gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil
Verbundvorhaben

Außergrammatische Phänomene in der Spontansprache: Gegenstandsbereich, Beschreibung, Merkmalinventar

A. Batliner, S. Burger, A. Kießling

L.M.-Universität München

F.–A.–Universität Erlangen–Nürnberg

Vm

Report 57
Februar 1994

Februar 1994

A. Batliner, S. Burger, A. Kießling

Institut für Deutsche Philologie
Ludwig–Maximilian Universität München
Schellingstr. 3
D–80799 München

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

Tel.: (089) 2180 - 2916
e-mail: `ue102ac@cd1.lrz-muenchen.de`

Gehört zum Antragsabschnitt: 3.11, 3.12, 6.4

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 C 6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Institut für Deutsche Philologie
Ludwig-Maximilians-Universität

Prof. Dr. H. Altmann

Schellingstr. 3
D–80799 München 40
(089) 2180–2916

Verbundprojekt VERBMOBIL

Entwicklung eines mobilen Systems
zur
Übersetzung von
Verhandlungsdialogen
in face-to-face Situationen

Lehrstuhl für Mustererkennung
(Inf. 5)
Universität Erlangen-Nürnberg

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

Martensstr. 3
D–91058 Erlangen
(09131) 85–7774

Außergrammatische Phänomene in der Spontansprache

Gegenstandsbereich, Beschreibung, Merkmalinventar

Anton Batliner, Susanne Burger, Andreas Kießling

Verbmobil Report Nr. 57, Feb. 94, LMU (V1.0)

Februar 1994

Gehört zum Antragsabschnitt: 3.12 Spontansprachliche Phänomene
6.4 Syntax und Satzprosodie

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter den Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 F/4 gefördert. Die Verantwortung für den Inhalt dieses Berichts liegt bei den Autoren.

Inhaltsverzeichnis

Vorbemerkung

Im vorliegenden Beitrag finden sich eine Beschreibung außergrammatischer Phänomene sowie ein detaillierter Vorschlag zur Labelung dieser Phänomene. Dieser Vorschlag setzt auf dem aktuellen Stand der VERBMOBIL-Basistransliteration auf, vgl. [IPD94]. Einige der von uns hier vorgeschlagenen Modifikationen können und sollten u.E. allerdings auch in die Basistransliteration übernommen werden. Unsere Labelsymbole sind anhand des aktuellen Stands der Basistransliteration so gewählt, daß sie eindeutig sind. Alle Labelsymbole sowie die generelle Strategie (Detailliertheit der Labelung, Analysetiefe) sollten natürlich noch eingehend diskutiert und im Verlaufe dieser Diskussion eventuell auch modifiziert werden. Diese Diskussion muß im Zusammenhang mit einer Diskussion zur Basistransliteration und zur prosodischen Labelung erfolgen; insbesondere sollten noch weitere Anforderungen aus der Anwendung spezifiziert werden. Für diesbezügliche Anmerkungen, Anregungen und Kritik sind wir dankbar.

Es wird gezeigt, daß mit diesem Labelsystem auch sehr komplexe außergrammatische Passagen bearbeitet werden können. Die Entscheidungen darüber, ob und in welchem Ausmaß solche komplexen Passagen im Szenario überhaupt zugelassen werden sollen, ob eine Verarbeitung versucht werden soll oder ob bei komplexeren Passagen das System mit Rückweisung und der Bitte um Wiederholung antworten soll, bleiben davon natürlich unberührt und müssen auf anderen Ebenen getroffen werden.

$$\textit{Habe nun} \left\{ \begin{array}{c} \textit{ach}! \\ \text{und das will heute schon was heißen} \\ \text{verflixt nochmal} \\ \text{ähm} \end{array} \right\} \textit{Philosophie},$$

{äh,} *Juristerei und Medizin,*
Und leider auch Theologie!
Durchaus studiert, mit heißem Bemühn.

1 Einleitung

Im VERBMOBIL-Arbeitspaket 6.4: *"Syntax und Satzprosodie"* geht es u.a. um die prosodische Form von Parenthesen, im Arbeitspaket 3.12: *"Behandlung spontansprachlicher Phänomene auf Äußerungsebene"* geht es zum einen um die Erfassung der prosodisch-rhythmischen Gesamtstruktur der Spontansprache im Gegensatz zur Lesesprache, zum anderen um die Erfassung und Behandlung der Prosodie "inkorrekter", aber sehr häufiger außergrammatischer Phänomene wie Verzögerungen (Häsitationen) und Versprecher mit anschließender Korrektur. Ziel ist dabei eine möglichst gute Erkennung dieser Phänomene, die bei der Worterkennung und dem Parsen Schwierigkeiten bereiten. Man beachte, daß in diesem Arbeitspaket nicht Phänomene bearbeitet werden, die prosodisch unauffällig bleiben, da sie sich ausschließlich in einer falschen oder in der Schriftsprache unüblichen Lexik, Syntax oder Semantik manifestieren; vgl. dazu im einzelnen [Tro94]. Diese Phänomene können nur in den anderen, jeweils zuständigen Modulen verarbeitet werden. Das im folgenden beschriebene Inventar von Phänomenen und Merkmalen wurde erstellt aufgrund der einschlägigen Literatur (s. Kapitel 7) sowie anhand eines eigenen spontansprachlichen Korpus mit Klärungsdialogen (vgl. Anhang 1).

An der wohl berühmtesten Interjektion der deutschen Literatur, vgl. das *"ach"* aus Goethes Faust oben in kursiv, sowie an möglichen paradigmatischen und syntagmatischen Alternativen, vgl. die nicht-kursiven Einfügungen im Zitat, soll die Problematik kurz skizziert werden:

- An der selben Position im Satz können "reguläre, grammatische" Phänomene (Parenthesen: *"und das will heute schon was heißen"*, Interjektionen: *"ach"* und Flüche: *"verflixt nochmal"*) sowie außergrammatische Phänomene (Verzögerungssignale wie *"ähm"*) auftreten. Eine **ausschließliche** Beschäftigung mit außergrammatischen Phänomenen erscheint uns also wenig sinnvoll.
- Ohne prosodische Information ist es nicht klar, ob es sich in der vorletzten Zeile des Zitats bei dem *"äh"* um eine Häsitation oder um eine Korrektur handelt: entweder zögert Faust nach *"Philosophie"*, weil ihm *"Juristerei"* nicht gleich einfällt, oder er hat

"Philosophie" gar nicht studiert und korrigiert zu *"Juristerei"*. (Bei einer Korrektur ist im Normalfall das korrigierende Wort — hier also *"Juristerei"* — ähnlich wie beim Kontrastakzent prosodisch stärker hervorgehoben, bei einer reinen Verzögerung bleibt dieses Wort jedoch prosodisch unauffällig.) Es ist also nötig, bei der Verarbeitung dieser Phänomene prosodische Information mit zu berücksichtigen.

- Die Terminologie ist nicht konsistent: Flüche z.B. gelten als "sekundäre" Interjektionen, *"ach"* und *"äh"* als (primäre) Interjektionen, obwohl die Funktion von *"äh"* sowohl als Korrektureinleiter als auch als Verzögerungssignal hauptsächlich diskurssteuernd ist, die von *"ach"* aber auch, wie bei Faust, Bedeutung transportieren kann; vgl. dazu auch [Wil88, S. 65ff]. Es wird daher im folgenden versucht, eine in sich konsistente Terminologie zu entwickeln.

In dieser Arbeit versuchen wir zuerst, aus einem anwendungsbezogenen Blickwinkel heraus einen Überblick über die einschlägigen Phänomene zu geben. Auf einen ausführlichen Literaturbericht wird dabei verzichtet und dafür auf die im Literaturverzeichnis angeführten Arbeiten verwiesen. Das Merkmalsystem wird — in modifizierter Form — teilweise von [BDSP93] übernommen und an die Konventionen der VERBMOBIL-Basistransliteration angepaßt; vgl. dazu [IPD94] und Anhang 2. Am Ende stehen Überlegungen darüber, wie die automatische Verarbeitung in einem spracherkennenden und -verstehenden System im einzelnen aussehen könnte und welcher Komplexitätsgrad der außergrammatischen Phänomene sinnvollerweise in einer ersten Phase zugelassen werden sollte.

Domänenspezifisches Material lag zum Zeitpunkt der Untersuchungen noch nicht in ausreichender Menge vor. Darüberhinaus besteht die Möglichkeit, daß sich bei einer eingeschränkten Domäne nur eine Teilmenge der möglichen außergrammatischen Phänomene beobachten läßt und daher die Beschreibungssprache zu wenig mächtig wird, um auch nicht beobachtete Phänomene, die aber in anderen Szenarien durchaus auftreten, modellieren zu können. Als Materialgrundlage dienen daher ein eigenes Korpus mit spontansprachlichen Äußerungen, in dem sich sehr komplexe außergrammatische Passagen finden, einige VERBMOBIL-Dialoge, sowie konstruierte Äußerungen. Das beschriebene Merkmalsystem ist dabei nicht so zu verstehen, daß es von vornherein in dieser elaborierten und damit auch zeitaufwendigen Form eingesetzt werden muß; es dient fürs erste dazu, zu demonstrieren, daß es grundsätzlich eine Beschreibungssprache auch für sehr komplexe außergrammatische spontansprachliche Phänomene gibt. Konzeptuell sinnvoll ist dabei sicherlich, von einer möglichst detaillierten, spezialisierten Beschreibung auszugehen, da eine nachträgliche Generalisierung automatisch erfolgen kann, wogegen bei einer umgekehrten Vorgehensweise eine nachträgliche Spezialisierung auch für einen menschlichen Bearbeiter einen großen Aufwand bedeutet und nicht mehr automatisiert werden kann.

2 Die Unterbrechung des normalen Satzgefüges

In der bisherigen Forschung wurden meist nur grammatisch wohlgeformte Äußerungen betrachtet, wie sie in elizitierter Lesesprache vorkommen. Dies gilt gleichermaßen für Linguistik, Phonetik sowie für spracherkennende und sprachverstehende Systeme. Anwendungen müssen aber auch spontansprachliche Äußerungen verarbeiten können, die oft außergrammatische Phänomene enthalten. Neben dieser Einteilung in **grammatische** vs. **außergrammatische** Phänomene[1] kann eine andere aufgestellt werden, nämlich die in **"grammatisches Satzgefüge"** vs. **"Unterbrechung des grammatischen Satzgefüges"**. "Unterbrechung" sagt fürs erste noch nichts darüber aus, ob danach in der Konstruktion weitergefahren oder ob eine neue Konstruktion (Neuansatz) begonnen wird. Normalerweise ist eine Unterbrechung durch mindestens ein zusätzliches Element — und sei dies "nur" eine ungefüllte Pause — gekennzeichnet. Nur beim Anakoluth fehlt dieses Element, vgl. das Beispiel aus [SB91, S. 56f] *"Können Sie mir sagen, wieviel ich für eine Fahrkarte nach Hamburg kostet?"* Solche Fälle, bei denen eine normalerweise vorhandene Markierung auf einer Ebene fehlt, sind durchaus üblich, vgl. das sog. Nullmorph.

Nimmt man den Begriff "Parenthese" in seiner wörtlichen Bedeutung als "Einfügung", so kann er als Oberbegriff für all die Phänomene verwendet werden, die den normalen Ablauf der Rede unterbrechen. Dazu gehören zum einen grammatische Phänomene wie die Parenthesen in ihrer üblichen Bedeutung als vom übrigen Satzgefüge strukturell unabhängige, eingefügte selbständige Ausdrücke; im weiteren Sinne ist aber alles satzwertige, also auch eine Interjektion, eine "Parenthese" im Sinne einer Einfügung. Zum anderen gehören jene außergrammatischen Phänomene dazu, die ebenfalls Unterbrechungen darstellen: Zögerungsphänomene, im prototypischen Fall ohne Auswirkung auf das übrige Satzgefüge, können bei der Verarbeitung einfach ausgeblendet werden. So ergibt sich etwa bei der Äußerung *"Ich will nach - äh - München."* nach der Ausblendung von *"äh"* ein grammatisch korrekter Satz. Die Verarbeitung von Korrekturen und Abbrüchen mit mehr oder weniger großer Auswirkung auf das übrige Satzgefüge ist dagegen komplizierter, da sie zum einen ausgeblendet werden müssen, zum anderen oft die Analyse neu aufsetzen muß. Bei der Äußerung *"Ich will nach München - äh - nach Ulm."* etwa ergibt sich nach der Ausblendung von *"nach München - äh"* ebenfalls ein grammatisch korrekter Satz; zu komplexen Korrekturen und Syntaxabbrüchen vgl. Abschnitt 4.6. Man beachte, daß mit einer einfachen Ausblendung allerdings nicht alle Fälle adäquat behandelt werden können, vgl. etwa die Äußerung:

"Gestern hab ich mit dem Willi- , äh, er hat gemeint, daß ..."

Wenn hier die abgebrochene Passage inklusive Verzögerungsmarkierer *"Gestern hab ich mit dem Willi- , äh"* voll ausgeblendet wird, dann erhält man ein anaphorisches Pronomen *"er"* ohne Antezedens; die Syntax ist dann zwar korrekt, eine Anapherauflösung aber nicht mehr möglich. (Aus diesem Grund wird weiter unten ein eigenes Label für die Beziehung

[1] Man beachte, daß hier unter dem Begriff *'Grammatik'* keine formale Grammatik im Sinne eines expliziten Regelwerkes verstanden wird, sondern eine Grammatik im Sinne einer beschreibenden Grammatik einer Einzelsprache, die sich normalerweise fast ausschließlich mit 'korrekten', wohlgeformten Äußerungen beschäftigt.

von Anapher und Antezedens eingeführt.)

Im Gegensatz zu [BDSP93], die nur "self-repairs", also Korrekturen, modellieren, sollen im weiteren auch Verzögerungsphänomene und "korrekte" Unterbrechungen wie Parenthesen berücksichtigt werden. Sinnvollerweise dürfte nämlich zumindest die anfängliche Verarbeitung (Extraktion prosodischer Merkmale, Klassifikation zur gegenseitigen Abgrenzung, etc.) sehr ähnlich sein. So finden sich auch bei Parenthesen die ansonsten für Korrekturen typischen Phänomene (Neuanfang des Trägersatzes, Wortwiederholungen) sowie Konstruktionsabbrüche. Analoges gilt für Interjektionen und insbesondere für eine Teilmenge davon, nämlich die der diskurssteuernden Gliederungspartikeln, vgl. dazu [Wil88] und die Beispiele weiter unten.

Postulierte Regularitäten bei solchen "irregulären" spontansprachlichen Phänomenen dürften beim jetzigen Kenntnisstand noch sehr korpus- (diskurs-, sprecher-) abhängig sein. Es kann also noch nicht das Ziel sein, eine komplette Taxonomie oder gar eine "formale Grammatik ungrammatischer Phänomene" aufzustellen; es geht uns hier vielmehr darum, eine möglichst einfache, allerdings gleichzeitig auch umfassende Beschreibungssprache zu entwickeln, mit der alle einschlägigen Phänomene erfaßt werden können. Eine Beschränkung auf genügend häufige und verarbeitbare Phänomene sollte erst in einem zweiten Schritt erfolgen.

3 Die Phänomene im Einzelnen

Die Beschreibung erfolgt auf zwei Ebenen, einer segmentalen (Wort-) Ebene und einer suprasegmentalen prosodischen Ebene. Die Wortebene kann als kategorial aufgefaßt werden, d.h. die Phänomene sind entweder vorhanden oder nicht vorhanden. Die prosodische Ebene hat kontinuierliche Ausprägungen, ihre Merkmale sind die auch sonst üblichen Merkmale F0-Verlauf, Dehnung und Intensität.[2] Das Beschreibungssystem trennt allerdings nicht strikt zwischen diesen beiden Ebenen. Dies entspricht auch der angestrebten Verarbeitung in einem Prosodiemodul, das zumindest teilweise auf der Worterkennung aufbaut, vgl. dazu Kapitel 5.

Im folgenden werden zuerst die Phänomene kurz beschrieben und einige Hypothesen über ihre (prosodische) Form erwähnt. Im nächsten Kapitel wird das Beschreibungssystem erklärt und mit Beispielen illustriert. Tabelle 1 gibt einen Überblick über die zu behandelnden Phänomene. Bei der Unterbrechung des grammatischen Satzgefüges unterscheiden wir zwischen **grammatischen** und **außergrammatischen** Phänomenen. Grammatische Phänomene sind **Parenthesen** und **Interjektionen**, außergrammatische Phänomene sind

[2] Die zu extrahierenden Merkmale dürften denen entsprechen, die bei der Bestimmung von Phrasengrenzen eingesetzt werden können, vgl. dazu [BKK*93]; dort wurde versucht, möglichst alle potentiell relevanten (kontinuierlichen) Merkmale aufzunehmen. Eine zusätzliche prosodisch/phonologische Ebene mit kategorialen Einheiten ist unserer Erfahrung nach nicht nötig; sie dürfte sogar zu einer Verschlechterung der Klassifikationsergebnisse führen, die auf den immer bei einer Quantisierung auftretenden Quantisierungsfehler zurückzuführen ist.

Korrekturen, **Syntaxabbrüche** und **Verzögerungen** (Häsitationen). Die Merkmale, mit denen diese Phänomene – die sich auf der "funktionalen Ebene" unterscheiden – markiert sind, werden auf der "terminalen Ebene" aufgeführt. Diese Merkmale sind zum Teil bei allen Phänomenen die gleichen: Dehnung, Wiederholung, leere (ungefüllte) Sprechpause sowie Füllwort (als gefüllte Pause). Keines der Merkmale **muß** vorhanden sein, sie **können** aber auch **mehrmals** vorhanden sein. Die Dehnung dürfte im Normalfall zwar nur einmal vorkommen, als finale Dehnung direkt vor der Unterbrechung. Man kann sich aber durchaus vorstellen, daß etwa sowohl das letzte als auch das vorletzte Wort vor der Unterbrechung gedehnt sind. Bei den Korrekturen finden sich **Einfügungen**, **Ersetzungen** und **Eliminierungen**. Abbrüche innerhalb eines Wortes (**Wortabbruch**) bzw. innerhalb einer Konstituente (**Phrasenabbruch**) finden sich gleichermaßen bei Korrekturen und Syntaxabbrüchen. Syntaxabbrüche sind per definitionem dadurch gekennzeichnet, daß nach der Unterbrechung ein **Neuansatz** folgt. Bei Parenthesen kann alles stehen; zu Einfügungen bzw. Eliminierungen bei Verzögerungen vgl. 4.3.4, zu Wortabbrüchen bei Verzögerungen vgl. 4.3.5.

<table>
<tr><td></td><td>"grammatische"
Phänomene</td><td colspan="3">"außergrammatische",
spontansprachliche Phänomene</td></tr>
<tr><td>Funktionale
Ebene</td><td>Parenthese
(inkl. Interjektion)</td><td>Korrektur</td><td>Syntaxabbruch</td><td>Verzögerung</td></tr>
<tr><td rowspan="6">Merkmale
der
terminalen
Ebene</td><td colspan="4">Dehnung, Markierung durch F0 und Energie
Wiederholung
leere (ungefüllte) Sprechpause</td></tr>
<tr><td colspan="4">Füllwort = gefüllte Pause</td></tr>
<tr><td colspan="3">oder Korrektureinleiter</td><td></td></tr>
<tr><td colspan="2">Einfügung
Ersetzung
Eliminierung</td><td></td><td>Einfügung

Eliminierung</td></tr>
<tr><td colspan="3">Wortabbruch, Phrasenabbruch</td><td rowspan="2">Wortabbruch</td></tr>
<tr><td>Satzabbruch
Neuansatz</td><td></td><td>Satzabbruch
Neuansatz</td></tr>
</table>

Tabelle 1: Unterbrechung des Satzgefüges

Die Unterscheidung zwischen Korrektur und Syntaxabbruch ist nicht immer von vornherein eindeutig, kann aber an prototypischen Beispielen leicht nachvollzogen werden; vgl. dazu Abschnitt 4.5. Interjektionen werden im folgenden unter Parenthesen subsumiert, da sie sich ähnlich wie kurze, formelhafte Parenthesen verhalten. Im Rahmen der Worterkennung können für beide Phänomene, ebenso wie für die Füllwörter, Wortmodelle aufgestellt werden, wogegen komplexere (längere, satzförmige) Parenthesen nicht als Ganzes modelliert, sondern als Kette von Wortmodellen dem Parser übergeben werden.

All diese Phänomene können i.a. an der gleichen Position auftreten, und ein Neuansatz kann auch nach einer Parenthese/Interjektion stehen, vgl. oben das einleitende Faustzitat bzw. die folgenden Kombinationsmöglichkeiten:

$$Das\ ist \left\{ \begin{array}{c} zumindest\ vermute\ ich\ das \\ Glaub'\ ich \\ leider \\ Teufel\ auch \\ \ddot{a}hm \end{array} \right\} \left\{ \begin{array}{c} eine\ schwierige\ Aufgabe \\ da\ haben\ wir\ uns\ was\ eingebrockt \end{array} \right\}$$

Man beachte, daß im Augenblick nur interessiert, welche Kombinationen grundsätzlich vorkommen können, nicht jedoch die Wahrscheinlichkeit der einzelnen Kombinationen. So mag etwa die Kombination: Abbruch, gefüllte Pause, Parenthese mit eingebetteter gefüllter Pause, Neuansatz sehr selten vorkommen; sie ist aber nicht unmöglich. (Vgl. zu solchen statistischen Untersuchungen etwa [Lev83].)

Verzögerung, Korrektur und Syntaxabbruch mit folgendem Neuansatz sind hauptsächlich **funktionale** Begriffe, Parenthese ist fürs erste ein **formaler** Begriff. Wie sich aber z.B. häufig bei Talkshows beobachten läßt, kann eine Parenthese auch die gleiche Funktion wie eine Verzögerung – nämlich die Verhinderung der Turn-Übergabe – besitzen, etwa bei: *"Das ist - und das muß endlich einmal mit aller Deutlichkeit gesagt werden - ..."*. Wir vermuten daher, daß in vielen Fällen zwar eindeutig entschieden werden kann, ob nun eine Parenthese, eine Verzögerung, eine Korrektur oder ein Neuansatz vorliegt. Letztlich gibt es aber keine definierenden Merkmale, sondern mehr oder weniger prototypische Formen der Verteilung. Die terminalen Elemente in Tabelle 1 (Dehnung, Pause, etc.) sind grundsätzlich formale Merkmale, wobei allerdings bei den Füllwörtern noch funktional unterschieden wird zwischen gefüllter Pause als Verzögerungssignal oder als Korrektureinleiter. Diese zusätzliche Unterscheidung läßt sich damit begründen, daß beide Arten oft unterschiedlich prosodisch – und damit formal – markiert sind und auch prosodisch unterschiedlich modelliert werden sollten, da die weitere syntaktische Verarbeitung auch unterschiedlich ausfällt. Bei Parenthesen können grundsätzlich alle Merkmale auftreten, oft sind aber nur (wenige) häsitationsspezifische vorhanden.

In Tabelle 2 wird der zeitliche Ablauf schematisch am Beispiel der Korrektur wiedergegeben. Die Darstellung orientiert sich zum Teil an [Lev83] und an [HN93]; sie ist wieder so zu verstehen, daß keines der aufgeführten Merkmale vorhanden sein **muß**, aber viele oder sogar alle, auch mehrmals, vorhanden sein **können**. Grundsätzlich gibt es drei Phasen: die 1. Phase **vor** der Unterbrechung, die 2. Phase als eigentliche Unterbrechungsphase und die 3. Phase **nach** der Unterbrechung. Im zeitlichen Ablauf ist bei der Korrektur die 1. Phase links von der Unterbrechung das Reparandum, und die 3. Phase rechts von der Unterbrechung die Korrektur. Bei Verzögerungen[3] tritt nur ein Teil der Elemente auf: Es gibt natürlich kein Reparandum, allerdings kann oft ebenfalls eine finale Dehnung beobachtet

[3]Verzögerungen sind hierarchisch auf der "tiefsten" Ebene; sie können bei allen anderen Phänomenen auftreten; vgl. dazu auch Tabelle 4

werden; zu Einfügungen bzw. Eliminierungen bei Verzögerungen vgl. 4.3.4. Analog verhalten sich Parenthesen und Interjektionen. Es gibt meist keine Korrekturphase, sehr wohl aber auch Wiederholungen, etwa: *"Das ist – und das setze ich mal voraus – das ist ..."*. Auch ein Neuansatz ist möglich: *"Habt Ihr – weil Du ganz zu Beginn dieser Runde sagtest, Optimismus ist da nicht viel mehr da – gibt es bei Euch noch diesen Hoffnungsschimmer?"* Die möglichen prosodischen Merkmale sind für alle drei Phänomene die gleichen, die Merkmale dürften aber je nach Phänomen unterschiedliche Ausprägungsgrade besitzen. Beim Neuansatz sind alle Merkmale vor und während der Unterbrechung möglich, es findet sich aber keines der anderen Merkmale auf der Wortebene danach (Wiederholung, Einfügung, Ersetzung).[4] Auch prosodisch dürfte der Neuansatz eher unauffällig bleiben; diese Annahme sollte allerdings noch empirisch bestätigt werden.

Normalerweise dürften Füllwort, Parenthese und Interjektion bei der Unterbrechung alternativ vorkommen, das muß aber nicht immer der Fall sein: Füllwörter können z.B. Parenthesen einleiten oder in sie eingebettet sein:

"Das ist – ähm, und darauf hab' ich schon gestern verwiesen – in diesem Zeitraum nicht möglich"

"Das ist – und darauf – das muß man mal mit aller Deutlichkeit sagen – hab' ich schon gestern verwiesen – in diesem Zeitraum nicht möglich"

Einbettungen müssen also rekursiv verarbeitet werden, vgl. auch Abschnitt 4.6.

	1. Phase Reparandum	2. Phase Unterbrechung	3. Phase Korrektur
Wortebene	Abbruch Eliminierung	Füllwort	Wiederholung Einfügung Ersetzung
Prosodie	Dehnung F0, Energie	leere Pause F0, Energie	F0, Energie

Tabelle 2: Der zeitliche Ablauf am Beispiel "Korrektur"

Füllwörter finden sich oft, aber nicht immer, an Konstituentengrenzen. Sie haben zwei unterschiedliche Funktionen: zum einen sind sie **Verzögerungssignal** (Häsitation), zum anderen **Korrektureinleiter**. In beiden Funktionen haben sie keine Eigenbedeutung, sind aber **diskurssteuernde Gliederungspartikeln**. Typische Häsitationen sind: *"äh", "äm"*; typische Korrektureinleiter sind etwa: *"äh", "äm", "ah", "nein", "Entschuldigung", "das heißt"*. *"äh"* und *"äm"* kommen in beiden Funktionen wohl am häufigsten vor und dürften eine Teilmenge von Korrektureinleitern bilden, die auch Häsitationen sein können. Ein Füllwort wie *"also"* – normalerweise ein Korrektureinleiter – ist auch als Verzögerungssignal vorstellbar, wogegen *"das heißt"* nur Korrektureinleiter sein kann. Bei der

[4] Genauer gesagt wird kein Wort des Neuansatzes entsprechend gelabelt, vgl. Abschnitt 4.5, auch wenn z.B. Wiederholungen vorkommen, vgl. etwa: *"Gestern haben wir – ach was – wir sind gestern weit genug gekommen."* Gelabelt wird nur eine ansonsten nicht auflösbare Antezedens/Anapher-Relation.

Worterkennung bietet es sich an, für die Füllwörter, die am häufigsten vorkommen, eigene Wortmodelle zu verwenden. (In unserem eigenen Material, vgl. Anhang 1, findet sich z.B. 72 mal *“äm”* und 51 mal *“äh”*. Die restlichen 24 sind sehr unterschiedlich – *“n”*, *“m”*, *“hm”*, *“pff”*, *“öh”*, *“öhm”* – und müßten fürs erste als ‘Papierkorbkategorie’ modelliert werden. Bei einer größeren Stichprobe können sicher noch für mehr Füllwörter Wortmodelle aufgestellt werden.) Häsitationen haben zumindest drei unterschiedliche, sich nicht ausschließende Funktionen: sie können Planungspausen indizieren, sie können eine Übernahme des Turn durch den Hörer verhindern, und sie können reine Sprecheridiosynkrasien darstellen. Sprecheridiosynkrasien können auch als nicht– bzw. nur partiell pathologische Form des Stotterns auftreten, etwa bei *“in die Schu–, Schu–, Schule”*. In solchen Fällen ist es letztlich unentscheidbar, ob es sich um eine Sprecheridiosynkrasie und/oder eine Verhinderung der Turn-Übernahme durch den Hörer im Sinne einer Verzögerung oder doch um eine Korrektur handelt.

Es wird sicher Fälle geben, bei denen Füllwörter nicht eindeutig als Häsitation oder Korrektureinleiter klassifiziert werden können oder wahrscheinlich beiden Funktionen zugeordnet werden müssen; vgl. auch die schwierige Abgrenzung zu Interjektionen bzw. zwischen Interjektionen und formelhaften Parenthesen wie *“glaub ich”*, etc. Für Interjektionen und formelhafte Parenthesen sollten ebenfalls Wortmodelle erstellt werden; Parenthesen ohne Wortmodell sind damit automatisch keine formelhaften Parenthesen. Formelhafte Parenthesen sind oft, analog zu Satzadverbien oder Modalpartikeln wie *“vielleicht”*, prosodisch ‘unauffällig’, d.h. intonatorisch integriert, im Gegensatz zu nicht formelhaften Parenthesen, vgl. das folgende Beispiel[5]:

$$\textit{Heute morgen hab' ich} \left\{ \begin{array}{c} \textit{glaub' ich} \\ \textit{vielleicht} \\ \textit{naja, ich würd es mal so bezeichnen} \end{array} \right\} \textit{eine Dummheit gemacht.}$$

Der **F0-Verlauf** von Füllwörtern ist normalerweise von der Umgebung abgesetzt (tiefer), wiederholte Elemente sind oft weniger stark betont (geringere Dauer, kein großer F0-Anstieg), wogegen korrigierte Elemente oft stärker betont sind (größere Dauer, höherer F0-Anstieg) als die Umgebung, vgl. [Wil88, S. 246-253, insb. 251]. Es wird im einzelnen zu untersuchen sein, inwieweit sich verläßliche Merkmale finden lassen, mit deren Hilfe normale Phrasengrenzen — und damit auch “reguläre” Parenthesen — von Häsitationen und Korrekturen unterschieden werden können.

[5] Der Intonationsverlauf des Satzes dürfte mit der formelhaften Parenthese *“glaub ich”* und mit dem Satzadverb *“vielleicht”* ähnlich aussehen; wenn die Modalpartikel *“vielleicht”* in einem Exklamativsatz verwendet wird, so ist aber ein anderer Intonationsverlauf etwa mit dem Satzakzent auf den thematischen Elementen *“habe”* oder *“ich”* wahrscheinlicher. Dagegen wird die längere Parenthese *“naja, ich würd es mal so bezeichnen”* prosodisch sicherlich vom Kontextsatz abgesetzt sein.

4 Das Beschreibungssystem

Das folgende Beschreibungssystem für außergrammatische Phänomene spontaner Sprache beruht auf der VERBMOBIL-Basistransliteration, den Vorschlägen aus [BDSP93] und zusätzlichen Erweiterungen unsererseits. [6] Diese Kombination läßt eine sehr detaillierte Labelung zu, erlaubt aber auch eine schrittweise Rückführung der Spezifikationen etwa im Verzicht auf unsere Erweiterungen oder in der ausschließlichen Benutzung der Basistransliteration. Man beachte allerdings, daß die Basistransliteration lediglich die Möglichkeit bietet, 'störende' Passagen herauszufiltern, um eine für die linguistische Verarbeitung korrekte Wortfolge zu erhalten. Um solche Passagen jedoch automatisch lokalisieren zu können, sind zunächst Untersuchungen zu ihrer Struktur nötig. Solche Untersuchungen sind durch die Verwendung der Basistransliteration allein nicht möglich. Durch die explizite Labelung der einzelnen Wörter innerhalb der betroffenen Passagen können Vertreter einer ganz bestimmten Phänomenstruktur aus der Datenbasis extrahiert und anschließend getrennt analysiert werden.

Sollte sich aus der Transliterationspraxis die Notwendigkeit ergeben, weitere Symbole einzuführen, so sollten diese möglichst so gewählt werden, daß sie auf die gegebenen Symbole abgebildet werden, d.h. immer im Sinne einer Überspezifikation. Man beachte, daß sich die Auswahl der Symbole auf diejenigen beschränken muß, die noch nicht durch die Basistransliteration besetzt sind.[7] Tabelle 3 gibt eine Übersicht über die bei der Labelung benutzten Symbole.

Das Beschreibungssystem ist so angelegt, daß es auch von Transkribenten angewandt werden kann, die weder ausgebildete Phonetiker noch ausgebildete Syntaktiker sind. Zum einen werden deshalb keine prosodischen Auffälligkeiten wie etwa "Kontrastakzent" oder "steigender Tonverlauf" gelabelt. Zum anderen beschränkt sich die Analyse grundsätzlich auf die Wortebene und dort auf die Oberfläche. In Zweifelsfällen, z.B. bei der Entscheidung, ob es sich um eine Korrektur oder um einen Syntaxabbruch handelt, müssen allerdings doch syntaktische Kriterien herangezogen werden, vgl. dazu weiter unten den Abschnitt über Zweifelsfälle (4.5).

Bei der expliziten und detaillierten Labelung von außergrammatischen Phänomenen kann noch auf keine größere Erfahrung zurückgegriffen werden. Das Beschreibungssystem ist als ein Vorschlag zu verstehen; Modifikationen bei den Symbolen bzw. der Art der Beschreibung sind daher möglich.

[6] Um einen leichten Vergleich mit den Konventionen des Handbuchs der VERBMOBIL-Basistransliteration ([IPD94]) zu ermöglichen, finden sich im Anhang die einschlägigen Passagen aus dem Handbuch, teilweise von uns kommentiert. Wir verzichten daher darauf, im Text zu vermerken, woher die jeweilige Notation kommt.

[7] So sind dort bereits die eckigen und geschweiften Klammern reserviert, die sich ansonsten z.B. für die Parenthesen angeboten hätten.

Labelung der größeren Einheiten:	
<...>	Zur Klammerung von leeren und gefüllten Pausen, Dehnungen, Korrektureinleitern, formelhaften Parenthesen
<<...>>	Zur Klammerung von längeren (satzwertigen) Parenthesen
=/.../=	Zur Klammerung der 1. Phase, sofern diese Phase aus Elementen besteht, die in der 3. Phase **wiederholt** (nicht korrigiert) werden, wobei /= den Zeitpunkt der Unterbrechung bezeichnet
+/.../+	Zur Klammerung der 1. Phase, sofern diese Phase **auch** aus Elementen besteht, die in der 3. Phase **korrigiert** werden, wobei /+ den Zeitpunkt der Unterbrechung bezeichnet
-/.../-	Zur Klammerung des kompletten Syntaxabbruchs mit anschließendem Neuansatz
&/.../&	Zur Klammerung von kompletten syntaktischen Konstituenten, wenn nur Teile daraus gestört sind.
Labelung der störenden Elemente:	
Nichtwörter:	
<A>	Atmen
<P>	Leere Pause
<ähm>	gefüllte Pausen oder Korrektureinleiter (insb. *"äh", "öh"*, etc.)
Wortabbrüche und Dehnungen:	
−−	Wortabbruch
<Z>	Dehnung ("Zögerung")
funktionale, größere Einheiten:	
<also>	Korrektureinleiter, formelhafte Parenthesen, Interjektionen (*"das heißt", "Schmarrn", "meine ich"*, etc.)
<<........>>	längere (satzförmige) Parenthesen (auch längere Flüche, etc.)
Wörter (terminale Ebene):	
@R	Ersetzung (repair)
@M	Wiederholung (match)
@D	Eliminierung (deletion), d.h. ein Wort vor der Unterbrechung wird danach nicht mehr wiederholt oder korrigiert
@I	Einfügung (insertion), d.h. ein vor der Unterbrechung nicht vorhandenes Wort wird danach eingefügt
@A	Beziehung von Anapher und Antezedens bei Syntaxabbrüchen
Sonderfälle der terminalen Ebene	
@L	Sonderfall der Ersetzung mit gleicher Lexik, gleicher Semantik, aber unterschiedlicher Morphologie in nicht-betonten Silben, z.B. *"einen"* vs. *"einem"*
@m	Wortzusammenziehungen (z.B. *"so einem"* vs. *"som"*)
@r	Wortzusammenziehungen mit gleichzeitiger Korrektur (z.B. *"so einen"* vs. *"som"*)

Tabelle 3: Übersicht über die bei der Labelung verwendeten Symbole

4.1 Unterbrechung der Syntax (1. Phase)

Mit einen Schrägstrich (/) und einem nachfolgendem Zeichen $\in \{-,+,=\}$ wird der Zeitpunkt markiert, an dem die Syntax unterbrochen wird; vgl. dazu die Labelung der größeren Einheiten in Tabelle 3. Dabei bezeichnet

- ein nachfolgendes Minuszeichen (/-) einen Syntaxabbruch, dem ein Neuansatz folgt
- ein nachfolgendes Pluszeichen (/+) eine anschließende Korrektur
- ein nachfolgendes Gleichheitszeichen (/=) eine anschließende Wiederholung.

Die korrigierende bzw. wiederholende Weiterführung erfolgt nach dem Unterbrechungszeichen oder nach den mit spitzen Klammern markierten Einfügungen. Es wird mit der Umkehrung der jeweiligen Unterbrechungssymbole (-/, +/ und =/) der Beginn der abgebrochenen Passage, des Reparandums bzw. des Teils, der nach der Unterbrechung wiederholt wird, gekennzeichnet, so daß die Möglichkeit des Herausfilterns störender Passagen besteht. Gegenüber der Basistransliteration wird auch der Anfang eines Syntaxabbruchs mit "-/" gelabelt: Wenn so gefiltert werden soll, daß eine "korrekte" Kette übrigbleibt, dann muß der Anfang des Syntaxabbruchs gesetzt werden. Eine Markierung des Anfangs ist wohl auch für die Verarbeitung auf den höheren Stufen (Semantik, Dialogstruktur) sinnvoll. Die Entscheidung ist allerdings nicht immer eindeutig; es sollte daher im weiteren untersucht werden, ob und inwiefern sich möglichst eindeutige Kriterien bestimmen lassen.

Im Falle mehrerer aufeinanderfolgender außergrammatischer Passagen wird der Beginn des ganzen Störungskomplexes mit +/, -/ bzw. =/ gekennzeichnet und jede Unterbrechung im Komplex mit /-, /+ bzw. /=. Ein späterer Filtervorgang bezieht dann alles zwischen Anfangskennzeichnung und dem letzten nachfolgenden /-, /+ bzw. /= ein. Anfangs- und Endzeichen werden ohne Leerstelle vor das erste bzw. nach das letzte Wort der Unterbrechung gesetzt; zu möglichen Modifikationen dieser Konventionen der Basistransliteration vgl. 4.5.

Beispiele:

Kompletter Abbruch/Neuansatz:
"-/aber der rechte/- also so daß se..."
" mir wäre ganz recht <P> -/sagen wir mal zwischen dem/- , der achzehnte oder neunzehnte November , wenn Sie da Zeit hätten . "

Korrektur:
"da schließen +/die hinteren/+ <äh> die roten Steine..."
"ja gut , +/das ist /+ das paßt mir sehr gut"

Wiederholung:
"=/von dem Haus weg/= von dem Haus weg"
"<A> <m> ja =/wie wär's denn/= wie wär's denn in der <Z> letzten Novemberwoche ?"

4.2 Unterbrechungsphase (2. Phase)

Wenn keines der mit Schrägstrich geklammerten Phänomene der ersten Phase vorhanden ist, so indizieren die Elemente in spitzen Klammern (leere Pausen, gefüllte Pausen, Korrektureinleiter und Parenthesen) die Unterbrechungsphase (vgl. dazu die Labelung der größeren Einheiten in Tabelle 3).

Dabei werden leere Pause mit <P> und Atmen mit <A> gelabelt. Füllwörter und Korrektureinleiter (z.B. *"äh, ähm"*, aber auch *"das heißt"*), werden ebenso wie formelhafte Parenthesen, Flüche und Interjektionen zwischen einfache spitze Klammern gesetzt. Nicht-formelhafte Parenthesen stehen zwischen doppelten spitzen Klammern; vgl. im einzelnen dazu die Labelung der störenden Elemente in Tabelle 3. Fallen Pausen, Füllwörter und Parenthesen in die 2. Phase, so werden sie nach dem entsprechenden Unterbrechungssymbol und einer Leerstelle notiert.

Füllwörter am Anfang eines Turn werden natürlich nicht von einem Unterbrechungssymbol eingeleitet, vgl. die nächsten beiden Beispiele:

"<äh> <phh> <P> <nee> <P> so zwanzig Zentimeter von mir weg ungefähr"

"<A> <ähm> <A> <Schmatzen> <ähm> <P> ja , wie sähe es denn bei Ihnen Anfang November aus ? "

"das sind +/vier/+ <P> <nee> <Schmarrn> drei..."

"-/aber der rechte/- <P> also so daß sie nach innen kucken"

"daß ich +/bis Dienstag /+ <äh> von Dienstag bis Freitag in Hamburg bin ."

"+/ich würde sagen/+ <also> ich würde vorschlagen"

"=/das wären/= <<also , ich dachte mir ,>> das wären sechs Termine "

Im Gegensatz zur Basistransliteration werden also nach dem vorliegenden Stand alle Elemente der Unterbrechungsphase mit Ausnahme der nicht-formelhaften Parenthesen gleich gelabelt. Diese Gleichbehandlung ist u.E. möglich, da alle diese Elemente in das Lexikon mit aufgenommen werden müssen, und da sie jeweils endlichen und auch überschaubaren Teilmengen zugeordnet werden können. Mit einer einfachen Abgleichung mit einer Liste dieser Teilmengen kann also z.B. entschieden werden, ob es sich bei dem Element in spitzen Klammern um eine Interjektion (*"ach"*), eine formelhafte Parenthese (*"glaub' ich"*), etc. handelt. Die Liste dieser Teilmengen kann von Experten aufgestellt und jeweils auf den neuesten Stand gebracht werden; die Entscheidung darüber muß also nicht von den Transliterierern getroffen werden.[8]

[8] Diese Listen sind so zu verstehen, daß z.B. *"das heißt"* ein möglicher Korrektureinleiter ist, aber nicht immer in dieser Funktion verwendet wird. Sollte sich z.B. aus der Praxis dennoch die Notwendigkeit einer expliziten Labelung ergeben, so kann eine solche natürlich vereinbart werden. Dafür bieten sich unterschiedliche Kombinationen von Klammerungen an, da alle Einzelklammern schon vergeben sind, also etwa "[< ... >]" versus "< ... >".

4.3 Phänomene der Wortebene (1. und 3. Phase)

Jedes Wort einer außergrammatischen Passage wird gelabelt. Korrekturen besitzen immer eine linke, zu korrigierende Seite (Reparandum) und eine rechte, korrigierende Seite; auch bei reinen Wiederholungen gibt es eine linke, zu wiederholende und eine rechte, wiederholende Seite. Alle Wörter in der relevanten Passage werden mit einem der vier Kennzeichnungen @M (Wiederholung), @R (Ersetzung), @I (Hinzufügung) oder @D (Eliminierung) gekennzeichnet. Die relevante Passage besteht aus der geklammerten 1. Phase vor der Unterbrechung sowie aus den zugeordneten Elementen der 3. Phase. Die linke Seite wird von der rechten Seite durch eine der beiden Trennungsmarkierungen (/+, /=) unterschieden. Um die Beziehung der Labels vor und nach der Trennmarkierung auszudrücken, werden R, M, I und D mit Indizes versehen. Bei Satzabbrüchen wird nichts auf der Wortebene gelabelt; Ausnahme ist die Anapher/Antezedens–Beziehung, vgl. Abschnitt 4.3.6.

Bei Dehnungen wird **nach** der gedehnten Stelle im Wort ein "Z" für "Zögerung" zwischen spitzen Klammern gesetzt (<Z>). Soll eine Dehnung innerhalb des Wortes beschrieben werden, so wird <Z> nach der Dehnung ohne Leerstellen in das Wort eingefügt: *"Lebens<Z>mittel"*. Finale Dehnung wird nach dem Wort zusammen mit einer Leerstelle gelabelt. Folgt allerdings nach der finalen Dehnung ein Unterbrechungszeichen, so wird "<Z>" **ohne** Leerstelle an das Wort angehängt. Zwischen "<Z>" und dem Unterbrechungszeichen steht ebenfalls keine Leerstelle.

Eine explizite Indizierung ist nötig, da z.B. eine Abfolge M1 M2 (bzw. R1 R2) in der 1. Phase einer Abfolge M2 M1 (bzw. R2 R1) in der 3. Phase entsprechen kann. Deshalb scheint es sinnvoll zu sein, in einer Anfangsphase mit Indizes zu labeln. Diese Indizes können, falls sie für eine bestimmte Anwendung nicht benötigt werden, leicht automatisch entfernt werden, wohingegen umgekehrt eine nachträgliche automatische Indizierung nicht möglich ist.

4.3.1 Wiederholungen

Wiederholungen treten oft alleine auf, oft aber auch in Verbindung mit Korrekturen. Als Wiederholung werden nur semantisch und morphologisch identische Wörter bezeichnet. Wiederholungen werden mit @M (von "matching") plus Index gekennzeichnet; das Label wird ohne Leerstelle vor das betreffende Wort gesetzt. Um eine Suche bei einer eventuellen Filterung zu vereinheitlichen, sollte der Index auch bei nur einem wiederholten Wort gesetzt werden. Gekennzeichnet werden das zu wiederholende Wort auf der linken Seite der Unterbrechung (/=) und das wiederholte Wort auf der rechten Seite der Unterbrechung. Treten mehrere @M mit dem gleichen Index auf, wurde das entsprechende Wort auch mehrmals wiederholt. Die zu wiederholende Passage steht zwischen =/ und /= ; zu Randtypen der Wiederholungen vgl. 4.3.4.

"so daß ich ungefähr =/@M1halb/= @M1halb auf die glatte Seite schau"

"=/@M1wieder @M2in/= @M1wieder @M2in Richtung Wand"

“steht dann dieses =/@M1rote @M2Haus/= @M1rote/= @M1rote @M2Haus”

“aber am besten wäre vielleicht doch =/@M1die/= <A> <ähm> @M1die Woche , die mit dem neunundzwanzigsten November beginnt .”

“die Woche vom ersten =/@M1bis @M2zum/= @M1bis @M2zum achten Januar ,”

4.3.2 Ersetzungen

Sehr häufig werden bei Korrekturen Wörter durch neue ersetzt. Diese Korrekturen werden mit @R (von “repair”) und einem Index bezeichnet. Auch hier muß der Index als Teil des Labels in jedem Fall gesetzt werden, also auch bei nur einem zu ersetzenden Wort. Gekennzeichnet werden sowohl das Reparandum vor dem Syntaxabbruch als auch die Ersetzung nach dem Syntaxabbruch. Wie bei der Wiederholung werden die Label vor das entsprechende Wort ohne Leerstelle gesetzt. Tritt ein @R öfters mit dem gleichen Index auf, so wurde das korrigierte Wort weiter korigiert. Das Reparandum steht zwischen +/ und /+, auch wenn während der Korrekturpassage Wörter wiederholt werden:

“und zwar so daß +/@R1die/+ <ähm> @R1der Dachfirst..”

“vor dem <äh> zehnten +/@R1September/+ <A> <ähm> ja <A> @R1August schaut ’s bei mir schlecht aus ,”

“ +/@M1wie @R1bu-/+ @M1wie @R1buchspa-/+ @M1wie @R1buchstabiert man das ?”

4.3.3 Sonderfälle bei Wiederholung und Ersetzung

Wenn nur morphologisch (in der unbetonten Endsilbe) korrigiert wurde, die Lexik sich aber nicht ändert, wird mit @L (für ‘Lexik’) plus Index gelabelt.

“+/@L1einen @L2hellblauen @L3flachen/+ @L1ein @L2hellblaues @L3flaches Rechteck”
“+/@L1ein @M1D-/+ @L1einen/+ @I1auch @I2so @L1’nen @M1Dachfirst”
“da hab’ ich +/@L1keinen/+ @L1keine fünf Tage Zeit ,”

Auf besondere Weise können Wortzusammenziehungen behandelt werden. Das zusammengezogene Wort wird mit @m plus Index gekennzeichnet, wenn sich die Morphologie nicht ändert, ansonsten mit @r plus Index. Wird die Zusammenziehung aufgelöst, dann wird auch das Label geteilt (@m1 = @M1 @M1):

“+/@r1son @M1Klotz/+ @r1som @M1Klotz”
“=/@m1son @M2Käse/= @M1so @M1ein @M2Käse”
“=/@r1son @M1Käse/= @R1dieser @M1Käse”

Möglicherweise ist allerdings diese spezielle Art der Labelung von Zusammenziehungen nicht nötig, da in der Basistransliteration nicht *“son”*, sondern *“so ’n”* transliteriert wird.

4.3.4 Einfügungen und Eliminierungen

Wörter, die nur auf einer Seite der Korrektur erscheinen, und Wörter, die weder eindeutig wiederholt noch korrigiert wurden, werden mit @I für Insertion (Einfügung) und @D für Deletion (Eliminierung) gelabelt, wobei @D nur auf der linken Seite der Unterbrechung und @I nur auf der rechten auftreten kann. Hier ist zwar ein Index eigentlich unnötig, kann aber aus Gründen der Systematik gesetzt werden, um ein späteres Herausfiltern dieser Elemente zu ermöglichen; zu Grenzfällen zwischen Korrektur und Syntaxabbruch vgl. den Abschnitt über die Zweifelsfälle (4.5).

Ein Randtyp der Wiederholung ist eine Wiederholung mit einer oder mehreren Einfügungen bzw. Eliminierungen. Möglich wäre für solche Fälle natürlich auch eine Labelung als Korrektur, d.h. mit "+/ ... /+". Wir entscheiden uns aber für eine Labelung als Wiederholung, da durch die zusätzliche Labelung von mindestens einem Wort als Eliminierung bzw. Einfügung diese Randtypen eindeutig bestimmbar sind.

"dann gibt es nämlich =/@M1zwei/= @I1immer @I2jeweils @M1zwei die gleich hoch sind"

"+/@L1ein @M1D-/+ @L1einen/+ @I1auch @I2so @L1'nen @M1Dachfirst"

"dann nimmst Du wieder eins +/@M1von @M2diesen @D1roten @R1Teilen/+ @M1von @M2diesen @R1Dachteilen"

"ja , am ersten <ähm> =/@M1am/= @I1und @M1am neunundzwanzigsten August hätte ich Zeit ."

"und zwar =/@M1im/= , @I1bei @I2uns <m> @M1im Ratzenkammerl , da ißt man sehr gut ."

"vom vierundzwanzigsten Februar bis Donnerstag , den elften , =/@M1wär' @D1bei @M2m−−/= @M1wär @M2mir jeder Termin recht ."

Das letzte Beispiel ist zusätzlich ein Randtyp zwischen Korrektur und Syntaxabbruch, da ja als Alternativen vorstellbar sind zum einen *"wäre bei mir jeder Termin recht"* und zum anderen *"wäre bei mir noch Platz im Kalender"* o.ä. . Die Entscheidung ist letztlich arbiträr und wird auch schon in der Basistransliteration getroffen.

4.3.5 Wortabbrüche

Um einen Wortabbruch zu kennzeichnen, werden dem Wortfragment zwei Minuszeichen (−−) ohne Leerstelle nachgestellt. Ob das Wort dann nach dem Abbruch vervollständigt oder korrigiert wird, ist durch das vorangestellte @M (bei Vervollständigung) oder @R (bei Korrektur) plus Index ersichtlich.

"mit =/@M1der @M2li−−/= @M1der @M2linken Kante"

"ja gut da wär' ein =/@M1W−−/= @M1Wochenende dabei "

“+/@M1von @L1diesem @M2ganzen @R1Geb−−/+ <äh> @M1von @L1dieser @M2ganzen @R1Stadt”

Trifft ein Wortabbruch mit einem Syntaxabbruch zusammen, wird wie folgt gelabelt:

“nein also die Fe−−/-”

Zur Diskussion über nicht entscheidbare Zweifelsfälle vgl. unsere Kommentare im Anhang 2.[9]

4.3.6 Anapherauflösung

Wie schon weiter oben erwähnt wurde, kann bei der Äußerung:
“Gestern hab ich mit dem Willi- , äh, also, er hat gemeint, daß ...”
nach dem kompletten Ausblenden der abgebrochenen Passage die Anapher nicht mehr aufgelöst werden. Es wird deshalb das Label @A für diese Beziehung von Anapher und Antezedens eingeführt:[10]

“-/Gestern hab ich mit dem @A1Willi/- <äh> <also> @A1er hat gemeint ”

Dieses Label ist nur dann zu setzen, wenn eine Anapherauflösung nötig ist, also nicht dann, wenn beim Neuansatz explizit das Antezedens wieder aufgenommen wird, bzw. wenn es gar kein Antezedens gibt. Wenn zur Anapherauflösung nicht nur ein einziges Wort als Antezedens, sondern eine komplexere Phrase nötig ist, so wird jedes Wort dieser Phrase, und zwar mit dem selben Index, gelabelt.[11]

“<A> ja gut , bei mir fällt allerdings dann die dreißigste Woche aus , -/aber wir könnten <Z> @A1an <Z> @A1der @A1neunundzwanzigsten @A1Woche zum Beispiel/- vielleicht können wir @A1da auch direkt zwei Termine machen , irgendwie wieder montags und dienstags wär' ja ganz geschickt , wenn das ginge . %also das wär' Montag der achtzehnte und Dienstag der neunzehnte Juli .”

[9] Wenn eine als ein abgebrochenes Wort gelabelte Segmentfolge als Eintrag in einem Vollformenlexikon existiert, so kann es sich grundsätzlich auch um ein nicht abgebrochenes Wort handeln. Oft dürfte aber der Kontext die Entscheidung für oder gegen Abbruch erleichtern.

[10] Man beachte, daß hier “also” sowohl als Korrektureinleiter als auch als eine normale “Satzauftakt”-Partikel interpretiert werden kann. Prosodisch ist allerdings eine unterschiedliche Markierung vorstellbar: Wenn “also” an “er” klitisch angeschlossen wird, dann wird es sich eher um einen Satzauftakt handeln; wenn es durch eine Pause abgetrennt ist, so ist es als Korrektureinleiter zu interpretieren.

[11] Das folgende Beispiel ist aus einer VERBMOBIL-Transliteration übernommen und illustriert gleichzeitig die Schwierigkeiten der Transliteration; u.E. liegt hier nur bei einer eindeutigen prosodischen Indizierung ein Abbruch vor.

4.4 Konstituentengrenzen

Um einen sinnvollen Anfangs- und Endpunkt für eine automatische Extraktion der Passagen festzulegen, werden bei Störungen, die nicht eine ganze syntaktische Konstituente betreffen, die Konstituentengrenzen vor dem Beginn der 1. Phase sowie nach dem Ende der 3. Phase mit &/ bzw. mit /& gelabelt. Wörter, die in diese Bereiche, aber nicht in die 1. oder 3. Phase fallen, werden nicht gelabelt. Der Anfangszeitpunkt wird nur dann gelabelt, wenn er **nicht** mit =/ oder mit +/ zusammenfällt. Der Endzeitpunkt wird nur dann gelabelt, wenn das ihm vorausgehende Wort keine Wortlabelung (@M, @R, ...) besitzt. Grundgedanke war, daß die Labelung genügend große, aber auch nicht zu große und damit irrelevante Bereiche abdeckt, die gegebenenfalls extrahiert und gezielt prosodisch, syntaktisch oder semantisch bearbeitet werden können.

“&/mit =/@M1der @M2li−−/= @M1der @M2linken Kante/&”
“also &/das <äh> mit der glatten Fläche/&”
“und zwar so daß +/@R1die/+ <ähm> @R1der Dachfirst/&”

Auf diese Weise können Ambiguitäten auf der Oberfläche aufgelöst werden, so etwa bei *die* und *der* die von Artikel vs. Relativpronomen, und es steht genügend viel Information für linguistische Untersuchungen zur Verfügung. Reicht eine solche Umgebung für eine ausführliche syntaktische Analyse nicht aus, so kann immer noch nach den in der Basistranskription notierten Satzzeichen gesucht werden.

4.5 Randtypen und Zweifelsfälle

Prototypische Fälle, die (auch von uns) gerne als Beispiele angeführt werden, sind eindeutig klassifizierbare Fälle. Zweifelsfälle betreffen Randtypen, wobei die Entscheidung für eine der alternativen Labelungen teils reversibel, teils irreversibel ist. Bei den folgenden Beispielen handelt es sich bei den beiden ersten um eindeutige Verzögerungen, beim letzten um eine eindeutige Korrektur:

1. *“&/dieser <Z> Termin/&”*
2. *“&/dieser <äh> Termin/&”*
3. *“=/@M1dieser/= @M1dieser Termin/&”*
4. *“=/@M1di−−/= @M1dieser Termin/&”*
5. *“+/@R1die/+ @R1dieser Termin/&”*
6. *“+/@R1die/+ @R1der Termin/&”*

Eine reine Wiederholung (3. Beispiel) wird von uns als Verzögerung klassifiziert; [Lev83] bezeichnet sie als "covert repair" und impliziert damit bereits eine Art (intendierte?) Korrektur. Dieser Unterschied mag für die Theoriebildung bedeutsam sein, stört aber nicht weiter, da die Labelung ausreichend spezifiziert. Beim 4. und 5. Beispiel gibt es zwei Probleme. Man muß sich entscheiden, ob es sich um den nicht abgebrochenen Artikel *die* handelt, der dann korrigiert wird zum Demonstrativpronomen (Labelung wie im 5. Beispiel), oder um das abgebrochene gleiche Demonstrativpronomen, das dann wiederholt wird (Labelung wie im 4. Beispiel). Diese Entscheidung ist fürs erste irreversibel. Ob man jedoch einen Abbruch plus Wiederholung wie im 4. Beispiel als Wiederholung im Sinne einer Verzögerung (mit "=/.../=" zu labeln) oder als Korrektur (mit "+/.../+" zu labeln) definiert, ist zweitrangig. Sinnvoll ist die erste Labelung, da sie durch die Kombination von Wiederholungsmarkierer und Abbruchlabel eindeutig spezifiziert und solche Fälle später beliebig klassifiziert werden können.

Bei der folgenden Äußerung sind mindestens drei verschiedene Arten der Labelung vorstellbar:

"nach hinten und 'n bißchen also ungefähr zwei Zentimeter nach links versetzt"

- Behandlung als Korrektur:
  "nach hinten und +/@R1'n @R2bißchen/+ <also> @R1ungefähr @R2zwei Zentimeter nach links versetzt"
- Behandlung als Abbruch:
  "nach hinten und -/'n bißchen/- <also> ungefähr zwei Zentimeter nach links versetzt"
- Interpretierung als reguläre Passage im Sinne einer Spezifizierung:
  "nach hinten und 'n bißchen also ungefähr zwei Zentimeter nach links versetzt"

Eine Behandlung als Korrektur, vgl. die erste Alternative, scheint wenig sinnvoll zu sein. Bessere Alternativen sind u.E. die zweite und die dritte, wobei die Entscheidung von der prosodischen Markierung von *'also'* abhängt: wenn *'also'* an die nachfolgende Konstituente prosodisch integriert ist und wie eine Rechtsversetzung (*"...und zwar zwei Zentimeter nach links versetzt"*) interpretiert werden kann, so ist die dritte Alternative zu wählen. Wenn *'also'* dagegen prosodisch abgesetzt ist und wie zum Beispiel *"nee, stimmt nicht"* verwendet wird, so ist die zweite Alternative zu wählen.[12]

Probleme mit der Labelung ergeben sich auch schon bei relativ einfachen Phänomenen wie etwa bei:

"das rote - grüne, grüne Haus"

Hier handelt es sich um die Verzahnung einer Wiederholung und einer Korrektur; bei einer

[12][Lev89, S. 459] bezeichnet eine 'echte' Korrektur als 'error repair' und eine Spezifizierung als 'appropriateness repair'. Diese eher semantische Unterscheidung wird von uns nicht getroffen, da wir uns bei der Analyse auf die syntaktische bzw. prosodische Ebene beschränken.

Links-Rechts–Verarbeitung liegt eine Überlappung bzw. Einbettung einer Korrektur in eine Wiederholung vor.

Diese Einbettung kann folgendermaßen symbolisiert werden: es werden sowohl die Wiederholung als auch die Korrektur geklammert. Die Einbettung ist somit erkennbar an kombinierten Klammerungssymbolen (=/+/) und an kombinierten Wortlabels (@R1@M1). Eine solch explizite Art der Labelung kann jedoch recht komplex und unübersichtlich werden.

"&/das =/+/@R1rote/+ <Z> @R1@M1grüne/= @M1grüne Haus/&"

Das letzte Beispiel weicht insofern von den Prinzipien der Basistransliteration ab, als mehr als ein Anfangssymbol aufeinander folgen – in diesem Fall an der gleichen Position. Im folgenden Beispiel, das Mehrfach-Korrekturen enthält, können dementsprechend auch zwei Korrekturanfangssymbole aufeinander folgen (*... +/@M1das +/@R1grüne ...*).

"ich möchte +/@M1das +/@R1grüne/+ <äh> @R1rote @R2Haus/+ <Quatsch> @M1das @R1blaue @R2Auto"

Eine explizite Indizierung der Klammersymbole analog zu der bei den Labelsymbolen auf Wortebene ist natürlich möglich, aber nicht unbedingt nötig. Eine automatische Filterung ist z.B. weiterhin möglich, wenn die Regel heißt: "Gehe vom Anfangssymbol nach rechts und beziehe alle darauffolgenden Passagen bis zum letzten Endsymbol (maximal bis zum Ende des Turn) vor dem nächsten Anfangssymbol in die Filterung mit ein." Der Vorteil dieser Alternative ist, daß damit auch die hierarchische Struktur der Einbettungen vollständig beschrieben und analysiert werden kann.

Die Äußerung *"die siehst du jetzt - die stehn jetzt"* könnte zwar auch als Korrektur gelabelt werden:

"+/@M1die @R1siehst @D1du @M2jetzt/+ @M1die @R1stehn @M2jetzt"

Im Sinne von [BDSP93] wird hier keine neue Idee eingeführt. Die syntaktische Konstruktion ändert sich allerdings, und daher ist hier die Interpretation als Syntaxabbruch vorzuziehen:

"-/die siehst du jetzt/- die stehn jetzt"

4.6 Komplexe außergrammatische Störungen

Als ein Beispiel für eine komplexe außergrammatische Störung soll die folgende Passage interpretiert werden, die unserem spontansprachlichen Korpus (vgl. Anhang 1) entnommen ist. Sie steht zuerst ohne, dann mit Labelung.

"und zwar nicht äm man kanns ja so darstellen daß m also du stellst sie praktisch so hin daß die blaue Seite daß die rote Seite ein Teil von der blauen Seite überdeckt das heißt die gan nee die ganze blaue Seite überdeckt"

"-/und zwar nicht/- <P> <äm> -/man kanns ja so darstellen daß/- <m> <P> also du stellst sie praktisch so hin +/@M1daß @M2die @R1blaue @M3Seite/+ @M1daß

@M2die @R1rote @M3Seite +/@R1ein @R2Teil @D1von @D2der @L3blauen @M1Seite @M2überdeckt/+ <P> <das heißt> =/@M3die @M4gan--/= <P> <nee> @M3@R1die @M4@R2ganze @L3blaue @M1Seite @M2überdeckt"

Die Abarbeitung durch den menschlichen Bearbeiter sowie eine – wenn auch weit in der Zukunft liegende – automatische Verarbeitung geschieht in vier Stufen:

1: *"-/und zwar nicht/- <P> <ähm>"*
= erster kompletter Abbruch mit Pause und Füllwort
würde komplett herausgefiltert werden.

2: *"-/man kanns ja so darstellen, daß/- <m> <P>"*
= zweiter kompletter Abbruch mit Füllwort und Pause
würde auch komplett herausgefiltert werden.

3: *"also du stellst sie praktisch so hin +/@M1daß @M2die @R1blaue @M3Seite/+ @M1daß @M2die @R1rote @M3Seite"*
= Korrektur
Herausgefiltert würde: *"+/daß die blaue Seite/+"*
Also bleibt stehn: *"Also du stellst sie praktisch so hin daß die rote Seite"*

4: *"+/@R1ein @R2Teil @D1von @D2der @L3blauen @M1Seite @M2überdeckt/+ <P> <das heißt> =/@M3die @M4gan--/= <P> <nee> @M3@R1die @M4@R2ganze @L3blaue @M1Seite @M2überdeckt"*
= Korrektur und Wiederholung, wobei die Wiederholung während des korrigierenden Teils geschieht.

4a: *"+/@R1ein @R2Teil @D1von @D2der @L3blauen @M1Seite @M2überdeckt/+ <P> <das heißt>"*
= Korrektur mit Pause und Korrektureinleiter
das ganze Reparandum würde einschließlich Pause und Korrektureinleiter herausgefiltert werden.

4b: *"=/@M3die @M4gan--/= <P> <ne> @M3@R1die @M4@R2ganze @L3blaue @M1Seite @M2überdeckt"*
= Wiederholung im korrigierenden Teil von 4a.
Hier stellt sich das Problem, daß Teile der Wiederholung ebenfalls Teile der Korrektur darstellen:
"@M3@R1die @M4@R2ganze":
4a: *"+/ein Teil von der/+ die ganze"*
4b: *"=/die gan--/= die ganze"*
Die doppelte Labelung, wobei im Sinne einer Klammerung zuerst das dem näherliegenden Abbruch dazugehörige, dann das dem vorletzten Abbruch entsprechende Element gelabelt wird, sieht unübersichtlich aus, erscheint aber doch besser, als eine Klammerung, bei der noch mehr Symbole verwendet werden würden und die in der linearen Darstellung noch schwerer zu durchschauen wäre. *"die gan-"* wird ebenfalls gefiltert.

Es bleibt damit als korrigierte Version übrig:

"Also du stellst sie praktisch so hin, daß die rote Seite die ganze blaue Seite überdeckt."

Es dürfte allerdings immer Fälle geben, bei denen eine solche detaillierte Labelung auf Wortebene nicht eindeutig durchgeführt werden kann. In einem solchen Fall bietet sich als pragmatische Lösung an, auf die Labelung der Wortebene zu verzichten und lediglich die betroffenen Teile zu klammern.

4.7 Ein Überblick über die Hierarchie der Phänomene sowie den zeitlichen Ablauf

Tabelle 4 gibt einen Gesamtüberblick und faßt die anderen drei Tabellen zusammen. Sie zeigt die Phänomene von Tabelle 1 in vertikaler Anordnung. Die zeitliche Abfolge ist horizontal, von links nach rechts, abzulesen. Diese Abfolge ist detaillierter dargestellt als in Tabelle 2. Die in Tabelle 3 besprochenen Labels sind an der einschlägigen Position aufgeführt.

Die vier Phänomene stehen in der folgenden hierarchischen Beziehung zueinander: Bei Parenthesen können alle drei anderen Phänomene auftreten. Korrekturen und Syntaxabbrüche sind per Setzung alternativ, bei ihnen kann jeweils auch eine Verzögerung auftreten. Eine Verzögerung kann auch alleine stehen. Selbsteinbettungen (Parenthesen in Parenthesen, Korrekturen in Korrekturen) sind möglich.

Tabelle 4 kann also zum einen als eine Art Labelvorlage für die Bearbeiter benutzt werden, zum anderen als Anleitung für eine automatische Fehlersuche: Wenn z.B. das Symbol "@I" zwischen "+/" und "/+" oder das Symbol "@D" außerhalb einer Klammerung gefunden werden, so ist das ein Indiz für eine fehlerhafte Labelung.

5 Überlegungen zur Verarbeitung und zur Einschränkung des Gegenstandsbereiches

Der vorliegende Abschnitt beschäftigt sich zum einen mit der Frage, wie eine solche Labelung spontansprachlicher Phänomene in automatischen Spracherkennungssystemen verwendet werden kann. Zum anderen soll beschrieben werden, welche Informationen von unterschiedlichen Spracherkennungsmoduln geliefert werden müssen, um eine vollautomatische Lokalisierung und Verarbeitung der Unterbrechungen in einem Gesamtsystem zu ermöglichen. Unsere bisherigen Überlegungen beziehen auch komplexe Störungen mit ein; allerdings dürfte man sich in einer ersten Phase wohl auf relativ einfache Phänomene beschränken.

<table>
<tr><th></th><th></th><th></th><th colspan="4">1. Phase</th><th colspan="2">2. Phase</th><th colspan="2">3. Phase</th><th></th></tr>
<tr><th></th><th>Phänomene</th><th></th><th>Start</th><th colspan="2"></th><th>Ende</th><th></th><th>Ende</th><th></th><th>Ende</th><th></th></tr>
<tr><td>Gram-
ma-
tisch</td><td>I. Parenthese</td><td rowspan="4">Konstit.
Anfang

&/</td><td colspan="2"></td><td rowspan="3">Wortabbruch</td><td></td><td colspan="2"><Parenthese>
<<Parenthese>></td><td colspan="2"></td><td rowspan="4">Konstit.
Ende

/&</td></tr>
<tr><td rowspan="3">Außer–
gram-
ma-
tisch</td><td>II. Korrektur</td><td>+/</td><td>Reparandum
(Eliminierung)
{@M, @D, @R,
@L, @m, @r}</td><td>/+</td><td><Korrektur–
einleiter></td><td></td><td>Reparans
(Ersetzung,
Ergänzung)
{@M, @I, @R,
@L, @m, @r}</td><td></td></tr>
<tr><td>III. Syntax-
abbruch</td><td>−/</td><td>Phrasen–
≈
Satzabbruch
{@A}</td><td>/−</td><td><Korrektur–
einleiter></td><td></td><td>Neuansatz
{@A}</td><td></td></tr>
<tr><td>IV. Verzögerung</td><td>= /</td><td>Dehnung {<Z>}
Eliminierung {@D}
Wiederholtes
{@M, @m}</td><td></td><td>/ =</td><td><gefüllte |
ungefüllte
Pause></td><td></td><td>Insertion {@I}
Wiederholendes
{@M, @m}</td><td></td></tr>
</table>

Tabelle 4: Der zeitliche Ablauf

Am wertvollsten ist ein Wissen darüber, an welchen Stellen eine bestimmte Art von Unterbrechung vorliegt, sicherlich für die automatische linguistische Analyse, insbesondere für das Parsen im Rahmen der syntaktischen Analyse.

Da Unterbrechungen - sieht man von der Dehnung ab - durch Elemente auf der Wortebene indiziert werden, ist es unbedingt erforderlich, diese bereits im Rahmen der Worterkennung zu modellieren und zu erkennen. In der Praxis bedeutet das, daß für kurze, formelhafte Einheiten (Füllwörter, Korrektureinleiter, Interjektionen und formelhafte Parenthesen) individuelle Wortmodelle erstellt und modelliert und somit vom Worterkennungsmodul erkannt werden; Grundlage für die Auswahl der Wortmodelle ist eine entsprechend gelabelte Stichprobe. Natürlich müssen diese Einheiten auch im Lexikon enthalten sein. In einem ersten Schritt sollte erfaßt werden, welche Worteinheiten modelliert werden müssen; am wichtigsten sind wohl vorerst Füllwörter und Korrektureinleiter wie *"äh", "äm", "das heißt", ...*). Der zweite Schritt ist das Training dieser Modelle mit einer adäquaten Stichprobe.

Ein weiterer wichtiger Punkt ist die Erkennung von Wortabbrüchen. Da die bisherigen Worterkenner lediglich die Wörter erkennen, die auch in ihrem Erkennungswortschatz vorhanden sind, müssen sie dahingehend modifiziert werden, daß sie zusätzlich auch Hypothesen generieren, die unvollständigen (also abgebrochenen) Wörtern entsprechen. Analog zur Behandlung unbekannter, neuer Wörter sollte das Ergebnis dieser Analyse in der Schnittstelle zur Linguistik durch sog. Dummy-Kanten repräsentiert werden (s. [NP94]).

Da Unterbrechungen häufig auch prosodisch markiert sind, wird eine parallele prosodische Verarbeitung benötigt, die auf der bereits erkannten Wortkette (bzw. dem Wortgraphen) aufsetzt; manchmal, z.B. bei Zögerungen, die lediglich durch Dehnung und/oder lange Pausen indiziert sind, gibt die Prosodie den einzigen Hinweis auf eine Unterbrechung. Beide Ebenen (Worterkennung und Prosodie) werden auf der Zeitachse von links nach rechts abgearbeitet und die Ergebnisse an die Syntax weitergegeben. Bei der Wortgraphen-Schnittstelle ist jedes Wort durch eine Kante repräsentiert; im bisher noch freien Slot für Zusatzinformationen sollen nun für jedes Wort die extrahierten prosodischen Informationen an die Syntax mit übergeben werden; im einzelnen (s. [NP94]). Da es sich dabei jeweils um 'unsichere' Klassifikationsergebnisse handelt, werden alle Entscheidungen mit Bewertungen bzw. Wahrscheinlichkeiten versehen.

Schnittstelle zur Linguistik ist also ein Wortgraph mit prosodischer Zusatzinformation. Besteht die Unterbrechung lediglich aus einer reinen Verzögerung (ungefüllte oder gefüllte Pause) ohne Auftreten eines anderen Phänomens wie Wiederholung oder Korrektur, können die entsprechenden Elemente bei der Verarbeitung durch die Linguistik einfach ausgeblendet werden. (Ob bereits das Prosodiemodul eine Differenzierung von Füllwörtern in 'reine' Verzögerungssignale oder Korrektureinleiter mit hinreichender Sicherheit liefern kann, muß eigens untersucht werden.) Eine analoge Strategie ist für einfache Wiederholungen (Wiederholung eines Wortes unmittelbar nach dem korrespondierenden Wort) evtl. auch mit dazwischengeschalteter Zögerung (lange, ungefüllte Pause oder Füllwort) denkbar; Voraussetzung dafür ist wiederum die korrekte Erkennung der Wiederholung durch Worterkennung und Linguistik.

Werden mehrere Wörter wiederholt bzw. sind wiederholte Wörter durch Einschübe voneinander getrennt, stehen also noch Wörter zwischen den beiden korrespondierenden Wortketten, so erhöht sich die Komplexität bereits erheblich.

Schwieriger als die Verarbeitung einfacher Wiederholungen ist die Verarbeitung selbst einfacher Korrekturen, z.B. wenn die Korrektur eines Wortes unmittelbar nach dem korrespondierenden Wort folgt – evtl. auch mit dazwischengeschalteter Zögerung. Hier genügt nicht mehr ein einfacher lexikalischer Vergleich, sondern es müssen die syntaktisch-semantischen Kategorien der korrespondierenden Wörter miteinander verglichen werden: ist etwa ein Adjektiv durch ein Adjektiv, ein Nomen durch ein Nomen, eine Zahl durch eine Zahl, ein Abfahrtsort durch einen Abfahrtsort ersetzt worden? Auch hier hängt die Komplexität in starkem Maße von der zeitlichen Entfernung der korrespondierenden Elemente ab.

Die Erkennung von Wortabbrüchen (Abbruch **innerhalb** eines Wortes) ist aus den oben genannten Gründen sicherlich kein triviales Problem. Ein geeignet dimensioniertes Worterkennungsmodul vorausgesetzt, kann im wesentlichen eine Behandlung analog zu den Korrekturen durchgeführt werden (Ersetzung des abgebrochenen durch das vervollständigte oder korrigierte Wort). Da ein Worterkenner im allgemeinen darauf trainiert ist, vollständige Worte seines Wortinventars zu erkennen, und die Betrachtung aller möglichen Abbruchstellen in allen Wörtern ein kombinatorisches Problem darstellt, kann es sinnvoll sein, parallel zu diesem mit einem Lauterkenner zu arbeiten und auf dem erzeugten Lautstring nach möglicherweise auftretenden Wortabbrüchen zu suchen.

Ein Syntaxabbruch (also ein Abbruch der Satzkonstruktion an einer Wortgrenze) kann – sieht man einmal von einer Verwendung eines Sprach-Modells bei der Worterkennung ab – nicht bereits von der Worterkennung als solcher identifiziert werden. Erkennt ein Parser einen Syntaxabbruch (ein Indiz hierfür kann z.B. ein mißlungener Parse sein) muß er – in analoger Weise wie bei der Verarbeitung von Mehrsatzäußerungen am Ende einer Äußerung – an der Stelle nach der Unterbrechung neu aufgesetzt werden.

Für einfache Wiederholungen und Korrekturen ist auch eine weitere Vorgehensweise denkbar, nämlich eine Nachbearbeitung der Worterkennungsergebnisse, also ein Filter für die an die Linguistik weitergegebene Information: Dieses Filter stellt auf lexikalischer Ebene z.B. Konstruktionen fest wie:

=/ @M1 /= @M1 oder
=/ @M1 /= <ähm> @M1 oder
=/ @M1 @M2 /= @M1 @M2 oder
+/ @R1 /+ @R1,

das Filter korrigiert diese Konstruktionen (d.h. eliminiert Wiederholungen und das Reparandum) und gibt die nun 'grammatischere' Wortkette an die Linguistik weiter. Diese Vorgehensweise setzt allerdings als Schnittstelle die beste Wortkette voraus.

Wie in den obigen Beispielen (vgl. Abschnitt 4.6) verdeutlicht wurde, finden sich in spontaner Sprache sehr komplexe Phänomene, bei denen selbst der Mensch Schwierigkeiten hat, den Überblick zu bewahren. Ab einer gewissen Komplexität erscheint deshalb für ein automatisches System lediglich das Suchen nach anwendungs- bzw. dialogschrittabhängigen Schlüsselwörtern (sog. **key-word-spotting** nach Datumsangaben, Orten, Uhrzeitangaben etc.) und das schrittweise Füllen und Ergänzen der jeweils zugrundeliegenden assertionalen Wissensdatenbank sinnvoll.

Zur Abgrenzung des Problembereichs sowie zu einer schrittweisen Anpassung der Verarbeitung auf den jeweiligen Komplexitätsgrad erscheint es nötig, in einer Vorstudie zu untersuchen, wie komplex die Phänomene im angestrebten VERBMOBIL-Szenario überhaupt sind. Ein solches 'Setting' (z.B. Vorgabe unterschiedlicher Termine an zwei Gesprächspartner, die sich einigen müssen) ermöglicht sicher weniger Spontaneität und damit weniger ausgeprägte außergrammatische Phänomene als das bei dem von uns untersuchten Korpus der Fall ist; vgl. dazu auch [Tro94].

Vorstellbar ist bei der Labelung und der automatischen Verarbeitung auch eine Beschränkung auf einfachere Phänomene etwa mit der Bedingung, daß nur dann detailliert auf Wortebene gelabelt wird, wenn maximal vier Wörter vor bzw. nach der Unterbrechung betroffen sind. Bei den restlichen Fällen bleibt dann nur die größere Einheit gelabelt, d.h. dort wird die Basistransliteration übernommen und nur um die explizite Markierung von Korrektureinleitern, Parenthesen und kompletten Konstituenten o.ä. ergänzt. Für eine detaillierte Labelung, wie sie hier vorgeschlagen wird, spricht allerdings, daß nur mit ihrer Hilfe gezielt Einzelphänomene, wie z.B. Wortabbrüche, für ein Training extrahiert werden können bzw. eine schrittweise Anpassung des Schwierigkeitsgrades möglich ist.

6 Schlußbemerkungen

Die vorliegende Arbeit soll eine Basis bilden für eine Beschäftigung mit außergrammatischen Phänomenen in der spontanen Sprache insbesondere im Rahmen einer automatischen Verarbeitung und Erkennung. Neben einem Überblick über die einschlägigen Phänomene enthält sie einen ausführlichen Vorschlag zur Labelung dieser Phänomene.[13] Die Arbeit steht – wie dies fast immer der Fall ist – "auf den Schultern " früherer Arbeiten, obwohl auf die reichhaltige bisherige Literatur nicht ausführlich eingegangen werden konnte. Unseres Wissens steht aber der Gesichtspunkt einer Aufbereitung des Materials für eine weiterführende (automatische) Verarbeitung nicht im Mittelpunkt früherer Arbeiten; ausgenommen sind dabei natürlich die zitierten Arbeiten aus jüngster Zeit, die einen expliziten Bezug auf die Anwendung aufweisen.

Das weitere Vorgehen kann man sich idealerweise auf zwei parallelen Schienen vorstellen: In der **Anwendung** konzentriert man sich zuerst auf relativ einfache Phänomene. Das Heraussuchen geeigneter Vertreter dieser 'einfachen' Klassen aus einer auf diese Weise transliterierten Stichprobe ist trivial, vgl. dazu die Beispiele in Teil 5. Schrittweise kann dann auf immer komplexere Phänomene übergegangen werden. In der **Grundlagenforschung** kann anhand des Auftretens der unterschiedlichen Label versucht werden, fundierte statistische Aussagen über die Verteilung und Häufigkeit der einzelnen Phänomene zu machen; ebenfalls kann versucht werden, die reine Taxonomie in ein mehr oder minder explizites Regelwerk mit constraints umzuformulieren, vgl. dazu z.B. [Lev83] und [Lev89, S. 460]. Diese Erkenntnisse können wiederum in die Anwendung eingehen, wo man sich zumindest in einer ersten Phase auf die Phänomene beschränken sollte, die häufig genug auftreten.

[13] Man beachte, daß damit natürlich nicht alle spontansprachlichen Phänomene abgedeckt sind, insbesondere nicht alle syntaktischen Besonderheiten, sondern nur diejenigen, die im Zusammenhang mit einer Unterbrechung des Satzgefüges auftreten; zu spontansprachlichen syntaktischen Besonderheiten vgl. [Tro94].

7 Literatur

[BDSP93] J. Bear, J. Dowding, E. Shriberg, P. Price: *A System for Labeling Self-Repairs in Speech*, Technical Note 522, SRI International, Februar 22 1993.

[B.J92] B.J. Baars (Hrsg.): *Experimental Slips and Human Error*, Plenum Press, New York and London, 1992.

[BKK*93] A. Batliner, R. Kompe, A. Kießling, E. Nöth, H. Niemann, U. Kilian: *The prosodic marking of accents and phrase boundaries: expectations and results*, in *Proc. NATO ASI Conference "New Advances and Trends in Speech Recognition and Coding"*, Bd. 2, Bubion, 1993, S. 89–92.

[BKN93] A. Batliner, A. Kießling, E. Nöth: *Die prosodische Markierung des Satzmodus in der Spontansprache – Methodologie und erste Ergebnisse*, ASL-Süd–TR–14–93/LMU, Februar 1993.

[Hie81] A. Hieke: *A Content-Processing View of Hesitation Phenomena*, *Language and Speech 24*, 1981, S. 147–160.

[HN93] J. Hirschberg, C. Nakatani: *A Speech-First Model for Repair Identification in Spoken Language Systems*, in *Proc. European Conf. on Speech Communication and Technology*, Bd. 2, Berlin, September 1993, S. 1173–1176.

[IPD94] IPDS (Kiel), IPSK (München), IKP (Bonn), IfN (Braunschweig): *Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL, V1.2*, Januar 1994.

[KL91] W. Kindt, U. Laubenstein: *Reparaturen und Koordinationskonstruktionen. Ein Beitrag zur Strukturanalyse des gesprochenen Deutsch. Textteil. Kolibri*, 1991.

[LB92] R. Lickley, E. Bard: *Processing Disfluent Speech: Recognizing Disfluency Before Lexical Access*, in *Int. Conf. on Spoken Language Processing*, Bd. 2, Banff, 1992, S. 935–938.

[Lev83] W. Levelt: *Monitoring and self-repair in speech*, *Cognition*, Bd. 14, 1983, S. 41–104.

[Lev89] W. Levelt: *Speaking: From Intention to Articulation*, MIT Press, Cambridge, Massachusetts, 1989.

[NP94] E. Nöth, B. Plannerer: *Schnittstellendefinition für den Worthypothesengraphen, Verbmobil–Memo–2–94*, Januar 1994.

[O'S92a] D. O'Shaughnessy: *Analysis of False Starts in Spontaneous Speech*, in *Int. Conf. on Spoken Language Processing*, Bd. 2, Banff, 1992, S. 931–934.

[O'S92b] D. O'Shaughnessy: *Recognition of Hesitations in Spontaneous Speech*, in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Bd. 1, San Francisco, 1992, S. 521–524.

[O'S93a] D. O'Shaughnessy: *Analysis and Automatic Recognition of False Starts in Spontaneous Speech*, in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Bd. 2, Minneapolis, USA, 1993, S. 724–727.

[O'S93b] D. O'Shaughnessy: *Locating Disfluencies in Spontaneous Speech: An Acoustic Analysis*, in *Proc. European Conf. on Speech Communication and Technology*, Bd. 3, Berlin, September 1993, S. 2187–2190.

[SB91] S. Schachtl, H. U. Block: *Syntaktische Beschreibung in Systemen zur Verarbeitung gesprochener Sprache*, ASL–TR–11–91/SIM, September 1991.

[SBD92] E. Shriberg, J. Bear, J. Dowding: *Automatic Detection and Correction of Repairs in Human-Computer Dialog*, in *DARPA Speech and Natural Language Workshop*, Arden House, N.Y., 1992, S. 6 Seiten.

[SL92] E. Shriberg, R. Lickley: *Intonation of Clause-Internal Filled Pauses*, in *Int. Conf. on Spoken Language Processing*, Bd. 2, Banff, 1992, S. 991–994.

[Tro94] H. Tropf: *Spontansprachliche syntaktische Phänomene: Analyse eines Korpus aus der Domäne "Terminabsprache"*, Januar 1994, Manuskript, Siemens AG, ZFE ST SN 54, München.

[Wie92] N. Wiedenmann: *Versprecher und die Versuche zu ihrer Erklärung. Ein Literaturüberblick*, Wissenschaftlicher Verlag, Trier, 1992.

[Wil88] E. Willkop: *Gliederungspartikeln im Dialog*, Iudicium Verlag, München, 1988.

8 Anhang 1: Das zugrundeliegende spontansprachliche Korpus

Die Sprecher des Korpus, aus dem die Beispiele genommen sind, waren vier Studenten (3 weibliche: C, X und A, ein männlicher: F), die jeweils paarweise an einer Sitzung teilnahmen; die Sprecher innerhalb eines Paares (C und X bzw. A und F) waren miteinander befreundet. Die zwei Sprecher saßen sich ohne Blickkontakt in einem Versuchsraum des Psychologischen Instituts in München gegenüber und gaben sich gegenseitig Anweisungen, was der Partner mit auf dem Tisch aufgebauten Klötzchen machen sollte. Die Sitzungen dauerten, inkl. Pausen, ca. zwei Stunden. Die Aufgaben waren so angelegt, daß sich kurze Klärungsdialoge mit häufigem Sprecherwechsel ergaben, nicht längere, raisonierende Passagen o.ä. Im Gegensatz etwa zu einem erzählenden Monolog oder zu einem freien Vortrag mit längeren Planungspausen ergab sich dabei eine "echt" spontane, lebhafte Unterhaltung, ohne daß den Versuchspersonen bewußt war, daß ihre Sprache und nicht, wie ihnen gesagt wurde, ihr kooperatives Verhalten untersucht wurde. Die gesamten Dialoge wurden transliteriert; für die weitere Verarbeitung berücksichtigten wir grundsätzlich nur Äußerungen, die eine genügend gute Signalqualität aufwiesen; diese wurden mit 12Bit Auflösung und einer Abtastfrequenz von 10 kHz digitalisiert. Für die Untersuchung außergrammatischer Phänomene wurde ein Teil-Korpus erstellt, dem zwei von den jeweils drei Spieldurchgängen zugrunde liegen. (Für andere Fragestellungen wurden nur Äußerungen ohne außergrammatische Phänomene ausgewählt, vgl. [BKN93].) Füllwörter (Verzögerungsmarkierer und Korrektureinleiter) wurden mit Kontext digitalisiert und gelabelt. Die Beispiele für komplexere Korrekturen wurden anhand der orthographischen Transliteration ausgewählt.

9 Anhang 2: Abgleichung mit dem Handbuch zur Basistransliteration

Der folgende Text ist, wie das Handbuch zur Basistransliteration, ein reiner ASCII-Text. Er enthält wort-wörtlich die relevanten Passagen aus dem Handbuch und ermöglicht so einen schnellen und einfachen Vergleich. (Anführungszeichen und Umlaute sind in TEX-Konvention übergeführt.) Er enthält darüber hinaus zwei unterschiedliche Arten von Kommentaren:

- metasprachliche Kommentare werden am Zeilenanfang eingeleitet mit: %comment
- In unserem Beschreibungssystem überführte objektsprachliche Transliterationen werden am Zeilenanfang eingeleitet mit: %modif; sie stehen unter den entsprechenden Zeilen aus dem Handbuch

```
3.6 Konventionen zu Interjektionen

Interjektionen wie ``au'', ``ah'', ``oh'', ``aha'', ``mhm'' (Bejahung),
``mm'' (Verneinung) werden ohne Zus"atze in der hier
verschrifteten Form in den orthographischen Text eingef"ugt. Die
meisten von ihnen sind bereits im Duden aufgelistet.

%comment Analog dazu werden in der Worterkennung daf"ur
%comment Wortmodelle aufgestellt.

3.7 Konventionen hinsichtlich nicht-lexikalischer Einheiten

In die standardisierte Orthographie werden folgende Zus"atze
zwischen < > ohne Leerstellen eingef"ugt:

(1) Benennungen nonverbaler Produktionen der Sprecher
    - H"asitationen (vokalische und/oder nasale Artikulationen,
      die den fl"ussigen Wortstrom unterbrechen): Symbolisierung
      durch Vokalzeichen + h/m oder durch (h)m.

      Beispiele:
      <"am>, <"ahm>, <"ah>, <"om> etc., <em> etc., <am> etc.,
      <hm>,  <m>
      - Lachen, R"auspern, Husten, Schmatzen etc. werden als
      solche benannt.

%comment Die Bezeichnung ``nonverbale Produktionen''
```

```
%comment f"ur Verz"ogerungsph"anomene wie <"ah> halten wir
%comment f"ur etwas ungl"ucklich, da diese Produktionen immer auch
%comment diskurssteuernde Funktion haben und i.a. den phonotaktischen
%comment Regeln des Deutschen folgen; dar"uberhinaus sind ``genuine''
%comment H"asitationen und Korrektureinleiter oft homophon, auch
%comment wenn sie sich in ihrer Prosodie unterscheiden.
%comment Alle Korrektureinleiter m"ussen explizit, d.h. in
%comment spitzen Klammern, gekennzeichnet werden, da sie
%comment genauso wie die korrigierte Passage herausgefiltert
%comment werden m"ussen. Beispiele sind etwa: <das hei"st>, <nee>.
%comment Eine explizite Unterscheidung der verschiedenen Arten ist
%comment u.E. nicht n"otig, kann aber gegebenenfalls durch
%comment eine unterschiedliche Art der Klammerung erreicht werden.

      Beispiele:
      <Lachen>, <R"auspern>, <Husten>, <Schmatzen>.


      - Z"ogernde Dehnung der vorangehenden Artikulation
        (St"orung des fl"ussigen Sprechablaufs):
        <Z"ogern>, abgek"urzt <Z>
        -- <Z> bezieht sich auf ein ganzes Wort bzw. seinen
           finalen Teil und wird mit Leerstelle dahinter
           gesetzt,
        -- <Z> bezieht sich auf das Wortinnere und wird hinter
           den gedehnten Teil ohne Leerstelle gesetzt,
      - Atmen wird ebenfalls als solches benannt und schlie"st
        h"aufig vorangehende und/oder nachfolgende Pause ein:
        <Atmen>, abgek"urzt <A>
      - Pausen werden durch <Pause>, abgek"urzt <P>, bezeichnet,
        wenn sie im Innern eines Sprecherbeitrags vorkommen und
        nicht mit <Atmen> gekoppelt sind.

(... )

3.8 Abbr"uche, Unterbrechungen, Wiederaufnahmen, Korrekturen

Wenn der Sprechvorgang vor Erreichen eines Abschlusses
abgebrochen wird, dann gibt es vier M"oglichkeiten der
Kontinuit"atsst"orung:

(1) Der Sprechakt wird nicht fortgesetzt, es wird m"oglicher-
    weise ein neuer Sprechakt zu einem neuen Gedanken begonnen.
    Hier handelt es sich um einen Abbruch.
%comment In unserem Labelsystem entspricht diesem Abbruch
```

%comment der ‘‘Syntaxabbruch’’. ‘‘Neuer Gedanke’’ ist ein vager
%comment Begriff und sollte durch ‘‘syntaktischer Neuansatz’’
%comment ersetzt werden. Nur wenn die Syntax sich "andert, handelt
%comment es sich um einen Syntaxabbruch mit folgendem Neuansatz,
%comment in den anderen F"allen um eine Korrektur.

(2) Der Sprechakt wird nach einer Pause und/oder einer H"asitation ohne Wiederholung oder Korrektur fortgesetzt. Hier handelt es sich um eine Unterbrechung.
%comment Wir sprechen hier von einer Verz"ogerung ohne Wiederholung,
%comment also ohne das Label ‘‘@M’’.

(3) Der Sprechakt wird nach dem Abbruch durch einfache Wiederaufnahme (eines Teils) des bereits Gesagten fortgesetzt. Hier handelt es sich um Wiederaufnahme.
%comment Wir sprechen hier von einer Verz"ogerung mit Wiederholung,
%comment also mit dem Label ‘‘@M’’. ‘‘Wiederaufnahme’’ scheint
%comment ein vager Begriff zu sein, da auch ‘‘gedankliche
%comment Wiederaufnahme’’ gemeint sein kann. Wir pl"adieren f"ur
%comment eine strikte Begrenzung auf ‘‘Wortwiederholung’’.

(4) Der Sprechakt wird durch korrigierende Wiederaufnahme fortgesetzt. Hier handelt es sich um Korrektur.
%comment Dieser Fall wird auch bei uns als Korrektur bezeichnet.

In allen diesen F"allen enth"alt das Sprachsignal neben dem syntaktisch-semantischen Abbrechen auch phonetische Indikatoren einer St"orung des Sprechflusses. Dies ist vor allem f"ur die F"alle (3) und (4) von Bedeutung. Einfache Wiederholung, auch mit modifizierendem Wiederaufgreifen des Gesagten ist zwar eine notwendige, aber noch keine hinreichende Bedingung f"ur die hier festgelegten Transliterationskategorien; denn in einer Dialogf"uhrung kann ein Sprecher einen einmal ausgesprochenen Gedanken entweder insistierend wiederholen oder ihn anschlie"send verwerfen, abwandeln, spezifizieren, ohne da"s dadurch die charakteristische St"orung sprachlicher Kontinuit"at entst"unde. Entscheidend ist, da"s es sich hier um eine gest"orte Ausf"uhrung eines Sprechplans handelt, w"ahrend in den anderen F"allen ein ungest"orter Plan wiederholt oder durch einen weiteren erg"anzt, korrigiert, verworfen wird.

So handelt es sich in der "Au"serung ‘‘ich m"ochte Sie um vier Uhr treffen <Pause> nein, ich m"ochte, da"s Sie mich um vier Uhr treffen’’ nur dann um eine Korrektur in der hier vorgenommenen Festlegung, wenn der Abbruch nach dem ersten Auftreten von

''treffen'' phonetisch durch Prosodie signalisiert ist, so da"s
vor <Pause> kein Punkt gesetzt werden k"onnte. Andernfalls
liegen zwei wohlgeformte S"atze vor, von denen der zweite den
zuerst vorgetragenen Gedanken modifiziert. Diese Sequenz ist
vergleichbar der inhaltlich korrespondierenden Formulierung ''ich
m"ochte Sie um vier Uhr treffen. Wenn ich mir das jetzt nochmal
"uberlege, w"are es mir lieber, wenn Sie mich um vier Uhr
treffen w"urden.''

Zur Symbolisierung dieser Kategorien gilt folgendes:

(1) Abbruch:
    An der Abbruchstelle wird ohne Zwischenraum /- eingef"ugt;
    hinter /- steht kein Interpunktionszeichen; z.B.
INH067: im Dezem/- <P> ach so , wir brauchen ja noch
%modif im Dezem--/- <P> ach so , wir brauchen ja noch
        einen im November.
%comment Die von uns eingef"uhrte Unterscheidung zwischen
%comment Wortabbruch (doppelter Bindestrich zusammen mit {/-, /+, /=})
%comment und Phrasenabbruch (ohne vorangehende Bindestriche)
%comment (Schr"agstrich mit Bindestrich) erm"oglicht es z.B.,
%comment gezielt nur nach Wortabbr"uchen zu suchen.

INH0107: <A> gut . dann/-

(2) Unterbrechung:
    Der Abbruch wird nicht gesondert markiert; er ist durch das
    Vorhandensein von <P> und/oder H"asitation bereits eindeutig
    indiziert; z.B.
    im Wort:
    Lebens<P>mittel
    an Wortgrenzen:
THW138: dann <P> in der zweiten H"alfte <P>
        <hm> <A> , also <P> ich kann vom dreizehnten <P> bis
        zum <Z> f"unfundzwanzigsten .

(3) Wiederaufnahme:
    An der Abbruchstelle wird ohne Zwischenraum /= angef"ugt; z.B.
    =/f<Z>/=f"ahrt
%modif =/@M1f<Z>--/= @M1f"ahrt
INH027: <A> und dann , wie w"ar's
        =/um <P>/= um den f"unfzehnten rum <P> ,...
%modif =/@M1um/= <P> @M1um den f"unfzehnten rum/& <P> ,..

```
(4) Korrektur:
    An der Abbruchstelle wird ohne Zwischenraum /+ angef"ugt;
INH027: Maria Himmel+/f"ahr/+fahrt
%modif  Maria Himmel+/@R1f"ahr--/+@R1fahrt
        oder %was <A> .
THW022: ja , dann <P> k"onnen wir ja wieder ein/- <A>
        +/vielleicht nehmen wir der *Einfachhalt/+ <Lachen>
%modif  +/@M1vielleicht @M2nehmen @M3wir @M4der @R1*Einfachhalt/+
%modif  <Lachen>
        vielleicht nehmen wir der Einfachheit halber den <Z>
%modif  @M1vielleicht @M2nehmen @M3wir @M4der @R1Einfachheit
%modif  halber/& den <Z> dritten ?

INH061: erst so +/ab achten/+ <P> acht--/+ <P> zwischen dem
        achten und dem achtzehnten <A> .
%modif  erst so +/@D1ab @M1achten/+ <P> @M1acht--/+ <P> @I1zwischen
%modif  @I1dem @M1achten und dem achtzehnten <A> .
```

In den Typen (3) und (4) wird durch /= bzw. /+ die Stelle des Abbruchs markiert, die (korrigierende) Wiederaufnahme erfolgt mit der n"achsten lexikalischen Einheit.
Gleichzeitig wird durch =/ bzw. +/ die linke Grenze des Sprechst"ucks markiert, das durch die Wiederholung ersetzt wird und das bei einem Herausfiltern des ''Rauschens'' getilgt werden kann. Diese Markierung wird durchgef"uhrt, obwohl sie in vielen F"allen arbitr"ar ist. Im Falle mehrfacher Abbr"uche und (korrigierender) Wiederaufnahmen vor dem endg"ultigen Abschlie"sen eines Satzes wird die linke Grenze des ganzen Abbruchkomplexes einmal durch =/ bzw. +/ gekennzeichnet, jede Abbruchstelle im Komplex mit /= bzw. /+. Ein sp"aterer Filtervorgang bezieht dann alles zwischen =/ bzw. +/ und dem letzten nachfolgenden /= bzw. /+ ein.
%comment Bei solchen komplexen F"allen, die eventuell eingebettete
%comment St"orungen enthalten und die iterativ abgearbeitet werden
%comment m"u"sen, kann eine Labelung auf Wortebene Schwierigkeiten
%comment bereiten.

In den F"allen (1) und (2) kann eine solche Ausklammerung nicht vorgenommen werden, da ja keine Wiederholung eintritt, die Hinweise auf eine Textgl"attung geben k"onnte. Im Fall (2) k"onnen einfach <P> und H"asitationen herausgefiltert werden. Im Fall (1) m"ussen Linguisten, die bereinigte Texte f"ur ihre Fragestellungen w"unschen, selbst festlegen, wieviel sie von der abgebrochenen "Au"serung wegfiltern wollen. Diese Entscheidung kann nicht bei der Erstellung der transliterierten Basisdateien

gef"allt werden. Die gesonderte Markierung der Abbr"uche durch
/- macht es Linguisten sehr einfach, die relevanten Textstellen
schnell herauszufinden und im Hinblick auf die jeweiligen
Fragestellungen vorzuverarbeiten.

Da in den meisten F"allen nicht eindeutig entschieden werden kann, ob vor
dem Abbruch ein Wort komplett oder nur teilweise ge"au"sert wurde, werden
die Symbole /-, /=, bzw. /+ ohne Leerstelle an der Abbruchstelle
eingef"ugt. Unmittelbar vor /-, /= bzw. /+ d"urfen nur <Z> oder <A>, wenn
dieses eindeutig als Ausatmen erkennbar ist, stehen; H"asitationen, Pausen,
Lachen usw. an der Abbruchstelle werden nach /-, /= bzw. /+ mit einer
Leerstelle angef"ugt. Wird <Z> vor /-, /= bzw. /+ gesetzt, ist darauf zu
achten, das <Z> ohne Leerstelle an die vorangehende Zeichenfolge angef"ugt
wird.
%comment Wir nehmen eher an, da"s in den meisten F"allen schon entschieden
%comment werden kann, ob ein Wort nur teilweise ge"au"sert wurde;
%comment eindeutiges Kriterium daf"ur ist, wenn die Zeichenfolge
%comment nicht als Wort im Lexikon verzeichnet ist. Da die Modellierung
%comment von abgebrochenen W"ortern ein noch nicht gel"ostes Problem
%comment darstellt, sollten diese auch schon in der Basistransliteration
%comment gelabelt werden. Bei Zweifelsf"allen m"u"ste dann entweder
%comment einfach eine Alternative gew"ahlt werden, oder diese
%comment Zweifelsf"alle werden mit einem zus"atzlichen Label als
%comment solche gekennzeichnet.

IV. Filterung der Basisdateien f"ur die Weiterverarbeitung

Die hier vorgelegten Transliterationskonventionen erlauben die
Entwicklung eines zentralen Filters mit genau festgelegten
Schalteroptionen f"ur die Erstellung abgeleiteter Korpora wie
Lexika zu den Dialogaufnahmen, Aussprachew"orterb"ucher,
Language-Modell-Korpus, Trainingskorpus, Testkorpus,
Referenzkorpus, Lexikon non-verbaler Dialogsteuerungsmittel mit
den Belegstellen etc.. So k"onnen beispielsweise alle < >-Teile
weggefiltert werden, um nur das lexikalische Material weiter zu
bearbeiten, gegebenenfalls auch noch die in =/ /= bzw. +/ /+
eingeschlossenen Stellen (auch_: :_), oder aber es wird unter
den < >-Markierungen differenziert und z.B. nur <; > und <# >
ausgesondert. Die Struktur eines solchen Filters und vor allem
der Umfang seiner Applikationsbereiche m"ussen noch festgelegt
werden. An seiner Entwicklung, Implementierung und Dokumentation
wird TP14 zentral beteiligt sein, da in TP14 der "uberwiegende
Teil der Datensammlung und -verwaltung erfolgt.

```
%comment Die vorliegende Arbeit ist also auch als Vorschlag f"ur einen
%comment Filter zu verstehen, der eine Doppelfunktion erf"ullt:
%comment 1.Bereinigung durch Wegfilterung der au"sergrammatischen
%comment Elemente; notwendig dazu ist eine leichte Modifikation
%comment der Basistransliteration
%comment 2.Explizite Labelung der au"sergrammatischen Teile f"ur
%comment linguistische Untersuchungen, Trainingskorpara, etc.
```

VI. Beispiel eines transliterierten Dialogs

```
;Dialog: G072
;Zuletzt bearbeitet am: 15.10.93

HAH000: ja , sch"onen guten Tag , dann meld' ich mich noch ein-
        mal . und zwar wollt' ich gerne mit Ihnen <P> zwei Termine im
        Mai vereinbaren <A> und auch gleich einen Vorschlag machen . und
        zwar w"urde mir sehr gut passen Anfang Mai gleich <P> am vierten
        das ist ein Mittwoch .

TIS001: <A> ja , da bin ich also voll mit einverstanden , da m<;T>

HAH002: entschuldigen Sie bitte , ich hab' Sie jetzt nicht ganz
        <Z> verstehen k"onnen .

TIS003: <"ahm> <P> =/ha/= ha/- verstehen Sie mich jetzt ? <P>
%modif  <"ahm> <P> =/@M1ha--/= @M1ha--/- verstehen Sie mich jetzt ? <P>
        also , <"ahm> ich wollt' gerade sagen , der Mittwoch ,
        der +/ist <%>/+ pa"st mir ausgezeichnet , und das k"onnen
%modif  &/der +/@R1ist /+ @R1pa"st mir ausgezeichnet/& , und das k"onnen
        wir ja dann <P> schon mal festklopfen . also ,
        <"ahm> Mittwoch ist Ihnen recht , wenn ich zu Ihnen komme ?
HAH004: das ist mir sehr recht , ja . eigentlich fast lieber ,
        als wenn ich %zu %Ihnen %mu"s , und <Z> wegen der Uhrzeit ,
        vielleicht auch wieder um <P> acht Uhr ?

TIS005: <A> ja , ich denke , acht Uhr ist gut .

HAH006: wunderbar , vielen Dank , sind wir uns da ja <Z> schon
        einig  <A> .

TIS007: ja , danke auch . dann denk' ich , sollten wir gleich
        mal den zweiten im Mai <Z> uns vornehmen . <%> <"ahm> wie sieht
        das aus <Z> in der Woche vor Pfingsten ?
```

HAH008: <"ah> kleinen Augenblick . <A>
+/das pa"st mir/+ das ist mir sehr unpassend
%modif +/@M1das @R1pa"st @M2mir/+ @M1das @R1ist @M2mir sehr unpassend/&
, nein , da mu"s ich zu einem Besuch nach Leipzig ,
=/das<Z>/= das ist leider nicht zu machen .
%modif =/@M1das<Z>/= @M1das ist leider nicht zu machen .
ich k"onnte eher vorschlagen , <A> direkt nach Muttertag , am
Dienstag , dem zehnten , das w"are dann <P> eine Woche sp"ater
. quasi .

TIS009: eine Woche sp"ater , meinen Sie jetzt , also <"ahm> ich hab'
jedenfalls die Woche vom <Z> f"unften bis zum zw"olften des<Z>/-
tut mir leid , da kann ich "uberhaupt nicht . <A> <"ahm>
%modif eine Woche sp"ater , meinen Sie jetzt , -/<also> <"ahm> ich hab'
%modif jedenfalls die Woche vom <Z> @A1f"unften @A1bis
%modif @A1zum @A1zw"olften des<Z>/-
%modif tut mir leid , @A1da kann ich "uberhaupt nicht . <A> <"ahm>
ich dachte jetzt an
die <"ahm> +/Woche nach/+ <"ah> Woche vor Pfingsten .
%modif &/die <"ahm> +/@M1Woche @R1nach/+ <"ah> @M1Woche @R1vor
Pfingsten/&. aber <"ahm> sonst ginge auch bei mir noch bis zum
Mittwoch , dem f"unfundzwanzigsten .

HAH010: ja . nee , das hab' ich ja gesagt , die Woche vor
Pfingsten pa"ste bei mir leider nicht <;stark verschliffen> ,
denn <P>
+/bis zum vierundzwanzig/+ bis zum f"unfundzwanzigsten/+
%modif +/@M1bis @M2zum @R1vierundzwanzig--/+ @M1bis @M2zum
%modif @R1f"unfundzwanzigsten/+
<"ahm> vierundzwanzigsten
%modif <"ahm> @R1vierundzwanzigsten ,
Dienstag , den vierundzwanzigsten Mai , das w"are mir recht .
wenn Ihnen das da auch passen w"urde ?

TIS011: das w"urde mir ganz ausgezeichnet passen .
=/das/= also das k"onnen wir dann <Z> festhalten ,
%modif =/@M1das/= @I1also @M1das k"onnen wir dann <Z> festhalten ,
%comment sollte das 'also' prosodisch abgesetzt sein, m"u"ste es mit
%comment spitzen Klammern als gef"ullte Pause markiert sein.
denke ich . kommen Sie diesmal zu mir dann ?

HAH012: ja , nat"urlich . das mach' ich . und dann auch wieder
so um acht Uhr , w"urd' ich vorschlagen .

TIS013: ja , pa"st mir wunderbar .

HAH014: ja , okay , dann <Z> k"onnten wir uns vielleicht auch
noch gleich einig werden wegen eines dritten Termins , im Juni .
und wenn ich Ihnen da gleich einen Vorschlag machen darf. <P>
mir w"urd's sehr gut passen <Z> am Dienstag zum Beispiel , dem
vierzehnten Juni . wenn Ihnen das da auch passen w"urde ,
vielleicht ?

TIS015: das w"urd' mir leider <P> gar nicht gut passen , weil
ich da <Z> unterwegs bin nach Braunschweig , dienstlich <A> ,
aber <"ahm> wo ich im Juni Zeit h"atte , ich kann Ihnen das ja
mal sagen , w"are <Z> Samstag , den achtzehnten bis Donnerstag ,
den dreiundzwanzigsten , und dann wieder ab <Z> Montag , dem
siebenundzwanzigsten bis Ende des Monats . vielleicht haben wir
da irgendwann Zeit ?

HAH016: ja , also , mir pa"st sowohl am Samstag , dem acht-
zehnten , als auch am <Z> Mittwoch , dem neunundzwanzigsten ,
oder Donnerstag , dem drei"sigsten , das <Z> "uberlasse ich dann
auch gerne Ihnen , welcher <P> Termin Ihnen da am liebsten ist .

TIS017: ich mein' , wenn Ihnen das recht ist , dann k"onnen wir
ja gleich sagen , dann machen wir 's am Samstag , dem acht-
zehnten , dann haben wir das vom Tisch , und dann <Z> komm' ich
bei Ihnen vorbei .

HAH018: ja , nat"urlich . wunderbar . kommen Sie zu mir , wieder
<P> acht Uhr , schlag' ich vor , %dann h"atten wir das auch .

TIS019: gut , bin ich mit einverstanden , dann ist das klar .

HAH020: danke sch"on <A> .