BMBF

Verbmobil
Verbundvorhaben

Partiturformat für die Darstellung unterschiedlicher Repräsentationsebenen von gesprochener Sprache

S. Atmanspacher
S.Burger
Chr. Draxler
A. Kipp
Chr. Scheer
F. Schiel
M.-B. Wesenick

Ludwig-Maximilians-Universität München

Vm

MEMO 90
September 1995

September 1995

Stephan Atmanspacher

Susanne Burger

Christoph Draxler

Andreas Kipp

Christian Scheer

Florian Schiel

Maria-Barbara Wesenick

Institut für Phonetik und sprachliche Kommunikation

Ludwig-Maximilians-Universität München

Schellingstraße 3/II

80799 München

Tel.: (089) 2180 - 2807

e-mail: schiel@sun1.phonetik.uni-muenchen.de

Gehört zum Antragsabschnitt: 14 VERBMOBIL/PHONDAT

Die vorliegende Arbei t wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 102 L/4 gefördert. Die Verantwortung für den Inhalt liegt bei den Autoren.

Inhaltsverzeichnis

1 Einführung - Das Partiturkonzept 4

2 Repräsentationsebenen 4

2.1 Vorschlagstranskription 4

2.2 Transliteration 5

2.3 Lautsegmentation 5

2.4 Wortsegmentation 6

2.5 Artikulatorische Geräusche 6

2.6 Nicht-artikulatorische Geräusche 6

2.7 Prosodie 6

3 Formate 6

3.1 SAM-Format 7

3.2 Partitur-Format 8

3.2.1 Vorschlagstranskription 8

3.2.2 Transliteration 8

3.2.3 Lautsegmentation 8

3.2.4 Wortsegmentation 9

3.2.5 Artikulatorische Geräusche 10

3.2.6 Nicht-artikulatorische Geräusche 10

3.2.7 Prosodie 10

4 Literatur 11

1 Einführung - Das Partiturkonzept

Nachfolgend stellen wir ein neues Konzept für die Darstellung der unterschiedlichen Repräsentationen von Sprachsignalen vor. Im Rahmen des Verbmobil-Projekts fallen bei den verschiedenen Partnern *(*symbolische*)* Sprachdaten an, die mit unterschiedlichen Strategien und Intentionen bei der Analyse von Sprachsignalen entstehen. Dazu gehören beispielsweise die Segmentation und Transkription des Sprachsignals, Transliteration und prosodische Etikettierung. Diese verschiedenartigen Daten mit unterschiedlichen Informationen über dasselbe Sprachmaterial sollen am Computer dargestellt und weitere Analysen mit Hilfe von Bearbeitungsfunktionen ermöglicht werden.

Es soll möglich sein, sowohl weniger umfangreiche, gezielt aufgenommene Sprachdaten individuell zu bearbeitet, als auch sehr große Datenmengen voll- oder halbautomatisch zu segmentieren und etikettieren. Darüberhinaus soll die Möglichkeit bestehen, die verschiedenen Beschreibungsebenen unabhängig voneinander darzustellen und zu bearbeiten, wobei jedoch bestehende symbolische oder zeitliche Relationen zwischen den Ebenen übersichtlich in Form einer Partitur deutlich werden.

Durch das Partitur-Format werden unterschiedliche Daten, die zum selben Sprachsignal gehören, vereint und zueinander in Beziehung gesetzt. Auf dem Bildschirm werden sie in mehreren Ebenen mit mehr oder weniger direktem Bezug zu einer Zeitachse übereinander angeordet. Zwischen den Ebenen bestehen zeitliche und/oder symbolische Bezüge. Jede Ebene ist eindeutig identifizierbar und ist in Wortschatz und Syntax eigenständig. Bei Dialogdaten, die auf mehreren getrennten Kanälen aufgenommen werden, ist pro Kanal eine eigene Partitur vorgesehen.

Die Partitur soll einfache Erweiterungen um andere Ebenen zulassen, ohne daß bestehende Anwendersoftware geändert werden muß.

Neben der Darstellung durch ein Oberflächenprogramm, soll das interne Format der Daten so gewählt werden, daß es an sich gut lesbar ist. Dadurch ist der Benutzer selbst von der Software für das Oberflächenprogramm unabhängig. Die Partitur bleibt auch im ASCII-Format lesbar und interpretierbar.

2 Repräsentationsebenen

Im folgenden werden die einzelnen Ebenen der Partitur beschrieben, für die zur Zeit entsprechende Daten vorliegen. Es wird erklärt, was jede Ebene beinhaltet und wie die jeweiligen Daten zustande kommen, in welchem Format sie vorliegen und welche Funktion sie im Partiturkonzept erfüllen bzw. wie sie mit anderen Ebenen in Beziehung stehen.

2.1 Vorschlagstranskription

Die Vorschlagstranskription umfaßt die kanonische Form von regulären Wörtern, sowie Referenztranskriptionen von abgebrochenen Wörtern, Unwörtern und Neuschöpfungen und ist nach den Konventionen des erweiterten SAM - phonetischen Alphabets formuliert[1].

Die kanonischen Formen von regulären Wörtern können einem gängigen SAMPA- Aussprachelexikon entnommen werden (PhonDat, Verbmobil-Lexika, Stock, CELEX, etc.). Die ”kanonischen Formen” von Wortabbrüchen, Unwörtern und Wortneuschöpfungen, wie sie in

1. Die im PhonDat-Projekt verwendete Version von SAMPA für das Deutsche ist abgedruckt in [2] Pompino-Marschall (Hrsg.) (1992)

Spontansprache vorkommen, müssen soweit es geht aus den kanonischen Formen von regulären Wörtern abgeleitet oder neu formuliert werden.

Bei der Segmentation relativ zu einer Referenzform[2] wird dem Transkribierer die sogenannte Vorschlagstranskription angeboten und kann von diesem als zutreffend akzeptiert oder verändert werden. Die Vorschlagstranskription dient bei allen realisationsphonetischen Untersuchungen als Referenz. Phonetische Prozesse in fließender Rede werden in Bezug auf die kanonischen Formen der Wörter beschrieben.

Im Partiturkonzept hat die Vorschlagstranskription die Funktion der Referenzebene, an der sich symbolische Bezüge zu anderen Ebenen festmachen. Neben den einzelnen Worttranskriptionen enthält die Ebene der Vorschlagstranskription eine Wortnumerierung der einzelnen Wörter in der Reihenfolge ihres Auftretens in der jeweiligen Äußerung. Über diese Numerierung werden Beziehungen zu anderen Ebenen verankert. Symbolische Bezüge können über die Wortnummer verfolgt werden, Zeitbezüge über ein grobes Alignment von Wortsegmentationen, bzw. über die Ableitung von Wortgrenzen über die Lautsegmentation.

2.2 Transliteration

Die Transliteration repräsentiert Sprachmaterial in orthographischer Form mit Satzzeichen und Zusatzinformationen, u.a. über artikulatorische Geräusche (Atmen, Räuspern) und nicht-artikulatorische Geräusche, sowie über spontansprachliche Phänomene (Abbrüche, Wiederholungen). Die Transliterationen werden aufgrund festgelegter Konventionen erstellt, haben ein eigenes Notationssystem mit eigenem Symbolinventar und eigener Syntax[3].

Die unterschiedlichen Arten von Zusatzinformationen sind eindeutig gekennzeichnet und identifizierbar und somit auch computerlesbar. Das macht es möglich, aus der Transliteration andere Ebenen abzuleiten und separat darzustellen.

Die Bestandteile der Transliteration, die sprachliche Äußerungung bezeichnen[4] werden in der Reihenfolge ihres Auftretens in der jeweiligen Äußerung numeriert, so daß eine Relation zur Vorschlagstranskription besteht. Symbolische Bezüge können über die Wortnummer auch zu anderen Ebenen hergestellt werden. Durch ein grobes automatisches Time-Alignment mit der Wortsegmentation oder über die Ableitung von Wortgrenzen aus einer Lautsegmentation bestehen zeitliche Relationen der zeitkonsumierenden Bestandteile zu anderen Ebenen. Die übrigen Inhalte bleiben ohne Zeitbezug.

2.3 Lautsegmentation

Die Ebene der Lautsegmentation enthält Daten über die zeitlichen Grenzen von identifizierten Lauten im Sprachsignal (Anfangs- und Endsample) und ein entsprechendes Label aus einem definierten Inventar. Segmentiert werden alle sprachlichen Äußerungen.

Der Benutzer soll die Möglichkeit haben, nach unterschiedlichen Strategien zu segmentieren, wobei die einzelnen Strategien und Segmentationskonventionen genau definiert sind.

Es soll möglich sein, wenn nötig definierte Markierungen einzuführen, sofern dafür in gängigen phonetischen Alphabeten keine Zeichen vorgesehen sind (z.B. Pausenklassen). In vorhandenen Korpora existieren Segmentationen sowohl mit überlappenden (WD-Material) als auch mit

2. Diese Segmentationsstrategie wurde im PhonDat-Projekt verfolgt. Zu den Segmentationskonventionen siehe [2] Pompino-Marshall (Hrsg.)(1992).

3. Zu den Konventionen und Notationssystem der Transliteration siehe:[1] K. Kohler et al. (1994)

4. Darunter werden alle zeitkonsumierenden Abschnitte lautsprachlichen Charakters gefasst: reguläre Wörter, Abbrüche, Wiederholungen, Unwörter, Neuschöpfungen, Häsitationen und Versprecher; nicht jedoch artikulatorische und nicht-artikulatorische Geräusche

bündigen Segmentgrenzen (PhonDat-II). Die verwendeten phonetischen Alphabete sind entweder das gesamte IPA-Inventar (WD) oder erweitertes SAMPA (PhonDat II).

Über die Lautsegmentation erfolgt auf symbolischer Ebene die Identifizierung von tatsächlich realisierten Lauten einer Äußerung (Transkription). Die Laute werden gleichzeitig eindeutig einem bestimmten Abschnitt im Zeitsignal zugeordnet (Segmentation). Um Bezüge zu anderen Ebenen zu ermöglichen, erhält jedes Segment die Nummer des Wortes, zu dem es gehört. Da überlappende Segmentgrenzen zugelassen werden, entsteht gegebenenfalls eine doppelte Wortzugehörigkeit. So wird z.B.eine grobe Zuordung von Wörtern der Transliteration zur Zeitachse möglich oder auch die Zuordnung von Segmenten zu entsprechenden Abschnitten in der Oszillogramm- oder Spektrogrammdarstellung des Sprachsignals.

2.4 Wortsegmentation

Die Ebenen der Wortsegmentation enthält die orthographische Form der Wörter einer Äußerung wie sie in der Transliteration vorliegt, Anfangs- und Endgrenzen von allen zeitkonsumierenden Elementen der Transliteration und für die eindeutige Beziehung zu anderen Ebenen eine Wortnumerierung. Wortgrenzen können überlappend sein.

Durch die Wortsegmentation erfolgt eine eindeutige Zuordnung von Wörtern auf die Zeitachse. Über die Orthographie, bzw. durch die Wortnummer bestehen symbolische Relationen zu anderen Ebenen, z. B. zur Transliteration und Vorschlagstranskription. Wortsegmentationen sind für die zeitliche Zuordnung auch von anderen Ebenen besonders dann nützlich, wenn keine Lautsegmentation vorliegt.

2.5 Artikulatorische Geräusche

Auf der Ebene der artikulatorischen Geräusche sind die als solche gekennzeichneten Elemente der Transliteration separat und völlig unabhängig von jeder anderen Ebene dargestellt.

2.6 Nicht-artikulatorische Geräusche

Auf der Ebene der nicht-artikulatorischen Geräusche sind die als solche gekennzeichneten Elemente der Transliteration separat und völlig unabhängig von jeder anderen Ebene dargestellt.

2.7 Prosodie

Auf der Ebene der Prosodie werden die prosodischen Etikettierungen auf einer Zeitachse dargestellt. Um eine Zuordnung zu Wörtern bzw. Silben deutlich zu machen, werden die zugehörigen grammatischen Einheiten ebenfalls auf dieser Ebene dargestellt. Die Daten dieser Ebene sind entsprechend: Zeitpunkt des prosodischen Etiketts, Anfangs- und Endsample des zugehörigen Wortes oder Silbe.[5]

Über die Nummer des zugehörigen Wortes bestehen symbolische Beziehungen zu den anderen Ebenen, über die Zeitachse weitere zeitliche Bezüge.

3 Formate

Das Format der Partitur wird an die Dateiformate des SAM-Projektes angelehnt. Die Daten der zu einem Sprachsignal gehörigen Partitur befinden sich zusammengefaßt in einer eigenen

5. siehe z.B. M.Reyelt, Anton Batliner [3]

Datei. Diese trägt den gleichen Namen wie die Signaldatei, jedoch mit anderer Extension. Am Anfang jeder Zeile der Datei steht ein dreistelliges Label (gefolgt von Doppelpunkt und white-space), welches den Inhalt und die Syntax derselben Zeile eindeutig festlegt.

Die verschiedenen Ebenen stehen in beliebiger Reihenfolge in einer Datei hintereinander, welche aus einem Header besteht und den verschiedenen Labelbodies der Ebenen mit entsrpechenden Daten.

3.1 SAM-Format

Da sich das Partitur-Format soweit es geht an das SAM-Format anlehnt, soll im folgenden das SAM-Format der Übersichtlichkeit halber kurz beschrieben werden. Im Header sind bestimmte Label für die jeweils nötigen Informationen reserviert. Es müssen nicht alle Label verwendet werden, neue Label können falls nötig definiert werden. Folgende Label sind im SAM Format für Header vorgesehen (die für die Partitur obligatorischen Label sind markiert):

LHD:	**Headername und Version**	Partitur 1.0
FIL:	SAM-Dateityp	
TYP:	Typ des SAM-Labelfiles	
DBN:	Korpusname	Verbmobil
VOL:	Nummer des “volumes”	4.0
DIR:	Directory im “volume”	/m023d
SRC:	Name des Sprachsignalfiles	m023d000.a16
SAM:	sampling rate	16000
BEG:	Anfang der gelabelten Sequenz	
END:	Ende der gelabelten Sequenz	
RED:	Aufnahmedatum	16.03.1995
RET:	Aufnahmedauer	0 22’ 53”
REP:	**Aufnahmeort**	München
SNB:	**Anzahl von bytes pro sample**	2
SBF:	**Numerisches Format**	10
SSB:	**Bits**	16
RCC:	Aufnahmebedingungen (Mikrophone etc)	
NCH:	**Anzahl Kanäle**	1
CMT:	Kommentar	Angaben zur Tonqualität
SPN:	**Sprecherkürzel**	awe
SPI:	Sprecherinformation	m, 1.5.73, 76 kg, Student, Hessen, Bayern
PCF:	Name der Protokolldatei	
PCN:	Protokollnummer	
CMT:	Kommentar	Sprecherbesonderheiten, Sprachfehler
EXP:	Segmentierer	awe
SYS:	Labellingsystem	
DAT:	Datum der Fertigstellung der Labelung	27.05.1995
SPA:	SAMPA Version	
CMT:	Kommentar	

Zwischen der Label-body Markierung LBD: und der End-of-file Markierung ELF: können im SAM-Format die folgenden, weiteren Label auftauchen:

LBR:	Taucht bei Files auf, die automatisch mit EROPEC erzeugt wurden. Auf LBR: können LB2 (zweiter Kanal), LBL (Laryngosignal im zweiten Kanal), LBN (Nasenluftstrom), LBA (Luftstrom), LBT (Zungenkontakt) und weitere beliebig definierbare Label folgen.
LBO:	SAM-Transliteration (mit eingefügten EXT: und CMT: Zeilen, s.u.)
LBB:	breite phonetische Transkription, manuell, halb-automatisch oder automatisch
LBA:	akustisch-phonetische Transkription

LBP:	prosodische Label
EXT:	Zeilenextension
CMT:	Kommentar
DSC:	diskontinuierlicher Aufnahmemodus (automatisches Einfügen von Pausen zwischen Äußerungen während der Aufnahme)

3.2 Partitur-Format

Da das Partitur-Format soweit es geht an SAM-Formate angelehnt ist, wird als Header der Partitur das SAM-Headerformat verwendet. Die unter 3.1 fettgedruckten Label sind für den Partitur-Header obligatorisch, die übrigen optional. Die Label, die im SAM-Bodyformat benutzt werden, reichen für die Partitur nicht aus. Solche SAM-Label, die gewissen Ebenen in der Partitur entsprechen (z.B. LBO für Transliteration) werden nicht übernommen, damit keine Mißverständnisse über den tatsächlichen Inhalt der Ebenen oder Probleme bei der weiteren technischen Verarbeitung entstehen.

Im folgenden wird die genaue Syntax der einzelnen Ebenen beschrieben.

3.2.1 Vorschlagstranskription

Die Daten werden spaltenweise angeordnet, wobei jede Zeile mit der Markierung für die Vorschlagstranskription "KAN: " beginnt (1. Spalte). Die Trennung der Spalten erfolgt durch White Space (Leerzeichen oder Tabulator). Die zweite Spalte enthält die Wortnummer, wobei die Zählung bei "0" beginnt, die dritte Spalte die kanonische Form des entsprechenden Wortes.

Beispiel:

```
KAN:    0       maIn
KAN:    1       na:m@
KAN:    2       QIst
KAN:    3       QaplaItn6
KAN:    4       Qa:
KAN:    5       be:
KAN:    6       QEl
KAN:    7       Qa:
KAN:    8       Qi:
KAN:    9       te:
```

3.2.2 Transliteration

Die Daten werden zeilenweise angeordnet, wobei bis zu 70 Zeichen in einer Zeile stehen können. Ist die Transliteration länger, was gewöhnlich der Fall ist, findet ein automatischer Umbruch durch Ersetzen eines Leerzeichens durch Carriage Return statt. Jede Zeile beginnt mit der Markierung "TRL: ". Die Numerierung der spachlichen Bestandteile der Transliteration erfolgt jeweils direkt hinter den betreffenden Einheiten, wobei die Zählung bei "0" beginnt und der Nummer ein "&" vorangestellt ist. Etwaige Turnnumerierungen aus den orginalen Transliterationen entfallen.

Beispiel:

```
TRL: <A> mein&0 Name&1 ist&2 <!1 is> <:<#Klicken> Ableitner&3:>
<A>, $A&4 $B&5 $L&6 ...
```

3.2.3 Lautsegmentation

Die Daten werden spaltenweise angeordnet, wobei jede Zeile mit einer Markierung für die Lautsegmentation beginnt. Wird die breite Transkription nach SAMPA verwendet, werden die

Zeilen der Ebene mit “SAM: “ markiert, wird nach IPA gelabelt (IPA Nummern), lautet die Zeilenmarkierung: “IPA: “. Die zweite Spalte enthält das Anfangssample eines Segments, die dritte Spalte die Segmentdauer (in Samples), die vierte Spalte Transkriptionssymbol(e) und die fünfte Spalte die Referenz auf die Wortnummer(n).

Der Bezug der Transkription zur Vorschlagstranskription kann in der vierten Spalte symbolisch dargestellt werden. Folgende Möglichkeiten sind vorgesehen[6]:

- Segment x der kanonischen Form wird Übernommen: x
- Segment x der kanonischen Form wird durch y ersetzt: x-y
- Segment x der kanonischen Form wird elidiert: x-
- nicht vorgesehenes Segment y wird eingefügt: -y

Soll ein Segment, das in der kanonischen Form vorhanden ist, übernommen werden, dabei jedoch durch suprasegmentale Merkmale wie z.B. Nasalität (SAMPA: “~”) oder Laryngalisierung (SAMPA: “q”) charakterisiert werden, kann dies durch Hinzufügen der entsprechenden diakritischen Zeichen des jeweiligen Transkriptionsalphabets erfolgen.

Z. B. in SAM:

- Segment x der kanonischen Form wird übernommen und modifiziert: x~ oder qx

Ein Segment kann entweder einem Wort, oder bei Segmentation mit möglicherweise überlappenden Grenzen zwei Wörtern zugeordnet werden, wobei die Referenznummern durch Komma (ohne Blank) abgetrennt werden. Ist die Zuordnung eines Segments nicht eindeutig, kann anstelle der Wortnummer die Markierung “-1” vergeben werden.

Beispiel:

SAM:	6022	352	m	0
SAM:	6374	912	aI	0
SAM:	7286	312	n	0,1
SAM:	7598	721	a:~	1
SAM:	8310	298	m	1
SAM:	8608	213	@	1
SAM:	8821	0	Q-	2
SAM:	8821	200	qI	2
SAM:	9021	359	s	2
SAM:	9380	0	t-	2
SAM:	...			

3.2.4 Wortsegmentation

Die Daten werden spaltenweise angeordnet, wobei jede Zeile mit der Markierung “WRD: “ für die Wortsegmentation beginnt. Die zweite Spalte enthält das Anfangssample eines Wortes, die dritte Spalte die Wortdauer (in Samples), die vierte Spalte die Orthographische Form des Wortes (identisch mit Eintrag in Transliteration, falls vorhanden) und die fünfte Spalte die Referenznummer auf die Vorschlagstranskription.

Beispiel:

WRD:	6022	1576	mein	0
WRD:	7286	1544	Name	1
WRD:	8821	559	ist	2
WRD: ...				

6. siehe Pompino-Marshall [2]

3.2.5 Artikulatorische Geräusche

Die Daten werden spaltenweise angeordnet, wobei jede Zeile mit der Markierung “ATG: ” beginnt. Die zweite Spalte enthält das Anfangssample eines Geräusches, die dritte Spalte die Dauer (in Samples). In der vierten Spalte erfolgt der Marker für das jeweilige Geräusch aus einer endlichen Menge.

Definierte Marker:

- Schmatzen
- Schlucken
- Räuspern
- Husten
- Lachen
- Geräusch (Produktion nicht identifizierbar)

Beispiel:

ATG:	5723	277	Atmen
ATG:	10422	305	Atmen
ATG: ...			

3.2.6 Nicht-artikulatorische Geräusche

Die Daten werden spaltenweise angeordnet, wobei jede Zeile mit der Markierung “NAG: ” beginnt. Die zweite Spalte enthält das Anfangssample eines Geräusches, die dritte Spalte die Dauer (in Samples). In der vierten Spalte erfolgt der Marker für das jeweilige Geräusch aus einer endlichen Menge.

Definierte Marker:

- Klicken (Klicken bei Knopfdruck)
- Klingeln
- Klopfen
- Mikrobe (Geräusche, die beim Berühren des Mikrophons entstehen)
- Mikrowind (Pusten ins Mikrophon)
- Rascheln
- Quietschen
- Rest (Geräusch nicht identifizierbar)

Beispiel:

NAG:	9731	80	Klicken
NAG: ...			

3.2.7 Prosodie

Die Daten werden spaltenweise angeordnet, wobei jede Zeile mit der Markierung “PRO: ” beginnt. Die zweite Spalte enthält den Zeitpunkt des Label in Samples die dritte Spalte die prosodische Labelung[7] und die vierte Spalte die Referenznummer auf die Vorschlagstranskription.

Beispiel (nach [3]):

PRO:	10166	TON: L+H*; FUN: PA1
PRO:	13273	BRE: B3; TON: H-H%1
PRO:	17222	TON: H*; FUN: NA3
PRO:	24206	BRE: B2; TON: L-5

7. nach [3]

PRO: 31988 TON: H*; FUN: PA10

4 Literatur

[1] Kohler, Klaus et al. (1994): “Handbuch zur Datenaufnahme und Transliteration im TP14 von VERBMOBIL - 3.0”, IPDS Kiel, Technisches Dokument Nr. 11.

[2] Pompino-Marschall, Bernd (Hrsg.)(1992): PHONDAT. Verbundvorhaben zum Aufbau einer Sprachsignaldatenbank für gesrpchenes Deutsch. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM) 30, 99 - 128.

[3] Reyelt, Matthias, Anton Batliner (1994):”Ein Inventar prosodischer Etiketten für VERBMOBIL”, Memo 33, Version 1.0