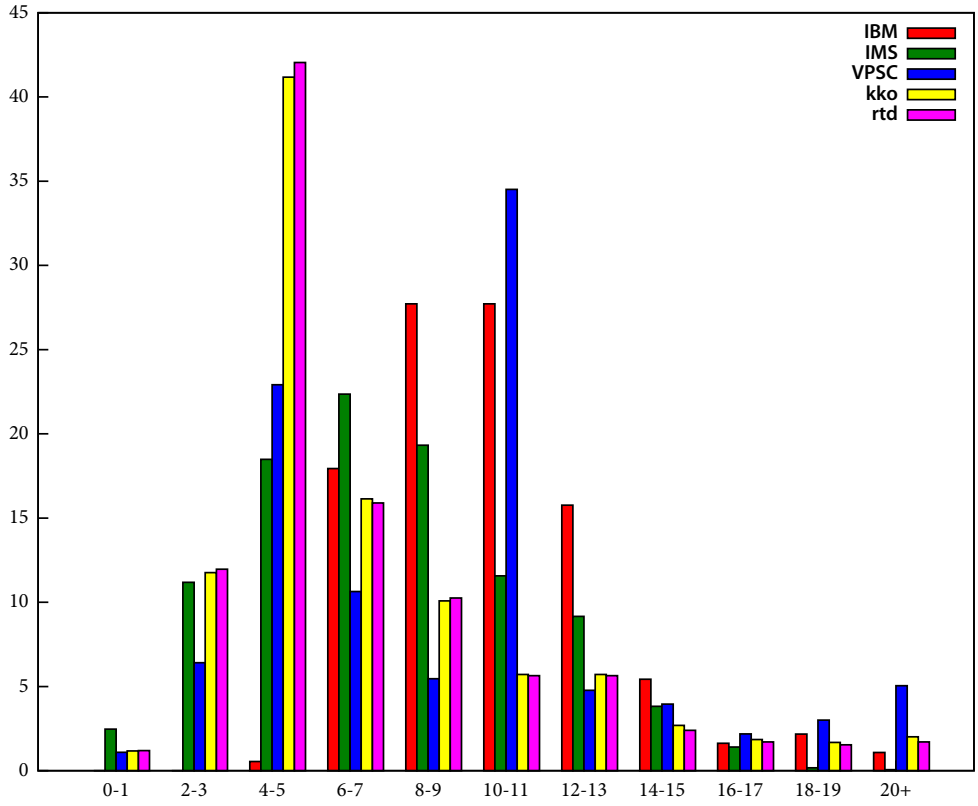
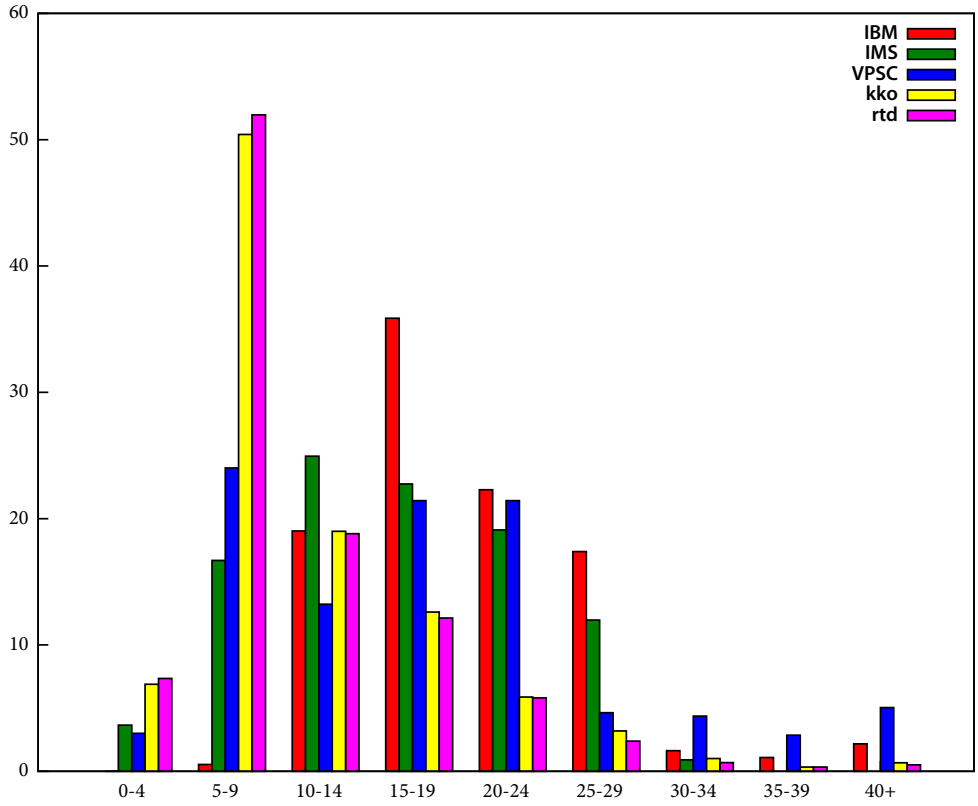


Abbildung 4.3. Relative Anzahl der Silben pro Äußerung in den Korpora



Deutlich zu erkennen ist, dass die meisten Äußerungen im Kiel-Korpus vier bis fünf Wörter beziehungsweise fünf bis neun Silben lang sind. Diese Häufung ergibt sich vor allem durch das phonetisch ausgeglichene Material aus PHONDAT 90. Die Länge der übrigen Äußerungen streut sehr viel stärker (Brinckmann 2004, S. 27).

Im VPSC fallen die Äußerungen des Frage/Antwort-Korpus mit ihrer immer gleichen Länge von elf oder zwölf Wörtern ins Auge. Die Silbenzahl dieser Äußerungen variiert etwas stärker, da nicht immer die gleichen Wörter realisiert werden. Die PHONDAT-90-Sätze im VPSC führen auch hier zu vielen Sätzen mit nur 5 oder 6 Wörtern und wenigen Silben. Durch die ebenfalls enthaltenen Zeitungstexte enthält das VPSC außerdem mehr lange Sätze als die anderen Korpora.

Sowohl im IBM- als auch im IMS-Korpus sind die Längen der Äußerungen in Wörtern und Silben eher normal verteilt, wobei die Äußerungen im IBM-Korpus im Schnitt etwas länger sind und ihre Länge am ehesten der Normalverteilung gleicht. Die kürzeste Äußerung im IBM-Korpus überhaupt ist schon 5 Wörter lang.

Bei den Akzentuierungen fällt auf, dass der Anteil akzentuierter Silben in den IBM-, IMS- und Wien-Korpora fast genau übereinstimmt. Das Kiel-Korpus hingegen hat einen deutlich höheren Anteil akzentuierter Silben als die übrigen Korpora. Dafür gibt es mehrere mögliche Gründe:

1. Die unterschiedlichen Labelling-Systeme: Auf der einen Seite GToBI mit einheitlich ~21,5 % akzentuierter Silben, auf der anderen Seite PROLAB mit einheitlich ~28 %.

2. Die unterschiedliche Segmentannotierung und die daraus resultierende mangelhafte Silbifizierung des Kiel-Korpus: Viele der reduzierten Silben am Wortende konnten im Kiel-Korpus nicht als eigene Silben gezählt werden, da bei ihnen kein vokalischer Silbenkern notiert war. Da reduzierte Silben niemals akzentuiert sind, erhöht sich der Anteil der akzentuierten Silben gegenüber den korrekt silbifizierten Korpora.

3. Ideolektale Unterschiede zwischen den Sprechern: Möglicherweise akzentuieren manche Sprecher mehr, andere weniger.

Um dies weiter zu untersuchen, wurde eine Statistik der Akzentuierungen nur der Marburg-Sätze erstellt. Die Marburg-Sätze kommen sowohl im Kiel-Korpus als auch im VPSC vor, wodurch der Einfluss des Textes auf die Prosodieannotierung ausgeschlossen werden kann.

Tabelle 4.3. Silben und Akzentuierungen in den *Marburg-Sätzen*

	VPSC	kko	rtd
Silben	825	761	738
Akzentuierungen	239	265	253
Anteil akzentuierter Silben	29,0 %	34,8 %	34,3 %

Aus Tabelle 4.3 ergibt sich, dass tatsächlich die ermittelten Silbenanzahlen in VPSC und Kiel-Korpus stark voneinander abweichen. Dies kann den deutlichen Unterschied im Anteil akzentuierter Silben zwischen den Korpora erklären.

Bereits die absolute Zahl von Akzenten im VPSC ist geringer als bei den beiden KCoRS-Sprechern. Ob dies auf Unterschiede zwischen den Labelling-Systemen, auf den Ideolekt und damit die individuelle Neigung der Sprecher zur Akzentuierung oder auf unterschiedliche Aufnahmebedingungen (zum Beispiel unterschiedliche Anweisungen zur Deutlichkeit an die Sprecher) hinweist, bleibt unklar.

Insgesamt fällt auf, dass die Silben in den Marburg-Sätzen relativ gesehen von allen Sprechern häufiger akzentuiert werden als im Durchschnitt der jeweiligen Korpora. Die Marburg-Sätze sind konstruierte Sätze, die nicht der Realität entsprechen: Im Schnitt sind sie fünf Wörter lang (bei geringer Varianz) und enthalten, je nach Sprecher, 7 bis 8 Silben. Außerdem sind alle gramatisch korrekt.

Der Anteil von Satzakzenten (also akzentuierten Wörtern im Satz) liegt in so kurzen Sätzen relativ hoch. Mindestens das Verb, meist zusätzlich Subjekt oder Objekt tragen einen Satzakzent. Da jedes

Wort im Schnitt nur eine oder zwei Silben lang ist ergibt sich pro Satzakzent ein hoher Anteil von Akzentuierungen. In den Sätzen kommen praktisch keine Füllwörter vor, die unakzentuiert bleiben würden.

Die relativ starke Gewichtung kurzer und konstruierter Sätze im Kiel-Korpus kann also ebenfalls zum beobachteten hohen Akzentuierungsanteil im Kiel-Korpus führen.

Kapitel 5. Systemaufbau und Merkmalsextraktion

Im praktischen Teil der Arbeit wurde ein System zur automatischen Erkennung und Annotierung der phonetischen Prosodiephänomene Akzentuierung und Phrasierung entwickelt. Dafür werden aus den Korpora wichtige Merkmale extrahiert. Zuvor trainierte Klassifizierer bilden dann die Grundlage für die Erkennung. Die bereits prosodieannotierten Korpora liefern die Grundlage für das Training der Klassifizierer und die Evaluierung ihrer Leistung.

5.1. Klassifizierung

Bei der Klassifizierung stellt sich die Frage, wie viele und welche Klassen unterschieden werden sollen. Silben sind nicht nur unterschiedlich stark akzentuiert, sie können am Ende eines Wortes oder bei schneller Sprechweise auch *reduziert* sein. Außerdem ist an der Akzentuierung meist einer von mehreren unterschiedlichen Akzenttönen beteiligt (vgl. Kapitel 2), die unterschieden werden.

Für Phrasengrenzen gilt entsprechend, dass es unterschiedlich stark realisierte Phrasengrenzen gibt, die durch unterschiedliche Phrasen- beziehungsweise Grenztöne angezeigt werden können. Hinzu kommt bei Phrasengrenzen im Tonsequenzmodell noch die phonologische Unterscheidung von intermediären Grenzen und Intonationsphrasengrenzen (vgl. Kapitel 2).

Technisch gesehen ist eine größere Zahl von Klassen mit einer stärkeren Aufteilung des Trainingmaterials sowie weniger Zufallstreffern und damit potentiell schlechteren Ergebnissen verbunden. Insbesondere die Unterscheidung zwischen Klassen mit ähnlichen Eigenschaften ist schwierig.

Die Wahl der Klassen muss zuvorderst in Übereinstimmung mit den zur Verfügung stehenden Daten getroffen werden. Die Nutzung unterschiedlicher und uneinheitlich annotierter Korpora erfordert bei der Klassifizierung einen Minimalkonsens:

- Für die Akzentuierung werden die beiden Klassen *akzentuiert* und *nicht akzentuiert* unterschieden.
- Für die Phrasierung werden die drei Klassen *Vollgrenze*, *Zwischengrenze* und *keine Grenze* unterschieden.[1]

Die beiden Phänomene Akzentuierung und Phrasierung werden in dieser Arbeit wegen ihrer unterschiedlichen Konstituenten getrennt betrachtet. Für beide wird jeweils ein Klassifizierer trainiert, der zwei beziehungsweise drei Klassen erkennt.

Eine Integration der Akzentuierungs- und Phrasierungsklassifizierung, ebenso wie die stärkere Beachtung des Kontextes ging über den Rahmen dieser Arbeit hinaus. Bespielsweise wäre ein N-Gramm-Modell auf Basis der einzelnen Klassifizierungsergebnisse möglich.

5.2. Datenhaltung

Die in den Korpora enthaltenen Äußerungen liegen als Audio-Dateien im Wave-Format oder in headerlosen Rohformaten vor. Die Korpora benutzen unterschiedliche Annotierungsformate. Entweder waren die einzelnen Annotierungsebenen in eigenen Dateien gespeichert, oder in einer Datei verbunden.

Die unterschiedlichen Annotierungsformate wurden so konvertiert, dass die einzelnen Ebenen jeweils in einzelnen Dateien stehen. Dies erlaubt eine einfache Kontrolle und die manuelle Nachbearbeitung von Zwischenergebnissen. Teilweise enthielten die Korpora nicht alle benötigten Annotierungsebenen. Sie wurden dann automatisch (Silbifizierung) oder teil-automatisch (Schrifttext), wie bereits in Kapitel 4 beschrieben, hinzugefügt.

[1] Im Kiel-Korpus werden die Grenzen nicht weiter unterschieden. Es gibt daher keine Zwischengrenzen.

Die verwendete Annotierung der Korpora enthält folgende Ebenen:

- Audio
- Schrifttext
- Sprechtext
- Silben
- Segment-Alignment
- Akzentuierungen (unabhängig von der Art der Akzentuierung)
- Phrasierungen (Vollgrenzen und Zwischengrenzen)

Die Verknüpfung der Ebenen untereinander erfolgt ausschließlich über die Zeit. Die erforderliche Hierarchiebildung (Segmente zu Silben, Silben zu Wörtern, ...) erfolgt erst zur Laufzeit. Dies erleichtert die unabhängige Darstellung der Annotierungsebenen mit dem Audiosignal[2].

Für die unterschiedlichen Annotierungsebenen wurden zwei Formate benutzt:

1. Das *xwaves*-Format gibt den Zeitpunkt des annotierten Ereignisses an. Damit eignet es sich für die Annotierung von Akzentuierungen und Phrasengrenzen.
2. Das *wavesurfer*-Format gibt den Zeitraum an, in dem das annotierte Ereignis gültig ist. In diesem Format werden Text, Silben und Segmente annotiert.

5.3. Merkmalsextraktion

Die verwendeten Merkmale wurden teilweise zur Laufzeit ermittelt, wie zum Beispiel Silbendauer aus der Dauer der zugehörigen Segmente. Aufwendigere Merkmale wurden separat berechnet und in zusätzlichen Annotierungsdateien gespeichert.

Die Einbindung externer Module zur Merkmalsextraktion war durch die zusätzlichen Annotierungsdateien besonders einfach. Es genügten kurze Skripte zur Anpassung der Eingangs- und Ausgangsdateien dieser Module. Auch konnten Fehler (beziehungsweise unerwünschtes Verhalten) der Module problemlos manuell korrigiert werden.

Phrasierung und Akzentuierung unterscheiden sich durch ihre Konstituenten. Die Stellen, für die Merkmale und anschließend ihre Klasse bestimmt werden müssen, unterscheiden sich also:

- Die Akzentuierung konstituiert sich auf Silbenebene. Die Merkmale werden deshalb für jede Silbe extrahiert und anschließend wird für jede Silbe bestimmt, ob sie akzentuiert ist oder nicht.
- Phrasierungen liegen immer an Wortgrenzen. Die Extraktions- und Klassifizierungsstellen liegen deswegen zwischen Wörtern.

5.4. Merkmale zur Akzentuierungserkennung

Wie in Kapitel 2 erläutert, äußert sich die Akzentuierung vor allem in den phonetischen Merkmalen Dauer, Qualität, Intensität und durch Akzenttöne im Tonhöhenverlauf. Diese entsprechen den akustischen Parametern Segmentdauern, Leistung, F0-Verlauf, sowie Parametern für die Lautqualität.

5.4.1. Silbenkern

Ein bisher nicht erwähntes sehr wichtiges Merkmal ist der Silbenkern. Das kategorielle Merkmal *Silbenkern* erhält für jeden Vokal und für jede Vokalkombination in Diphthongen eine Kategorie.

Warum ist der Silbenkern wichtig? Silben mit reduziertem Vokal als Silbenkern (`/@/` oder `/6/`) können nicht akzentuiert werden. Falls solche Silben akzentuiert gesprochen werden, so wird statt des reduzierten der entsprechende vollwertige Vokal (`/E/` beziehungsweise `/a/` oder `/e6/`) gesprochen (Gibbon 1998, S. 80).

[2] Hierfür, sowie für die Prosodieannotierung des IBM-Korpus, wurde das freie Programm Wavesurfer (http://www.speech.kth.se/wavesurfer/) benutzt.

In den Korpora führt das automatische Alignment beziehungsweise die Segmentierung mithilfe einer kanonischen Transkription dazu, dass tatsächlich einige akzentuierte Silben mit reduziertem Silbenkern vorkommen. Dies liegt daran, dass die *Hyperkorrektur* (Gibbon 1998, S. 80) dieser reduzierten Vokale in vollwertige Vokale nicht korrekt annotiert ist.

Auch darüber hinaus ist die Verteilung von akzentuierten und unakzentuierten Silben abhängig vom Silbenkern. In den Korpora sind Silben mit langen Vokalen deutlich überproportional häufig akzentuiert. Greenberg zeigt ähnliche Ergebnisse für das Englische (Greenberg 2005, S. 124).

Der Silbenkern bildet außerdem die Basis für die im Folgenden vorgestellten akustischen Merkmale. Verschiedenartige Silbenkerne haben unterschiedliche inhärente Tonhöhen, Intensitäten und Dauern (vgl. Abschnitt 2.4). So ist ein Diphthong im Schnitt deutlich länger als ein einfacher Silbenkern. Das Merkmal Silbenkern ist deswegen für die Klassifizierer sinnvoll, weil es die Aufspaltung des Trainingsmaterials in homogene Teilmengen unterstützt.

Viele der folgenden Merkmale werden nicht nur roh angegeben. Zusätzlich erfolgt eine *Normierung* des Werts relativ zum Mittelwert des Merkmals über alle Vorkommen des jeweiligen Silbenkerns.

Die automatische Merkmalsauswahl bestimmt dann, welche Kombination aus rohem Merkmal, normiertem Merkmal und Silbenkernmerkmal die besten Ergebnisse liefert.

5.4.2. Dauermerkmale

Akzentuierungen zeigen sich durch Längung. Sie schlagen sich also in der Dauer der realisierten Segmente und Silben nieder. Tamburini (2003) zeigt, dass die Längung der Silbe sich in gleicher Weise auch im Silbenkern zeigt.

Der Silbenkern hat gleichzeitig den Vorteil, dass das Ergebnis nicht durch etwaige Silbifizierungsfehler verschlechtert wird und die Dauer eigentlich extrasilbischer Konsonanten nicht von der Silbendauer abgezogen werden muss.
Folgende Merkmale werden benutzt:

- *Silbendauer*: Die Dauer der gesamten ermittelten Silbe.
- *Silbenkerndauer*: Die Dauer des Silbenkerns. Bei Diphthongen besteht dieser aus mehreren Segmenten. Dies schließt Schwa-Diphthonge durch R-Vokalisierung mit ein.
- *normierte Silbenkerndauer*: Die Dauer des Silbenkerns relativ zur mittleren Dauer aller Vorkommen des gleichen Silbenkerns im Korpus zur Berücksichtigung der Mikroprosodie.
- *Silbenkernanteil*: Die Dauer des Silbenkerns relativ zur Gesamtdauer der Silbe.

5.4.3. Einfache akustische Merkmale

Grundfrequenz und Schalldruckleistung der Audiosignale werden mit dem *Snack*-Toolkit[3] alle 10 ms ermittelt. Für jede Silbe wird dann der Median aller innerhalb des Silbenkerns liegenden Messwerte ermittelt. Durch zusätzliche Normierung der Merkmale wird die Mikroprosodie berücksichtigt.

- *Grundfrequenz*: Median der Messwerte innerhalb des Silbenkerns.
- *Schalldruckleistung*: Median der Messwerte innerhalb des Silbenkerns.
- *normierte Grundfrequenz*, *normierte Leistung*: Die Mikroprosodie beeinflusst die inhärente Grundfrequenz und Schalldruckleistung der Laute. Sie werden unter anderem durch Öffnungsgrad und Lippenrundung der Vokale sowie den konsonantischen Kontext bestimmt. Grundfrequenz und Schalldruckleistung werden für diese beiden Merkmale deswegen relativ zum Mittelwert aller Vorkommen des gleichen Silbenkerns im Korpus normiert.
- *Grundfrequenz im Kontext*: Zusätzlich zu der beschriebenen Normierung wird die Grundfrequenz auch relativ zur mittleren Grundfrequenz der Äußerung angegeben. Einzelne Äußerungen sind durch die Sprecher insgesamt etwas höher oder tiefer realisiert als die übrigen. Dafür kann die individuelle Performanz des Sprechers, Fortsetzung der Aufnahmen nach län-

[3] http://www.speech.kth.se/snack/

gerer Pause oder ein bewusster *Registerwechsel* durch den Sprecher verantwortlich sein. Dieses Merkmal ist gegenüber solchen Effekten robust.

5.4.4. Lautqualität

Die Längung akzentuierter Silben führt dazu, dass mehr Zeit für die Einstellung der Artikulationsorgane zur Verfügung steht. Der größere Artikulationseinsatz in akzentuierten Silben führt ebenfalls zu einer höheren Genauigkeit bei der Einstellung der Artikulationsorgane.

Die genauere Artikulation akzentuierter Silben schlägt sich in der Vokalqualität nieder. Die Formanten von akzentuierten Vokalen sind stärker ausgeprägt und liegen näher an den *kanonischen Formanten* für einzeln realisierte Laute, wie sie zum Beispiel Neppert (1999, S. 147) zitiert. Nichtakzentuierte Vokale werden hingegen stärker *zentralisiert*, das heißt ihr Formantenspektrum wird in Richtung Schwa-Laut verschoben (Neppert 1999, S. 150).

Als Merkmale werden die Formanten *F1*, *F2*, *F3* und *F4* jeweils in der Mitte der Silbenkerne berechnet (ebenfalls mithilfe des *Snack*-Toolkits). Anstatt einer Mittelwert- oder Medianbildung über den gesamten Silbenkern wird hier die zeitliche Mitte des Silbenkerns benutzt, da sich hier die stationäre Phase des Vokals befindet. Hier sind die Formanten am wenigsten durch Koartikulationseffekte beeinflusst und kommen der kanonischen Form am nächsten.[4]

Natürlich hängen die Formanten vom Typ des Silbenkerns ab. Deswegen werden die Formant-Merkmale zusätzlich normiert. Die Merkmale *normierter F1*, *normierter F2*, *normierter F3* und *normierter F4* geben die Formanten relativ zu den durchschnittlichen Formanten für den jeweiligen Silbenkern wieder.

5.4.5. Tonhöhenverlauf

Akzentuierungen äußern sich vor allem durch die intonatorischen Akzenttöne im Tonhöhenverlauf. Mit den oben erwähnten einfachen Grundfrequenzmerkmalen ist der Tonhöhenverlauf nur unzureichend beschrieben. Die Grundfrequenz im Silbenkern ist außerdem nicht unbedingt relevant, da die Akzenttöne auch schon vor und nach der akzentuierten Silbe realisiert werden können.

Zur Beschreibung des Tonhöhenverlaufs werden von verschiedenen Autoren unterschiedliche Merkmale benutzt. Batliner et at (2001) nutzen unter anderem den Frequenzumfang innerhalb der fraglichen Silbe, sowie darauf basierende Merkmale. Tamburini (2003) benutzt eine Parametrisierung des Grundfrequenzverlaufs mit dem *TILT*-Modell (Taylor 1998).

5.4.5.1. Parametrisierte intonatorische Ereignisse

Diese Arbeit nutzt *parametrisierte intonatorische Ereignisse* (PaIntE, Möhler 1998, Möhler und Conkie 1998) zur Beschreibung des Tonhöhenverlaufs. PaIntE beschreibt auf der Basis der Grundfrequenz den Tonhöhenverlauf im Bereich von Akzentuierungen mit wenigen Parametern.

Der Tonhöhenverlauf wird durch eine Funktion approximiert, die aus zwei Sigmoiden besteht. Ein Sigmoid lässt sich durch vier Parameter bestimmen: Grundlinie, zeitliche Ausrichtung, Amplitude und Steigung. Im PaIntE-Modell haben beide Sigmoide dieselbe Grundlinie sodass sich ein kontinuierlicher Kurvenverlauf ergibt. Außerdem ist die zeitliche Ausrichtung aneinander gekoppelt.

Die Kopplung der beiden Sigmoide bewirkt, dass die PaIntE-Funktion durch sechs Parameter beschrieben werden kann, die den Tonhöhenverlauf repräsentieren. Eine feste zeitliche Verzögerung zwischen den Sigmoiden bewirkt einen Gipfel im Funktionsverlauf.

Die Approximation erstreckt sich auf die vorhergehende, aktuelle und folgende Silbe. Dies ist der Bereich, in dem Akzentuierungen realisiert werden. Außerdem wird der eigentliche Grundfrequenzverlauf zunächst geglättet und in stimmlosen Bereichen linear interpoliert.

[4] Für Diphthonge gilt dies natürlich nicht. Welches aber der „richtige" Extraktionspunkt für die Bestimmung der Formanten wäre, ist unklar. Schließlich zeichnen sich Diphthonge dadurch aus, dass sich ihr Formantspektrum über die Zeit zwischen zwei Zielpunkten verschiebt. Deshalb wird auch für diphthongische Silbenkerne die zeitliche Mitte als Referenzpunkt benutzt.

Gleichung 5.1. PaIntE-Funktion

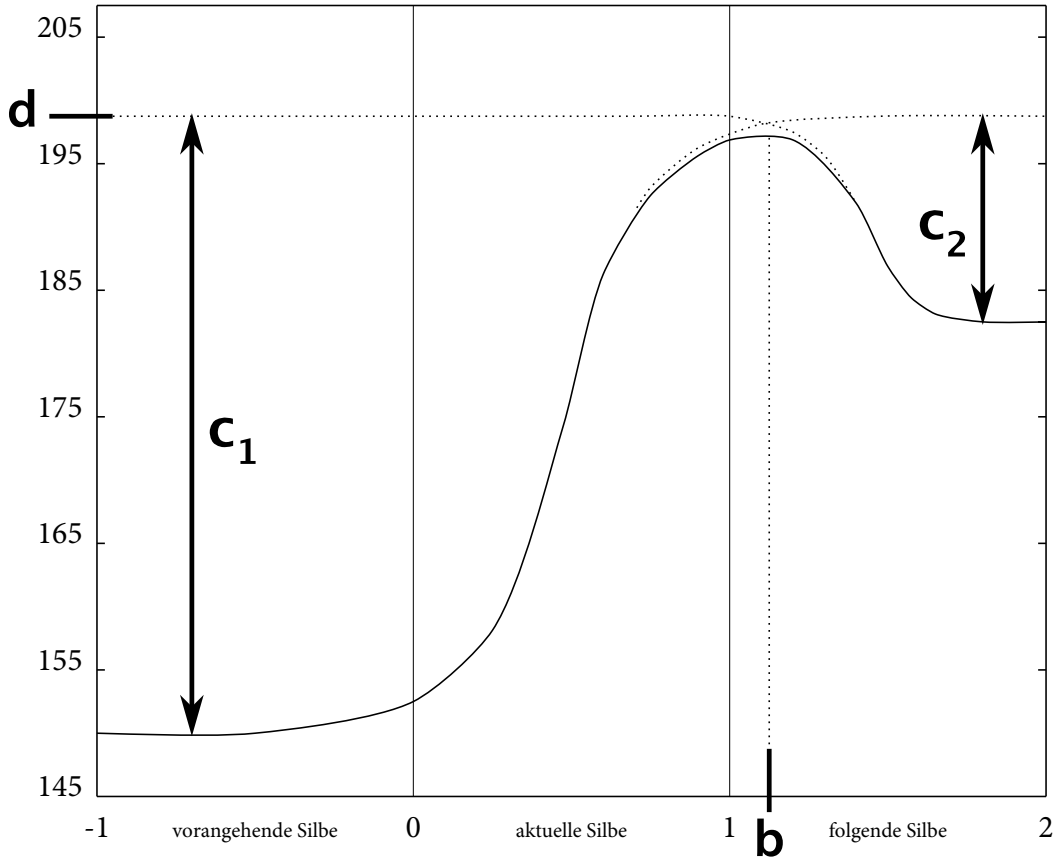
$$x(t) = d - c_1 / (1 + exp(-a_1 (b - t) + \lambda)) - c_2 / (1 + exp(-a_2 (b - t) + \lambda))$$

Gleichung 5.1 zeigt die zu approximierende Funktion und Abbildung 5.1 stellt die Funktion dar. Den sechs freien Parametern kann folgende Bedeutung zugeordnet werden, die jeweils auch aus Abbildung 5.1 ersichtlich sind:

- a_1 und a_2: Steigung der steigenden beziehungsweise fallenden Flanke,
- b: zeitliche Ausrichtung der Funktion,
- c_1 und c_2: Höhe der steigenden beziehungsweise fallenden Flanke,
- d: maximale Höhe der Funktion.

Der feste Parameter λ bestimmt die Verzögerung zwischen den beiden Sigmoiden und damit die Höhe und Breite des entstehenden Gipfels. Er wird so gewählt, dass der Gipfel nicht unter 96 % der Amplituden der Sigmoiden fällt.

Abbildung 5.1. PaIntE-Funktion



Neben den genannten sechs Parametern wird noch der *quadratische Fehler* zwischen der approximierenden Funktion und dem ursprünglichen Grundfrequenzverlauf als Merkmal benutzt.

5.4.5.2. Normalisierung der Parameter

Die Diskursstruktur hat einen starken Einfluss auf die Realisierung von Akzenttönen (Mayer 1999, nach Möhler 2001). Manche Phrasen werden eher gleichförmig, andere sehr dynamisch gesprochen. Die Merkmale zur Tonhöhe (c_1, c_2 und d) werden deswegen in der aktuellen PaIntE-Version (Möhler 2001) innerhalb der Phrasen normalisiert.

Der Stimmumfang der Phrase wird nach oben begrenzt durch den größten d-Wert innerhalb der Phrase und nach unten durch den niedrigsten Wert von $d - max(c_1,c_2)$. Die Werte der Tonhöhenmerkmale

werden dann auf diesen Bereich normalisiert. Ein Wert von $d = 0.5$ bezeichnet nach der Normalisierung einen halbhohen Akzent (Möhler 2001).

Die zeitliche Anordnung einer Akzentuierung relativ zur Silbe ist für die Akzenttonwahrnehmung besonders wichtig. Sie ist nicht linear, sondern hängt von den Konstituenten der Silbe ab. Möhler (2001) spricht von einer *Verankerung* der Akzenttöne innerhalb der Silbenstruktur. Deswegen werden die zeitlichen Merkmale (b, a_1 und a_2) auf die Silbenstruktur normalisiert.

Die Silbe wird unterteilt in die Konstituenten *stimmloser Silbenanlaut*[5], *sonoranter Silbenkern* und *Silbenauslaut*. Den drei Konstituenten werden die Intervalle [0; 0,5[, [0,5; 0,8] und [0,8; 1] zugeordnet. Der b-Wert wird dann innerhalb des entsprechenden Bereiches linear abgebildet.

Die Normalisierung für b-Werte innerhalb der vorhergehenden oder folgenden Silbe erfolgt entsprechend in die Intervalle [-1; 0] und [1; 2]. Die Skala in Abbildung 5.1 zeigt die Normalisierung. a_1 und a_2 werden passend zur Normalisierung von b angepasst.

Die Dauer der drei Silben auf die sich die PaIntE-Approximierung erstreckt hat einen direkten Einfluss auf den quadratischen Fehler: Je länger die drei Silben sind, desto größer der Fehler. Deswegen wird neben dem quadratischen Fehler zusätzlich der *mittlere quadratischer Fehler* berechnet, der den quadratischen Fehler durch die Dauer der drei zugrundeliegenden Silben teilt.

PaIntE ist ursprünglich dafür vorgesehen, den Tonhöhenverlauf im Bereich akzentuierter Silben zu beschreiben. In dieser Arbeit wird es jedoch dafür verwendet, den Akzentuierungsstatus aller Silben zu ermitteln. Dies führt zu dem Problem, dass jede Akzentuierung mehrfach und in unterschiedlichen Silbenfenstern erkannt und parametrisiert wird: Ein einfacher Akzentton H* wird deswegen zunächst als steigender Akzent mit spätem Gipfel (L*H), im folgenden Silbenfenster als Akzent mit mittlerem Gipfel (H*) und im nächsten Fenster möglicherweise noch als fallender Akzent mit frühem Gipfel (HL*[6]) beschrieben.

Für jede zu klassifizierende Silbe stehen deswegen die PaIntE-Merkmale der vorangehenden Silbe, der aktuellen Silbe und der folgenden Silbe zur Verfügung. Für tatsächlich akzentuierte Silben sollten sich die Merkmalsbündel der vorangehenden und aktuellen oder der aktuellen und folgenden Silbe nur in ihrem b-Wert um 1 unterscheiden, da sie sich auf denselben Gipfel beziehen.

Die Implementierung von PaIntE beruht auf *Festival*[7] (Taylor et al 1998). Als Eingabedaten für Festival werden Silbengrenzen benötigt, weshalb sie wie in Kapitel 4 beschrieben erzeugt wurden. Festival unterstützt zudem nur vokalische Silbenkerne, was die Silbifizierung wie in Kapitel 4 beschrieben einschränkt und die Ergebnisse für das Kiel-Korpus beeinträchtigt.

Nicht-phonetische Merkmale

Die bisher behandelten Merkmale sind rein phonetisch. Darüber hinaus sind aber noch weitere Merkmale für die Akzentuierungsvorhersage nützlich und wichtig, da prosodische Merkmale allein nicht ausreichen, um Akzentuierungen mit hinreichend hoher Genauigkeit vorherzusagen.

5.4.6. Silbenmerkmale

In Kapitel 2 wurde festgestellt, dass nur akzenttragende Silben akzentuiert sein können. Der Akzent ist deshalb ein sehr wichtiges Merkmal für die Vorhersage von Akzentuierungen.

Auf ein umfangreiches Akzentmodell, das für Wörter die akzenttragenden Silben bestimmt, wurde dennoch verzichtet. Bei der Vorbereitung der Annotierung des IBM-Korpus wurde das Akzentmodell eines TTS-Systems benutzt. Bei der anschließenden Handkorrektur ergaben sich Probleme durch die unterschiedliche Silbifizierung. Dem TTS-System steht nur die Textform zur Verfügung, weshalb teilweise die falsche Silbenzahl ermittelt wurde.

[5] Wenn der Anlaut aus einem Konsonantencluster besteht, dann zählt nur der stimmlose Anteil der Konsonantenclusters zum stimmlosen Anlaut. Eventuell übrige Konsonanten werden zum sonoranten Kern gezählt.
[6] Dieser Akzentton ist in GToBI nicht vorgesehen.
[7] http://www.cstr.ed.ac.uk/projects/festival/

Die Benutzung eines Akzentmodells hätte es zudem unmöglich gemacht, ein automatisch prosodieannotiertes Korpus zur Kontrolle oder Verbesserung der Akzentmodelierung zu nutzen. Der Verzicht erlaubt es, die Vorhersagen eines Akzentmodells zu prüfen, ohne dass sich eine zyklische Abhängigkeit und damit zu optimistische Schlüsse ergeben.

Dennoch wurden einige einfache Merkmale für die Akzentuierungsvorhersage benutzt:

- *Silbenposition*: Absolute Position der Silbe im Wort.
- *Silbenanzahl*: Die Silbenanzahl erlaubt, die absolute Position der Silbe im Wort besser einzuschätzen.
- *relative Silbenposition*: Die Kombination der Merkmale Silbenposition und Silbenanzahl.

5.4.7. Wortart

Die Verteilung der Akzentuierungen über die Wortarten ist sehr ungleichmäßig. Funktionswörter sind praktisch nie akzentuiert[8]. Auch die Inhaltswörter sind unterschiedlich häufig akzentuiert.

Darüber hinaus hilft die Wortart dabei, die unzureichende Bestimmung des Wortakzents zu verbessern. Während beispielsweise Verben im Infinitiv meist einen initialen Wortakzent tragen („GEHen“, „LAUfen“, ...), liegt er bei Partizipien fast immer auf der zweiten Silbe („geGANgen“, „geLAUfen“, ...).

Die Wortarten wurden durch das Programm *TreeTagger*[9] (Schmid 1995) bestimmt, welches Entscheidungsbäume zur Bestimmung der Wortarten benutzt. Die Wortarten werden aus dem STTS-Tagset (Schiller et al 1995) vergeben. Das Merkmal *Wortart* ist also kategorial.

Das STTS-Tagset teilt die Wörter in 54, sehr fein untergliederte Wortarten ein. Um den Klassifizierern die Untergliederung zu erleichtern, wurden die Wortarten Brinckmann (2004) folgend zusätzlich im Merkmal *vereinfachte Wortart* in zwölf Kategorien zusammengefasst: Satzzeichen, Nomen, Verb, Adjektiv, Adverb, Präposition, Pronomen, Artikel, Konjunktion, Partikel, Zahlwort und sonstige.

5.4.8. Worthäufigkeit

Die *Worthäufigkeit* ist ein Indikator für den Informationsgehalt eines Wortes. Sehr häufige Wörter tragen wenig zum Inhalt einer Aussage bei. Es gibt also eigentlich keinen Grund, solche Wörter zu akzentuieren. Funktionswörter gehören zu den häufigsten Wörtern und sind fast immer unakzentuiert.

Die Häufigkeit der einzelnen Wörter wurde einer Liste aus der Arbeit von Vera Demberg (2006) entnommen. Die Häufigkeitsliste basiert auf dem taz-Korpus.

Die Verwendung einer einfachen Häufigkeitsliste führt dazu, dass kein Kontext berücksichtigt werden kann. Zum Beispiel sind in der Domäne Fahrplanauskunft insgesamt seltene Wörter sehr häufig, was ihren Informationsgehalt innerhalb der Äußerung beeinflusst. Eine umfassende Informationsgehaltsvorhersage für jedes Wort kann in dieser Arbeit natürlich nicht geleistet werden.

5.4.9. Phrasenmerkmale

Die Realisierung der Akzentuierung sollte stark von der Position innerhalb der Phrase abhängen. Die Deklination führt dazu, dass sowohl die Grundlinie als auch die Dachlinie innerhalb der Phrase abfallen. Dies wirkt sich auf das Grundfrequenz-Merkmal als auch auf die PaIntE-Tonhöhenmerkmale aus.

Zum Phrasenende nimmt häufig auch die Intensität ab, sodass eine weniger starke Hervorhebung in Schalldruckleistung und Qualitätsmerkmalen zur Wahrnehmung einer Akzentuierung ausreicht. Gleichzeitig zeigt die Dehnung am Phrasenende keine Akzentuierung an, obwohl die Dauermerkmale dafür sprechen. Dehnung und Längung sind akustisch schwer zu unterscheiden.

[8] Akzentuierte Funktionswörter sind meist emphatisch akzentuiert.
[9] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Die Intensität der Äußerung ist am Phrasenanfang häufig stark, ohne dass eine Akzentuierung wahrgenommen wird. Teilweise ist hier auch der Tonhöhenverlauf sehr prägnant. Dieses Phänomen wird, je nach theoretischer Grundlage, als *phraseninitiale Kontur* (Kohler 1992b) oder als *phraseninitialer Grenzton* (Grice und Baumann 2000) bezeichnet. Erstere Bezeichnung erscheint mir einleuchtender, da meiner Erfahrung nach eben kein verortbarer Ton wahrgenommen wird, sondern das Phönomen über einen nicht klar begrenzten Bereich wirkt.

Phraseninitiale Phänomene werden in dieser Arbeit allerdings nicht weiter beachtet, da sie nur teilweise in den Korpora annotiert sind.

Die Nutzung von Phrasenmerkmalen bei der Akzentuierungsklassifizierung setzt voraus, dass diese Phrasenmerkmale zur Verfügung stehen. Dies ist in einem unannotierten Korpus zunächst nicht der Fall. Die Reihenfolge der Annotierungsschritte ist dann darauf festgelegt, dass zuerst die Phrasierung und nachfolgend die Akzentuierung annotiert würden. Eine Mischung der beiden Schritte (wie zunächst in der Arbeit beabsichtigt) wäre dadurch unmöglich.

In dieser Arbeit werden deshalb zwar Phrasenmerkmale ermittelt, im Anwendungsfall jedoch als mögliche Merkmale außeracht gelassen. Sie werden nur benutzt, um den Wert der Phrasierungsinformation für die Akzentuierungsannotierung zu bewerten. Neben den eigentlichen Phrasengrenzen stellen auch Satzzeichen sinnvolle Information zur Verfügung, die hier ebenfalls unter Phrasenmerkmale subsummiert sind. Satzzeichen haben den Vorteil, dass sie schon ohne eine Phrasierungsannotierung bekannt sind und unabhängig von ihr benutzt werden können.

Folgende Phrasenmerkmale werden ermittelt:

- *Abstand zur linken Phrasengrenze* in Wörtern
- *Abstand zur rechten Phrasengrenze* in Wörtern
- *Relative Phrasenposition* auf einer Skala von 0 (Phrasenanfang) bis 1 (Phrasenende)
- *Phrasentyp*: Zwischen- oder Vollgrenze
- *Abstand zum vorangehenden Satzzeichen*
- *Abstand zum folgenden Satzzeichen*
- *Relative Position zwischen den Satzzeichen*
- *Folgendes Satzzeichen*

5.5. Merkmale zur Phrasierungserkennung

Phonetisch äußert sich die Phrasierung, wie in Kapitel 2 erläutert, durch Grenztöne, Dehnung und Pausen. Neben den Grenztönen spielt auch die Deklination eine Rolle, ob sie nun ein phonetisches Zeichen oder ein artikulatorisches Artefakt ist. Außerdem werden Intensitätsunterschiede behandelt, die sich durch die über den Phrasenverlauf abnehmende Stimmspannung ergeben.

Eine signalbasierte Phrasierungserkennung führt nur zu unbefriedigenden Ergebnissen. Reyelt (1995) ermittelt eine Übereinstimmung zwischen unterschiedlichen Annotatoren für Phrasengrenzen von etwa 75 % (Reyelt 1995, S. 6).

Wird das Signal so verzerrt, dass der Inhalt nicht mehr verständlich ist, so sinkt die Übereinstimmung der Annotatoren mit ihrer ursprünglichen Annotierung auf etwa 50 % (Reyelt 1995, S. 8). Reyelt vermutet, dass die Phrasierungswahrnehmung stark durch die syntaktische Oberflächenstruktur beeinflusst ist.

Pfitzinger und Reichel (2006) nutzen in ansonsten gleichen Experimenten zur Phrasierungserkennung einmal textbasierte und einmal signalbasierte Merkmale und erhalten für letztere deutlich schlechtere Klassifizierungsergebnisse.

Strom und Widera (1996) untersuchen ebenfalls die Leistung von Annotatoren für verzerrte Sprache. Auch sie kommen zu dem Ergebnis, dass die prosodische Perzeption in einem starken Zusammenhang mit dem Verständnis des Gesprochenen einhergeht und eben dieses für die Prosodieannotierung notwendig ist: „Therefore we believe that detection with pure prosodic features cannot be substantially improved.“ (Strom und Widera 1996)

Neben signalbasierten Merkmalen werden deswegen auch einige textbasierte Merkmale für die Klassifizierer bereitgestellt.

5.5.1. Pausen

Ein Hauptmerkmal der Phrasierung liegt in der Einfügung kurzer Pausen in den Sprechstrom an externen offenen Junkturen. Das Pausenmerkmal beschreibt, ob im Korpus zwischen zwei Wörtern eine Sprechpause vorliegt oder nicht.

Zwischen den Korpora bestehen beträchtliche Unterschiede in der Pausenannotierung. Gerade bei den automatisch segmentierten Korpora hängt die Pausenhäufigkeit von der Neigung der Spracherkenner ab, am Übergang zwischen Wörtern eher gedehnte Laute zuzulassen oder Pausen einzufügen.

Gerade vor Plosiven sind in den automatisch segmentierten Korpora sehr selten Pausen annotiert auch wenn die Stille vor der Explosionsphase des Plosivs relativ lang ist. Das kategoriale Merkmal *Pause* erhält deswegen die drei Kategorien *ja*, *nein* und *Plosiv*.

Neben dem kategorialen Merkmal wird noch das kontinuierliche Merkmal *Pausendauer* genutzt, dass bei Pausen und Plosiven jeweils die Länge der Pause beziehungsweise des Plosivs angibt. Leider ist nicht in allen Korpora die stumme Phase und die Explosionsphase des Plosivs einzeln annotiert. Deswegen wird jeweils die Gesamtdauer der Plosive angegeben.

5.5.2. Finale Dehnung

Neben der Pause ist die Dehnung das zweite hauptsächliche Merkmal der Phrasierung (Kohler 1983, nach Reyelt 1995).

Dehnungen vor Phrasengrenzen zeigen sich vor allem zum Ende des letzten Wortes vor der Phrasengrenze. Dafür werden die Dauern der einzelnen Segmente der letzten Silbe gemessen. Aus der Dauer der Segmente relativ zur durchschnittlichen Dauer dieser Segmente bestimmt sich, ob die Laute langsamer oder schneller als durchschnittlich gesprochen worden sind. Das Merkmal *finale Dehnung* setzt sich nun aus dem arithmetischen Mittel der relativen Dauer für die einzelnen Laute der letzten Silbe zusammen.

Das beschriebene Merkmal berücksichtigt noch keine Schwankungen in der Sprechgeschwindigkeit: Für einen insgesamt langsam gesprochenen Satz ist jede Silbe langsamer gesprochen als in einem schneller gesprochenen. Entsprechend würde das Merkmal jeweils auf eine Phrasengrenze hindeuten.

Das Merkmal *relative finale Dehnung* beschreibt die Dehnung im Vergleich zur relativen Dehnung der ersten Silbe nach der potentiellen Phrasengrenze. Es ist also nur dann ausgeprägt, wenn die nachfolgende Silbe wieder schneller gesprochen wird als die vorangehende und nähert sich somit an das perzeptiv lokale Sprechtempo (Pfitzinger 1999 nach, Pfitzinger und Reichel 2006) an.

5.5.3. Intensität

Durch den mikroprosodischen Effekt der Deklination nimmt die Intensität über den Phrasenverlauf ab. Insbesondere direkt vor einer Phrasierung wird die Stimme schwächer. Eine neue Phrase zeigt sich häufig durch einen neuen Stimmeinsatz. Dieser Stimmeinsatz bewirkt einen Intensitätssprung zwischen der Silbe, die der Grenze vorangeht, und der ihr folgenden.

Die Intensität wird im folgenden durch die Schalldruckleistung repräsentiert. Folgende Merkmale werden extrahiert:

- *finaler Leistungsabfall*: Abfall der Leistung in der letzten Silbe entsprechend der finalen Dehnung.
- *relativer finaler Leistungsabfall*: Relativer Leistungsabfall entsprechend der relativen finalen Dehnung um Leisesprechstellen zu berücksichtigen.

- *Leistungssprung*: Unterschied zwischen den mittleren Leistungen des Silbenkerns vor und des Silbenkerns nach der potentiellen Phrasengrenze.
- *relativer Leistungssprung*: Der Sprung zwischen den auf die jeweiligen Silbenkerntypen normierten Leistungen.

5.5.4. Tonhöhenverlauf

Der Tonhöhenverlauf im Bereich von Phrasierungen wird phonologisch durch Grenztöne bestimmt. Nach einer Phrasengrenze setzt die Stimme neu ein. Dieser Neueinsatz erzeugt zusammen mit der über die Phrase verlaufenden Deklination einen Sprung im Tonhöhenverlauf am Phrasenende vor der Phrasengrenze und am Phrasenanfang nach der Grenze.

Es können also die zwei Phänomene Grenzton und Deklinationssprung gemessen werden.

Die Grenztöne[10] sind vor allem an schwachen Phrasengrenzen (Zwischengrenzen) wichtig, da diese nach meiner Erfahrung teilweise ausschließlich durch einen hohen Grenzton ohne Rücksetzung der Tonhöhe und ohne Pause oder deutliche Dehnung angezeigt wird[11].

Als Merkmale für die Grenztöne wurden deswegen die oben bereits erläuterten *PaIntE-Parameter* der vorletzten und letzten Silbe vor der potentiellen Phrasengrenze sowie der unmittelbar folgenden Silbe benutzt.

5.5.4.1. Regression des Grundfrequenzverlaufs

Eine Stichprobenuntersuchung am IBM-Korpus ergab, dass sich Grenztöne deutlich von Akzenttönen unterscheiden. Zunächst liegen sie immer am Wortende, im Bereich normalerweise unakzentuierter Silben. Außerdem ist der Frequenzanstieg zum Gipfel hin gleichmäßiger als bei Akzenttönen. Letzterer ähnelt etwa einer Glocke (siehe Abbildung 5.1), während Grenztöne (H- und H%) eher einem Dreieck gleichen.

Die Dauer des Frequenzanstiegs vor einer hohen Phrasengrenze lag im IBM-Korpus bei etwa 150 – 200 ms. Außerdem lag die Spitze des Gipfels von hohen Grenztönen immer sehr nah (< 10 ms) an der Wortgrenze.

Pierrehumbert schreibt dazu: „H* and H% are equally H tones but they differ in how they are associated with the text." (Pierrehumbert 1980, S. 29, nach Hirst und di Cristo 1998, S. 13) Ich möchte dem hinzufügen, dass sich die hohen Töne auch in ihrer Realisierung und Assoziierung am Sprachsignal unterscheiden.

Zusätzlich zu den PaIntE-Parametern wurden deswegen in Anlehnung an (Haase 2000) noch Parameter aus einer linearen Regression des Grundfrequenzverlaufs im Bereich vor und nach der möglichen Phrasengrenze ermittelt, die speziell für die Bestimmung hoher Grenztöne an Zwischengrenzen gewählt wurden.

Beim Sprechen können Gesten im Tonhöhenverlauf nur in stimmhaften Bereichen realisiert werden. Die Regression erstreckt sich deswegen auf den Zeitbereich bis zur letzten Stimmhaftigkeit die vom Algorithmus zur Grundfrequenzbestimmung festgestellt wurde. Für die Regression nach den möglichen Phrasengrenze gilt entsprechend, dass sie mit der Stimmhaftigkeit beginnt.

Die Regressionsgerade $x(t) = a \cdot t + b$ ist durch zwei Parameter (*Steigung* a und *Achsenabschnitt* b) beschrieben. Zusätzlich wird noch der *mittlere quadratische Fehler* der Regression angegeben.

Die Parameter werden jeweils im Bereich vor und im Bereich nach der möglichen Grenze ermittelt. Dadurch ist es möglich, Unterschiede in Steigung und Achsenabschnitt zu vergleichen. Die Merkmale *Wendung* und *Sprung* werden aus der Differenz der Merkmale Steigung beziehungsweise Achsenabschnitt vor und nach der möglichen Phrasengrenze bestimmt.

[10] Die ToBI-Unterscheidung zwischen Phrasen- und Grenztönen wird im folgenden nicht benutzt, sondern verallgemeinernd von Grenztönen gesprochen.
[11] Meine Beobachtung am IBM-Korpus.

Die Parameter wurden jeweils einmal über 150 ms und einmal über 200 ms ermittelt. Diese Bereiche sind nur genügend zur Bestimmung kurzzeitiger Tonphänomene.

Gerade an starken Phrasengrenzen ist ein Neueinsatz der Stimme und eine damit verbundene Rücksetzung der Tonhöhe zu erwarten. Um den Neueinsatz der Stimme zu modellieren ist ein Fenster von 150 oder 200 ms jedoch zu kurz. Deswegen wurden die Regressionsparameter zusätzlich über Zeiträume von 2000 ms vor und nach der möglichen Grenze ermittelt.

5.5.5. Einfache textbasierte Merkmale

Eine nur signalbasierte Phrasierungserkennung ist – wie oben beschrieben – ungenügend. Deswegen wurden auch zur Phrasierungsannotierung viele der oben bereits beschriebenen textbasierten Merkmale benutzt. Besonders wichtig ist hierbei die Interpunktion.

- *Wortart vor der potentiellen Grenze*
- *vereinfachte Wortart vor der potentiellen Grenze*
- *Wortart nach der potentiellen Grenze*
- *vereinfachte Wortart nach der potentiellen Grenze*
- *Worthäufigkeit vor der potentiellen Grenze*
- *Worthäufigkeit nach der potentiellen Grenze*
- *Abstand zum vorangehenden Satzzeichen*
- *Abstand zum folgenden Satzzeichen*
- *Relative Position zwischen den Satzzeichen*
- *Folgendes Satzzeichen*

5.5.6. Syntaktische Merkmale

Syntaktische Einschnitte bilden ein wichtiges Merkmal für die Phrasierung. Die vorangehende und folgende Wortart liefert indirekt bereits Informationen über den syntaktischen und korrelierend damit den prosodischen Zusammenhalt. Eine Syntaxanalyse bietet noch eine genauere Grundlage.

Eine – mehr oder weniger aufwendige – Syntaxanalyse wird häufig zur Phrasierungs*vorhersage* in TTS-Systemen gebraucht (vgl. Atterer 2005, Schweitzer 1999). Hier soll eine Syntaxanalyse auch zur Phrasierungs*erkennung* benutzt werden.

Dafür wird ein Parser benutzt, der automatisch einen Syntaxbaum der gesprochenen Sätze erstellt. Aus dem Syntaxbaum werden dann Merkmale über die Zusammengehörigkeit zwischen zwei Wörtern extrahiert.

Der benutzte Dependenzformalismus (Foth et al 2000) erstellt einen Baum von typisierten Abhängigkeiten zwischen Wörtern. Die Abhängigkeit besteht dabei direkt zwischen einzelnen Wörtern. Anders als in generativen Grammatiken werden keine Konstituenten gebildet.

Die Wurzel des Dependenzbaumes ist immer das finite Verb des Satzes. Die Kanten zwischen Wörtern sind jeweils mit ihrer grammatischen Funktion gekennzeichnet. So ist beispielsweise das Verb über eine Subjekt-Kante mit einem Nomen verbunden und dieses über eine Artikel-Kante mit seinem Artikel.

Die Konstruktion des Baumes wird durch *Beschränkungen* (engl. constraints) bestimmt, die die möglichen Zusammenhänge zwischen allen Wörtern einschränken. So wird zum Beispiel modelliert, dass Adjektive von Nomen, Adverbien von Verben abhängen.

Die Beschränkungen entsprechen also grammatischen Regeln und die Menge der Beschränkungen ergibt die Grammatik der Sprache. Die benutzte Grammatik ist in (Foth 2006) beschrieben.

Nur wenige Regeln der Sprache sind absolut. Häufig gibt es mehrere konkurrierende Regeln. Außerdem sollen auch grammatisch unkorrekte Sätze geparst werden. Dies wird durch *gewichtete Beschränkungen* ermöglicht, denen jeweils *Kosten* zugeordnet sind.

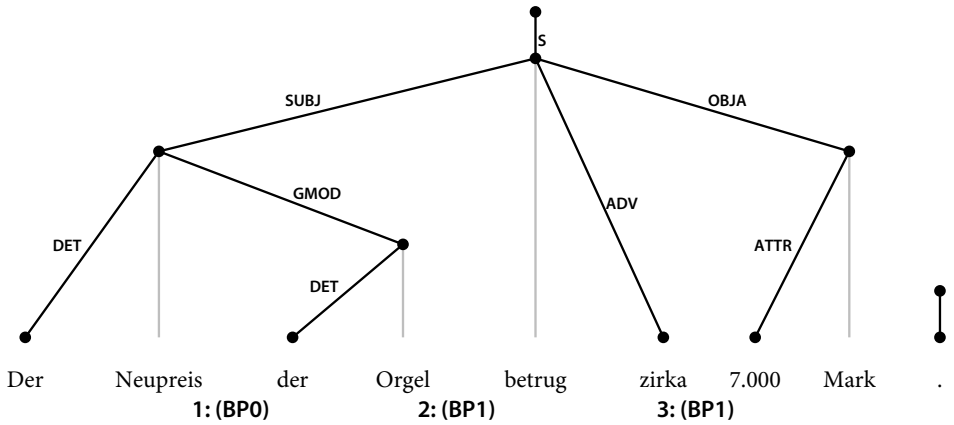
Es wird dann nicht mehr gefordert, dass der konstruierte Baum alle Beschränkungen erfüllt. Stattdesen wird der Baum mit den geringsten Gesamtkosten durch verletzte Beschränkungen gesucht. Dadurch ist es auch möglich, ungrammatischen und damit in sich widersprüchlichen Sätzen eine „beste" grammatische Beschreibung zuzuordnen.

Das verwendete Programm, *wcdg*[12] (Foth et al 2003) implementiert unterschiedliche Methoden zur Suche nach dem besten Baum.

Hier wird die *frobbing*-Methode verwendet (Foth 1999). Sie beginnt mit einer beliebigen Parsing-Hypothese und transformiert diese heuristisch, bis keine Verbesserung mehr erreicht wird.

Die frobbing-Suche ist nicht vollständig, aber deutlich schneller als eine vollständige Suche. Sie erzeugt auch nicht immer einen Baum, sondern manchmal zyklische Graphen. In diesen Einzelfällen wurden die Bäume von Hand korrigiert, um die weitere Verarbeitung zu ermöglichen.

Abbildung 5.2. Beispiel einer Dependenzanalyse



Aus dem Dependenzbaum werden nun für die Phrasierungserkennung hoffentlich hilfreiche Merkmale gewonnen. Ich habe mich für die drei Merkmale Abstand sowie linke und rechte Pfadinschrift entschieden.

Der *CDG-Abstand* ist die Anzahl der Kanten, die zwischen den beiden links und rechts von der möglichen Phrasengrenze liegenden Wörtern liegt. Dafür wird zunächst der Pfad vom linken Wort bis zur Wurzel verfolgt. Dann wird der Pfad vom rechten Wort in Richtung Wurzel verfolgt, bis er sich mit dem anderen Pfad vereint.

Der Abstand ist dann die Summe der auf beiden Pfaden bereits zurückgelegten Kanten. Es ist auch möglich, dass eines der beiden Wörter direkt auf dem Pfad des anderen Wortes liegt. Für die Berechnung des Abstandes ändert sich dadurch aber nichts.

Für die *Pfadinschriften* wird die jeweils oberste Kanteninschrift vor dem Zusammentreffen der Pfade des linken beziehungsweise rechten Wortes benutzt. Die Idee ist, dadurch die grammatische Funktion der Konstituente, zu der das jeweilige Wort gehört, möglichst gut zu beschreiben.

Hängt das eine Wort vom anderen ab, so gibt es keine Kanteninschrift, die vor dem Zusammentreffen der Pfade liegt. Dann wird für dieses Wort die *leere Kanteninschrift* (""") benutzt. Die Information der Abhängigkeit bleibt also erhalten.

In Abbildung 5.2 ist ein Beispiel der Dependenzanalyse angegeben. Für drei Stellen sind in Tabelle 5.1 die berechneten CDG-Merkmale zur Verdeutlichung angegeben. Natürlich werden die Merkmale auch für alle anderen Wortgrenzen berechnet.

[12]http://nats-www.informatik.uni-hamburg.de/

Tabelle 5.1. Beispielhafte Merkmale aus dem Dependenzbaum in Abbildung 5.2

	linkes Wort	rechtes Wort	Abstand	linke Pfadinschrift	rechte Pfadinschrift
1	Neupreis	der	2	”“	GMOD
2	Orgel	betrug	2	SUBJ	”“
3	zirka	7.000	3	ADV	OBJA

5.5.7. Akzentuierung

Die Akzentuierung sollte keinen Einfluss auf die Phrasierungserkennung haben. Die Grundannahme der Superpositionstheorie, dass die Phrasierung der Akzentuierung vorausgeht, wird auch in der Tonsequenztheorie nicht angezweifelt.

Die Aufnahme von Akzentuierungsmerkmalen für die Phrasierungserkennung dient hier lediglich dazu, den wechselseitigen Einfluss von Akzentuierung und Phrasierung genauer zu untersuchen und den Zusammenhang zwischen Akzentuierung und Phrasierung genauer zu verstehen. Die verwendeten Akzentuierungsmerkmale sind der *Abstand zur linken Akzentuierung* und *Abstand zur rechten Akzentuierung*.

Kapitel 6. Sprecherabhängige Experimente

In diesem Kapitel werden Klassifizierer für die einzelnen Korpora (beziehungsweise getrennt für die beiden Hauptsprecher im Kiel-Korpus) erstellt. Die Ergebnisse der vorangestellten Merkmalsauswahl werden diskutiert und verglichen. Es folgt der Versuch einer Bewertung der Merkmale.

Für die endgültige Evaluierung wurde jeweils 1/10 der Korpora als Testmaterial abgetrennt. Auf den verbliebenen Daten wurden in einer 10-fachen Kreuzvalidierung Merkmalsauswahlen vorgenommen. Die Experimente werden je einmal für einen Naïve-Bayes-Klassifizierer und einmal für den C4.5-Klassifizierer durchgeführt.

Die Ergebnisse der Merkmalsauswahlen müssen wegen der Kreuzvalidierung interpretiert werden, um die tatsächlich „beste" Merkmalsmenge für ein Korpus und einen Klassifizierer zu finden. Gleichzeitig liefert die Kreuzvalidierung eine bessere Bewertungsgrundlage für die Qualität einzelner Merkmale und ihrer Kombinationen, sowie eine Bewertung welcher Klassifizierungsalgorithmus geeigneter ist.

Die ausführlichen Ergebnisse der Merkmalsauswahl stehen in Anhang A.

Bei den Ergebnissen der Merkmalsauswahlen insgesamt überrascht die starke Streuung zwischen den Schichten der Kreuzvalidierungen. Erwartet worden war, dass alle Schichten – von einigen Ausreißern abgesehen – jeweils dieselben Merkmale auswählen. Dies war nicht der Fall. Nur selten wurde ein Merkmal von allen zehn Schichten ausgewählt.

Die Anzahl der ausgewählten Merkmale zwischen den Schichten schwankte zwischen und auch innerhalb der Korpora stark, wie aus Tabelle 6.1 hervorgeht. Häufig war die Leistung in den verschiedenen Schichten ähnlich gut, obwohl die zugrundeliegenden Merkmalsmengen sich sehr unterschieden.

6.1. Akzentuierungen

Tabelle 6.1. Anzahl der durchschnittlich zur Akzentuierungserkennung ausgewählten Merkmale

		IBM	IMS	VPSC	kko	rtd
J48	Mittelwert	11.2	12.6	10.5	7.5	10.3
	Standardabweichung	4.92	3.44	1.65	4.17	2.21
NB	Mittelwert	16.3	11.2	12.7	7.0	11.1
	Standardabweichung	3.30	1.93	3.89	2.71	2.38

Für die Akzentuierung schwankte die Zusammensetzung der gewählten Merkmalsmenge stark. Dies deutet darauf hin, dass viele ähnliche Merkmale gleich gut waren und jeweils eines eher zufällig als jeweils bestes ausgewählt wurde. Da es keine eindeutigen Ergebnisse gab, wurde eine Bewertung der unterschiedlichen Merkmale versucht und wurden auf Grundlage der Kreuzvalidierung manuell Merkmalsmengen zusammengestellt.

6.1.1. Gemeinsamkeiten

Allen Korpora gemein war die überwiegende Auswahl des Merkmals *Silbenkern*. Auch die *Silbenposition* wurde als Indikator für die Wortakzentposition überwiegend gewählt.

Die Merkmale *Wortart* und *vereinfachte Wortart* wurden auch sehr häufig, teilweise auch beide gewählt. Auch die Auswahl für den Naïve-Bayes-Algorithmus wählte teilweise beide Wortart-Merkmale aus. Dies überrascht, da die Merkmale stark korreliert sind und eigentlich das vereinfachte Merkmal für den Naïve-Bayes-Algorithmus keinen zusätzlichen Nutzen haben dürfte.

6.1.2. Unterschiede zwischen den Klassifizierungsalgorithmen

Der Naïve-Bayes- und der C4.5-Algorithmus unterscheiden sich in der Art ihrer Urteilsbildung: C4.5 *trifft Unterscheidungen*. Naïve-Bayes *sammelt Evidenz*. Dies zeigt sich auch in der jeweiligen Auswahl von Merkmalen.

Formantmerkmale werden im C4.5-Algorithmus nur selten in mehreren Schichten benutzt. Trotz der generellen Korrelation von Akzentuierung und Formantmerkmalen ist der Unterschied der Formanten zwischen akzentuierten und nicht akzentuierten Silben nur selten groß genug, um auf dieser Basis generelle Entscheidungen treffen zu können.

Der Naïve-Bayes-Algorithmus nutzt Formantmerkmale stärker. Ihm nutzen die feinen Unterschiede der Formanten zwischen akzentuierten und nicht akzentuierten Silbenkernen mehr, da er besser in der Lage ist, die einzelnen Unterschiede aufzusummieren und weniger am Problem der Datenunterteilung leidet.

Ebenso nutzt C4.5 Merkmale zur Phrasenposition fast nie, während Naïve-Bayes fast durchgehend den Abstand zur nächsten Phrasengrenze verwendet. Insgesamt überrascht der geringe Einfluss der Phrasenposition und zeigt, dass die Superpositionstheorie zur Prosodiebeschreibung nicht benötigt wird.

Die genutzte Konfiguration des Naïve-Bayes-Algorithmus nimmt an, dass kontinuierliche Merkmale normalverteilt sind. Sehr deutlich zeigt sich, dass die *PaIntE-Ausrichtung* nicht normalverteilt ist.

Während die PaIntE-Ausrichtung in Entscheidungsbäumen eines der primären Merkmalen ist, weil es die Unterscheidung zwischen früher und später Akzentuierung erlaubt, kann der Naïve-Bayes-Algorithmus diese Information nicht auswerten.

Insbesondere nutzen die Entscheidungsbäume auch wie vorausgesehen häufig die PaIntE-Ausrichtung der vorherigen oder folgenden Silbe, um beispielsweise den späten Akzentton der aktuellen Silbe vom frühen Akzentton der folgenden zu unterscheiden.

Für Naïve-Bayes werden anstatt der PaIntE-Ausrichtung häufiger als für C4.5 die *PaIntE-Flankenhöhen* benutzt. Möglicherweise können diese also einen Teil der wegfallenen Information der PaIntE-Ausrichtung ersetzen.

6.1.3. Unterschiede zwischen den Korpora

Die Segment-Annotierung der Vokale im Kiel-Korpus ist sehr ungenau. Eine Reduktion gespannter Vokale in ungespannte wird nicht notiert (vgl. Kapitel 4). Dies erklärt, warum das ansonsten durchgehend genutzte Merkmal *Silbenkern* im Kiel-Korpus weniger häufig genutzt wird, da das notierte Phonem nicht immer Aufschluss über die zu erwartenden Lauteigenschaften erlaubt.

Reduzierte Vokale sind im Wien-Korpus mit der Dauer von 1 ms annotiert. Es überrascht wenig, dass das Merkmal *Silbenkerndauer* immer ausgewählt wird, da reduzierte Vokale immer in nicht-akzentuierten Silben stehen.

Die Worthäufigkeit wurde in einigen Korpora gewählt, in anderen nicht. Es bleibt unklar, ob dies durch Unterschiede zwischen den Sprechern oder zwischen den zugrundeliegenden Texten begründet ist.

6.1.4. Unterschiede zwischen den Sprechern

Die ausgewählten Merkmale unterscheiden sich deutlich zwischen den Sprechern. Teilweise ist dies sicher den Korpusunterschieden geschuldet. Daneben gibt es aber weitergehende Unterschiede.

Der Sprecher *kko* im Kiel-Korpus fällt durch sehr regelmäßige Sprache auf. Tatsächlich ermittelt die Kreuzvalidierung für den Naïve-Bayes-Klassifizierer in einer Schicht nur die drei Merkmale *Silbenposition*, *Silbenanzahl* und *vereinfachte Wortart* und erreicht ein F-Maß von 83 %.

Alle drei gewählten Merkmale sind textbasiert. Die beste Schicht für den C4.5-Klassifizierer ermittelt dagegen 19 Merkmale (darunter viele PaIntE-Merkmale) und erreicht ein noch höheres F-Maß von 86 %.

Dies ist natürlich ein Zeichen von Überanpassung, da die anderen Schichten nur je 5 – 8 Merkmale nutzen. Andererseits ist es auch ein Zeichen für die Regelmäßigkeit der Sprache: Durch mehr Merkmale konnte die Akzentuierung des Sprechers immer noch besser beschrieben werden, ohne dass die neuen Merkmale an anderen Stellen die Erkennung verschlechtert hätten.

Die Sprecherin *rtd* im Kiel-Korpus benötigt ähnlich viele Merkmale wie die anderen Sprecher. Es handelt sich bei der regelmäßigen Aussprache von *kko* also nicht um eine Eigenheit des Kiel-Korpus.

Die Sprecherin *rtd* und der Sprecher des Wien-Korpus zeichnen sich durch das Merkmal *Schalldruckleistung* aus, das überwiegend ausgewählt wurde. Im IBM- und IMS-Korpus spielt sie praktisch keine Rolle, dafür aber die *Grundfrequenz* im Silbenkern.

Dies deutet darauf hin, dass Akzentuierungen nicht von allen Sprechern gleichartig realisiert werden. Anscheinend überwiegt bei einigen Sprechern die Akzentuierung durch Intensität, bei anderen die durch Tonhöhenvariation.

6.1.5. Klassifizierungsergebnisse

Auf Grundlage der Analyse der Kreuzvalidierungen wurden jeweils auch manuelle Merkmalsauswahlen getroffen. Diese sollten vermeintliche Fehler der automatischen Auswahl korrigieren und das Gesamtresultat über alle Schichten der Kreuzvalidierungen widerspiegeln.

Die manuell ausgewählten Merkmale sind in den Tabellen in Anhang Anhang A jeweils hervorgehoben.

Auf Basis der manuellen Auswahlen und der Auswahl der jeweils besten Schicht der Kreuzvalidierung wurden die Klassifizierer neu trainiert und ihre Leistung an der Testmenge überprüft. Die Ergebnisse sind in Tabelle 6.2 aufgelistet.

Tabelle 6.2. Ergebnisse der Akzentuierungserkennung

		IBM		IMS		VPSC		kko		rtd	
		auto	**man**	**auto**	**man**	**auto**	**man**	**auto**	**man**	**auto**	**man**
J48	precision	74	75	71	69	78	79	86	78	78	74
	recall	67	67	63	63	73	75	80	80	73	71
	F-Maß	70	71	67	66	76	77	83	79	75	72
NB	precision	56	57	48	48	61	57	79	75	79	75
	recall	85	86	84	86	91	95	70	76	78	79
	F-Maß	68	68	61	61	74	71	74	75	78	77

Die manuelle Merkmalsauswahl zeigt keine Vorteile gegenüber der Merkmalsauswahl der jeweils besten Schicht der Kreuzvalidierung. Teilweise konnte die Leistung etwas gesteigert werden, jedoch fiel sie teilweise auch deutlich ab.

Die Leistung der beiden *Klassifizierungsalgorithmen* bei der Akzentuierungserkennung unterscheidet sich kaum. Im Schnitt ist C4.5 etwas besser. Die Merkmalsgrundlage für die beiden Klassifizierungsalgorithmen unterscheidet sich dabei durchaus.

Hinsichtlich der Korpora zeigt sich, dass die Erkennungsrate im IMS- und im IBM-Korpus etwas schlechter ist als in den anderen Korpora. Dies ist möglicherweise der größeren Bandbreite des enthaltenen Textmaterials geschuldet.

Wie erwartet ist die Akzentuierungserkennung für den Sprecher *kko* am besten. Der oben ermittelte Naïve-Bayes-Klassifizierer mit nur drei Merkmalen fällt allerdings auf der unabhängigen Testmenge deutlich zurück. Der C4.5-Klassifizierer mit 19 Merkmalen ist auf der Testmenge hingegen beinahe genauso gut wie bei der Merkmalsauswahl. Dies spricht erneut für die überraschend gleichmäßige Aussprache von *kko*.

6.2. Phrasierungen

Die oben gemachten Bemerkungen zur Uneindeutigkeit der Ergebnisse der Merkmalsauswahlen für die Akzentuierungen gilt für die Phrasierungen genauso. Die Unterschiede zwischen den Klassifizierungsalgorithmen gelten ebenso für die Phrasierungserkennung.

Allerdings unterscheiden sich die Korpora bei den Phrasierungen stärker. Dadurch sind die Ergebnisse zwischen den Korpora noch schwerer vergleichbar als dies bei den Akzentuierungen der Fall war.

Die Unterschiede zwischen den Korpora fielen bei der Akzentuierungserkennung weniger stark auf. Die Ergebnisse zwischen den Korpora unterschieden sich zwar sowohl in den verwendeten Merkmalen und in der Leistung, die Akzentuierungsannotierung über die Korpusgrenzen hinweg scheint aber ähnlich zu sein. Die Phrasierungsannotierung hingegen unterscheidet sich stark zwischen den Korpora.

6.2.1. VPSC

Die auffälligste Eigenheit hat das Wien-Korpus: Jede Vollgrenze koinzidiert mit einer Pause und an jeder Pause steht eine starke Phrasengrenze. Die Klassifizierung der Vollgrenzen ist also trivial.

Vermutlich führt die syntaktische Regelmäßigkeit des Frage-Antwort-Teilkorpus im Wien-Korpus dazu, dass für die Erkennung der Zwischengrenzen fast ausschließlich textbasierte Merkmale verwendet werden. *Dehnung* und *Leistungsabfall* sind neben der *Pause* die einzigen verwendeten phonetischen Merkmale.

6.2.2. KCoRS

Das Kiel-Korpus ist nur mit einer Sorte Phrasengrenzen annotiert. Die Klassifizierung unterscheidet sich also grundsätzlich von den anderen Korpora, da schwache und starke Phrasierungen in einer Klasse zusammengefasst sind und nur zwei Klassen unterschieden werden.

Im Wien-Korpus sind zwar ebenfalls nur zwei Klassen nicht-trivial zu unterscheiden. Die triviale Klasse enthält jedoch alle starken Phrasierungen und die verbleibenden Phrasengrenzen sind alle schwach. Die Klasse ist deswegen in sich homogener als die Klasse aller Phrasierungen im Kiel-Korpus. Dies ist wahrscheinlich der Grund dafür, dass die Ergebnisse im Kiel-Korpus gegenüber dem Wien-Korpus deutlich abfallen.

Außerdem sind die Äußerungen im Kiel-Korpus vergleichsweise kurz, sodass insgesamt weniger Phrasierungen und somit weniger Beispiele zum Training der Klassifizierer vorkommen.

Die ausgewählten Merkmale unterscheiden sich zwischen den einzelnen Sprechern. Die für die Phrasierungserkennung jeweils besten Merkmale sind also von Sprecher zu Sprecher verschieden.

An textbasierten Merkmalen wird vor allem der *Abstand zum folgenden Satzzeichen* und die *vorangehende und folgende Wortart* genutzt. Letztere in geringerem Maße. Bei der Sprecherin *rtd* wurde zudem die *CDG-Distanz* genutzt.

Von den akustischen Merkmalen wird vor allem die *finale Dehnung* und der *finale Leistungsabfall* genutzt. Die *Pause* spielt eine untergeordnete Rolle. Unter den Merkmalen zum Tonhöhenverlauf werden verschiedene Merkmale, vor allem die *Painte-Ausrichtung* ausgewählt.

6.2.3. IBM- und IMS-Korpus

Die beiden übrigen Korpora sind die einzigen, in denen tatsächlich drei nicht-triviale Klassen zu unterscheiden sind: *starke*, *schwache* und *keine* Phrasengrenze.

6.2.3.1. Annotierung in den Korpora

Die Annotierung von starken und schwachen Phrasierungen scheint sich aber trotzdem zwischen den Korpora zu unterscheiden, wie schon aus Tabelle 4.2 hervorging. Die hier relevanten Teile sind noch einmal in Tabelle 6.3 zusammengefasst.

Tabelle 6.3. Vergleich der Phrasierungsannotierung im IBM-, IMS- und Wien-Korpus

	IBM	IMS	VPSC
Anzahl starker Phrasengrenzen*	2812	3091	1006
Anzahl schwacher Phrasengrenzen	3095	1099	588
Schwache Grenzen pro starke Grenze	1.1	0.36	0.58
Intermediäre Phrasen pro Intonationsphrase	2.1	1.36	1.58

*Die äußerungsfinalen Phrasengrenzen sind hier mitgezählt.

Aus den Anzahlen der starken und schwachen Phrasengrenzen ergibt sich der mittlere Anteil schwacher Phrasengrenzen pro starker Phrasengrenze.

Bei der ToBI-Annotierung bestimmen die starken Phrasengrenzen das Ende von *Intonationsphrasen* (IP) und die schwachen Phrasengrenzen das Ende der innerhalb von Intonationsphrasen liegenden *intermediären Phrasen* (vgl. Kapitel 2).

Die Anzahl der Unterteilungen von Intonationsphrasen in intermediäre Phrasen unterscheidet sich deutlich zwischen IBM- und IMS-Korpus. Die zum Vergleich angegebenen Zahlen des Wien-Korpus liegen zwischen denen der beiden genannten Korpora.

Das Textmaterial beider Korpora wurde automatisch aus großen Textkorpora extrahiert. Es kann wohl ausgeschlossen werden, dass sich die zugrundeliegenden Texte so stark unterscheiden, dass die unterschiedliche Phrasierungsannotierung zustandekommt.

Ein klares Bild, welches die „richtige“ Annotierung ist, ergibt sich nicht. Es kann nur festgehalten werden, dass sich bereits die Annotierung von Phrasierungen zwischen den beiden Korpora unterscheidet. Daher fallen Aussagen über die einzelnen Sprecher, wie sie bei der Akzentuierung möglich waren noch schwerer.

6.2.3.2. Textbasierte Merkmale

Als textbasierte Merkmale sind vor allem der *Abstand zum nächsten Satzzeichen* sowie die *Wortarten* vor und nach der möglichen Grenze wichtig. Die Wortart nach der Grenze wird dabei häufig in ihrer vereinfachten Form benutzt.

Der *Typ des nächsten Satzzeichens* ist nur im IBM-Korpus wichtig. Dies entspricht der Einschätzung aus der Prosodieannotierung, dass in diesem Korpus die Phrasierungen hauptsächlich den Satzzeichen entsprechend annotiert wurden.

Mir scheint es von Vorteil, dem Annotator nicht den ursprünglichen Text sondern eine um Satzzeichen bereinigte Form zu präsentieren, damit er sich mehr auf die tatsächliche akustische Realisierung stützt und nicht Phrasierungen annotiert, die „vorhanden sein müssten“.

Im IBM-Korpus wird zusätzlich die *rechte CDG-Pfadinschrift* benutzt, um über die Wortarten hinausgehende syntaktische Information zu integrieren.

6.2.3.3. Phonetische Merkmale

Pause und *Pausendauer* werden in beiden Korpora häufig aber nicht durchgehend ausgewählt. Die *Dehnung* am Wortende vor einer Phrasierung wurde nur im IBM-Korpus ausgewählt.

Intensitätsmerkmale wurden in beiden Korpora benutzt. Sowohl der *Leistungsabfall* am Phrasenende als auch der *Leistungssprung* zur nächsten Silbe wurden häufig ausgewählt.

In beiden Korpora werden Merkmale des Tonhöhenverlaufs häufig ausgewählt. Die *PaIntE-Ausrichtung* scheint besonders wichtig, sowie für das IBM-Korpus die *Höhe der fallenden Flanke* und für das IMS-Korpus die *Steigung der fallenden Flanke*.

Der fallenden Flanke nach dem Phrasen- beziehungsweise Grenzton kommt also eine zentrale Bedeutung zu. Die Ausrichtung des Gipfels erlaubt die Abgrenzung von Akzenttönen: Bei Grenztönen liegt der Gipfel nah an der Wortgrenze.

Die berechneten *Regressionsparameter* ergänzten die PaIntE-Merkmale zum Tonhöhenverlauf. Für das IBM-Korpus war vor allem die *Steigung in den folgenden 2000 ms* wichtig. Wenn die Regressionsgerade in diesem Bereich nicht oder nur sehr schwach abfiel, dann wurde eine Phrasierung erkannt.

Im IMS-Korpus wurde die *mittlere Abweichung in den folgenden 150 ms* der Grundfrequenz von der Regressionsgeraden häufig ausgewählt.

Die Auswahl dieser beiden Merkmale könnte mit *phraseninitialen Konturen* zusammenhängen. Diese sorgen sowohl dafür, dass die Regressionsgerade zu Anfang einer Phrase eher steigen als fallen, als auch dafür, dass durch Tonbewegungen Abweichungen von der Regressionsgeraden entstehen.

Im IMS-Korpus wurde häufig der *Abstand von der letzten Akzentuierung* ausgewählt. Umgekehrt war bei der Merkmalsauswahl für Akzentuierungen bereits der Abstand zur nächsten Phrasierung häufig ausgewählt worden. Diese Wechselbeziehung betrifft in dieser Stärke allerdings nur das IMS-Korpus. Sie ist also entweder der Annotierung geschuldet oder ein sprecherspezifisches Merkmal.

6.2.4. Klassifizierungsergebnisse

Wie schon für die Akzentuierung wurden auf Grundlage der Kreuzvalidierungen manuell Merkmale ausgewählt und die Leistung der entstehenden Klassifizierer mit der der jeweils besten Schichten auf unabhängigem Testmaterial verglichen.

Die Ergebnisse sind in Tabelle 6.4 zusammengefasst. In den einzelnen Tabellenzellen gibt die linke Zahl die Leistung (in Prozent) für Zwischengrenzen an, die rechte Zahl die für Vollgrenzen.

Tabelle 6.4. Ergebnisse der Phrasierungserkennung

		IBM		IMS		VPSC		kko		rtd	
		auto	**man**	**auto**	**man**	**auto**	**man**	**auto**	**man**	**auto**	**man**
J48	precision	64 / 75	69 / 71	53 / 78	52 / 68	93 / 100	95 / 100	71	69	59	72
	recall	52 / 53	44 / 55	44 / 57	20 / 64	71 / 100	64 / 100	38	44	43	43
	F-Maß	57 / 62	54 / 62	48 / 66	29 / 66	81 / 100	77 / 100	49	54	50	54
NB	precision	44 / 38	44 / 43	30 / 41	20 / 63	84 / 100	70 / 100	38	38	34	27
	recall	72 / 73	75 / 78	40 / 82	84 / 51	63 / 100	71 / 100	80	84	77	73
	F-Maß	55 / 50	55 / 55	34 / 55	32 / 56	72 / 100	71 / 100	51	52	47	39

Bei der Phrasierungserkennung sind die Entscheidungsbäume deutlich erfolgreicher als die Naïve-Bayes-Klassifizierer. Möglicherweise liegt dies daran, dass die Klassen in sich nicht besonders homogen sind und je nach Kontext unterschiedlich realisiert werden.

So geht am Ende einer Frage die Grundfrequenz meist nach oben, am Ende einer Aussage allerdings nach unten. Dieses Phänomen ist allerdings regelmäßig mit dem Satzzeichen verbunden. Manche

Kombinationen von vorangehender und folgender Wortart zeigen eine Phrasierung an, dafür wird also eine Kombination der einzelnen Merkmale benötigt. Der Entscheidungsbaum ist in der Lage, je nach Kontext zu unterscheiden. Dem Naïve-Bayes-Klassifizierer gelingt diese Unterscheidung nicht.

Die manuelle Merkmalsauswahl zeigt gegenüber der jeweils besten Schicht der Kreuzvalidierung bei C4.5 selten Vorteile. Die Leistung von NB kann durch die manuelle Auswahl zwar gesteigert werden, bleibt aber ungenügend.

Die Erkennungsleistung im Wien-Korpus ist erwartungsgemäß am höchsten. Zunächst müssen nur zwei Klassen (*keine* und *schwache* Grenze) wirklich unterschieden werden. Die bessere Leistung gegenüber dem Kiel-Korpus erklärt sich dann daraus, dass die Klassen in sich homogener sind, da die starken Grenzen in eine weitere (aber trivial zu bestimmende) Klasse ausgelagert sind.

Kapitel 7. Sprecherübergreifende Experimente

In diesem Kapitel werden die Experimente zur sprecherübergreifenden Prosodieerkennung beschrieben. Die benutzten Korpora unterschieden sich in ihrer Zusammensetzung, in ihrer Segmentierung und in ihrer Prosodieannotierung teils deutlich voneinander. Es war deswegen nicht möglich, alle zur Verfügung stehenden Daten zum Training eines verallgemeinernden Klassifizierers zu benutzen.

Aus diesem Grund konnten nur Experimente mit jeweils einem Teil der Daten durchgeführt werden. Das Kiel-Korpus enthält Daten vieler unterschiedlicher Sprecher. Alle Daten innerhalb des Kiel-Korpus sind gleichartig annotiert. Dadurch wird es möglich, die Leistung der sprecherunabhängigen Erkennung mit einer Vielzahl unterschiedlicher Sprecher zu messen.

Die drei übrigen Korpora (IBM, IMS und VPSC) ähneln sich in Aufbau und Annotierung. Deswegen wurden auch zwischen diesen Korpora Experimente zur sprecherunabhängigen Klassifizierung durchgeführt.

Da sich bisher keine deutlichen Unterschiede zwischen automatisch und manuell ausgewählten Merkmalen zeigte, wird hier auf die manuelle Auswahl und den Vergleich zu den automatisch ermittelten Merkmalen verzichtet.

7.1. Sprecherunabhängige Erkennung im Kiel-Korpus

Das Kiel-Korpus enthält insgesamt 3876 Äußerungen von 53 Sprechern (vgl. Kapitel 4). Allerdings werden dieselben Sätze mehrfach gesprochen: Das Textmaterial beschränkt sich auf 603 Sätze.

Die gleichen Sätze werden meist auch durch unterschiedliche Sprecher gleich akzentuiert und phrasiert. Die Klassifizierer sollten nicht übermäßig auf genau die im Korpus vorkommenden Sätze trainiert werden. Deswegen wurde jeder Satz nur genau einmal verwendet.

Die *Buttergeschichte* und *Nordwind und Sonne* wurden nicht verwendet, da sich Probleme bei ihrer Verarbeitung durch PaIntE ergaben. Das zur Verfügung stehende Material beschränkt sich also auf 598 Äußerungen von 25 Sprechern.

Die Äußerungen der einzelnen Teilkorpora wurden jeweils zufällig auf die Sprecher verteilt, die dieses Teilkorpus sprachen. Für die meisten Sprecher ergeben sich zehn, für die beiden Hauptsprecher je 99 und für drei andere Sprecher je 66 Äußerungen. Diese Verteilung ergibt sich auch aus Tabelle 4.1.

Mit dem Material wurde im Auslassverfahren jeweils auf den Daten von 24 Sprechern die automatische Merkmalsauswahl durchgeführt und anschließend die Leistung auf dem Material des verbliebenen Sprechers evaluiert.

Die Ergebnisse der Merkmalsauswahl und die Klassifizierungsergebnisse sind in Anhang B aufgelistet.

7.1.1. Akzentuierungen

Der Naïve-Bayes-Algorithmus liefert bei der sprecherunabhänngigen Akzentuierungserkennung gegenüber C4.5 die besseren Ergebnisse. Im Schnitt erreichen beide Algorithmen ein hohes F-Maß von 70 % (NB) beziehungsweise 69 % (C4.5). Der beste NB-Klassifizierer erreicht für den Sprecher *k09* ein F-Maß von fast 82 %.

Die Ergebnisse für die auch einzeln untersuchten Sprecher *kko* und *rtd* sind erwartungsgemäß schlechter als beim Training auf sprechereigenem Material. Für *kko* fällt die Leistung (F-Maß) von 83 % auf 78 % zurück, für *rtd* von 78 % auf 73 %. Der Rückgang beträgt also jeweils nur etwa 5 %.

Die sprecherunabhängige Akzentuierungsvorhersage kann unter diesen Umständen als gelungen betrachtet werden. Die schlechteren Ergebnisse für *rtd* als für *kko* liegen wahrscheinlich an einer weniger regelmäßigen Akzentuierung durch die Sprecherin.

Bei den Merkmalen ergibt sich ein ähnlich durchmischtes Bild wie bei der sprecherabhängigen Merkmalsauswahl. Häufig werden *Wortart*, *Silbenkern*, verschiedene *Dauermerkmale* des Silbenkerns und der ganzen Silbe, sowie die *Silbenposition* ausgewählt.

PaIntE-Merkmale (Ausrichtung, Flankensteigung) werden ebenfalls häufig ausgewählt. Überraschend ist die Auswahl der Formanten, die anscheinend über Sprechergrenzen hinweg hinreichend stabil sind, um bei der Akzentuierungsvorhersage zu helfen.

7.1.2. Phrasierungen

Die Ergebnisse der sprecherübergreifenden Phrasierungserkennung überzeugen nicht. Dies liegt wahrscheinlich an dem zu kleinen Anteil von Phrasierungen im Kiel-Korpus. Dadurch reichen die Trainingsdaten für erfolgreiches Lernen nicht aus.

Der starke Datenmangel beim Training zeigt sich auch in der Anzahl ausgewählter Merkmale. Der Entscheidungsbaum funktioniert nach dem Prinzip *teile und herrsche* und leidet deswegen noch stärker am Datenmangel. Entsprechend werden für ihn im Schnitt nur etwa 12, für den Naïve-Bayes-Klassifizierer fast 16 Merkmale ausgewählt.

Eine Analyse der Merkmalsauswahl ist trotz der schlechten Ergebnisse der Klassifizierung sinnvoll. Sie spiegelt Muster innerhalb der sprecherübergreifenden Trainingsmengen wider, unabhängig von den Klassifizierungsergebnissen auf unabhängigem Testmaterial.

Unter den textbasierten Merkmalen werden *Abstand zu Satzzeichen*, *Art des Satzzeichens* und die *Wortart* vor und nach der möglichen Grenze gewählt. Der Naïve-Bayes-Klassifizierer nutzt auch die *rechte CDG-Pfadinschrift*.

Beide Klassifizierer nutzen die *Dehnung* fast durchgängig. Ansonsten unterscheiden sie sich bei den phonetischen Merkmalen stärker. C4.5 wählt neben der Dehnung nur noch den *Sprung* der Regressionsgeraden über 2000 ms. Wahrscheinlich reichen die Trainingsdaten für C4.5 nicht aus, um noch weitere Merkmale auszuwählen.

Der Naïve-Bayes-Klassifizierer wählt mehrere zusätzliche akustische Merkmale, darunter verschiedene *PaIntE-Merkmale*, den *Leistungsabfall* in der letzten Silbe und den *quadratischen Fehler* der Regressionsgeraden über die folgenden 2000 ms.

7.2. Sprecherunabhängige Erkennung in den übrigen Korpora

Die drei übrigen Korpora sind sich untereinander ähnlich: Die Annotierung von Akzentuierungen und Phrasierungen basiert auf GToBI, die Segmentierung ist ähnlich und die Textauswahl auch.

Entsprechend dem im vorangegangenen Abschnitt skizzierten Vorgehen wurde deswegen zwischen den drei Korpora im Auslassverfahren Experimente zur sprecher- und korpusübergreifenden Prosodieerkennung durchgeführt. Die Ergebnisse stehen ebenfalls in Anhang B.

7.2.1. Akzentuierungen

Die Ergebnisse der sprecherunabhängigen Erkennung fallen zwar gegenüber der sprecherabhängigen Erkennung um einige Prozentpunkte zurück, liegen aber immer noch im Schnitt bei 65 %.

Die Leistung ist im Vergleich zum Kiel-Korpus schlechter. Dies kann daran liegen, dass nur auf Material von jeweils zwei Sprechern und nicht von 24 trainiert wurde. Hinzu kommen die Unterschiede in Segmentierung und Prosodieannotierung zwischen den Korpora.

Die gewählten Merkmale entsprechen den ansonsten häufig benutzten: Silbenkern, Silbendauer, relative Silbenkerndauer, Silbenposition und Wortart. Für C4.5 wird die PaIntE-Ausrichtung und für Naïve-Bayes die PaIntE-Flankensteigung gewählt. Diese beiden Merkmale scheinen vergleichsweise sprecherinvariant zu sein. Ebenfalls werden Formanten und die mittlere Grundfrequenz des Silbenkerns ausgewählt.

Das letztgenannte Merkmal macht deutlich, dass alle drei Korpora von Männern gesprochen wurden. Diese Ergebnisse zeigen also, dass die Akzentuierungserkennung sprecherunabhängig und zu einem gewissen Maße auch korpusunabhängig ist. Sie zeigen aber nicht, ob eine geschlechtsunabhängige Erkennung funktioniert.

7.2.2. Phrasierungen

Wie erwartet unterscheidet sich die Phrasierung (beziehungsweise ihre Annotierung) zwischen den Korpora zu stark als dass sinnvolle Ergebnisse bei der sprecherunabhängigen Phrasierungserkennung erzielt werden könnten.

Es zeigt sich allerdings, dass korpusübergreifend die *Dehnung* das robusteste Merkmal ist. Der Naïve-Bayes-Algorithmus zeigt etwas bessere Ergebnisse als C4.5. Wahrscheinlich läuft er weniger stark Gefahr, korpusspezifische Eigenheiten zu lernen, die im Testmaterial nicht auftauchen.

Kapitel 8. Anwendung in der Sprachsynthese

Die bis hier gemachten Erfahrungen zur automatischen Prosodieerkennung werden in diesem Kapitel in einem Text-to-Speech-System (TTS-System) praktisch angewendet. Ein bisher nicht annotiertes Sprachsynthesekorpus der IBM wird automatisch prosodieannotiert werden. Dieses Korpus ist im Aufbau identisch mit dem in der Arbeit bisher behandelten IBM-Korpus.

Auf Grundlage der Prosodieannotierung werden TTS-Module zur symbolischen und akustischen Prosodiegenerierung neu trainiert. Die Qualität der bisherigen Sprachsynthese mit dieser Stimme wird mit der neu trainierten in einem Perzeptionstest verglichen.

Einführend wird das benutzte TTS-System vorgestellt.

8.1. Aufbau des TTS-Systems

Der hier vorgestellte Prototyp, der auf dem IBM-TTS-System (Donovan et al 2001) aufbaut, gliedert sich in zwei getrennte Programmteile. Das *Front-End* führt die symbolische Verarbeitung des Eingabetextes durch. Es normalisiert den Text, setzt die Graphemfolge in eine Lautfolge um, zieht Silbengrenzen zwischen den Lauten, weist den Silben Akzentarten zu, setzt Phrasierungen und bestimmt die Akzentuierung von Wörtern.

Die Textnormalisierung verläuft durch sprachspezifische Regeln, welche Abkürzungen, Daten, Uhrzeiten, Zahlen und so weiter erkennen und für die weitere Verarbeitung geeignet umsetzen. Graphem-Phonem-Umsetzung, Silbifizierung, Wortakzentzuweisung, Phrasierung und Akzentuierung der Wörter erfolgen durch Kaskaden von datengetriebenen Modulen (z. B. Entscheidungsbäumen). Nur die beiden letztgenannten Module werden neu trainiert.

Die Ausgabe des Front-Ends besteht aus Angaben über die zu realisierenden Phoneme, die Zusammenfassung der Phoneme zu Silben, die Akzentuierung der Silben sowie die Phrasierung der Äußerung.

Die akustische Verarbeitung erfolgt im *Back-End*. Es entscheidet über akustische Eigenschaften der zu realisierenden Einheiten und führt die Einheitenauswahl und Synthese auf Grundlage der Sprechdaten im Korpus aus. Zunächst entscheidet das Back-End über die geeignete Umsetzung von Phrasierungen und Akzentuierungen durch eine Kombination von Lautdauer und Grundfrequenzverlauf. Diese Ziel-Eigenschaften werden durch Regressionsbäume[1] bestimmt. Zusätzlich werden für Phrasierungen Pausen eingefügt.

Dann folgt die eigentliche konkatenative Synthese: Das Back-End wählt geeignete Einheiten von kurzen Abschnitten gesprochener Sprache aus dem Synthesekorpus aus und rekombiniert sie zu der zu generierenden Äußerung. Die Einheiten können jeweils auf Drittelphonebene miteinander verknüpft werden (Eide et al 2003). Die Dreiteilung der Laute ergibt sich aus dem bei der Segmentierung durch Forced-Alignment verwendeten HMM-Spracherkenner, der pro Laut drei HMM-Zustände vorsieht (Donovan und Woodland 1999).

Die zu verkettenden Einheiten müssen zwei Anforderungen genügen: Sie sollen (1) möglichst gut den vorgegebenen akustischen Eigenschaften entsprechen, und (2) an ihren Übergängen möglichst gut zueinander passen. Entsprechend ergeben sich zwei Arten von Kosten: *Einheitenkosten*, die die Abweichung der einzelnen Einheiten von ihren Ziel-Eigenschaften beschreiben und *Verkettungskosten*, die die Güte des Übergangs zwischen zwei benachbarten Einheiten bewerten. Beide Kosten werden durch komplizierte Kostenfunktionen berechnet.

Die beste Auswahl von zu verkettenden Einheiten wird durch den Viterbi-Algorithmus getroffen, der die Summe der Verkettungs- und Einheitenkosten minimiert. Bei der Verkettung der Sprachsignale

[1] *Regressionsbäume* (Witten und Frank 2005, S. 76f.) ähneln Entscheidungsbäumen, jedoch ermitteln sie aus den zur Verfügung gestellten Merkmalen keine Klassenzugehörigkeit, sondern den Wert einer kontinuierlichen Zielvariablen.

werden die Einheiten auf die vorgesehene Dauer und Grundfrequenz skaliert. Der verwendete Algorithmus ist unveröffentlich und ähnelt dem PSOLA-Verfahren (Moulines und Charpentier 1990, nach Donovan et al 2001).

8.1.1. Bisheriges Training der symbolischen Prosodieerzeugung

Für das verwendete Synthesekorpus stand bisher keine Prosodieannotierung zur Verfügung. Die Entscheidungsbäume für die Vorhersage von Phrasierungen und Akzentuierungen wurden deswegen mit sprecherfremdem Material trainiert. Natürlich können bei der Klassifizierung nur textbasierte Merkmale genutzt werden, weil die akustische Realisierung bei der Synthese erst noch bestimmt werden muss.

Das Phrasierungstraining nutzte die annotierten Phrasengrenzen des Korpus, das auch im Rahmen dieser Arbeit (als IBM-Korpus) eingehend untersucht wurde. Dem Entscheidungsbaum zum Lernen der drei Phrasengrenzklassen (keine Grenze, Zwischengrenze und Vollgrenze) stehen als Merkmale Wortart, Silbenzahl, Distanzen zum linken und rechten Satzzeichen und Art dieses Satzzeichens zur Verfügung. Die Merkmale werden jeweils für das vorvorhergehende, vorhergehende, aktuelle, nächste und übernächste Wort ermittelt und zur Verfügung gestellt. Die Merkmalsauswahl wird allein dem Entscheidungsbaum überlassen.

Das Akzentuierungstraining nutzte als Trainingsmaterial die Ausgabe eines syntaxgesteuerten, regelbasierten Verfahrens. Dieses Verfahren ist rein text-basiert und damit sprecherunabhängig. Es unterscheidet insgesamt sieben Grade der Akzentuierung auf Wortebene. Diese sieben Grade werden als Klassen und nicht als geordnete Werte oder Regression gelernt. Zur Verfügung stehen die Merkmale Wortart, Silbenzahl, Distanzen zur linken und rechten Phrasengrenze und Typ der Phrasengrenze, ebenfalls in einem Fenster vom vorvorhergehenden bis zum übernächsten Wort.

Für die Merkmale die auf Phrasengrenzen basieren, wird das Ergebnis der zuvor erfolgten Phrasengrenzklassifizierung benutzt und nicht die tatsächlichen, annotierten Phrasengrenzen. Auch bei der Synthese steht nur die Klassifizierung durch das Phrasierungsmodul zur Verfügung und nicht eine absolut korrekte Phrasierung. Der Umgang mit Fehlern des Phrasierungsmoduls wird deswegen im Akzentuierungsmodul gleich mitgelernt.

Die Zuordnung der Akzentuierungen auf Silbenebene ergibt sich aus der Kombination der gelernten Akzentuierungen auf Wortebene und der unabhängig vom Kontext zugewiesenen Wortakzente für dieses Wort.

8.1.2. Bisheriges Training akustischer Parameter

Die Regressionsbäume für die akustischen Parameter Grundfrequenzverlauf und Lautdauer, die schließlich die Prosodie repräsentieren, werden auf der Grundlage der symbolischen Ausgabe des Front-Ends aus dem Synthesekorpus ermittelt. Die Grundfrequenz wird für jede Silbe bestimmt. Merkmale für den Klassifizierer sind Akzentuierung, Wortart, Position in der Phrase und Phrasentyp, jeweils in einem Fenster von der vorvorhergehenden bis zur übernächsten Silbe.

Die Lautdauer wird für jeden Laut einzeln bestimmt. Der Klassifizierer kann Lautmerkmale (Phonem, Artikulationsart, Stimmhaftigkeit, Silbenstellung), die Akzentuierung der enthaltenden Silbe, die Position der Silbe im Wort und des Wortes in der Phrase für seine Vorhersagen nutzen.

8.2. Training der Klassifizierer und Prosodieannotierung

Das Synthesekorpus des beschriebenen TTS-Systems wird nun geeignet prosodieannotiert. Wie in den bisherigen Abschnitten werden dafür zwei Typen von möglichen Phrasengrenzen zwischen Wörtern unterschieden. Anschließend wird die automatische Annotierung zum Training der Module des TTS-Systems genutzt.

Akzentuierungen werden vom TTS-System auf Wortebene trainiert, nicht auf Silbenebene wie in dieser Arbeit.

Die akustisch zu akzentuierende Silbe und die Stärke der Akzentuierung ergibt sich im TTS-System aus der Kombination der Akzentuierung des Wortes im Satz (Satzakzent) mit der Akzentuierung der Silben im Wort (Wortakzent). Für den Wortakzent wird ein Entscheidungsbaum mit einem Lexikon trainiert.

Damit die automatisch erzeugte Akzentuierungsannotierung (auf Silbenebene) durch das TTS-System (auf Wortebene) benutzt werden kann, wird die Annotierung vergröbert: Jedes Wort wird als akzentuiert markiert, das mindestens eine akzentuierte Silbe enthält.

8.2.1. Auswahl des Trainingsmaterials

Die vorangegangenen Kapitel haben gezeigt, dass sich die Korpora hinsichtlich ihrer Aufbereitung und Annotierung teils deutlich unterscheiden. Die Nutzung der Daten aller Korpora als Trainingsmaterial hätte deswegen wahrscheinlich zu schlechteren Ergebnissen geführt als eine gezielte Auswahl. Andererseits gibt es auch große sprecherspezifische Unterschiede, weshalb die Nutzung von Daten mehrerer Sprecher Vorteile bringen sollte.

Während sich die IBM- und IMS-Korpora für die Akzentuierungsvorhersage nur wenig voneinander unterschieden, waren die Unterschiede bei Phrasengrenzen auch zwischen diesen beiden Korpora groß. Das VPSC ist mit den beiden genannten Korpora zumindest hinsichtlich der Akzentuierungsannotierung vergleichbar. Allerdings unterscheidet sich das Segment-Alignment im VPSC in reduzierten Silben deutlich davon (vgl. Kapitel 4 und Kapitel 6).

Für das Phrasierungstraining wird deshalb nur das IBM-Korpus, für das Akzentuierungstraining zusätzlich das IMS-Korpus benutzt.

8.2.2. Auswahl der Klassifizierer und ihrer Merkmale

Der Naïve-Bayes-Algorithmus wurde als Basis für die zu erstellenden Klassifizierer gewählt, da er sich bisher als robuster gegenüber unterschiedlichen Verhältnissen in unterschiedlichen Teilen des Trainingsmaterials gezeigt hat.

Das zu annotierende Korpus ist von einer Frau gesprochen, während die Sprecher sowohl des IBM- als auch des IMS-Korpus Männer sind. Die zu wählenden Merkmale für die Klassifizierer müssen also geschlechterunabhängig – oder noch besser: spezifisch für eine Frauenstimme – sein.

Für einen kleinen Teil des zu annotierenden Korpus stand bereits eine Prosodieannotierung zur Verfügung.[2] Dieses annotierte Material sollte gewinnbringend genutzt werden.

Innerhalb des Kiel-Korpus wurde deswegen ein Experiment zur gesteuerten Merkmalsauswahl durchgeführt. Die Ergebnisse waren vielversprechend.

8.2.2.1. Gesteuerte Merkmalsauswahl

Bei der *gesteuerten Merkmalsauswahl* unterscheidet sich die Holdout-Menge systematisch vom Trainingsmaterial. Üblicherweise wird die Holdout-Menge vom vorhandenen Trainingsmaterial abgetrennt und entspricht deshalb in ihren Eigenschaften dem Trainingsmaterial.

Die Merkmalsauswahl optimiert dann die Merkmalsmenge so, dass die Leistung auf der Holdout-Menge maximal wird. Damit ist die Hoffnung verbunden, dass diese Merkmale auch auf unabhängigem, dem Trainingsmaterial entsprechenden Material optimal sind.

Für die Klassifizierung von Material eines anderen Sprechers sind mitunter ganz andere Merkmale wichtig als im Trainingsmaterial. Die Idee besteht nun darin, die Merkmalsauswahl durch die Nutzung von Holdout-Material des Zielsprechers zu steuern.

[2] An dieser Stelle geht mein Dank erneut an Stella Müller!

Die gesteuerte Merkmalsauswahl verbindet den Vorteil einer großen (sprecherfremden) Trainingsmenge mit dem Vorteil der Auswahl sprecherabhängiger Merkmale. Insbesondere sorgt sie hier dafür, dass nur geschlechterunabhängige Merkmale ausgewählt werden, obwohl das Trainingsmaterial nur von Männern stammt.

8.2.2.1.1. Experiment im Kiel-Korpus

Die Idee der gesteuerten Merkmalsauswahl wurde im Kiel-Korpus auf ihre Tauglichkeit überprüft. Ziel war eine möglichst gute Erkennung für die Sprecherin *rtd* unter Nutzung nur der von Männern aufgenommenen Teile des Korpus.

Die Leistung der Akzentuierungserkennung nach einfacher Merkmalsauswahl lag bei 69 %. Die gesteuerter Merkmalsauswahl verbessert die Akzentuierungserkennung auf 74 %. Damit liegt das Ergebnis sogar noch besser als wenn im Training auch Material von Sprecherinnen benutzt wurde.

8.2.2.1.2. Auswahl für die Anwendung

Entsprechend dem Vorgehen im Kiel-Korpus werden die von der IBM-Sprecherin vorhandenen annotierten Daten als Holdout-Menge für eine gesteuerte Merkmalsauswahl benutzt.

Für Akzentuierungen wurden die Merkmale *Silbenkern*, *Silbendauer*, *normierte Silbenkerndauer*, PaIntE *C2-1*, *C1_0*, *C2+1*, *2. Formant*, *normierter 3.* und *4. Formant*, *Silbenposition*, *Wortart* und *Worthäufigkeit*.

Für Phrasengrenzen wurden die Merkmale *vorangehende Wortart*, *folgende Worthäufigkeit*, PaIntE *A2-1*, *A2_0*, *C1+1*, *C2+1*, *Leistungssprung*, *normierter Leistungssprung*, *relative Dehnung*, Regression über 200 ms mit *Steigung*, *Sprung*, *quadratischer Fehler* und *Steigungsunterschied* über 2000 ms.

Der Abstand zu Satzzeichen wird bei der Phrasierung nicht ausgewählt, was zunächst überrascht. Es deckt sich jedoch mit der Aussage der Annotatorin, dass die Sprecherin ihre Phrasengrenze sehr häufig anders als der männliche IBM-Sprecher setze und sich dabei nur sehr wenig an Satzzeichen orientiert[3].

Für das abschließende Training nach der Merkmalsauswahl wurde auch das bereits annotierte Material der Sprecherin benutzt um die Erkennungsleistung zu optimieren. Dadurch blieb allerdings kein unabhängiges Testmaterial mehr übrig, weshalb die endgültige Leistung der Erkennung nicht quantifiziert werden kann.

Die informelle Überprüfung von nicht-handannotiertem Material zeigte aber für Akzentuierungen gute Ergebnisse, für Phrasierungen hingegen sehr schlechte.

8.3. Erneutes Training der TTS-Module für die Prosodiegenerierung

Die symbolische Vorhersage von Phrasengrenzen und Akzentuierungen im Front-End wurde anschließend mit dem automatisch annotierten Korpus trainiert. Für eine Fehlerabschätzung dieses Trainings wurde das Korpus in 80 % Trainings- und 20 % Testdaten unterteilt.

Informelle Tests ergaben schlechte Ergebnisse für das Phrasierungsmodul. Das trainierte Modul zeigte eine starke Übergenerierung von Phrasengrenzen an vollkommen unerwarteten Stellen. Insgesamt ist die schlechte Leistung wenig erstaunlich, da das Modul ausschließlich Textmerkmale benutzt, während die Merkmalsauswahl (siehe oben) gerade gezeigt hatte, dass die Phrasierung der Sprecherin sich nur sehr schlecht mit Textmerkmalen erfassen lässt. Das einzig gemeinsame Merkmal der beiden Klassifizierer (dem zur automatischen Annotierung des Korpus und dem im Phrasierungsmodul) ist die Wortart vor der möglichen Phrasengrenze.

Zusammen mit dem schon schlechten Ergebnis der Phrasierungsannotierung und der noch schlechteren Leistung des darauf trainierten Phrasierungsmoduls entschied ich, anstatt des trainierten wei-

[3] Dies betrifft natürlich nicht die äußerungsfinalen Phrasengrenzen, die immer mit einem Satzzeichen zusammenfallen aber nicht durch den Klassifizierer vorhergesagt werden müssen.

terhin die bisher benutzten, vom anderen Sprecher übernommenen Phrasierungsdaten zu verwenden und nur die Akzentuierung neu zu trainieren. Diese entsprechen zwar möglicherweise nicht der Sprecherin, waren für das TTS-System aber gut lernbar und hatten ihre Praxistauglichkeit im bisherigen System schon bewiesen.

Der Test des trainierten Akzentuierungsmoduls zeigte eine Leistung von etwa 0,9 (F-Maß). Nur die akzentuierten Wörter, nicht aber die akzentuierten Silben wurden vorhergesagt. Außerdem war das Trainingsmaterial für das Akzentuierungsmodul bereits das Ergebnis eines Klassifizierers, sodass es keine Ausreißer enthielt. Insofern war die gute Leistung zu erwarten.

Die akustischen Module für die Lautdauer- und Grundfrequenzverlaufsbestimmung wurden wie bisher auf Grundlage der Ausgabe der – jetzt neu trainierten – symbolischen Prosodiemodule berechnet. Die im letzten Schritt gemachten Fehler setzen sich also im Training der akustischen Module fort.

Weitere Fehler entstehen dadurch, dass zunächst die Akzentuierungsgenauigkeit für das Training des Akzentuierungsmoduls von Silben auf Wörter herabgesetzt und anschließend durch die im Lexikon angegebene Wortakzentuierung wieder präzisiert wurde. Hierbei entstehen unweigerlich Fehler, die sich direkt beim Training der akustischen Module bemerkbar machen.

8.4. Evaluierung

Die Evaluierung ermittelt, ob und wie sich die Sprachsynthese gegenüber der ursprünglichen Version verändert hat. Hierfür wurden neun für mögliche Anwendungen repräsentative Äußerungen ausgewählt und in einem doppelt verblindeten Perzeptionstest die alternativen Synthesen paarweise bewertet.

8.4.1. Auswahl der Testäußerungen

Mit der Auswahl der Testäußerungen sollten die beiden Haupteinsatzgebiete von TTS-Systemen, Dialog-Systeme und Vorlesesysteme, abgedeckt werden. Für Dialog-Systeme sind Fragen, Anweisungen und Hinweise/Auskünfte charakteristisch. Vorlese-Systeme äußern überwiegend lange und kurze Aussagesätze, aber auch Fragen.

Die Zahl der Testäußerungen musste klein bleiben. Die Perzeptionstests sollten in wenigen Minuten durchführbar sein, um die Teilnehmer nicht übermäßig zu strapazieren. Gleichzeitig sollten die Paare von Testäußerungen sich auch tatsächlich in ihrer Prosodie unterscheiden. Deshalb wurden zunächst 70 Äußerungen aus sieben Kategorien (für *Dialoge*: *Frage*, *Anweisung*, *Hinweis/Auskunft*, für *Vorlesen*: *Frage*, *kurzer Satz*, *langer Satz*) aus dem IMS-Korpus ausgewählt. Die Äußerungen wurden jeweils mit dem alten und dem neuen System synthetisiert und dann die Herkunft der beiden Stimuli innerhalb der Synthesepaare randomisiert.

Aus den 70 Synthesepaaren wurden dann für den Perzeptionstest geeignete ausgewählt. Sätze mit gravierenden Synthesefehlern wurden zunächst aussortiert, dann solche, die sich nicht voneinander unterschieden. Vier der Synthesepaare waren bitgenau identisch, viele andere unterschieden sich nicht oder nur sehr gering[4]. Aus den verbliebenen Äußerungen habe ich neun Paare ausgewählt, deren Stimuli sich möglichst stark unterschieden.

Die ausgewählten Äußerungen sind in Anhang C aufgelistet.

8.4.2. Durchführung der Perzeptionstests

Am Perzeptionstest nahmen 21 Personen (13 Männer und 8 Frauen) teil. Einige der Teilnehmer waren keine Muttersprachler des Deutschen. Diese Teilnehmer sprachen aber meiner und ihrer eigenen Einschätzung nach hinreichend gut Deutsch, um die Prosodie bewerten zu können.

Die Stimuli wurden in einer Powerpoint-Datei präsentiert, die das Abspielen der Stimuli durch einfaches Klicken auf Lautsprechersymbole erlaubte. Die Teilnehmer wurden gebeten, vor dem Abspie-

[4] Mein auditiver Eindruck.

len der Stimuli die ebenfalls schriftlich präsentierte Äußerung durchzulesen. Auf diese Weise war der Inhalt der Äußerung dem Teilnehmer sowohl beim Hören des ersten als auch des zweiten Stimulus bekannt.

Die Teilnehmer konnten die Stimuli beliebig häufig und in beliebiger Reihenfolge abspielen und sollten dann auf einem Bewertungsbogen angeben, welcher der beiden *viel besser* oder *etwas besser* war, oder ob sie *beide gleich* waren. Kriterium sollte ihr „Gesamteindruck unter besonderer Beachtung der Natürlichkeit und prosodischen Qualität“ sein. Soweit notwendig, habe ich „prosodische Qualität“ dem Teilnehmer erläutert. Das Ausfüllen des Bewertungsbogens dauerte fünf bis zehn Minuten.

Möglicherweise war es ein Fehler, die Teilnehmer nicht darauf hinzuweisen, dass nicht immer rechts die neuen und links die alten Syntheseergebnisse (oder anders herum) zu hören waren. Mindestens zwei der Teilnehmer (die am Test per E-Mail teilnahmen) vermuteten dies. Dies legt den Schluss nahe, dass sie und vielleicht weitere Teilnehmer sich dadurch in ihren Entscheidungen haben beeinflussen lassen.

8.4.3. Ergebnis des Perzeptionstests

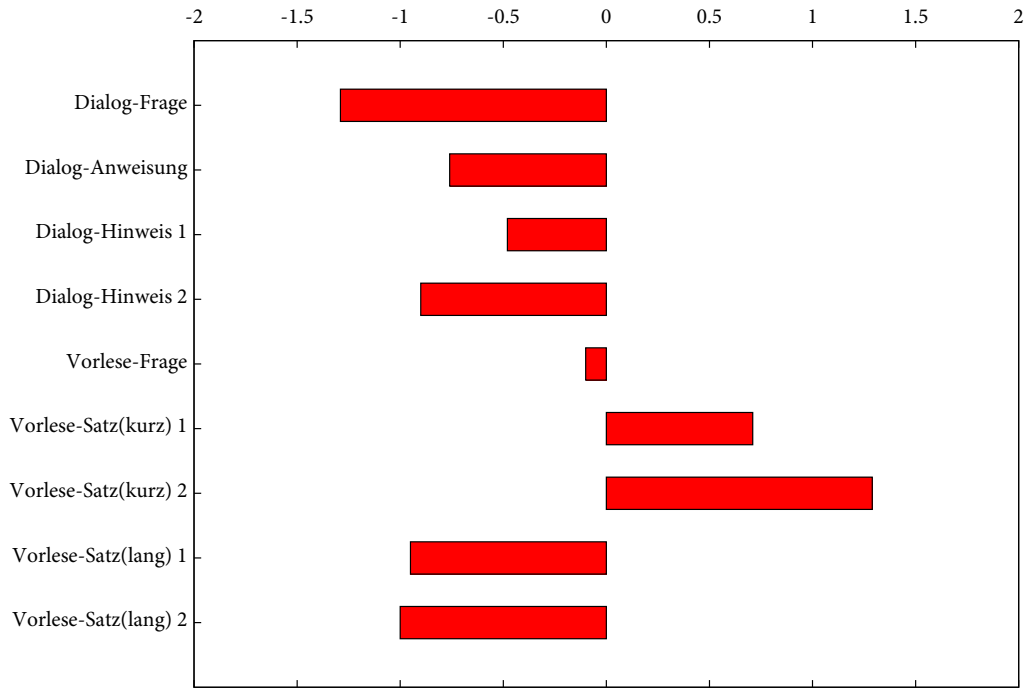
Zur Berechnung der durchschnittlichen Bewertung wurde die Zuordnung von alt und neu wieder hergestellt und die einzelnen Bewertungen wie in Tabelle 8.1 angegeben in Zahlenwerte umgerechnet.

Tabelle 8.1. Umrechnung der Bewertungskategorien

alt *viel besser*	**alt** *etwas besser*	*beide gleich*	**neu** *etwas besser*	**neu** *viel besser*
−2	−1	0	1	2

Abbildung 8.1 gibt den Mittelwert der Bewertung für die einzelnen Äußerungen an. Die mittlere Bewertung liegt bei −0,39 auf der Skala von −2 bis +2. Die ursprüngliche Synthese ist also als deutlich besser eingestuft worden.

Abbildung 8.1. Durchschnittliche Bewertung der Testäußerungen



Dieses sehr deutliche Ergebnis relativiert sich bei der Betrachtung der beiden angenommenen Einsatzgebiete des TTS-Systems. Während die Durchschnittswertung für Dialog-Äußerungen bei −0.86 liegt,

liegt das Ergebnis fürs Vorlesen fast ausgeglichen bei −0,01. Die beiden kurzen Vorlesesätze wurden sogar deutlich besser bewertet als die langen und die Frage.

Das automatisch prosodieannotierte Synthesekorpus, mit dessen Hilfe die Prosodiemodule des TTS-Systems trainiert wurden, bestand ausschließlich aus Sätzen, die Zeitungstexten entnommen waren. Dies sind typische Lesetexte. Dialog-Ausschnitte oder ähnliches waren nicht im annotierten Korpus enthalten. Dies korreliert mit der besonders schlechten Leistung der neu trainierten Sprachsynthese bei Dialog-Äußerungen. Gleichzeitig zeigt es die Möglichkeit der Verbesserung in diesem Bereich auf, wenn entsprechende Daten automatisch prosodieannotiert werden.

Die größten prosodischen Unterschiede zwischen der alten und der neu trainierten Sprachsynthese liegen, auch nach Meinung der Teilnehmer, in der Phrasierung. Dies überrascht, da das Phrasierungsmodul letztendlich nicht auf dem neu annotierten Material trainiert werden sollte. Entweder ist an dieser Stelle beim Training ein Fehler geschehen, oder aber das neu trainierte Akzentuierungsmodul beeinflusst die Realisierung der Phrasierung so stark, dass unterschiedliche Fehlphrasierungen desselben Moduls unterschiedlich zur Geltung kommen.

An Phrasengrenzen fügt das Back-End kurze Pausen in das Signal ein. Die Länge dieser Pausen wird nicht trainiert, sondern durch Regeln bestimmt. Pausierungsfehler wirken sich in mindestens zwei der Äußerungen deutlich stärker in der neu trainierten Synthese aus als in der alten. An diesen Stellen geht der generierte Grundfrequenzverlauf an der Zwischengrenze in der neuen Synthese nach oben, während die alte monoton verläuft und nur die Laute gedehnt sind.

Die Pause nach der ansteigenden Zwischengrenze (entsprechend dem ToBI-Phrasenton `H-`) wirkt sehr unnatürlich. Wenn die Pause manuell entfernt wird, klingt (meiner Meinung nach) die neu trainierte Synthese an diesen Stellen deutlich besser.

Das insgesamt schlechtere Abschneiden der neu trainierten Synthese kann also auch darauf zurückgeführt werden, dass die neu trainierten Module zwar besser zur Sprecherin passen, jedoch schlechter in das Gesamtsystem integriert sind. Vor allem die Feineinstellung vieler Parameter wurde vom alten System direkt übernommen anstatt auf die neuen Gegebenheiten abgestimmt.

8.5. Fazit

Durch die Nutzung der automatischen Prosodieannotierung des Korpus zum Training der Prosodiemodule des TTS-Systems geht die Synthesequalität spürbar zurück. Hierfür gibt es mehrere mögliche Gründe.

Fehlerquellen

Die ursprüngliche Synthese benutzt nicht zwei Akzentuierungsniveaus, sondern sieben. Die Aufteilung in nur zwei Niveaus bedeutet, dass jeder Fehler in der Akzentzuweisung zur Zuweisung der ganz falschen Klasse führt. Bei sieben Niveaus fällt ein Fehler von ein oder zwei Niveaus weniger stark ins Gewicht. Die Unterteilung in zuviele Akzentuierungsniveaus verstärkt allerdings das Problem der Datenknappheit beim Training der Module für die akustischen Parameter.

Anscheinend ist die Aufteilung in nur zwei Akzentuierungsniveaus nicht ausreichend. Meiner Erfahrung nach sinnvoll wäre eine Einteilung in (1) reduzierte Silben, (2) normale, nicht akzentuierte Silben, (3) normal akzentuierte Silben und (4) emphatisch akzentuierte Silben. Letztere treten beispielsweise beim Buchstabieren auf und könnten bei entsprechendem Trainingsmaterial auch ein Schlüssel zu lebendigeren Dialog-Systemen sein.

Bei der Auswahl des Klassifizierers und der Merkmale für die Annotierung des Korpus wurde nicht auf die weitere Anwendung geachtet. Insbesondere wurde ein Naïve-Bayes-Klassifizierer gewählt, da dieser minimal bessere Ergebnisse lieferte als der Entscheidungsbaum. Die im TTS-System auf der anderen Seite trainierten Module nutzen Entscheidungsbäume. Die Klassifizierungsalgorithmen unterscheiden sich auch in ihrem Fehlerverhalten. Es wäre möglicherweise besser, den (zunächst schlechteren) Entscheidungsbaum bei der Korpusannotierung zu benutzen, da sich die entstehenden Muster durch die TTS-Module besser erlernen lassen und somit die Gesamtperformanz steigt.

Gleiches gilt für die durch die Merkmalsauswahl ausgewählten Merkmale: Die beste Annotierung nützt nichts, wenn sie nicht durch die entsprechenden TTS-Module repliziert werden kann.

Der Rückgriff auf das bereits trainierte symbolische Phrasierungsmodul könnte auch eine Fehlerquelle dargestellt haben. Zum einen heißt die bessere Trainierbarkeit der vom anderen Sprecher übernommenen Phrasierungsdaten noch nicht, dass sie der Phrasierung dieser Sprecherin entsprechen. Zum anderen wäre der (zugegeben relativ zufälligen) Phrasierung des neu trainierten Moduls automatisch beim Training des akustischen Moduls weniger Gewicht zugemessen worden. Dadurch wären Phrasierungsfehler weniger stark aufgefallen, als es bei der Evaluierung der Fall war.

Um die Leistung der symbolischen TTS-Module abschätzen zu können, wurden nur 80 % des annotierten Materials zu ihrem Training benutzt. Im Anschluss an die Leistungsabschätzung wäre ein erneutes Training auf dem gesamten Material von Vorteil gewesen.

Das Training der akustischen Module des TTS-Systems erfolgte mithilfe der trainierten symbolischen Module. Dadurch pflanzen sich die unausweichlichen Fehler beim Training des symbolischen Teils fort. Es wurden aber zwei völlig unterschiedliche Dinge trainiert. Einerseits das symbolische Modul für Junktur und Satzakzent. Andererseits das akustische Modul für die Realisierung von Phrasierungen und Akzentuierungen. Letzteres benötigt möglichst präzise Informationen darüber, welche Silben akzentuiert waren und wo Phrasierungen vorliegen. Das ist völlig unabhängig vom im Einsatz vorgeschalteten symbolischen Modul. Es geht nicht darum, die Fehler des vorgeschalteten Moduls mitzulernen. Das symbolische Modul wurde für das Training des akustischen Moduls nur deswegen verwendet, weil dadurch die Konvertierung des annotierten Korpus in ein weiteres Austauschformat unnötig wurde. Die direkte Konvertierung des (silbenannotierten) Korpus würde sich sicherlich lohnen.

Insgesamt werden die Anwendungsergebnisse durch vielfältige, ungünstige Rahmenbedingungen mitbestimmt, die im Rahmen der Arbeit nicht verändert werden konnten. Dennoch zeigt diese Anwendung die prinzipielle Nutzbarkeit der automatischen Prosodieannotierung für die Sprachsynthese.

Evaluierung

Die Vorauswahl der Stimuluspaare erfolgte, damit den Teilnehmern überhaupt unterschiedliche Beispiele zur Bewertung vorgelegt wurden. Es war aber ungünstig, nur extreme Beispiele in die Evaluierung mit einzubeziehen. Dadurch lag für die Teilnehmer auf der Hand, welcher Stiumulus besser oder schlechter war.

Die Vorauswahl wurde so durch die Evaluierung in erster Linie bestätigt, anstatt neue Informationen zu liefern. Es wäre günstiger gewesen, sich deutlich voneinander unterscheidende Stimuli auszuwählen, für die aber nicht auf Anhieb klar ist, welcher der bessere wäre. Der Nutzen einer Evaluierung liegt eigentlich gerade in der Klärung solcher Streitfälle.

Prosodisch motivierte Sprecherauswahl

Abschließend ergibt sich ein möglicher Schluss für die Sprecherauswahl eines neu aufzunehmenden Synthesekorpus: Ein guter Sprecher und eine gute Prosodieannotierung seiner Sprechdaten nützen nichts, wenn die ideolektale Prosodierealisierung des Sprechers nicht gut genug vom TTS aus dem Sprechtext vorhergesagt werden kann. Es muss also darauf geachtet werden, dass der Sprecher möglichst konsistent phrasiert (möglichst an Satzzeichen). Die Akzentuierung sollte ebenfalls gleichmäßig und konsistent sein. Bei Wörtern mit mehreren möglichen Betonungen („Marzipan“) sollte vom Sprecher die Akzentuierung gewählt werden, die auch vom TTS realisiert wird.

Ein solches Korpus ist dann selbstverständlich nicht mehr zur Prosodieforschung geeignet, da es nicht die natürliche Prosodie des Sprechers widerspiegelt. Außerdem eignet es sich wahrscheinlich nicht für besonders lebendige Sprachsynthese und kann zukünftige Verbesserungen im System nur bedingt umsetzen. Innerhalb der Beschränkungen des zugrundeliegenden TTS-Systems wäre es aber das ideale Korpus.

Kapitel 9. Zusammenfassung, Fazit und Ausblick

Im Rahmen dieser Arbeit wurde die automatische Annotierung von Akzentuierungen und Phrasierungen in Sprachsynthesekorpora untersucht.

Dazu wurden mehrere Korpora auf ihre Prosodieannotierung untersucht und verglichen. Für ein weiteres Korpus wurde eine minimale Prosodieannotierung durchgeführt.

Es wurde ein System erstellt, dass für die Akzentuierungs- und Phrasierungserkennung wichtige Merkmale aus den Korpora auf Silben- beziehungsweise Wortebene extrahiert.

Unter den zur Verfügung stehenden Merkmalen wurden automatisch die besten Merkmale für zwei unterschiedliche Klassifizierungsalgorithmen, C4.5 und Naïve-Bayes, ausgewählt. Dafür wurde die Umhüllungsmethode benutzt.

Auf den Korpora wurden sprecherspezifisch und -übergreifend Merkmalsauswahlen vorgenommen und Klassifizierer trainiert. Die Ergebnisse der Merkmalsauswahl und die Ergebnisse der Klassifizierung wurden vorgestellt und diskutiert.

Die sprecherabhängige Akzentuierungserkennung erreicht ein F-Maß von bis zu 84 %. Sie kann damit als erfolgreich betrachtet werden. Auch sprecherübergreifend zeigt sich eine gute Leistung.

Die Phrasierungserkennung erreicht keine so guten Werte. Die korpusübergreifende Phrasierungserkennung scheitert an zu großen Unterschieden zwischen den ToBI-annotierten Korpora beziehungsweise an zu wenig Trainingsmaterial im Kiel-Korpus.

Beide verwendeten Klassifizierungsalgorithmen liegen in ihrer Leistung in etwa gleich auf. Möglicherweise ist der Naïve-Bayes-Klassifizierer für die sprecherübergreifende Erkennung besser geeignet.

Ein klares Bild, welche Merkmale geeignet und welche ungeeignet sind, ergibt sich nicht. Gerade die sprecherübergreifende Erkennung gestaltet sich wegen Unterschieden zwischen den Sprechern und vor allem wegen Unterschieden zwischen den Korpora als schwierig.

Abschließend wurde eine mögliche Anwendung der automatischen Prosodieerkennung in der Sprachsynthese vorgestellt und prototypisch durchgeführt.

Mit der gesteuerten Merkmalsauswahl wurde ein Verfahren vorgestellt um mit wenigen annotierten Daten für einen Zielsprecher spezifisch Merkmale zur Prosodieerkennung auszuwählen.

Es zeigt sich, dass zumindest im vorliegenden Fall die Leistung des TTS-Systems durch die automatische Prosodieannotierung sich nicht verbessert.

Es ist aber zweifelhaft, ob dies auf Fehler der automatischen Prosodieannotierung zurückzuführen ist, oder ob nur die Anwendung nicht von ihr profitiert. Zudem mussten viele Rahmenbedingungen akzeptiert werden, die den Nutzen der automatischen Prosodieannotierung im verwendeten TTS-System weiter einschränken.

Schlüsse für die Prosodieannotierung

Für die Prosodieannotierung eines Korpus können einige Schlüsse gezogen werden. Die Annotierung von Akzentuierungen und Phrasierungen sollte in einem Schritt erfolgen. Die unabhängige Annotierung von Phrasierungen und Akzentuierungen führt zu Verwechslungsfehlern von Zwischengrenzen und Akzentuierungen.

Sehr hilfreich bei der Prosodieannotierung ist eine Visualisierung des Audiosignals (Spektrogramm und Grundfrequenzverlauf). Dadurch wird die Erinnerung an den genauen Höreindruck unterstützt und die Notwendigkeit des Mehrfachhörens reduziert.

Aus dem Schrifttext sollten für die Phrasierungsannotierung die Satzzeichen entfernt werden. So ist gewährleistet, dass diese den Annotator nicht vom eigentlichen Hören ablenken und die Phrasierungsannotierung verfälscht wird.

Das vorgeschlagene Vorgehen verträgt sich gut mit der Annotierung von Akzentuierungen im Zeitsignal. Die zunächst am IBM-Korpus durchgeführte Annotierung auf Wortebene ist für die Akzentuierungsannotierung nicht ausreichend. Die direkte Markierung der akzentuierten Silbe bedeutet außerdem keinen nennenswerten Mehraufwand.

Das im Rahmen der Arbeit prosodieannotierte Korpus enthält noch Potential für die Verbesserung der Sprachsynthese. Insbesondere könnte es (sowie die anderen Korpora) benutzt werden, um die Variabilität der Wortakzentuierung zu untersuchen.

Ein Vergleich der automatischen Prosodieannotierung mit der Handannotierung kann dabei helfen, letztere von Fehlern und Unstimmigkeiten zu bereinigen.

Eine weitere mögliche Anwendung wäre, die Versuche Zehnpfennings (2005) zur Einheitenauswahl auf Grundlage von Akzentuierungen mit einem größeren und automatisch prosodieannotierten Synthesekorpus zu wiederholen. Möglicherweise würde dies den Datenmangel in Zehnpfennings Modell reduzieren und damit seine Leistung verbessern.

9.1. Überlegungen zur Prosodieannotierung

Meines Erachtens nach ist die verwendete Prosodieannotierung für die in der Arbeit verfolgten Ziele ungenügend. Sie ist überwiegend phonologisch orientiert; in dieser Arbeit wird aber gerade eine phonetische Unterscheidung benötigt.

Die binäre Unterscheidung in akzentuierte und nicht-akzentuierte Silben in GToBI reicht für die phonetische Beschreibung der Silben nicht aus. Außerdem wird einseitig nur der Tonhöhenverlauf als das alleinige Phänomen der Akzentuierung angesehen.

Besser erscheint mir die mehrstufige (1–4) Unterscheidung in PROLAB. Allerdings würde ich in einer phonetischen Akzentuierungsannotierung die Stufen *-1: reduziert*, *0: nicht akzentuiert*, *1: akzentuiert* und *2: emphatisch* unterscheiden.

Die Akzentuierungsannotierung ist bereits gut standardisiert, was sich durch die übereinstimmenden Ergebnisse bei der Korpusanalyse und bei der Akzentuierungserkennung zeigt. Bei der Phrasierungsannotierung ist dies nicht der Fall, auch wenn mehrere Korpora scheinbar standardisiert GToBI-annotiert sind.

Die phonologisch motivierte Unterscheidung zweier Phrasierungsebenen durch GToBI (*intermediäre Phrasen* und sie umfassende *Intonationsphrasen*) ist phonetisch nicht zielführend.

Eine Unterscheidung stärkerer und schwächerer Phrasierungen ist sicherlich notwendig. Die phonologische Unterscheidung ist aber für die phonetische Beurteilung nicht immer hilfreich.

Viele der GToBI-Zwischengrenzen werden nur aufgrund zum Beispiel syntaktischer Zusammenhänge vom Menschen annotiert und bei einer Delexikalisierung des Sprachsignals nicht mehr gesetzt (Strom und Widera 1996).

Den Wert einer solchen Annotierung für die Vorhersage von Phrasierungen halte ich für zweifelhaft, da offensichtlich keine prosodischen Signalparameter die Wahrnehmung der Phrasierung bestimmen. Wenn es keine (hinreichend konsistenten) akustischen Merkmale der Zwischengrenzen gibt und der Phrasierungseindruck erst beim Hörer entsteht, dann ist es auch unmöglich und zudem überflüssig, diese bei der Sprachsynthese gezielt zu erzeugen.

Besonders die Phrasierungsannotierung steht im Spannungsfeld zwischen sprachlichem Zeichen und kontinuierlichem Parameter. Deshalb glaube ich, dass die Angabe der *Junktur* auf einer – potentiell mehrdimensionalen – Skala zwischen Wörtern besser für die Sprachverarbeitung geeignet ist, als die Unterteilung in zwei oder mehr Phrasierungsklassen.

Es wäre beispielsweise interessant, auf Basis des Trainingsmaterials keine Klassifizierung sondern eine lineare Regression zu trainieren und auf dieser Basis die Junktur zwischen den Wörtern weiter zu untersuchen. Die Angabe der Junkturstärke anstatt der Phrasierungsklasse im Korpus würde Zweifelsfälle der jetzigen Annotierung klären helfen.

Anhang A. Ergebnisse der sprecherabhängigen Merkmalsauswahl

Die Tabellen in diesem Anhang geben die Ergebnisse der automatischen Merkmalsauswahl für die einzelnen Sprecher wieder. Sie wurde getrennt für die Klassifizierungsalgorithmen C4.5 (in den Tabellen *J48*) und Naïve-Bayes (*NB*) mit zehnfach stratifiziert geschichteter Kreuzvalidierung durchgeführt. Die Tabellen enthalten außerdem (in der Spalte *all*) jeweils die Zusammenfassung der Kreuzvalidierung.

Die Ergebnisse der automatischen Merkmalsauswahl für die einzelnen Sprecher werden ausführlich in Kapitel 6 erläutert. Der beste der zehn Klassifizierer jeder Kreuzvalidierung ist jeweils hervorgehoben. Ebenso sind die aufgrund der Kreuzvalidierungen manuell ausgewählten Merkmale in der *all*-Spalte hervorgehoben.

Tabelle A.1. Automatische Merkmalsauswahl für Akzentuierungen im IBM-Korpus

	j48											**nb**										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**
nucleus	+	+	+	+	+	+	+	+	+	+	**10**	+	+	+	+	+	+	+	+	+	+	**10**
sylduration									+		1	+	+						+	+		**4**
absnucleusduration		+									1						+					1
relnucleusduration							+	+			2											
nucleusproportion	+							+	+	+	**4**	+					+					2
Painte_A1-1	+							+	+		3									+		1
Painte_A2-1					+	+		+			3											
Painte_B-1																						
Painte_C1-1	+			+	+						3	+				+			+		+	**4**
Painte_C2-1												+	+	+	+	+	+		+		+	**8**
Painte_D-1	+				+	+					3	+					+	+		+		**4**
Painte_SE-1				+							1											
Painte_MSE-1	+										1										+	1
Painte_A1_0		+		+			+			+	4											
Painte_A2_0																						
Painte_B_0	+		+		+	+		+	+		**6**					+						1
Painte_C1_0												+	+	+	+	+	+	+	+	+	+	**10**
Painte_C2_0	+										1	+			+			+		+		**4**
Painte_D_0	+										1		+				+		+	+		**4**
Painte_SE_0																						
Painte_MSE_0	+										1			+								1
Painte_A1+1				+	+						2								+			1
Painte_A2+1		+			+		+	+			**4**	+							+		+	3
Painte_B+1				+							1											
Painte_C1+1												+	+			+	+	+			+	**6**
Painte_C2+1	+				+						2					+						1
Painte_D+1			+	+			+				3											
Painte_SE+1																	+				+	2
Painte_MSE+1																				+	+	2
medianF0	+	+		+					+		**4**	+	+	+	+	+	+	+	+	+	+	**10**
relativeMedianF0	+	+	+		+		+	+		+	**7**	+	+	+	+	+	+	+	+	+	+	**10**
F0inContext																						
middleF1	+								+		2	+			+	+	+					**4**
middleF2	+	+	+		+					+	**5**	+						+				2
middleF3												+		+		+	+				+	**5**
middleF4	+	+						+			**3**						+					1
relativeMiddleF1	+										1											
relativeMiddleF2	+										1			+				+		+	+	**4**
relativeMiddleF3					+						1					+						1
relativeMiddleF4	+										1											
medianpower																						
relativeMiddlePower																						
sylInWord	+	+	+	+		+	+	+	+	+	**9**	+		+	+	+	+	+		+	+	**8**
relSylInWord		+			+	+	+				**4**		+						+			2
numOfSyllablesInWord		+			+						2			+								1
simplePOS	+				+	+				+	4			+	+		+		+			4
POS		+	+	+			+	+	+		**6**	+	+		+	+		+		+	+	**7**
wordFrequency												+	+	+	+	+	+	+	+	+	+	**10**
punctuationdistanceleft				+							1			+								1
punctuationdistanceright					+						1	+		+	+	+	+	+	+		+	**8**
positionbetweenpunctuation								+			1			+		+		+	+	+	+	**6**
followingpunctuation	+										1						+				+	2
phraseboundarydistanceleft																						
phraseboundarydistanceright																	+			+		2
phraseboundarytype	+										1	+		+		+	+		+			5
positioninphrase																						
features	**23**	**12**	**7**	**11**	**15**	**7**	**9**	**12**	**9**	**7**	**11**	**20**	**11**	**16**	**12**	**18**	**21**	**14**	**16**	**16**	**19**	**16**
precision	**0.71**	**0.73**	**0.73**	**0.74**	**0.73**	**0.67**	**0.74**	**0.75**	**0.73**	**0.73**	**0.73**	**0.56**	**0.56**	**0.56**	**0.52**	**0.56**	**0.56**	**0.55**	**0.56**	**0.56**	**0.56**	**0.55**
recall	**0.74**	**0.72**	**0.74**	**0.74**	**0.75**	**0.68**	**0.74**	**0.73**	**0.73**	**0.66**	**0.72**	**0.89**	**0.86**	**0.88**	**0.90**	**0.88**	**0.87**	**0.86**	**0.88**	**0.86**	**0.88**	**0.88**
f-measure	**0.73**	**0.73**	**0.74**	**0.74**	**0.74**	**0.68**	**0.74**	**0.74**	**0.73**	**0.69**	**0.73**	**0.68**	**0.68**	**0.68**	**0.66**	**0.68**	**0.68**	**0.67**	**0.68**	**0.68**	**0.68**	**0.68**

Tabelle A.2. Automatische Merkmalsauswahl für Phrasengrenzen im IBM-Korpus

	j48											nb										
	1	2	3	4	5	6	7	8	9	10	all	1	2	3	4	5	6	7	8	9	10	all
distancetopunctuationleft			+	+							2		+	+	+				+			4
distancetopunctuationright	+		+	+	+	+		+		+	7	+	+	+	+	+	+	+	+	+	+	10
positionbetweenpunctuation			+	+							2	+									+	2
nextpunctuation	+		+	+	+	+		+		+	7	+	+	+	+	+	+	+	+	+	+	10
POSbefore	+	+	+	+			+			+	6	+	+	+	+	+	+	+	+	+	+	10
POSafter	+	+	+	+	+	+	+		+		8			+	+	+	+	+		+		6
simplePOSbefore		+			+	+	+	+	+		6		+									1
simplePOSafter																						
wordFreqbefore										+	1											
WordFreqafter																	+	+	+	+		4
cdgdistance					+	+					2	+				+		+				3
cdgleftlabel					+						1											
cdgrightlabel	+	+	+					+		+	5	+	+						+		+	4
pause			+	+	+	+	+		+		6											
silenceDuration			+	+	+	+	+		+		6											
Painte_A1-1	+			+			+				3											
Painte_A2-1	+			+				+			3								+	+		2
Painte_B-1		+	+		+	+	+		+		6					+						1
Painte_C1-1												+				+		+	+		+	5
Painte_C2-1		+		+	+		+	+			5			+			+	+			+	4
Painte_D-1	+	+					+				3		+				+					2
Painte_SE-1				+							1						+		+		+	3
Painte_MSE-1		+		+		+					3											
Painte_A1_0	+		+								2											
Painte_A2_0					+				+		2				+	+			+	+	+	5
Painte_B_0			+								1	+	+	+	+	+	+	+	+	+	+	10
Painte_C1_0				+	+			+			3								+	+		2
Painte_C2_0			+								1	+							+			2
Painte_D_0			+		+	+					3					+						1
Painte_SE_0							+				1								+	+		2
Painte_MSE_0			+	+						+	3				+	+				+	+	4
Painte_A1+1			+	+			+		+		4											
Painte_A2+1								+			1							+				1
Painte_B+1				+				+			2						+					1
Painte_C1+1			+								1			+	+		+		+	+	+	6
Painte_C2+1			+	+				+			3											
Painte_D+1	+		+								2											
Painte_SE+1	+		+								2							+				1
Painte_MSE+1	+	+									2						+					1
PowerJump	+	+		+			+				4						+		+			2
RelativePowerJump		+	+						+		3											
finalLenghtening			+			+	+	+	+		5						+					1
relativefinalLenghtening						+				+	2	+	+		+			+				4
finalPowerDecrease		+		+							2											
relativefinalPowerDecrease			+		+		+				3										+	1
linefit150intercept				+	+						2		+			+						2
linefit150slope												+	+						+			3
linefit150mse	+		+						+		3											
nextlinefit150intercept		+						+			2		+									1
nextlinefit150slope																		+				1
nextlinefit150mse				+							1		+									1
intercept150difference		+			+				+		3				+	+						2
slope150difference		+	+			+	+				4								+			1
linefit200intercept					+			+			2	+										1
linefit200slope	+										1				+		+			+		3
linefit200mse								+			1		+									1
nextlinefit200intercept									+		1		+			+						2
nextlinefit200slope								+	+		2					+						1
nextlinefit200mse	+			+							2									+		1
intercept200difference	+										1				+							1
slope200difference	+									+	2					+	+		+			3
linefit2000intercept			+								1			+		+	+			+		4
linefit2000slope	+				+						2							+				1
linefit2000mse													+		+		+					3
nextlinefit2000intercept	+							+			2	+			+				+			3
nextlinefit2000slope		+	+	+	+	+		+			6								+		+	2
nextlinefit2000mse			+		+						2											
intercept2000difference			+			+	+				3	+										1
slope2000difference	+		+	+							3									+		1
accentuationleftinsyllables																			+			1
accentuationrightinsyllables			+			+					2			+						+		2
features	21	16	31	25	20	16	16	17	13	8	18	14	16	10	15	17	18	14	22	17	14	15
BP2-prec	0.64	0.62	0.73	0.71	0.62	0.62	0.90	0.68	0.82	0.60	0.70	0.43	0.43	0.39	0.42	0.31	0.32	0.41	0.36	0.38	0.38	0.38
BP2-rec	0.71	0.38	0.77	0.77	0.71	0.71	0.33	0.66	0.40	0.67	0.61	0.79	0.85	0.82	0.84	0.79	0.76	0.88	0.87	0.84	0.83	0.83
BP2-f	0.67	0.47	0.75	0.74	0.66	0.66	0.48	0.67	0.54	0.64	0.63	0.56	0.57	0.53	0.56	0.45	0.45	0.56	0.51	0.52	0.53	0.52
BP1-prec	0.63	0.61	0.71	0.68	0.71	0.77	0.72	0.62	0.71	0.61	0.68	0.46	0.44	0.45	0.43	0.47	0.45	0.46	0.41	0.42	0.41	0.44
BP1-rec	0.50	0.56	0.59	0.56	0.55	0.52	0.58	0.57	0.53	0.51	0.55	0.75	0.77	0.74	0.75	0.69	0.76	0.77	0.76	0.71	0.84	0.75
BP1-f	0.56	0.58	0.65	0.61	0.62	0.62	0.64	0.59	0.61	0.55	0.60	0.57	0.56	0.56	0.55	0.56	0.57	0.58	0.54	0.53	0.55	0.55

Tabelle A.3. Automatische Merkmalsauswahl für Akzentuierungen im IMS-Korpus

	j48											**nb**										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**
nucleus		+		+	+	+	+	+	+	+	8		+	+		+		+	+		+	6
sylduration	+					+					2	+	+	+	+	+	+	+	+	+	+	10
absnucleusduration		+	+	+	+	+	+	+	+	+	9	+	+	+	+	+	+	+	+	+		9
relnucleusduration		+		+							2										+	1
nucleusproportion																	+					1
Painte_A1-1	+									+	2						+		+	+	+	4
Painte_A2-1	+						+			+	3	+	+		+							3
Painte_B-1	+							+			2											
Painte_C1-1	+								+	+	3		+						+			2
Painte_C2-1	+	+							+		3				+							1
Painte_D-1								+	+		2					+						1
Painte_SE-1	+							+	+		3											
Painte_MSE-1						+		+		+	3						+			+		2
Painte_A1_0		+	+		+	+	+		+	+	7											
Painte_A2_0								+		+	2				+							1
Painte_B_0		+	+		+		+		+	+	6	+										1
Painte_C1_0	+			+	+			+			4	+		+	+	+	+	+			+	7
Painte_C2_0			+								1					+						1
Painte_D_0							+				1		+									1
Painte_SE_0																						
Painte_MSE_0																						
Painte_A1+1	+		+		+						3											
Painte_A2+1		+									1											
Painte_B+1			+								1											
Painte_C1+1	+			+				+		+	4											
Painte_C2+1				+	+						2		+	+					+			3
Painte_D+1										+	1		+		+		+			+	+	5
Painte_SE+1					+			+		+	3											
Painte_MSE+1					+					+	2											
medianF0	+		+		+	+				+	5	+			+			+			+	4
relativeMedianF0															+							1
F0inContext																						
middleF1					+				+		2											
middleF2		+	+	+		+			+	+	6	+							+	+	+	4
middleF3					+						1		+		+						+	3
middleF4									+		1											
relativeMiddleF1																+				+		2
relativeMiddleF2				+				+			2					+						1
relativeMiddleF3									+		1		+								+	2
relativeMiddleF4	+			+				+			3								+		+	2
medianpower				+			+				2											
relativeMiddlePower								+			1											
sylInWord						+		+			2	+							+			2
relSylInWord	+	+	+	+						+	5	+	+	+	+	+	+	+		+	+	9
numOfSyllablesInWord						+			+		2				+			+				2
simplePOS				+							1					+						1
POS	+										1		+		+		+			+		4
wordFrequency								+	+	+	3	+	+	+	+	+	+	+	+	+	+	10
punctuationdistanceleft																						
punctuationdistanceright								+		+	2						+	+	+			3
positionbetweenpunctuation		+								+	2											
followingpunctuation																			+			1
phraseboundarydistanceleft																						
phraseboundarydistanceright		+		+	+		+				4				+							1
phraseboundarytype																	+					1
positioninphrase																						
features	**14**	**11**	**9**	**13**	**13**	**9**	**8**	**16**	**14**	**19**	**12**	**10**	**13**	**8**	**14**	**11**	**12**	**9**	**12**	**10**	**13**	**11**
precision	**0.78**	**0.69**	**0.71**	**0.72**	**0.71**	**0.75**	**0.68**	**0.76**	**0.67**	**0.72**	**0.72**	**0.51**	**0.47**	**0.51**	**0.51**	**0.48**	**0.51**	**0.52**	**0.49**	**0.47**	**0.49**	**0.49**
recall	**0.72**	**0.73**	**0.70**	**0.70**	**0.69**	**0.75**	**0.68**	**0.71**	**0.71**	**0.72**	**0.71**	**0.85**	**0.90**	**0.88**	**0.83**	**0.91**	**0.86**	**0.86**	**0.89**	**0.87**	**0.91**	**0.88**
f-measure	**0.75**	**0.71**	**0.71**	**0.71**	**0.70**	**0.75**	**0.68**	**0.73**	**0.69**	**0.72**	**0.71**	**0.63**	**0.61**	**0.64**	**0.63**	**0.63**	**0.64**	**0.65**	**0.63**	**0.61**	**0.63**	**0.63**

Tabelle A.4. Automatische Merkmalsauswahl für Phrasengrenzen im IMS-Korpus

	j48											nb										
	1	2	3	4	5	6	7	8	9	10	all	1	2	3	4	5	6	7	8	9	10	all
distancetopunctuationleft				+							1	+		+				+	+			4
distancetopunctuationright		+	+				+	+	+	+	6	+									+	2
positionbetweenpunctuation					+						1			+								1
nextpunctuation								+			1	+										1
POSbefore	+		+			+					3	+	+	+	+	+	+	+	+	+	+	10
POSafter	+		+								2			+		+		+			+	4
simplePOSbefore	+	+		+			+				4	+										1
simplePOSafter				+		+		+	+	+	5	+						+		+		3
wordFreqbefore				+							1	+	+	+	+	+	+	+	+	+	+	10
WordFreqafter				+						+	2				+							1
cdgdistance		+									1	+		+		+	+	+			+	6
cdgleftlabel													+	+				+	+			4
cdgrightlabel													+					+			+	3
pause		+						+			2	+				+			+	+	+	5
silenceDuration								+		+	2				+	+	+	+	+			5
Painte_A1-1				+					+	+	3											
Painte_A2-1	+		+				+				3											
Painte_B-1						+	+		+		3								+			1
Painte_C1-1				+							1			+								1
Painte_C2-1					+	+	+				3							+				1
Painte_D-1																						
Painte_SE-1				+							1											
Painte_MSE-1	+			+							2											
Painte_A1_0					+				+		2											
Painte_A2_0		+	+	+	+	+	+	+	+	+	9	+	+			+			+	+	+	6
Painte_B_0				+			+	+	+		4											
Painte_C1_0										+	1	+	+		+	+		+				5
Painte_C2_0	+							+		+	3	+	+	+	+	+	+	+	+	+	+	10
Painte_D_0					+					+	2						+	+				2
Painte_SE_0				+						+	2					+						1
Painte_MSE_0				+		+					2											
Painte_A1+1				+	+					+	3	+				+	+	+	+			5
Painte_A2+1	+										1							+				1
Painte_B+1												+		+								2
Painte_C1+1												+			+	+	+		+	+	+	7
Painte_C2+1	+		+								2					+						1
Painte_D+1					+						1										+	1
Painte_SE+1				+				+			2					+						1
Painte_MSE+1						+		+			2						+					1
PowerJump	+	+		+				+			4											
RelativePowerJump		+						+			2				+		+			+	+	4
finalLenghtening					+						1	+				+	+					3
relativefinalLenghtening					+			+		+	3	+		+	+							3
finalPowerDecrease	+	+						+			3	+						+	+	+		4
relativefinalPowerDecrease	+	+		+					+		4		+	+	+			+			+	5
linefit150intercept					+		+	+			3					+	+		+	+		4
linefit150slope				+							1			+		+						2
linefit150mse																			+			1
nextlinefit150intercept			+				+		+	+	4								+			1
nextlinefit150slope		+						+		+	3	+							+			2
nextlinefit150mse	+	+		+		+	+	+	+	+	8	+				+	+	+	+	+		6
intercept150difference	+					+					2	+										1
slope150difference	+	+		+	+					+	5										+	1
linefit200intercept		+		+		+					3	+	+									2
linefit200slope		+									1			+	+						+	3
linefit200mse		+	+						+		3											
nextlinefit200intercept		+				+		+	+		4						+	+				2
nextlinefit200slope													+		+	+	+	+	+		+	7
nextlinefit200mse	+	+		+							3								+			1
intercept200difference				+		+					2						+					1
slope200difference	+	+			+					+	4											
linefit2000intercept		+		+							2		+				+	+				3
linefit2000slope									+		1			+		+			+	+		4
linefit2000mse	+			+						+	3				+	+		+			+	4
nextlinefit2000intercept																+	+		+	+		4
nextlinefit2000slope	+							+			2	+			+		+					3
nextlinefit2000mse	+	+		+		+					4			+		+		+				3
intercept2000difference		+	+	+						+	4					+						1
slope2000difference									+		1	+										1
accentuationleftinsyllables					+					+	2	+		+		+		+	+	+		6
accentuationrightinsyllables	+		+					+			3	+					+					2
features	20	21	10	27	13	13	10	20	14	20	16	26	11	17	14	25	20	24	22	14	17	19
BP2-prec	0.68	0.72	0.70	0.85	0.87	0.77	0.71	0.79	0.69	0.76	0.75	0.57	0.68	0.55	0.70	0.63	0.64	0.59	0.63	0.68	0.57	0.62
BP2-rec	0.73	0.77	0.70	0.70	0.69	0.67	0.72	0.71	0.72	0.75	0.72	0.77	0.71	0.83	0.61	0.73	0.70	0.64	0.70	0.59	0.60	0.69
BP2-f	0.71	0.74	0.70	0.77	0.77	0.72	0.72	0.75	0.70	0.75	0.73	0.65	0.70	0.66	0.65	0.68	0.67	0.62	0.66	0.63	0.59	0.65
BP1-prec	0.60	0.51	0.58	0.55	0.56	0.51	0.48	0.60	0.60	0.45	0.54	0.32	0.33	0.30	0.33	0.36	0.34	0.33	0.25	0.28	0.30	0.31
BP1-rec	0.52	0.48	0.62	0.53	0.59	0.47	0.48	0.47	0.49	0.48	0.51	0.73	0.86	0.71	0.86	0.85	0.67	0.83	0.79	0.88	0.86	0.81
BP1-f	0.55	0.50	0.60	0.54	0.57	0.49	0.48	0.53	0.54	0.46	0.53	0.44	0.47	0.42	0.47	0.51	0.45	0.47	0.38	0.42	0.45	0.45

Tabelle A.5. Automatische Merkmalsauswahl für Akzentuierungen im VPSC

	j48											**nb**										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**
nucleus	+		+	+	+	+			+	+	7	+		+	+	+	+	+	+	+	+	9
sylduration			+	+			+	+			4		+	+	+		+	+		+		6
absnucleusduration	+	+			+	+	+	+	+	+	8		+					+			+	3
relnucleusduration							+				1			+			+	+	+			4
nucleusproportion								+			1											
Painte_A1-1																	+					1
Painte_A2-1				+	+						2						+					1
Painte_B-1		+					+			+	3		+							+		2
Painte_C1-1																			+			1
Painte_C2-1			+	+	+	+			+		5											
Painte_D-1																		+		+		2
Painte_SE-1					+	+	+				3						+	+				2
Painte_MSE-1																	+					1
Painte_A1_0												+	+		+		+	+	+	+	+	8
Painte_A2_0		+									1							+				1
Painte_B_0			+	+		+	+				4					+				+	2
Painte_C1_0					+	+		+	+	+	5		+							+		2
Painte_C2_0												+					+					2
Painte_D_0															+			+	+			3
Painte_SE_0			+								1											
Painte_MSE_0																						
Painte_A1+1		+		+							2					+		+				2
Painte_A2+1	+						+				2	+						+				2
Painte_B+1			+					+		+	3			+			+	+				3
Painte_C1+1							+			+	2											
Painte_C2+1		+					+			+	3											
Painte_D+1					+		+				2			+			+	+			+	4
Painte_SE+1					+						1	+		+		+				+		4
Painte_MSE+1			+								1											
medianF0												+	+		+			+	+	+	+	7
relativeMedianF0	+						+			+	3							+				1
F0inContext																						
middleF1	+					+					2									+		1
middleF2										+	1									+		1
middleF3										+	1		+							+	+	3
middleF4	+										1	+										1
relativeMiddleF1			+						+		2									+		1
relativeMiddleF2								+	+		2				+					+	+	3
relativeMiddleF3																						
relativeMiddleF4						+			+		2		+									1
medianpower	+		+		+	+		+			5				+	+	+	+	+	+	+	7
relativeMiddlePower				+							1					+						1
sylInWord		+	+	+	+	+		+			6					+			+		+	3
relSylInWord				+			+		+		3	+	+		+		+	+		+		6
numOfSyllablesInWord												+					+	+		+		4
simplePOS			+			+	+		+		4								+			1
POS		+		+					+		3	+	+		+	+	+	+		+	+	8
wordFrequency			+	+				+			3	+			+	+			+		+	5
punctuationdistanceleft															+							1
punctuationdistanceright			+								1											
positionbetweenpunctuation	+				+						2						+		+			2
followingpunctuation																		+				1
phraseboundarydistanceleft																						
phraseboundarydistanceright	+										1	+	+	+								3
phraseboundarytype		+									1								+			1
positioninphrase																						
features	9	8	13	11	11	11	13	9	10	10	10	12	11	7	11	9	16	20	12	17	12	12
precision	0.74	0.78	0.81	0.84	0.78	0.82	0.81	0.75	0.81	0.78	0.79	0.59	0.62	0.51	0.60	0.55	0.62	0.67	0.58	0.66	0.62	0.60
recall	0.74	0.74	0.82	0.83	0.81	0.84	0.82	0.82	0.85	0.84	0.81	0.94	0.92	0.91	0.97	0.90	0.95	0.92	0.93	0.94	0.94	0.93
f-measure	0.74	0.76	0.82	0.84	0.79	0.83	0.81	0.78	0.83	0.81	0.80	0.72	0.74	0.65	0.74	0.68	0.75	0.77	0.71	0.78	0.74	0.73

Tabelle A.6. Automatische Merkmalsauswahl für Phrasengrenzen im VPSC

	j48											**nb**										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**
distancetopunctuationleft										+	1	+								+		2
distancetopunctuationright	+	+				+	+	+	+	+	7							+				1
positionbetweenpunctuation					+						1							+				1
nextpunctuation																						
POSbefore	+	+	+		+	+		+	+	+	8	+	+	+		+	+		+	+	+	8
POSafter		+									1					+				+		2
simplePOSbefore															+			+	+			3
simplePOSafter	+			+							2				+						+	2
wordFreqbefore				+					+		2	+							+			2
WordFreqafter			+							+	2	+			+				+		+	4
cdgdistance														+								1
cdgleftlabel	+										1						+					1
cdgrightlabel			+	+			+				3				+	+		+			+	4
pause	+	+	+	+	+	+	+	+	+	+	10	+	+	+	+	+	+	+	+	+	+	10
silenceDuration														+			+					2
Painte_A1-1			+								1	+	+		+	+					+	5
Painte_A2-1																			+			1
Painte_B-1	+										1									+		1
Painte_C1-1																						
Painte_C2-1																						
Painte_D-1						+					1											
Painte_SE-1															+	+	+			+		4
Painte_MSE-1																+	+					2
Painte_A1_0																+						1
Painte_A2_0																			+			1
Painte_B_0			+								1					+						1
Painte_C1_0							+				1						+					1
Painte_C2_0															+							1
Painte_D_0																+						1
Painte_SE_0																						
Painte_MSE_0																						
Painte_A1+1																						
Painte_A2+1														+		+						2
Painte_B+1	+										1											
Painte_C1+1						+			+		2										+	1
Painte_C2+1																				+		1
Painte_D+1																				+		1
Painte_SE+1																						
Painte_MSE+1																						
PowerJump																						
RelativePowerJump				+							1						+		+			2
finalLenghtening										+	1	+				+				+		3
relativefinalLenghtening	+	+							+	+	4				+							1
finalPowerDecrease															+	+	+					3
relativefinalPowerDecrease				+							1	+			+			+				3
linefit150intercept																						
linefit150slope																+	+					2
linefit150mse					+						1				+					+	+	3
nextlinefit150intercept																			+			1
nextlinefit150slope																+	+					2
nextlinefit150mse																						
intercept150difference													+									1
slope150difference																						
linefit200intercept															+						+	2
linefit200slope																						
linefit200mse															+							1
nextlinefit200intercept																						
nextlinefit200slope																				+		1
nextlinefit200mse																						
intercept200difference																				+		1
slope200difference																				+		1
linefit2000intercept																+						1
linefit2000slope																						
linefit2000mse																+						1
nextlinefit2000intercept																+	+					2
nextlinefit2000slope																						
nextlinefit2000mse	+										1	+				+						2
intercept2000difference																						
slope2000difference															+							1
accentuationleftinsyllables																				+		1
accentuationrightinsyllables																+						1
features	**9**	**5**	**6**	**6**	**4**	**5**	**4**	**3**	**6**	**7**	**5**	**9**	**4**	**5**	**15**	**20**	**12**	**6**	**9**	**14**	**9**	**10**
BP2-prec	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.96**	**1.00**	**1.00**	**0.89**	**0.92**	**0.92**	**1.00**	**1.00**	**1.00**	**1.00**	**0.97**
BP2-rec	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
BP2-f	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**1.00**	**1.00**	**0.94**	**0.96**	**0.96**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**
BP1-prec	**0.85**	**0.93**	**0.89**	**0.81**	**0.93**	**0.93**	**0.87**	**0.84**	**0.95**	**0.86**	**0.89**	**0.68**	**0.93**	**0.72**	**0.69**	**0.70**	**0.72**	**0.84**	**0.44**	**0.56**	**0.70**	**0.70**
BP1-rec	**0.76**	**0.76**	**0.89**	**0.74**	**0.77**	**0.74**	**0.77**	**0.71**	**0.77**	**0.83**	**0.77**	**0.81**	**0.74**	**0.87**	**0.91**	**0.91**	**0.79**	**0.79**	**0.77**	**0.91**	**0.83**	**0.83**
BP1-f	**0.80**	**0.83**	**0.89**	**0.77**	**0.84**	**0.82**	**0.82**	**0.77**	**0.85**	**0.85**	**0.82**	**0.74**	**0.82**	**0.79**	**0.78**	**0.79**	**0.76**	**0.82**	**0.56**	**0.70**	**0.76**	**0.75**

Tabelle A.7. Automatische Merkmalsauswahl für Akzentuierungen des Sprechers *kko* im KCoRS

	j48											**nb**										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**
nucleus				+	+	+					3	+	+	+								3
sylduration					+				+	+	3		+								+	2
absnucleusduration	+	+					+				3					+				+		2
relnucleusduration						+			+		2					+						1
nucleusproportion									+	+	2				+		+					2
Painte_A1-1													+	+		+						3
Painte_A2-1													+									1
Painte_B-1								+			1											
Painte_C1-1					+		+				2		+			+			+			3
Painte_C2-1																	+				+	2
Painte_D-1					+			+		+	3											
Painte_SE-1					+						1											
Painte_MSE-1																						
Painte_A1_0						+		+			2		+	+		+	+					4
Painte_A2_0														+								1
Painte_B_0	+	+		+		+		+			5											
Painte_C1_0			+								1		+						+			2
Painte_C2_0					+						1					+						1
Painte_D_0		+									1					+						1
Painte_SE_0			+		+						2						+					1
Painte_MSE_0								+			1	+										1
Painte_A1+1		+								+	2			+		+				+		3
Painte_A2+1		+									1					+					+	2
Painte_B+1																						
Painte_C1+1					+						1			+								1
Painte_C2+1	+	+							+		3			+		+	+					3
Painte_D+1					+						1					+					+	2
Painte_SE+1						+		+			2						+					1
Painte_MSE+1					+			+			2						+					1
medianF0		+				+			+		3		+			+						2
relativeMedianF0	+										1		+									1
F0inContext																						
middleF1														+				+		+		3
middleF2					+		+	+	+		4	+										1
middleF3								+			1		+		+							2
middleF4													+	+	+	+	+		+			6
relativeMiddleF1																						
relativeMiddleF2	+							+			2											
relativeMiddleF3																						
relativeMiddleF4																						
medianpower	+	+						+			3	+										1
relativeMiddlePower										+	1						+		+			2
sylInWord	+		+	+	+		+	+	+	+	8									+		1
relSylInWord		+				+		+			3	+	+	+	+	+	+	+	+	+	+	10
numOfSyllablesInWord					+						1					+		+		+	+	4
simplePOS				+		+		+	+		4		+									1
POS	+	+	+				+	+		+	6	+	+	+	+	+	+	+	+	+	+	10
wordFrequency			+					+			2	+		+			+					3
punctuationdistanceleft																						
punctuationdistanceright				+	+						2	+										1
positionbetweenpunctuation					+				+		2											
followingpunctuation	+										1				+							1
phraseboundarydistanceleft																						
phraseboundarydistanceright	+	+			+	+	+		+		6			+								1
phraseboundarytype																						
positioninphrase																						
features	**10**	**11**	**5**	**5**	**16**	**9**	**6**	**16**	**10**	**7**	**9**	**8**	**14**	**13**	**6**	**16**	**12**	**4**	**6**	**7**	**7**	**9**
precision	**0.78**	**0.84**	**0.84**	**0.80**	**0.81**	**0.88**	**0.83**	**0.82**	**0.86**	**0.85**	**0.83**	**0.64**	**0.80**	**0.64**	**0.68**	**0.76**	**0.70**	**0.75**	**0.75**	**0.73**	**0.81**	**0.73**
recall	**0.81**	**0.86**	**0.82**	**0.80**	**0.84**	**0.84**	**0.77**	**0.85**	**0.76**	**0.78**	**0.81**	**0.92**	**0.86**	**0.93**	**0.80**	**0.88**	**0.89**	**0.81**	**0.86**	**0.86**	**0.83**	**0.87**
f-measure	**0.79**	**0.85**	**0.83**	**0.80**	**0.82**	**0.86**	**0.80**	**0.84**	**0.81**	**0.81**	**0.82**	**0.76**	**0.83**	**0.76**	**0.73**	**0.81**	**0.78**	**0.78**	**0.80**	**0.79**	**0.82**	**0.79**

Tabelle A.8. Automatische Merkmalsauswahl für Phrasengrenzen des Sprechers *kko*

	j48											**nb**										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**
distancetopunctuationleft						+	+	+			3								+			1
distancetopunctuationright		+		+	+	+				+	5	+		+								2
positionbetweenpunctuation				+	+	+	+	+			5					+					+	2
nextpunctuation												+				+					+	3
POSbefore	+	+	+				+		+		5	+	+	+		+	+	+	+	+	+	9
POSafter			+	+							2		+			+		+			+	4
simplePOSbefore										+	1				+	+					+	3
simplePOSafter																						
wordFreqbefore				+				+	+		3											
WordFreqafter		+									1	+				+						2
cdgdistance					+		+				2					+		+				2
cdgleftlabel		+						+			2	+	+			+						3
cdgrightlabel				+							1			+		+				+	+	4
pause		+		+	+		+			+	5		+			+	+					3
silenceDuration	+				+		+				3											
Painte_A1-1			+								1					+						1
Painte_A2-1	+										1					+						1
Painte_B-1	+		+	+					+		4											
Painte_C1-1																				+		1
Painte_C2-1									+		1									+	+	2
Painte_D-1				+							1	+	+		+	+	+	+	+	+	+	9
Painte_SE-1																						
Painte_MSE-1																						
Painte_A1_0				+	+	+		+	+		5	+					+			+		3
Painte_A2_0			+		+						2		+									1
Painte_B_0	+		+	+			+	+	+	+	7											
Painte_C1_0												+									+	2
Painte_C2_0												+					+					2
Painte_D_0		+									1		+			+					+	3
Painte_SE_0																						
Painte_MSE_0							+		+		2											
Painte_A1+1		+	+		+					+	4	+	+						+			3
Painte_A2+1					+						1				+		+			+		3
Painte_B+1					+						1											
Painte_C1+1				+			+				2					+			+	+	+	4
Painte_C2+1						+		+			2		+									1
Painte_D+1												+		+		+	+			+		5
Painte_SE+1		+			+			+			3											
Painte_MSE+1				+		+				+	3					+						1
PowerJump																+	+					2
RelativePowerJump																						
finalLenghtening	+					+	+			+	4	+	+	+			+	+	+			6
relativefinalLenghtening					+				+		2					+	+		+	+		4
finalPowerDecrease									+		1			+						+		2
relativefinalPowerDecrease		+									1	+				+	+			+		4
linefit150intercept		+									1											
linefit150slope										+	1	+				+			+			3
linefit150mse		+									1											
nextlinefit150intercept																	+					1
nextlinefit150slope								+	+		2	+				+	+					3
nextlinefit150mse	+						+				2	+										1
intercept150difference	+				+				+		3									+		1
slope150difference												+	+		+	+						4
linefit200intercept					+						1											
linefit200slope																		+			+	2
linefit200mse												+							+		+	3
nextlinefit200intercept									+	+	2				+	+						2
nextlinefit200slope												+		+			+					3
nextlinefit200mse																	+					1
intercept200difference			+	+					+		3						+					1
slope200difference																+	+					2
linefit2000intercept						+					1											
linefit2000slope																						
linefit2000mse							+				1						+					1
nextlinefit2000intercept	+					+					2	+		+		+	+		+	+		6
nextlinefit2000slope					+						1		+					+			+	3
nextlinefit2000mse												+				+					+	3
intercept2000difference																						
slope2000difference							+	+			2											
accentuationleftinsyllables									+		1											
accentuationrightinsyllables																						
features	**9**	**11**	**8**	**13**	**15**	**9**	**13**	**10**	**14**	**9**	**11**	**21**	**12**	**8**	**5**	**27**	**19**	**7**	**10**	**14**	**15**	**13**
BP2-prec	**0.56**	**0.63**	**0.70**	**0.68**	**0.84**	**0.79**	**0.85**	**0.78**	**0.84**	**0.71**	**0.74**	**0.38**	**0.36**	**0.29**	**0.21**	**0.42**	**0.37**	**0.33**	**0.30**	**0.29**	**0.36**	**0.33**
BP2-rec	**0.63**	**0.73**	**0.53**	**0.70**	**0.70**	**0.63**	**0.76**	**0.62**	**0.70**	**0.67**	**0.67**	**0.93**	**1.00**	**0.87**	**0.90**	**0.93**	**1.00**	**0.97**	**0.90**	**0.90**	**0.93**	**0.93**
BP2-f	**0.59**	**0.68**	**0.60**	**0.69**	**0.76**	**0.70**	**0.80**	**0.69**	**0.76**	**0.69**	**0.70**	**0.54**	**0.53**	**0.44**	**0.34**	**0.58**	**0.54**	**0.50**	**0.45**	**0.44**	**0.52**	**0.49**

Tabelle A.9. Automatische Merkmalsauswahl für Akzentuierungen der Sprecherin *rtd* im KCoRS

	j48											**nb**										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**all**
nucleus	+					+			+	+	4	+				+	+		+			4
sylduration		+	+	+	+	+		+	+	+	8		+	+		+					+	4
absnucleusduration									+		1			+		+		+		+		4
relnucleusduration					+						1											
nucleusproportion							+				1		+			+		+			+	4
Painte_A1-1																+						1
Painte_A2-1		+		+							2											
Painte_B-1	+		+			+		+			4											
Painte_C1-1	+			+			+	+	+		5	+	+	+			+				+	5
Painte_C2-1						+					1			+			+	+				3
Painte_D-1							+				1										+	1
Painte_SE-1						+					1		+	+								2
Painte_MSE-1								+			1					+					+	2
Painte_A1_0										+	1						+					1
Painte_A2_0		+									1											
Painte_B_0			+	+						+	3											
Painte_C1_0	+						+			+	3	+	+	+	+			+	+	+	+	8
Painte_C2_0							+				1		+	+		+		+	+	+	+	7
Painte_D_0						+			+	+	3											
Painte_SE_0																						
Painte_MSE_0								+			1											
Painte_A1+1				+							1		+	+	+							3
Painte_A2+1						+					1							+				1
Painte_B+1																						
Painte_C1+1		+						+			2	+	+	+			+				+	5
Painte_C2+1							+			+	2							+				1
Painte_D+1		+								+	2	+				+	+					3
Painte_SE+1		+		+	+						3											
Painte_MSE+1																						
medianF0																	+	+	+			3
relativeMedianF0							+				1											
F0inContext																						
middleF1			+						+	+	3		+									1
middleF2									+	+	2											
middleF3		+			+						2				+			+				2
middleF4	+						+				2							+				1
relativeMiddleF1		+	+		+						3							+	+			2
relativeMiddleF2																						
relativeMiddleF3				+							1						+					1
relativeMiddleF4					+	+					2			+					+			2
medianpower		+	+	+	+	+		+		+	7					+						1
relativeMiddlePower	+					+					2			+	+							2
sylInWord	+		+		+		+				4						+	+				2
relSylInWord							+		+		2	+	+	+	+			+	+	+		7
numOfSyllablesInWord				+					+	+	3		+							+		2
simplePOS	+						+		+		3	+	+	+					+	+	+	6
POS	+										1	+	+		+		+	+	+	+		7
wordFrequency				+			+				2						+					1
punctuationdistanceleft					+	+					2		+							+		2
punctuationdistanceright				+		+	+				3					+					+	2
positionbetweenpunctuation												+							+			2
followingpunctuation	+						+				2	+			+					+		3
phraseboundarydistanceleft																						
phraseboundarydistanceright						+		+			2						+	+		+		3
phraseboundarytype																						
positioninphrase																						
features	**10**	**9**	**7**	**11**	**9**	**13**	**14**	**8**	**10**	**12**	**10**	**10**	**14**	**13**	**7**	**10**	**12**	**15**	**10**	**10**	**10**	**11**
precision	**0.84**	**0.65**	**0.75**	**0.77**	**0.73**	**0.76**	**0.82**	**0.73**	**0.82**	**0.79**	**0.77**	**0.76**	**0.77**	**0.81**	**0.76**	**0.57**	**0.64**	**0.78**	**0.70**	**0.79**	**0.60**	**0.72**
recall	**0.84**	**0.71**	**0.75**	**0.80**	**0.76**	**0.80**	**0.78**	**0.67**	**0.77**	**0.80**	**0.77**	**0.85**	**0.84**	**0.82**	**0.80**	**0.89**	**0.93**	**0.83**	**0.83**	**0.79**	**0.87**	**0.85**
f-measure	**0.84**	**0.68**	**0.75**	**0.78**	**0.75**	**0.78**	**0.80**	**0.70**	**0.79**	**0.80**	**0.77**	**0.80**	**0.81**	**0.82**	**0.78**	**0.70**	**0.76**	**0.80**	**0.76**	**0.79**	**0.71**	**0.77**

Tabelle A.10. Automatische Merkmalsauswahl für Phrasengrenzen der Sprecherin *rtd*

	j48											nb										
	1	2	3	4	5	6	7	8	9	10	all	1	2	3	4	5	6	7	8	9	10	all
distancetopunctuationleft	+	+									2	+				+		+		+		4
distancetopunctuationright	+	+	+	+		+		+	+	+	8	+				+				+	+	4
positionbetweenpunctuation			+		+						2									+		1
nextpunctuation									+		1				+						+	2
POSbefore		+			+		+				3	+	+		+	+			+	+		6
POSafter		+		+		+				+	4									+		1
simplePOSbefore				+			+				2			+			+	+			+	4
simplePOSafter				+				+		+	3											
wordFreqbefore	+	+					+				3											
WordFreqafter									+		1											
cdgdistance					+		+		+		3				+	+		+				3
cdgleftlabel			+								1				+							1
cdgrightlabel	+				+						2	+					+	+	+		+	5
pause					+				+		2	+			+	+		+			+	5
silenceDuration							+				1		+			+						2
Painte_A1-1	+	+			+					+	4		+								+	2
Painte_A2-1	+								+		2	+			+							2
Painte_B-1																						
Painte_C1-1						+					1				+							1
Painte_C2-1				+							1	+		+	+	+	+		+			6
Painte_D-1		+							+		2	+	+		+	+		+	+	+		7
Painte_SE-1																						
Painte_MSE-1													+		+	+		+		+	+	6
Painte_A1_0	+		+								2											
Painte_A2_0										+	1				+	+			+			3
Painte_B_0			+		+			+		+	4											
Painte_C1_0		+						+			2					+						1
Painte_C2_0						+					1	+										1
Painte_D_0	+	+						+			3											
Painte_SE_0					+						1					+						1
Painte_MSE_0									+		1	+	+		+	+				+		5
Painte_A1+1		+	+								2				+	+				+		3
Painte_A2+1					+						1				+	+			+			3
Painte_B+1																						
Painte_C1+1				+					+		2			+	+	+						3
Painte_C2+1		+						+		+	3	+										1
Painte_D+1		+									1				+							1
Painte_SE+1	+								+	+	3	+		+	+	+						4
Painte_MSE+1	+										1	+		+			+		+			4
PowerJump																						
RelativePowerJump												+	+		+							3
finalLenghtening			+				+		+		3			+							+	2
relativefinalLenghtening			+								1							+				1
finalPowerDecrease												+			+	+						3
relativefinalPowerDecrease												+			+	+	+					4
linefit150intercept			+								1	+			+							2
linefit150slope			+	+	+						3			+	+							2
linefit150mse				+			+			+	3		+							+		2
nextlinefit150intercept										+	1			+								1
nextlinefit150slope	+							+	+		3			+	+	+			+	+		5
nextlinefit150mse														+	+	+						3
intercept150difference								+			1		+									1
slope150difference									+		1	+			+	+	+					4
linefit200intercept						+	+				2	+		+	+				+	+		5
linefit200slope	+					+					2						+					1
linefit200mse		+		+				+			3	+										1
nextlinefit200intercept			+			+			+		3							+	+			2
nextlinefit200slope								+			1	+				+				+	+	4
nextlinefit200mse	+			+							2						+					1
intercept200difference				+						+	2					+						1
slope200difference																+		+				2
linefit2000intercept												+								+	+	3
linefit2000slope				+							1				+							1
linefit2000mse													+			+						2
nextlinefit2000intercept						+					1					+		+	+			3
nextlinefit2000slope													+					+	+			3
nextlinefit2000mse				+	+					+	3	+				+	+					3
intercept2000difference		+						+	+		3					+			+			2
slope2000difference	+										1											
accentuationleftinsyllables		+									1							+				1
accentuationrightinsyllables																+				+		2
features	**14**	**15**	**11**	**13**	**11**	**8**	**8**	**11**	**15**	**12**	**11**	**23**	**11**	**11**	**27**	**30**	**9**	**13**	**13**	**15**	**10**	**16**
BP2-prec	**0.72**	**0.64**	**0.79**	**0.90**	**0.71**	**0.58**	**0.82**	**0.73**	**0.88**	**0.79**	**0.76**	**0.46**	**0.45**	**0.55**	**0.50**	**0.41**	**0.35**	**0.54**	**0.44**	**0.43**	**0.42**	**0.46**
BP2-rec	**0.74**	**0.83**	**0.77**	**0.77**	**0.65**	**0.62**	**0.68**	**0.79**	**0.65**	**0.68**	**0.72**	**0.89**	**0.94**	**0.91**	**0.94**	**0.88**	**0.79**	**0.97**	**1.00**	**0.91**	**0.73**	**0.90**
BP2-f	**0.73**	**0.72**	**0.78**	**0.82**	**0.68**	**0.60**	**0.74**	**0.76**	**0.75**	**0.73**	**0.73**	**0.61**	**0.61**	**0.69**	**0.65**	**0.56**	**0.48**	**0.69**	**0.61**	**0.58**	**0.54**	**0.60**

Anhang B. Ergebnisse der sprecherübergreifenden Merkmalsauswahl

Die Tabellen in diesem Anhang geben die Ergebnisse der automatischen Merkmalsauswahl für die Experimente mit mehreren Sprechern wider.

Sie wurde getrennt für die Klassifizierungsalgorithmen C4.5 (in den Tabellen *J48*) und Naïve-Bayes (*NB*) im Auslassverfahren durchgeführt. Aus satztechnischen Gründen stehen die Ergebnisse für C4.5 und Naïve-Bayes in getrennten Tabellen.

Die Tabellen enthalten außerdem (in der Spalte *all*) jeweils eine Zusammenfassung der Ergebnisse des Auslassverfahrens.

Die Ergebnisse der Merkmalsauswahl für sprecherunabhängige Erkennung werden in Kapitel 7 erläutert.

Tabelle B.1. Automatische Merkmalsauswahl für Akzentuierungen im KCoRS (J48)

	k01	k02	k03	k04	k05	k06	k63	k64	k65	k66	dlm	hpt	uga	kko	rtd	k07	k08	k09	k10	k11	k12	k67	k68	k69	k70	all
nucleus	+					+		+				+	+	+	+	+	+		+		+	+		+		13
sylduration			+			+		+		+	+		+			+								+	+	9
absnucleusduration	+	+	+	+	+		+				+	+		+	+	+		+			+		+	+		15
relnucleusduration					+																		+		+	3
nucleusproportion	+					+		+		+				+		+					+			+		8
Painte_A1-1		+											+				+									3
Painte_A2-1			+		+											+									+	4
Painte_B-1				+		+																+	+			4
Painte_C1-1									+						+					+		+				4
Painte_C2-1										+				+	+	+						+				5
Painte_D-1								+							+							+	+			4
Painte_SE-1	+														+								+	+	+	5
Painte_MSE-1						+																		+	+	3
Painte_A1_0		+				+			+									+							+	5
Painte_A2_0				+		+	+		+			+	+		+									+		8
Painte_B_0						+	+	+		+	+		+	+			+		+			+				10
Painte_C1_0	+					+			+									+								4
Painte_C2_0															+											1
Painte_D_0						+																				1
Painte_SE_0	+				+						+														+	4
Painte_MSE_0															+						+					2
Painte_A1+1							+								+	+	+							+		5
Painte_A2+1																+	+					+				3
Painte_B+1			+				+					+										+				4
Painte_C1+1								+								+										2
Painte_C2+1						+		+																	+	3
Painte_D+1					+									+					+	+			+			5
Painte_SE+1											+			+								+				3
Painte_MSE+1						+		+						+		+						+	+			6
medianF0			+																							1
relativeMedianF0								+																		1
F0inContext										+												+				2
middleF1			+					+		+	+				+											5
middleF2			+						+						+						+					4
middleF3				+		+		+			+				+							+				6
middleF4			+								+					+										3
relativeMiddleF1		+				+										+	+									4
relativeMiddleF2	+						+																			2
relativeMiddleF3																			+							1
relativeMiddleF4						+									+									+		3
medianpower		+		+				+							+							+				5
relativeMiddlePower		+	+	+											+											4
sylInWord	+	+		+	+	+	+	+	+					+		+	+				+		+	+	+	15
relSylInWord	+		+							+			+		+		+	+	+	+		+		+		11
numOfSyllablesInWord				+						+		+					+		+					+		6
simplePOS		+	+	+	+				+	+					+						+	+	+			10
POS	+	+		+		+	+	+	+			+	+	+		+	+	+	+	+		+		+	+	18
wordFrequency				+																						1
punctuationdistanceleft										+					+				+			+	+	+		6
punctuationdistanceright												+		+								+			+	4
positionbetweenpunctuation			+	+																+		+	+	+		6
followingpunctuation																										
phraseboundarydistanceleft																										
phraseboundarydistance-right										+					+	+	+					+				5
phraseboundarytype																										
positioninphrase																										
features	10	9	12	12	7	17	8	14	8	11	8	7	7	11	20	15	11	5	8	5	7	20	11	15	11	10.8
precision	0.86	0.68	0.86	0.81	0.76	0.84	0.71	0.78	0.81	0.74	0.58	0.67	0.72	0.80	0.73	0.73	0.50	0.90	0.54	0.70	0.85	0.73	0.81	0.68	0.77	0.74
recall	0.70	0.59	0.52	0.74	0.62	0.73	0.63	0.58	0.61	0.80	0.44	0.55	0.73	0.77	0.73	0.73	0.59	0.69	0.79	0.62	0.65	0.64	0.72	0.59	0.69	0.66
f-measure	0.78	0.63	0.65	0.77	0.68	0.78	0.67	0.67	0.69	0.77	0.50	0.60	0.72	0.78	0.73	0.73	0.54	0.78	0.64	0.65	0.74	0.68	0.76	0.63	0.73	0.69

Tabelle B.2. Automatische Merkmalsauswahl für Akzentuierungen im KCoRS (NB)

	k01	k02	k03	k04	k05	k06	k63	k64	k65	k66	dlm	hpt	uga	kko	rtd	k07	k08	k09	k10	k11	k12	k67	k68	k69	k70	all
nucleus	+	+	+	+	+	+	+			+	+	+	+			+	+	+			+	+			+	17
sylduration	+					+	+	+	+		+			+				+		+			+	+	+	12
absnucleusduration		+	+			+	+				+	+	+	+		+						+		+	+	12
relnucleusduration		+		+	+			+		+	+	+		+	+							+				10
nucleusproportion				+	+			+		+	+	+	+						+		+	+			+	11
Painte_A1-1	+				+			+			+						+					+				6
Painte_A2-1					+			+			+			+	+							+			+	7
Painte_B-1											+															1
Painte_C1-1					+					+			+	+			+	+								6
Painte_C2-1					+					+							+						+		+	5
Painte_D-1				+																					+	2
Painte_SE-1																										
Painte_MSE-1																										
Painte_A1_0	+	+	+	+			+	+	+					+			+				+	+		+	+	13
Painte_A2_0			+	+									+			+	+									5
Painte_B_0																										
Painte_C1_0			+		+	+				+		+	+	+	+			+		+		+	+		+	13
Painte_C2_0	+		+													+		+								4
Painte_D_0																+							+			2
Painte_SE_0																				+						1
Painte_MSE_0	+							+																		2
Painte_A1+1							+		+			+	+				+					+		+		7
Painte_A2+1	+			+			+	+	+	+				+							+	+		+		10
Painte_B+1																										
Painte_C1+1				+							+		+				+							+		5
Painte_C2+1	+			+	+											+			+					+		6
Painte_D+1				+	+		+	+		+								+								6
Painte_SE+1				+						+			+			+									+	5
Painte_MSE+1				+	+				+				+		+											5
medianF0	+									+	+								+		+		+	+		7
relativeMedianF0																							+			1
F0inContext		+									+	+						+				+				5
middleF1																	+								+	2
middleF2			+	+					+	+				+				+				+				7
middleF3				+						+	+										+	+		+		6
middleF4	+		+														+									3
relativeMiddleF1													+													1
relativeMiddleF2			+		+									+		+										4
relativeMiddleF3			+	+		+	+			+		+	+		+							+	+			10
relativeMiddleF4	+		+		+			+	+					+		+				+		+	+	+		11
medianpower															+					+		+	+			4
relativeMiddlePower			+									+														2
sylInWord		+			+				+									+							+	5
relSylInWord	+		+	+		+	+	+		+			+	+	+	+	+		+	+	+	+	+	+		18
numOfSyllablesInWord	+		+					+	+	+						+		+								7
simplePOS	+	+		+		+					+		+					+		+		+				9
POS			+	+	+		+	+		+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	20
wordFrequency		+	+	+	+		+		+	+		+	+	+			+					+		+	+	14
punctuationdistanceleft					+	+			+	+	+		+	+								+				8
punctuationdistanceright			+			+							+								+	+				5
positionbetweenpunctuation			+	+						+												+				4
followingpunctuation								+	+	+					+		+	+				+			+	8
phraseboundarydistanceleft																										
phraseboundarydistance-right		+		+			+						+									+				5
phraseboundarytype																										
positioninphrase																										
features	14	9	18	21	17	9	12	14	12	20	14	11	19	15	9	12	14	13	5	8	9	25	11	13	15	13.6
precision	0.72	0.54	0.64	0.57	0.66	0.80	0.68	0.79	0.64	0.69	0.46	0.49	0.60	0.64	0.77	0.58	0.40	0.87	0.46	0.65	0.63	0.74	0.68	0.63	0.54	0.63
recall	0.85	0.91	0.78	0.91	0.88	0.73	0.90	0.79	0.82	0.88	0.79	0.91	0.86	0.90	0.68	0.73	0.73	0.77	0.67	0.77	0.58	0.80	0.86	0.91	0.90	0.81
f-measure	0.78	0.67	0.71	0.70	0.75	0.76	0.77	0.79	0.72	0.77	0.58	0.63	0.71	0.74	0.72	0.64	0.52	0.82	0.55	0.70	0.60	0.77	0.76	0.74	0.68	0.70

Tabelle B.3. Automatische Merkmalsauswahl für Phrasierungen im KCoRS (J48)

	k01	k02	k03	k04	k05	k06	k63	k64	k65	k66	dlm	hpt	uga	kko	rtd	k07	k08	k09	k10	k11	k12	k67	k68	k69	k70	all
distancetopunctuationleft	+						+	+				+			+								+		+	7
distancetopunctuationright	+	+	+	+	+	+		+		+	+	+	+	+	+	+		+	+		+	+	+		+	20
positionbetweenpunctuation		+				+			+	+			+					+					+	+		8
nextpunctuation					+				+									+							+	4
POSbefore	+			+			+					+				+	+	+		+	+	+			+	11
POSafter		+				+	+						+			+					+		+	+		8
simplePOSbefore		+	+	+						+	+		+	+				+	+			+	+			11
simplePOSafter		+			+			+	+		+															5
wordFreqbefore		+										+			+		+		+		+	+		+		8
WordFreqafter	+			+					+			+			+					+		+	+		+	9
cdgdistance		+						+								+			+							4
cdgleftlabel					+																					1
cdgrightlabel						+		+												+						3
pause		+			+						+		+						+				+			6
silenceDuration				+		+	+																			3
Painte_A1-1					+				+		+							+								4
Painte_A2-1								+												+						2
Painte_B-1		+					+									+					+					4
Painte_C1-1											+						+						+		+	4
Painte_C2-1													+	+							+					3
Painte_D-1			+		+												+									3
Painte_SE-1																										
Painte_MSE-1		+						+				+														3
Painte_A1_0							+			+										+						3
Painte_A2_0					+			+		+								+						+		5
Painte_B_0		+																+				+		+		4
Painte_C1_0												+					+		+	+						4
Painte_C2_0		+		+			+																	+		4
Painte_D_0				+																+						2
Painte_SE_0			+	+						+		+				+			+					+	+	8
Painte_MSE_0		+												+					+							3
Painte_A1+1											+		+				+			+						4
Painte_A2+1	+						+	+			+				+						+					6
Painte_B+1		+								+	+															3
Painte_C1+1														+												1
Painte_C2+1																										
Painte_D+1		+																								1
Painte_SE+1		+		+	+			+		+	+		+	+												8
Painte_MSE+1		+		+	+	+				+													+		+	7
PowerJump																				+						1
RelativePowerJump				+	+	+		+												+						5
finalLenghtening		+	+	+	+			+	+	+		+			+			+				+				11
relativefinalLenghtening	+	+				+	+		+							+	+		+		+					9
finalPowerDecrease								+												+						2
relativefinalPowerDecrease													+													1
linefit150intercept											+			+									+			3
linefit150slope						+		+													+					3
linefit150mse						+	+				+	+				+								+		6
nextlinefit150intercept							+						+		+											3
nextlinefit150slope					+																					1
nextlinefit150mse									+											+				+		3
intercept150difference	+								+		+													+		4
slope150difference				+																						1
linefit200intercept													+										+			2
linefit200slope																				+	+				+	3
linefit200mse	+		+	+		+				+			+										+			7
nextlinefit200intercept								+								+				+	+			+		5
nextlinefit200slope		+		+					+						+					+	+				+	7
nextlinefit200mse									+											+						2
intercept200difference																+	+		+	+						4
slope200difference				+			+																		+	3
linefit2000intercept			+	+			+	+		+		+							+	+	+	+				10
linefit2000slope										+					+											2
linefit2000mse	+															+								+		3
nextlinefit2000intercept								+												+						2
nextlinefit2000slope																+										1
nextlinefit2000mse																										
intercept2000difference							+													+	+					3
slope2000difference	+		+																					+		3
accentuationleftinsyllables			+																							1
accentuationrightinsyllables																			+							1
features	**10**	**20**	**9**	**17**	**13**	**11**	**14**	**17**	**11**	**13**	**13**	**11**	**12**	**7**	**9**	**12**	**8**	**9**	**12**	**20**	**14**	**8**	**12**	**13**	**11**	**12.2**
BP2-prec	**0.00**	**1.00**	**0.00**	**0.25**	**0.00**	**0.00**	**0.50**	**0.00**	**0.00**	**0.00**	**0.81**	**0.51**	**0.70**	**0.36**	**0.57**	**0.00**	**0.17**	**0.00**	**0.50**	**0.67**	**1.00**	**0.00**	**0.33**	**0.00**	**1.00**	**0.33**
BP2-rec	**0.00**	**1.00**	**0.00**	**0.25**	**0.00**	**0.00**	**0.50**	**0.00**	**0.00**	**0.00**	**0.34**	**0.51**	**0.30**	**0.10**	**0.21**	**0.00**	**0.25**	**0.00**	**0.33**	**0.40**	**0.33**	**0.00**	**0.13**	**0.00**	**0.17**	**0.19**
BP2-f	**0.00**	**1.00**	**0.00**	**0.25**	**0.00**	**0.00**	**0.50**	**0.00**	**0.00**	**0.00**	**0.48**	**0.51**	**0.42**	**0.16**	**0.30**	**0.00**	**0.20**	**0.00**	**0.40**	**0.50**	**0.5**	**0.00**	**0.18**	**0.00**	**0.29**	**0.23**

Tabelle B.4. Automatische Merkmalsauswahl für Phrasierungen im KCoRS (NB)

	k01	k02	k03	k04	k05	k06	k63	k64	k65	k66	dlm	hpt	uga	kko	rtd	k07	k08	k09	k10	k11	k12	k67	k68	k69	k70	all
distancetopunctuationleft	+	+		+	+	+			+				+	+							+			+		10
distancetopunctuationright		+						+		+			+	+						+						6
positionbetweenpunctuation							+							+							+					3
nextpunctuation		+				+			+	+		+	+	+		+	+						+	+	+	12
POSbefore	+	+	+	+		+	+	+	+		+	+	+			+		+	+	+		+	+	+	+	19
POSafter				+		+						+														3
simplePOSbefore		+	+		+					+											+					5
simplePOSafter						+																		+		2
wordFreqbefore						+																				1
WordFreqafter												+		+	+										+	4
cdgdistance			+	+												+										3
cdgleftlabel							+														+					2
cdgrightlabel	+			+	+	+			+			+	+	+		+			+							10
pause		+	+				+				+	+	+	+	+			+				+		+	+	12
silenceDuration														+			+			+						3
Painte_A1-1												+	+													2
Painte_A2-1							+																		+	2
Painte_B-1																										
Painte_C1-1		+	+			+					+	+				+					+					7
Painte_C2-1						+					+	+				+		+	+	+	+					8
Painte_D-1			+	+									+								+					4
Painte_SE-1																										
Painte_MSE-1	+															+										2
Painte_A1_0			+		+	+	+					+		+				+								7
Painte_A2_0								+			+			+	+			+						+	+	7
Painte_B_0			+		+		+			+	+	+		+	+		+	+						+	+	12
Painte_C1_0			+	+		+	+	+		+			+	+	+									+		10
Painte_C2_0						+												+				+				3
Painte_D_0		+		+									+								+		+			5
Painte_SE_0	+	+					+			+		+												+		6
Painte_MSE_0		+	+				+	+	+			+		+			+							+	+	10
Painte_A1+1			+		+	+	+						+	+						+						7
Painte_A2+1			+											+									+	+		4
Painte_B+1			+			+		+	+										+	+		+				7
Painte_C1+1	+		+		+			+	+	+	+			+	+					+					+	11
Painte_C2+1		+	+			+										+								+		5
Painte_D+1									+		+			+	+	+										5
Painte_SE+1												+	+	+	+					+						5
Painte_MSE+1													+													1
PowerJump																										
RelativePowerJump									+	+		+		+						+					+	6
finalLenghtening		+	+		+			+				+	+		+		+	+	+	+	+	+	+	+	+	16
relativefinalLenghtening			+											+					+							3
finalPowerDecrease						+	+	+				+		+		+		+	+			+		+		10
relativefinalPowerDecrease				+					+			+			+		+	+			+		+	+	+	10
linefit150intercept								+		+		+		+	+				+							6
linefit150slope										+				+			+		+	+				+		6
linefit150mse		+										+									+					3
nextlinefit150intercept								+						+					+					+		4
nextlinefit150slope	+				+			+											+							4
nextlinefit150mse						+	+	+		+			+	+						+						7
intercept150difference	+		+											+	+											4
slope150difference			+			+								+					+			+				5
linefit200intercept							+						+	+						+						4
linefit200slope			+	+					+		+			+												5
linefit200mse						+						+														2
nextlinefit200intercept	+									+				+			+			+		+		+		7
nextlinefit200slope		+				+		+	+					+												5
nextlinefit200mse						+								+												2
intercept200difference								+		+				+					+							4
slope200difference			+					+		+				+		+			+			+				7
linefit2000intercept																										
linefit2000slope								+						+		+										3
linefit2000mse						+						+		+			+			+						5
nextlinefit2000intercept			+																	+		+				3
nextlinefit2000slope			+	+								+							+	+		+		+		7
nextlinefit2000mse		+	+			+	+				+	+		+	+	+		+			+		+			12
intercept2000difference					+									+	+			+							+	+	6
slope2000difference		+					+																			+	3
accentuationleftinsyllables								+	+										+							+	4
accentuationrightinsyllables			+											+	+								+	+		5
features	**9**	**16**	**25**	**11**	**10**	**23**	**16**	**18**	**13**	**14**	**10**	**24**	**16**	**39**	**15**	**13**	**9**	**12**	**16**	**17**	**12**	**11**	**8**	**21**	**15**	**15.7**
BP2-prec	**0.00**	**0.00**	**0.14**	**0.17**	**0.00**	**0.00**	**0.25**	**0.00**	**0.14**	**0.10**	**0.33**	**0.26**	**0.25**	**0.15**	**0.21**	**0.17**	**0.19**	**0.00**	**0.00**	**0.33**	**1.00**	**0.00**	**0.47**	**0.00**	**0.27**	**0.18**
BP2-rec	**0.00**	**0.00**	**1.00**	**0.25**	**0.00**	**0.00**	**1.00**	**0.00**	**1.00**	**1.00**	**0.53**	**0.79**	**0.72**	**0.66**	**0.45**	**1.00**	**0.75**	**0.00**	**0.00**	**0.60**	**0.33**	**0.00**	**1.00**	**0.00**	**0.50**	**0.46**
BP2-f	**0.00**	**0.00**	**0.25**	**0.20**	**0.00**	**0.00**	**0.40**	**0.00**	**0.25**	**0.18**	**0.41**	**0.40**	**0.37**	**0.25**	**0.29**	**0.29**	**0.30**	**0.00**	**0.00**	**0.43**	**0.50**	**0.00**	**0.64**	**0.00**	**0.35**	**0.22**

Tabelle B.5. Automatische Merkmalsauswahl für Akzentuierungen in den übrigen Korpora

	j48				**nb**			
	ibm	**ims**	**vpsc**	**all**	**ibm**	**ims**	**vpsc**	**all**
nucleus	+	+	+	3	+	+	+	3
sylduration	+		+	2	+	+	+	3
absnucleusduration			+	1				
relnucleusduration		+		1	+		+	2
nucleusproportion			+	1				
Painte_A1-1								
Painte_A2-1	+		+	2	+			1
Painte_B-1	+			1				
Painte_C1-1							+	1
Painte_C2-1						+		1
Painte_D-1					+		+	2
Painte_SE-1			+	1				
Painte_MSE-1								
Painte_A1_0								
Painte_A2_0								
Painte_B_0	+	+	+	3				
Painte_C1_0	+			1	+	+	+	3
Painte_C2_0					+			1
Painte_D_0	+			1				
Painte_SE_0								
Painte_MSE_0	+			1		+		1
Painte_A1+1	+			1				
Painte_A2+1								
Painte_B+1	+			1				
Painte_C1+1								
Painte_C2+1								
Painte_D+1						+		1
Painte_SE+1							+	1
Painte_MSE+1								
medianF0		+	+	2			+	1
relativeMedianF0		+	+	2		+		1
F0inContext								
middleF1	+			1			+	1
middleF2		+		1	+		+	2
middleF3	+			1				
middleF4								
relativeMiddleF1	+		+	2				
relativeMiddleF2			+	1		+		1
relativeMiddleF3		+		1		+		1
relativeMiddleF4			+	1				
medianpower								
relativeMiddlePower								
sylInWord	+		+	2	+	+		2
relSylInWord	+	+		2	+		+	2
numOfSyllablesInWord	+	+	+	3				
simplePOS	+	+		2				
POS	+		+	2	+	+	+	3
wordFrequency								
punctuationdistanceleft		+		1				
punctuationdistanceright	+	+		2	+			1
positionbetweenpunctuation		+		1				
followingpunctuation			+	1	+	+		2
phraseboundarydistanceleft								
phraseboundarydistanceright								
phraseboundarytype								
positioninphrase								
features	**19**	**13**	**16**	**16**	**13**	**12**	**12**	**12**
precision	**0.62**	**0.53**	**0.69**	**0.61**	**0.48**	**0.56**	**0.66**	**0.57**
recall	**0.67**	**0.77**	**0.69**	**0.71**	**0.71**	**0.63**	**0.45**	**0.60**
f-measure	**0.64**	**0.63**	**0.69**	**0.65**	**0.58**	**0.59**	**0.54**	**0.57**

Tabelle B.6. Automatische Merkmalsauswahl für Phrasierungen in den übrigen Korpora

	j48				nb			
	ibm	**ims**	**vpsc**	**all**	**ibm**	**ims**	**vpsc**	**all**
distancetopunctuationleft		+		1			+	1
distancetopunctuationright		+	+	2		+		1
positionbetweenpunctuation		+		1		+	+	2
nextpunctuation		+	+	2		+	+	2
POSbefore	+	+	+	3	+	+	+	3
POSafter	+	+	+	3	+		+	2
simplePOSbefore								
simplePOSafter			+	1		+		1
wordFreqbefore								
WordFreqafter							+	1
cdgdistance	+	+	+	3	+			1
cdgleftlabel							+	1
cdgrightlabel						+	+	2
pause	+	+	+	3	+			1
silenceDuration		+	+	2			+	1
Painte_A1-1	+			1				
Painte_A2-1								
Painte_B-1							+	1
Painte_C1-1	+			1		+		1
Painte_C2-1	+			1			+	1
Painte_D-1	+			1				
Painte_SE-1						+	+	2
Painte_MSE-1								
Painte_A1_0								
Painte_A2_0	+		+	2	+			1
Painte_B_0								
Painte_C1_0						+	+	2
Painte_C2_0								
Painte_D_0							+	1
Painte_SE_0							+	1
Painte_MSE_0	+			1				
Painte_A1+1	+			1				
Painte_A2+1								
Painte_B+1								
Painte_C1+1	+		+	2		+		1
Painte_C2+1	+			1				
Painte_D+1								
Painte_SE+1	+			1			+	1
Painte_MSE+1								
PowerJump								
RelativePowerJump								
finalLenghtening					+	+	+	3
relativefinalLenghtening								
finalPowerDecrease								
relativefinalPowerDecrease								
linefit150intercept								
linefit150slope					+	+		2
linefit150mse								
nextlinefit150intercept								
nextlinefit150slope		+		1				
nextlinefit150mse								
intercept150difference								
slope150difference								
linefit200intercept					+			1
linefit200slope								
linefit200mse							+	1
nextlinefit200intercept	+	+		2				
nextlinefit200slope								
nextlinefit200mse								
intercept200difference						+		1
slope200difference	+			1		+		1
linefit2000intercept			+	1				
linefit2000slope								
linefit2000mse					+			1
nextlinefit2000intercept								
nextlinefit2000slope								
nextlinefit2000mse					+			1
intercept2000difference	+			1		+		1
slope2000difference								
accentuationleftinsyllables								
accentuationrightinsyllables								
features	**17**	**11**	**11**	**13**	**10**	**15**	**18**	**14**
BP2-prec	**0.10**	**0.62**	**0.05**	**0.26**	**0.12**	**0.52**	**0.39**	**0.34**
BP2-rec	**0.20**	**0.12**	**0.07**	**0.13**	**0.30**	**0.38**	**0.55**	**0.41**
BP2-f	**0.13**	**0.19**	**0.06**	**0.13**	**0.17**	**0.44**	**0.45**	**0.35**
BP1-prec	**0.17**	**0.24**	**0.27**	**0.22**	**0.40**	**0.29**	**0.22**	**0.30**
BP1-rec	**0.02**	**0.23**	**0.25**	**0.17**	**0.24**	**0.38**	**0.49**	**0.37**
BP1-f	**0.04**	**0.23**	**0.26**	**0.18**	**0.30**	**0.33**	**0.31**	**0.31**

Anhang C. Testäußerungen im Perzeptionstest

Tabelle C.1. Übersicht der bei der Evaluierung benutzten Äußerungen

	Typ	Wortlaut
1.	Dialog: Frage	Könnten sie das bitte wiederholen?
2.	Dialog: Anweisung	Bitte berühren Sie den roten Knopf um das Gespräch zu beenden.
3.	Dialog: Hinweis	Leider läuft der Film „Herr der Ringe“ heute nicht.
4.	Dialog: Hinweis	Das ist der Fußweg vom Schloß zum Friedhof.
5.	Vorlesen: Frage	Sind wir in der Lage, unsere Werkzeuge auch wieder wegzulegen?
6.	Vorlesen: kurzer Satz	Er sah sich im Karstadt-Kaufhaus um.
7.	Vorlesen: kurzer Satz	Fürsorglich nimmt er den Freund in die Arme.
8.	Vorlesen: langer Satz	Unter anderem flog Lufthansa von Friedrichshafen nach Oldenburg.
9.	Vorlesen: langer Satz	Die jüdische Kultur mit der Geige zu identifizieren ist unbegründet und falsch.

Literaturverzeichnis

Atterer, M. (2005). *Experiments on the Prediction of Prosodic Phrasing for German Text-to-Speech Synthesis*. Dissertation. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Batliner, A., J. Buckow, R. Huber, V. Warnke, E. Nöth und H. Niemann (2001). „Boiling Down Prosody for the Classification of Boundaries and Accents in German and English". In: *Proceedings of Eurospeech 2001*. Aalborg, Dänemark. S. 2781-2784.

Breiman, L., J. H. Friedman, R. A. Olshen und C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth, Monterey, Kalifornien.

Brinckmann, C. (2004). *The 'Kiel Corpus of Read Speech' as a Resource for Speech Synthesis*. Diplomarbeit. Universität des Saarlandes, Saarbrücken.

Bußmann, H. (1990). *Lexikon der Sprachwissenschaft*. 2. Ausgabe. Kröner, Stuttgart.

Chomsky, N. und M. Halle (1968). *The Sound Pattern of English*. Harper and Row, New York.

Demberg, V. (2006). *Letter-to-Phoneme Conversion for a German Text-to-Speech System*. Diplomarbeit. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

van Dommelen, W.A. (1992). „Segmentieren und Etikettieren im Kieler PHONDAT-Projekt". In: K. J. Kohler (Hrsg). *Arbeitsberichte des Insituts für Phonetik der Universität Kiel (AIPUK)*. Nr. 26. S. 197–224.

Donovan, R.E. und P. C. Woodland (1999). „A Hidden Markov-Model-Based Trainable Speech Syntheser". In: *Computer Speech and Language*. Jahrgang 13, Heft 3. S. 223–242.

Donovan, R.E., A. Ittycheriah, M. Franz, B. Ramabhadran, E. Eide, M. Viswanathan, R. Bakis, W. Hamza, M. Picheny, P. Gleason, T. Rutherfoord, P. Cox, D. Green, E. Janke, S. Revelin, C. Waast, B. Zeller, C. Günther und J. Kunzmann (2001). „Current Status of the IBM Trainable Speech Synthesis System". In: *Proceedings of the 4th ESCA Tutorial and Research Workshop on Speech Synthesis*. Schottland.

Duda, R. und P. Hart (1973). *Pattern Classification and Scene Analysis*. John Willey, New York.

Eide, E., A. Aaron, R. Bakis, R. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith und M. Viswanathan (2003). „Recent Improvements to the IBM Trainable Speech Synthesis System". In: *Proceedings of the ICASSP 2003*. Band 1. Hong Kong. S. 708–711.

Foth, K. (1999). *Transformationsbasiertes Constraint-Parsing*. Diplomarbeit. Fachbereich Informatik, Universität Hamburg.

Foth, K. (2006). *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*. Technische Dokumentation. Fachbereich Informatik, Universität Hamburg.

Foth, K., I. Schröder und W. Menzel (2000). „A Transformation-based Parsing Technique With Anytime Properties". In: *Proceedings of IWPT-2000*. Trento. S. 89–100.

Foth, K., St. Hamerich, I. Schröder, M. Schulz und T. By (2003). *[X]CDG User Guide*. Version 1.3. Technische Dokumentation. Fachbereich Informatik, Universität Hamburg.

Gibbon, D. (1998). „Intonation in German". In: D. Hirst und A. Di Cristo (Hrsg). *Intonation Systems. A Survey of Twenty Languages*. Cambridge University Press. S. 78–95.

Greenberg, S. (2005). „From Here to Utility". In: W. Barry. W. A. van Dommelen (Hrsg). *The Integration of Phonetic Knowledge in Speech Technology*. Springer, Berlin. S. 107–132.

Grice, M. und S. Baumann (2000). „Deutsche Intonation und GToBI". In: *Linguistische Beiträge*. 181. S. 1–33.

Grice, M. und R. Benzmüller (1998). „Tonal affiliation in German falls and fall-rises". Poster presented at the 5th Conference on Laboratory Phonology. York.

Grice, M., M. Reyelt, R. Benzmüller, J. Mayer und A. Batliner (1996). „Consistency in Transcription and Labelling of German Intonation with GToBI". In: *Proceedings of the Fourth ICSLP*. Band 3. Philadelphia. S. 1716–1719.

Günther, C. (1999). *Prosodie und Sprachproduktion*. Niemeyer, Tübingen.

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

Haase, M. (2000). *Nutzung prosodischer Merkmale zur Satzsegmentierung*. Diplomarbeit. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

IPA, International Phonetic Association (Hrsg) (1999). *Handbook of the International Phonetic Association. A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

IPDS (1994). *The Kiel Corpus of Read Speech*. CD-ROM. Institut für Phonetik und Digitale Sprachverarbeitung (IPDS), Universität Kiel.

John, G.H. (1997). *Enhancements to the data mining process*. Dissertation. Stanford University.

John, G.H. (1995). „Robust Decision Trees: Removing Outliers from Databases". In: *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Montreal. S. 174–179.

Jurafsky, D. und J. H. Martin (2007). *Speech and Language Processing*. 2. Ausgabe. Prentice Hall, New Jersey.

Kohavi, R. (1995). „A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". In: *Proceedings of the IJCAI*. Morgan Kaufmann, San Mateo. S. 1137–1145.

Kohavi, R. und G. H. John (1997). „Wrappers for Feature Subset Selection". In: *Artificial Intelligence*. 97. S. 273–324.

Kohler, K.J. (1983). „Prosodic Boundary Signals in German". In: *Phonetica*. 40. S. 89–134.

Kohler, K.J. (1991). „A model of German intonation". In: K. J. Kohler (Hrsg). *Arbeitsberichte des Insituts für Phonetik der Universität Kiel (AIPUK)*. Nr. 25. S. 295–360.

Kohler, K.J. (1992a). „Erstellung eines Textkorpus für eine phonetische Datenbank des Deutschen". In: K. J. Kohler (Hrsg). *Arbeitsberichte des Insituts für Phonetik der Universität Kiel (AIPUK)*. Nr. 26. S. 11–39.

Kohler, K.J. (1992b). „Prosodisches Transkriptionssystem für die Etikettierung von Sprachsignaldaten". In: K. J. Kohler (Hrsg). *Arbeitsberichte des Insituts für Phonetik der Universität Kiel (AIPUK)*. Nr. 26. S. 239–250.

Kohler, K.J. (1992c). K. J. Kohler (Hrsg). *Arbeitsberichte des Insituts für Phonetik der Universität Kiel (AIPUK)*. Nr. 26. Vorwort. Institut für Phonetik und Digitale Sprachverarbeitung (IPDS), Universität Kiel. S. 7–10.

Kohler, K.J. (1995). „ToBIG and PROLAB: Two prosodic transcription systems for German compared". Report at the Workshop on Prosodic Labelling, ICPhS. Stockholm.

Langley, P., W. Iba und K. Thompson (1992). „An Analysis of Bayesian Classifiers". In: *Proceedings of the National Conference on Artificial Intelligence*. S. 223–228.

Ludwig-Mayerhofer, W. (o. J.). „Auswahl, geschichtete". In: *ILMES. Internet-Lexikon der Methoden der empirischen Sozialforschung*. Online: http://www.lrz-muenchen.de/~wlm/ilmes.htm (Stand: 2007-02-12).

Mayer, J. (1999). „Prosodische Merkmale von Diskursrelationen". In: *Linguistische Berichte*. 177. S. 65–86.

Mengel, A. (2000). *Deutscher Wortakzent: Symbole, Signale*. Books on Demand, Hannover.

Moulines, E. und F. Charpentier (1990). „Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones". In: *Speech Communication*. 9. S. 453–467.

Möhler, G. (1998). *Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese*. Dissertation. Intitut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Möhler, G. (2001). *Improvements of the PaIntE model for F0 parametrization*. Technical Report (Draft Version). Intitut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Möhler, G. und A. Conkie (1998). „Parametric modeling of intonation using vector quantization". In: *SSW3-1998*. ISCA Archive. Blue Mountains, Australien. S. 311–316.

Neppert, J.M. H. (1999). *Elemente einer Akustischen Phonetik*. 4. Ausgabe. Buske, Hamburg.

Neubarth, F., K. Alter, H. Pirker, E. Rieder und H. Trost (2000). „The Vienna Prosodic Speech Corpus: Purpose, Content and Encoding". In: *KONVENS 2000 / Sprachkommunikation, Vorträge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS), 6. ITG-Fachtagung "Sprachkommunikation"*. Ilmenau. S. 191–195.

Nöth, E. (1991). *Prosodische Information in der automatischen Spracherkennung. Berechnung und Anwendung*. Niemeyer, Tübingen.

Peters, B. und K. J. Kohler (2004). *Trainingsmaterialien zur prosodischen Etikettierung mit dem Kieler Intonationsmodell KIM*. Technical Report. Institut für Phonetik und Digitale Sprachverarbeitung (IPDS), Universität Kiel.

Pfitzinger, H.R. (1999). „Local speech rate perception in German speech". In: *Proceedings of ICPhS*. Band 2. San Francisco. S. 893–896.

Pfitzinger, H.R. und U. D. Reichel (2006). „Text-based and Signal-based Prediction of Break Indices and Pause Durations". *Proceedings of ICSP*. Band 1. Dresden. S. 133–136.

Pierrehumbert, J.B. (1980). *The Phonology and Phonetics of English Intonation*. Dissertation. Massachusetts Institue of Technology.

Pétursson, M. (1978). „Akustische und physiologische Aspekte der Junktur in den Konsonantengruppen s+Verschlußlaut. Ein Beitrag zur allgemeinen Problematik der Junktur und zur Junktur im modernen Isländischen". In: *Hamburgische Phonetische Beiträge*. 25. Buske, Hamburg. S. 363–399.

Qin, S. und V. Fischer (2004). „A comparison of statistical methods and features for the prediction of prosodic structures". In: *Proceedings of INTERSPEECH 2004*. Jeju Island, Korea. S. 1877–1880.

Quinlan, R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann, San Francisco.

Reyelt, M. (1995). *Untersuchungen zur Konsistenz prosodischer Etikettierungen*. Verbmobil-Report 77. Institut für Nachrichtentechnik, Technische Universität Braunschweig.

Reyelt, M., M. Grice, R. Benzmüller, J. Mayer und A. Batliner (1996). „Prosodische Etikettierung des Deutschen mit ToBI". In: D. Gibbon. *Natural Language and Speech Technology, Results of the third KONVENS conference*. Mouton de Gruyter, New York. S. 144–155.

Russel, S. und P. Norvig (2003). *Artificial Intelligence. A Modern Approach*. 2. Ausgabe. Prentice Hall, New Jersey.

Schiller, A., S. Teufel und C. Thielen (1995). *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technical Report. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart und Seminar für Sprachwissenschaft, Universität Tübingen.

Schmid, H. (1995). „Improvements in Part-of-Speech Tagging with an Application to German". In: *Proceedings of the ACL SIGDAT-Workshop*. Dublin.

Schröder, M. und S. Breuer (2004). „XML Representation Languages as a Way of Interconnecting TTS Systems". In: *Proceedings of INTERSPEECH 2004*. Jeju Island, Korea. S. 1889-1892.

Schweitzer, A. (1999). *Bestimmung der Intonation mit Hilfe von Wortklassen*. Diplomarbeit. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Schweitzer, A., N. Braunschweiler, T. Klankert, B. Möbius und B. Säuberlich (2003). „Restricted Unlimited Domain Synthesis". In: *Proceedings of Eurospeech 2003*. Band 2. Genf. S. 1321–1324.

Sotscheck, J. (1976a). „Methoden zur Messung der Sprachgüte I: Verfahren zur Bestimmung der Satz- und Wortverständlichkeit“. In: *Der Fernmeldeingenieur*. Jahrgang 30, Heft 10. S. 1–31.

Sotscheck, J. (1976b). „Methoden zur Messung der Sprachgüte II: Verfahren zur Bestimmung der Satz- und Wortverständlichkeit“. In: *Der Fernmeldeingenieur*. Jahrgang 30, Heft 12. S. 1–33.

Spranger, K. (2001). *Zeitliche Alignierung der F0-Kontur als Funktion der Silbenstruktur im Deutschen*. Studienarbeit. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Strom, V. und C. Widera (1996). „What's in the "Pure" Prosody?“. In: *Proceedings of the Fourth ICSLP*. Band 3. Philadelphia. S. 1497–1500.

Tamburini, F. (2003). „Automatic Prosodic Prominence Detection in Speech using acoustic Features. An Unsupervised System“. In: *Proceedings of Eurospeech 2003*. Band 1. Genf. S. 129–132.

Taylor, P. (1998). „The TILT Intonation Model“. In: *Proceedings of the ICSLP 1998*. Band 4. Sidney. S. 1383-1386.

Taylor, P., A. Black und R. Caley (1998). „The Architecture of the Festival Speech Synthesis System“. In: *Proceedings of the Third ESCA Workshop in Speech Synthesis*. Jenolan Caves, Australia. S. 147–151.

Thon, W. (1992). „Struktur eines Datenverarbeitungssystems für das Kieler PHONDAT-Projekt: Von der Aufnahme ASL-PHONDAT 92 zur Datenanalyse“. In: K. J. Kohler (Hrsg). *Arbeitsberichte des Insituts für Phonetik der Universität Kiel (AIPUK)*. Nr. 26. S. 111–174.

Thon, W. und W. A. van Dommelen (1992). „PHONDAT 90: Rechnerverarbeitbare Sprachaufnahmen eines umfangreichen Korpus des Deutschen“. In: K. J. Kohler (Hrsg). *Arbeitsberichte des Insituts für Phonetik der Universität Kiel (AIPUK)*. Nr. 26. S. 41–80.

Wahlster, W., N. Reithinger und A. Blocher (2001). „SmartKom: Multimodal Communication with a Life-Like Character“. In: *Proceedings of Eurospeech 2001*. Aalborg, Dänemark. S. 1547-1550.

Wells, J.C. (o. J.). *SAMPA computer readable phonetic alphabet*. Online: http://www.phon.ucl.ac.uk/home/sampa/ (Stand: 2007-03-21).

Witten, I.H. und E. Frank (2000). *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.

Zehnpfenning, A. (2005). *Verbesserung der Intonation bei der Sprachsynthese durch datengetriebene Einheitenauswahl*. Diplomarbeit. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.