

Transliteration spontanprachlicher Daten - Lexikon der Transliterationskonventionen - VERBMOBIL II

Susanne Burger
Universität München
Version 1 (April 1997)

Vorwort

Die für Verbmobil I festgelegten Transliterationskonventionen ([1], [2]) wurden für die zweite Phase des Verbmobil-Projektes teilweise erweitert und verändert.

Dies geschah vor allem im Rahmen des Daten-Workshops vom 29.01.-30.01.97 an der Uni München. Zuvor hatte sich der Arbeitskreis Transliteration mit der Überarbeitung der Transliterationen auf Basis einer neuen Version des Transliterations-Lexikons, die sich vor allem die Parsbarkeit der Transliterations-Konventionen zur Aufgabe gemacht hat, befasst. Das Ergebnis des Arbeitskreises wurde im Workshop präsentiert, diskutiert und protokollarisch festgehalten.

Teilnehmer des Arbeitskreises Transliteration:

Anton Batliner (Uni Erlangen)
Susanne Burger (Uni München)
Anja Geumann (Uni München)
Henrik Heine (Uni Hamburg)
Christiane Hofbauer (Uni München)
Susanne Jekat (Uni Hamburg)
Andreas Kipp (Uni München)
Heinz Kirchmann (DFKI Kaiserslautern)
Harald Längen (Uni Bielefeld)
Matthias Reyelt (TU Braunschweig)
Christian Scheer (Uni München)
Johannes Schwinn (DFKI Kaiserslautern)
Andreas Witt (Uni Bielefeld)

Zu den im Transliterationslexikon definierten Konventionen wird es einen Parser und ein Filterprogramm mit diversen Optionen geben.

Inhalt

Allgemeines zur Transliteration von Spontansprache in VERBMOBIL

1. Spontansprache in VERBMOBIL
2. Transliteration von Spontansprache

3. Objekte der Symbolisierung
4. Grundanforderungen
5. Grenzen der Transliteration
6. Neues im Vergleich zum alten Transliterationslexikon
7. Benutzung

Lexikon der Transliterationskonventionen

A. Struktur eines Transliterationsfiles

1. Schematischer Überblick
2. Globales Fileformat
 - 2.1. Header
 - 2.2. Transliteration
 - 2.2.1. Turns
 - 2.2.1.1. Turnname
 - 2.2.1.2. Sprachenkennzeichnung bei mehrsprachigen Dialogen
 - 2.2.1.3. Turnbody
 - 2.2.1.4. Globale Kommentare

B. Transliteration der Turnelemente

1. Lexikalische Einheiten
 - 1.1. Wörter des Dictionaries
 - 1.2. Interjektionen
 - 1.3. Wortreduktionen
 - 1.4. Komposita
 - 1.5. Klassifizierte lexikalische Einheiten
 - 1.5.1. Buchstabierung und Abkürzungen
 - 1.5.2. Namen
 - 1.5.3. Zahlen
 - 1.5.4. Neologismen
 - 1.5.5. fremdsprachige Wörter
 - 1.6. Lexikalische Einheiten mit artikulatorischen Besonderheiten
 - 1.6.1. Zögerung
 - 1.6.2. schwerverständliche Wörter
 - 1.6.3. Abgebrochene lexikalische Einheiten
 - 1.6.3.1. artikulatorischer Abbruch
 - 1.6.3.2. Artikulatorische Unterbrechung lexikalischer Einheiten
 - 1.6.3.3. Technischer Abbruch
 - 1.7. Aussprachekommentare
2. syntaktisch-semantische Gliederung
 - 2.1. Interpunktion
 - 2.2. agrammatische Phänomene
 - 2.2.1. Wiederholung / Korrektur
 - 2.2.2. False Starts (oder Neustart)
3. nonverbale artikulatorische Produktionen
 - 3.1. Atmen
 - 3.2. Häsitationen
 - 3.3. Unverständliche Sprachproduktionen
 - 3.4. nonverbale artikulatorische Geräusche
4. Geräusche und technische Artefakte
5. Sprechpausen

6. akustische Überlagerung
 - 6.1. Sprecherüberlagerung
 - 6.2. Geräuschüberlagerung
7. lokale Kommentare
8. Sonderkommentare
 - 8.1. Code-Wörter
 - 8.2. szenario-bedingte Pause

C. Tabellen

1. Reihenfolge der Markierungssymbole bei lexikalischen Einheiten
2. Liste aller verwendbarer Symbole

Literatur

Allgemeines zur Transliteration von Spontansprache in VERBMOBIL

1. Spontansprache in VERBMOBIL

Für das Projekt VERBMOBIL werden große Mengen an Dialogen oder Multiparty-Konversationen zwischen verschiedenen Sprechern aufgenommen, die während des Gesprächs verschiedene Aufgaben aus verschiedenen Szenarien, etwa die Terminabsprache für ein geschäftliches Treffen oder die Planung einer Reise lösen sollen.

Das Ergebnis sind mehrkanalig aufgezeichnete spontansprachliche Daten, die als Grundlage zu Forschung und Entwicklung im Bereich der Spracherkennung, Sprachsynthese und der automatischen Übersetzung in andere Sprachen im VERBMOBIL-Projekt dienen.

2. Transliteration von Spontansprache

Mit Hilfe der orthographischen Transliteration sollen in einem ersten breiten Verschriftungsschritt Dialogaufnahmen allen Verbmobil-Partnern symbolisch verfügbar gemacht werden. Das heißt, Projektmitarbeiter hören die Aufnahmen der Dialoge ab und verschriften diese auf Wortebene.

Zusätzlich zur Orthographie kommen dabei die Transliterationskonventionen zum Einsatz,
- da in der Spontansprache Phänomene auftreten, die in der Schriftsprache nicht vorkommen, wie Satzabbrüche, Korrekturen und Wiederholungen von Äußerungen, Reduktionen oder Häsitationen,
- da in den VERBMOBIL-Dialogen technische Artefakte auftreten, wie technische Aufnahmeabbrüche oder Mikrophongeräusche,
- da Dialogsituationen zu Sprecherüberlagerungen führen können.

3. Objekte der Symbolisierung

Folgende Grob-Kategorien bilden die zu verschriftenden Elemente innerhalb eines Dialogbeitrags:

1. lexikalische Einheiten
2. syntaktisch-semantische Gliederung
3. nonverbale artikulatorische Produktionen
4. Geräusche

5. Pausen
6. akustische Überlagerung
7. Kommentare
8. Sonderkommentare

4. Grundanforderungen

Die Grundanforderungen, die VERBMOBIL an die Transliterationen stellt, sind:

a, automatische Weiterverarbeitung

- einheitliche File-Struktur
- informative und einheitliche Turnnamen
- konsistente Transliterationen
- eindeutige Symbolisierungen
- ASCII-Zeichen
- parsbare Transliterations-Konventionen

b, inhaltliche Anforderungen:

- Protokollierung aller hörbaren Dialog-Elemente
- syntaktisch-semantische Marker
- Kennzeichnung von Sprecher- und Geräuschüberlagerungen
- Kennzeichnung bestimmter Wortkategorien (Namen, Zahlen) und Defekte
- möglichst noch lesbare Text-Files

c, Transliterationsvorgang

- einfache Anwendbarkeit der Konventionen
- eingängliche und verständliche Konventionen, auch für Nicht-Experten

5. Grenzen der Transliteration

Breite Verschriftung heißt jedoch auch, daß die hörbaren Ereignisse nur protokolliert, nicht jedoch genau beschrieben werden. Die Orthographie kann keine lautliche Beschreibung der gesprochenen Äußerungen liefern, für Geräusche und nonverbale Produktionen stehen dem Transliterierer nur Kategorien zur Verfügung. Bei besonders auffälligen Aussprachen oder Begebenheiten kann durch einen sogenannten Aussprachekommentar oder lokalen Kommentar der Weiterverarbeitung angezeigt werden, daß hier etwas Außerordentliches aufgetreten ist.

Die Annotation auf Ebene der breiten Verschriftung leistet

- keine phonologische Verschriftung
- keine phonetische Transkription
- keine zeitliche Zuordnung zu den Signaldaten.

6. Neues im Vergleich zum alten Transliterations-Lexikon

Die Konventionen aus der ersten Phase [1], [2] deckten die Grundvorstellungen durchaus ab, waren aber durch zu viele geklammerte Elemente nicht parsbar, was bei der Überführung in andere Datenformate, bei der Entwicklung von geeigneten Filtern, aber auch bei der Fehlersuche in den

Transliterationen Probleme verursachte. Während der ersten Projekt-Phase stellte sich heraus, daß einige Konventionen in der vorliegenden Form zu Inkonsistenzen führten, andere nicht gebraucht wurden oder Bedarf zu weiteren Konventionen vorhanden war.

Deshalb wurden schon während der ersten Projekt-Phase immer wieder Änderungen am Transliterations-Handbuch vorgenommen oder entwickelten sich Verschriftungs-Konventionen, die in keiner Form schriftlich festgelegt waren.

Zu Beginn der zweiten Projekt-Phase wurden die entstandenen Probleme und Erfahrungswerte gesammelt.

Die neuen Konventionen basieren auf den alten Konventionen und wurden mit den Wünschen und Vorschlägen der Partner, den Erfahrungen aus der bisherigen Transliteration und aus den Bedingungen, die eine terminierende Grammatik der Transliterations-Konventionen stellt, ergänzt bzw. verändert.

Im Vergleich zu den Transliterations-Konventionen von Verbmobil I sind folgende Konventionen neu oder verändert:

1. neues Textformat
2. neuer Turnname
3. veränderter Marker für fremdsprachliche Turns
4. neue Apostroph-Regelung
5. Buchstabiersequenzen
6. Symbol für Namen
7. Symbol für Zahlen
8. Symbol für fremdsprachliche Wörter
9. neue Markierung von schwerverständlichen Wörtern
10. Symbolisierung technischer Wort- und Turnabbrüche / Turnunterbrechungen
11. neue Klammerung bei Phrasenabbrüchen
12. neue Konventionen bei Sprecherüberlagerung
13. neue Konventionen für Geräuschüberlagerung
14. Markierungsmöglichkeiten für szenariobedingte Pausen und systemsteuernde Codewörter

7. Benutzung

Das Transliterations-Lexikon löst das Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL -3.0 [1] und das Lexikon zur Transliteration [2] VERBMOBIL I aus der ersten Projekt-Phase ab.

Es soll in erster Linie den transliterierenden Mitarbeitern als Schulungsunterlage und als Nachschlagewerk dienen, aber auch beim Lesen und Weiterverarbeiten der Transliterationen zum "decodieren" von Nutzen sein.

Zielsprache des Lexikons ist deutsch, abgesehen von manchen Konventionen für deutsche lexikalische Einheiten kann es jedoch auch von anderen Sprachen als Standard zur Verschriftung von Spontansprache genutzt werden.

Die nur das Deutsche betreffenden Lexikoneinträge sind gesondert mit **deutsch**** markiert.**

Struktur des Transliterations-Lexikons:

Das Lexikon beschreibt die

- Struktur und das Format einer Transliterationsdatei
- die Konventionen für die Transliteration der einzelnen Turnelemente

- und bietet Übersichtstabellen für alle verwendbaren Symbole.

Die Ordnung der Transliterations-Konventionen richtet sich nach den in die Turnelement-Kategorien eingeteilten Objekten der Symbolisierung.

Lexikoneinträge zu den Turnelementen:

Ein Lexikoneintrag findet sich unter der entsprechenden Element-Kategorie.

Der Eintrag selbst enthält

- Namen des Markers, Symbols oder Ereignisses

- Symbol

- Beispiele

- Definition

- Transliterationskonvention

Die Beispiele stammen entweder direkt aus den VERBMOBIL-Transliterationen oder sind entsprechend konstruiert.

HTML-Format:

Das Lexikon wurde im HTML-Format erstellt. So kann mit Hilfe der Links schnell und einfach auf Begriffe, Symbole, deren Bedeutung und Anleitungen zu bestimmten Konventionen zugegriffen werden.

Bei Benutzung des Lexikons im HTML-Format kann die Liste aller verwendeten Symbole mit der Maus angeklickt werden und so schnell der entsprechende Lexikoneintrag mit Anweisung und Verwendungsbeispiel gefunden werden.

Innerhalb der Lexikoneinträge können sich weitere Links zu entsprechenden Symbolen oder in den Beispielen verwendeten Konventionen befinden.

Alle **Hypertext-Links** arbeiten **innerhalb des Dokuments**, so daß beim Ausdrucken keine Information verloren geht.

Lexikon der Transliterationskonventionen

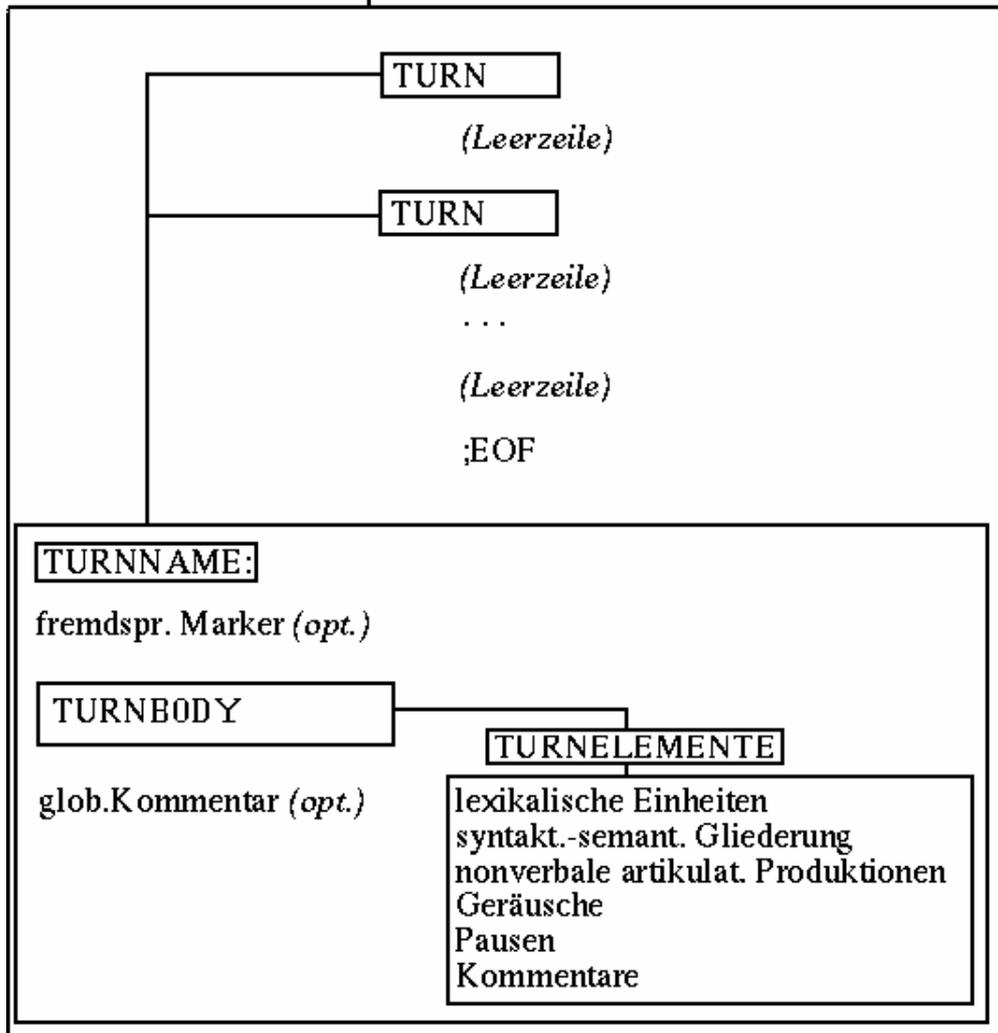
A. Struktur eines Transliterationsfiles

1. Schematischer Überblick

HEADER

; (Leerzeile)

TRANSLITERATION



2. Globales Fileformat

Ein Transliterations-File besteht aus dem

Header

und der

Transliteration der Dialogbeiträge.

Header und Transliteration sind durch eine **mit Semicolon beginnende Leerzeile** voneinander getrennt.

Die Transliteration ist in **Turns** unterteilt.

Zwischen den einzelnen Dialog-Turns steht **eine Leerzeile** als **Turntrenner**.
Das Ende der Transliteration wird mit
einer **Leerzeile nach dem letzten Turn**
und
einer **mit ;EOF beginnenden Endzeile** angezeigt.

Die Turns bestehen aus
Turnname,
gegebenenfalls einer **Kennung für die Sprache**, in der der Dialogbeitrag gesprochen wurde,
dem **Turnbody** mit den den **Turnelementen**,
und optional einem **globalen Kommentar** zum Turn.

2.1. Header

Beispiel:

```
; CDR: 12.00
; TRV: 12.02
; Dialog N057K
; zuletzt bearbeitet am 23.5.94
; Tonqualit"at: (allgemeine Kommentare zu
; Sprechern oder zur Aufnahmequalit"at des
; Dialogs)
;
(m123d000_AAP_120002DD1xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx: so .... )
```

Jede Headerzeile beginnt mit **Semicolon, gefolgt von Leerzeichen**.

In der ersten Headerzeile steht nach "CDR:" und Leerzeichen die CD-Rom Version

```
; CDR: 12.00
```

in der zweiten Headerzeile steht nach "TRV:" und Leerzeichen die Transliterations-Update-Version

```
; TRV: 12.02
```

in der dritten Headerzeile steht nach "Dialog:" und Leerzeichen der Dialogname (= auch Name des entsprechenden Directories, unter dem auf der CD-Rom die Signalfiles des Dialogs zu finden sind.)

```
; Dialog: M123D
```

in möglichen weiteren Headerzeilen stehen Kommentare, die sich auf den gesamten Dialog beziehen.

```
; zuletzt bearbeitet am 17.10.97
; Tonqualit"at: ganzer Dialog sehr leise
; .....
```

Header und Transliteration sind durch eine **mit Semicolon beginnende Leerzeile** von einander getrennt.

2.2. Transliteration

Beispiel:

```
m123d000_AAP_120002DD1xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx: so
, guten Tag , mein Name ist <!1 is'> ~J"ansch. <"ah> wir hatten
bereits telefoniert<Z> , mein Name ~J<Z>"ansch , $J $"A $N $S $C
$H , wegen <:<#Mikrobe> eines:> <:<#Mikrobe> Arbeitstreffens:> .

m123d001_BBP_120002DD1xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx: gr"u"s
Gott , mein Name ist ~G<Z>"urtner , <A> <"ahm> $G $"U $R $T $N $E
$R
. <A> <"ahm>
;Brummen "uber gesamtem Turn

m123d002_AAP_120002DD1xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx: ja ,
<:<#Mikrobe> ich:> <:<#Mikrobe> kuck' :> <:<#Mikrobe> jetzt:> mal
nach bei mir , wann ich <A> einen <!1 ein'> Termin frei h"atte .
<A> das <:<#Mikrowind> erste:> w"are <Schmatzen> in der Woche
oder die Tage vom #vier-zehnten Juli bis zum #acht-zehnten
<:<#Mikrobe> Juli:> . <A> ginge das bei <:<#> Ihnen:> ?

m123d003_BBP_120002DD1xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx: ...

;EOF
```

Der Transliterationsteil beinhaltet die Niederschrift des gesamten Dialoges.
Jeder Sprecherbeitrag wird dabei in einem gekennzeichneten Turn festgehalten.

Zwischen den Turns steht eine Leerzeile.

Die Transliteration endet
mit einer Leerzeile nach dem letzten Turn und
;EOF als Ende-Markierung.

2.2.1. Turns

Ein Turn oder Sprecherbeitrag beginnt mit dem
Turnnamen,
gefolgt von einer Leerstelle.

Danach kann bei mehrsprachigen Dialogen eine **Kennung der Hauptsprache** des folgenden
Turnbodies stehen, wieder gefolgt von einer Leerstelle.

Im **Turnbody** werden alle hörbaren Ereignisse, syntaktisch-semantische Markierungen und
Kommentare in der Regel mit einer Leerstelle voneinander getrennt protokolliert.

Am **Zeilenende innerhalb eines Turns** steht
Zeilenbruch,

die **neue Zeile beginnt mit einer Leerstelle**.

Nach dem **letzten Turnelement im Turn** steht **Leerstelle** , gefolgt von **Zeilenumbruch**.

Dem Turn kann ein **globaler Kommentar** folgen. Dieser beginnt in der nächsten Zeile nach dem vorausgegangenen Turn und endet mit Zeilenumbruch.

2.2.1.1. Turnname

Beispiel:

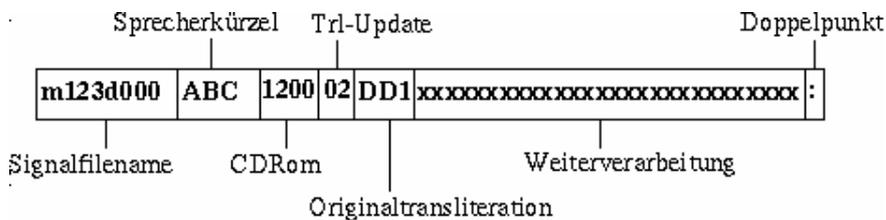
```
m123d000_AAP_120002DD1xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx: hallo .
```

Definition:

Jeder Turn beginnt mit dem Turnnamen. Dieser Name dient als Identifikator, wenn nach einzelnen transliterierten Turns gesucht werden soll.

Im Turnnamen wird der Bezug zum Signalfilename und zur Sprecherdatenbank geschaffen. Weiter können von der Datenweiterverarbeitung Codes im Turnnamen gesetzt werden, um Bearbeitungsversionen der Transliteration zu kennzeichnen.

Transliteration:



Signalfilename:

Der Signalfilename ist der Name des Signalfiles auf der CD-Rom, ohne Extension.

Signalfilenamen bestehen aus

Dialog-Directory-Namen mit Kleinbuchstaben, und Turnnummer innerhalb des Dialoges.

x000x000_ (9 Zeichen)

z.B. m123d000_

Sprecherkürzel:

Jeder Sprecher wird mit einem individuellen Sprecherkürzel markiert. Dieses Sprecherkürzel besteht aus drei Großbuchstaben (keine Umlaute), die den Sprecher in der Sprecherdatenbank eindeutig identifizieren.

XXX_ (4 Zeichen)

z.B. ABC_

CDRom-Version:

entspricht der CDRom-Ausgabenummer (2 Zeichen) und Version (2 Zeichen) aus dem Header.

0000 (4 Zeichen)

z.B. 1200

Transliterations-Update-Version:

entspricht der Transliterations-Update-Versionsnummer aus dem Header (CDRom-Ausgabe bleibt gleich).

00 (2 Zeichen)

z.B. 02

Originaltransliteration:

Die Originalfilemarkierung kennzeichnet, ob es sich um die Originaltransliteration handelt. In den ersten beiden Zeichen kann die Originaltransliteration zusätzlich Informationen codieren. Das letzte der drei Zeichen **muß 1 für Original** sein.

XX1 (3 Zeichen)

z.B. DD1

Weiterverarbeitung:

Weitere 28 Stellen stehen zur Verfügung, um Bearbeitungsversionen zu kennzeichnen.

Diese Stellen werden von der Originaltransliteration mit x aufgefüllt.

(Anzahl und Position der Stellen, die von weiterverarbeitenden Projektpartnern verwendet werden, werden zentral verwaltet.)

Zur Codierung sind folgende Zeichen zugelassen:

a-z (keine Umlaute)

A-Z (keine Umlaute)

0-9

_ (Subline)

Der Turnname endet mit **Doppelpunkt** und **Leerstelle** (oder Zeilenumbruch-Leerstelle).

2.2.1.2. Sprachenkennzeichnung bei mehrsprachigen Dialogen

Symbol: <*tXXX>

wobei XXX:

ENG = englisch

JAP = japanisch

GER = deutsch (bei nicht-deutschen Dialogen oder mehrsprachigen Dialogen)

auch

FRA =französisch

ITA = italienisch

SPA = spanisch

etc.

Beispiel:

```
m123d000_AAP_120002ED1xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx: <*tENG> good morning , ~John . how are you ?
```

```
m123d000_AAP_120002ED1xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx: <*tGER> guten Tag , Herr ~Miller . danke , es geht mir gut .
```

Definition:

Bei mehrsprachigen Dialogen wird nach jedem Turnnamen eine Kennzeichnung für die Sprache, in der der folgende Dialogbeitrag geführt wurde, gesetzt.

Transliteration:

Nach der Leerstelle des Turnnamens steht `<*tXXX>` und eine weitere Leerstelle (oder Zeilenumbruch-Leerstelle).

Anmerkung:

Für den Dialogbeitrag gelten die orthographischen Regeln der jeweiligen Sprache.

2.2.1.3. Turnbody

Nach dem Turnnamen, im Turnbody, werden alle hörbaren Ereignisse, wie lexikalische Einheiten und Geräusche, und zusätzliche Marker wie syntaktische Markierungen oder Kommentare, transliteriert.

Diese Ereignisse sind die **Turnelemente**.

Format des Turnbodies:

a, ASCII-Kodierung

Die Kodierung der Transliteration erfolgt in **7-bit ASCII**. Für die Umlaute und "ß" wird im Ablieferungsformat die **TEX-Schreibweise** verwendet, also "a, "U, "s etc.

b, Element-Trennung

Zwischen den Turnelementen steht

ein **Leerzeichen**

oder **am Zeilenende** ein **Zeilenumbruch**, gefolgt von einem **Leerzeichen**.

c, Worttrennung

Worttrennungen werden **nicht** durchgeführt; es wird immer das ganze Wort in die nächste Zeile geschrieben.

2.2.1.4. Globale Kommentare

Symbol: ;.....

Beispiele:

```
m123d001_BBP_120002DD1xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx: gr"u"s
  Gott , mein Name ist ~G<Z>"urtner , <A> <"ahm> $G $"U $R $T $N $E
  $R
  . <A> <"ahm>
```


Entsprechend der Sprache, in der der Dialogbeitrag geführt wurde, gelten bei der Transliteration die jeweiligen für diese Sprache gültigen orthographischen Regeln. Das gilt auch für die in der Transliteration gesondert markierten Wörter, soweit sie noch als der entsprechenden Sprache zugehörig erkennbar sind.

Zusätzlich gilt für die Verbmobil-Transliterationen eine Wortliste, in der bei verschiedenen möglichen Schreibweisen eine einheitliche festgelegt wird.

Unabhängig von der Orthographie kann nach gezögerten Lauten <Z> für Zögerung an die entsprechende Position des Wortes gesetzt werden.

****deutsch****

Für die Transliteration deutscher Wörter gilt als orthographische Referenz der "alte" Duden, 20. Auflage.

Die deutsche Rechtschreibreform wird noch nicht beachtet!

Groß- und Kleinschreibung:

Der Duden gilt auch hinsichtlich Groß- und Kleinschreibung gemäß der Wortklasse.

ABER: Am Äußerungsanfang werden nur Substantive großgeschrieben. In allen anderen Fällen wird die Dudenkonvention zur Großschreibung am Anfang eines Satzes nicht beachtet. Sowohl die nominalen Formen der Anredepronomen "Sie" und "Ihnen", als auch die possessive Form "Ihr" (mit allen Flexionen) werden groß geschrieben, um sie von den Pronomen für die dritte Person zu trennen. Da eine solche Trennung bei "du" nicht erforderlich ist, wird es immer klein geschrieben.

****deutsch****

Sondermarkierungen:

Alle Markierungssymbole werden ohne Leerstelle je nach entsprechender Konvention direkt davor, innerhalb der lexikalischen Einheit oder direkt dahinter transliteriert.

Lexikalische Einheiten können nur einmal bezüglich ihrer Wortklasse (Namen, Zahlen) markiert werden (keine Doppelklassifizierung).

Wenn artikulatorische Besonderheiten symbolisiert werden (Abbrüche, Zögerung), können diese Markierungen auch an klassifizierte lexikalische Einheiten angehängt werden, solange noch erkennbar ist, um welche lexikalische Einheit es sich handelt.

Akustische Überlagerungen werden auch bei lexikalischen Einheiten mit Sondermarkierung transliteriert.

Siehe auch: C.1. Reihenfolge der Markierungssymbole

Elementumgebung:

Lexikalische Einheiten können von Geräuschen überlagert oder sprecherüberlagert sein.

Lexikalischen Einheiten kann ein Aussprachekommentar folgen.

Agrammatische Phänomene treten nur in Verbindung mit lexikalischen Einheiten auf.

Die Interpunktion kann nur nach lexikalischen Einheiten (oder kommentierten lexikalischen Einheiten) stehen.

Turnende:

Nach einer lexikalischen Einheit (oder lexikalischen Einheit mit Kommentar) am Turnende **muß** stehen:

Punkt oder

Fragezeichen oder

technischer Turnabbruch (bei technischem Abbruch des Turns oder wenn der Sprecher seinen Dialogbeitrag selbst abbricht)

1.1. Wörter eines Dictionaries

Symbol:

(nicht markiert)

Beispiele:

```
m123d000_AAP_120002DD1xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx: so
, guten Tag , mein Name ist <!1 is'> ~J"ansch . <"ah> wir hatten
bereits telefoniert<Z> wegen <:#Mikrobe> eines:> <:#Mikrobe>
Arbeitstreffens:> .
```

```
m123d001_BBP_120002DD1xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx: gr"u"s
Gott , mein Name is<Z>t <!1 is'> <*T>t
```

Definition:

Wörter eines Dictionaries sind Wörter,

- die in der transliterierten Form in die Verbmobil-Wortliste eingehen und in eine anderer Sprache übersetzt werden können,

- die gut verständlich und nicht verstümmelt, nicht neu erfunden oder nicht fremdsprachlich sind.

Transliterationskonvention:

Entsprechend der Sprache, in der der Dialogbeitrag geführt wurde, gelten bei der Transliteration die jeweiligen für diese Sprache gültigen orthographischen Regeln.

Für Wörter eines Dictionaries gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

Anmerkungen:

****deutsch****

Vom deutschen Duden etwas abweichend werden behandelt:

Reduzierte Wortformen,

Interjektionen,

Komposita,

Abkürzungen und Buchstabierungen,

Namen,

Zahlen

Dialektale Wörter:

Finden sich in einem Dialog dialektale Wörter,

- die keine dialektale Aussprache von im Duden aufgelisteten Wörtern darstellen,

- aber auch noch nicht in den Duden aufgenommen wurden,

- wie: "ja mei" , "grüß Gott" , "moin, moin" , "Herrschaftszeiten" , "fei" ,

so werden diese ohne besondere Kennzeichnung orthographisch verschriftet und optional mit Aussprachekommentar und / oder Kommentar zur Herkunft oder Bedeutung versehen.

****deutsch****

1.2. Interjektionen

Symbol:

(nicht markiert)

Beispiele:

... oh, das pa"st mir gar nicht . mm . ja nun , ne ,
dann brauchen wir einen anderen Termin

Definition:

Interjektionen sind Ausrufe der Überraschung wie "au" , "ah" , "oh" , "ui" , "he" ,

Bejahung wie "mhm" ,

oder Verneinung "mm" ,

Einwürfe oder Bestätigungen wie "ne" , "gell" , "aha" .

Transliterationskonvention:

Interjektionen werden ohne Zusätze in der in der Definition verschrifteten Form in den orthographischen Text eingefügt. Die meisten von ihnen sind bereits im Duden aufgelistet. Für Interjektionen gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

Anmerkung:

Verneinendes "m´m" wird mit "mm" verschriftet, im Gegensatz zu bejahendem "mhm".

1.3. Wortreduktionen

****deutsch****

Symbol:

´ (teilweise mit Apostroph)

Beispiele:

... ich möcht´ einen Termin ausmachen . wie w"ar´ ´s am Dienstag ?
...
... dann fahren wir mit dem <!2 mit´m> Flugzeug
... ich hab´ ´nen Vorschlag . ´n m"oglicher Termin ...

Definition:

Wortreduktionen, die in die Transliteration mit eingehen, sind:

- End-e Reduzierung
- Reduktion unbestimmter Artikel
- Verschmelzung von Präposition und reduziertem bestimmten Artikel
- reduziertes Pronomen der 3. Pers. Sing. (es)

Transliterationskonvention:

a, Mit Apostroph werden transliteriert:

- End-e Reduzierungen:

Für fehlendes End-e steht Apostroph.

Aber: wurde die Endsilbe -en nicht gesprochen, wird sie trotzdem verschriftet und ihr Fehlen durch einen Aussprachekommentar angezeigt.

hab´

m"u"st´

wollt´

(aber: wir haben <!1 hab´> da keine Zeit)

- unbestimmte Artikel:

Bei reduzierten unbestimmten Artikeln wird mit Apostroph verschriftet.

Aber: wurde z.B. statt einen nur "´n" gesagt, steht "´nen <!1 ´n> ", d.h. der Apostroph steht immer nur für fehlendes /ei/.

´nen

´ner

´n

´nem

´ne

´nes..

- reduziertes Pronomen der 3. Pers. Sing. : es

´s

Diese Wortformen stehen als jeweils eigene lexikalische Einheiten zwischen Leerstellen,

auch wenn zwei Apostrophe aufeinandertreffen.

w"ar´ ´s

hab´ ´nen ..

b, Reduktion von Präposition und bestimmtem Artikel

Führt die Verschmelzung von Präposition und bestimmtem Artikel zur Reduktion der Silbenanzahl, so werden diese lexikalischen Einheiten zusammengeschrieben (nach Duden).

Präp. +das	Präp. +dem	Präp. +den	Präp. +der
ans	-	-	-
aufs	-	-	-
durchs	-	-	-
fürs	-	fürn	-
hinters	hinterm	hintern	-
ins	im	-	-
übers	überm	übern	-
unters	unterm	untern	-
vors	vorm	vorn	-
	zum	-	zur

c, Sonstige Reduktionen

In allen anderen Fällen wird die Verschmelzung durch einen Aussprachekommentar angezeigt. In diesem Fall sollte auch die Präposition im Aussprachekommentar mitverschrieben werden, da ja die Verschmelzung von Präposition mit bestimmtem Artikel angezeigt werden soll.

mit dem <!2 mit ´m>
nach dem <!2 nach ´m>..

Für Wortreduktionen gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

****deutsch****

1.4. Komposita

****deutsch****

Symbol:

- (teilweise mit Bindestrich)

oder auch

-- (Doppelbindestrich bei Kompositumsergänzung)

Beispiele:

... wir treffen uns in der Filiale von ~\$O-\$K-\$B ...
... wenn wir die Acht-Uhr-Maschine noch erreichen ...
... der Filialen-Abteilungsleiter holt uns von der \$U-Bahn-Station
ab ...

Definition:

Komposita sind aus mehreren Wörtern gebildete zusammengesetzte Wörter.

Transliterationskonvention:

Einfache Komposita werden nach Duden verschriftet.

Bei mehr als Zwei-Wort-Komposita, ungewöhnlichen Zusammensetzungen und Kombinationen mit buchstabierten Einheiten, Namen oder Zahlen stehen Bindestriche zwischen den Kompositateilen.

Acht-Uhr-Maschine

M"anner-Selbsterfahrungs-Gruppe

Hauptbahnhofs-Eingangshalle

Berlin-Video

zusammengesetzte Namen:

Die aus mehreren Wörtern bestehenden Namen wie

~Zur-blauen-Traube

~Bu"s-und-Bettag

~Heilig-Drei-K"onig

werden mit Bindestrichen zusammengeschrieben und als Namen markiert. Die Groß/Kleinschreibung der Namensbestandteile wird in diesem Fall beibehalten.

zusammengesetzte Zahlen:

Die einzelnen Bestandteile von zusammengesetzten Zahlen werden ebenfalls mit Bindestrichen transliteriert:

#acht-und-zwanzig

#neun-zehn-hundert #drei-und-zwanzig

Abkürzungen:

Bei Abkürzungen, die aus buchstabierten Einheiten bestehen, werden zwischen den buchstabierten Einheiten Bindestriche gesetzt:

~\$I-\$B-M

\$U-\$S-\$A

\$U-Bahn

Um verschiedene Schreibweisen zu vermeiden, kann in Zweifelsfällen die Verbmobil-Wortliste zu Rate gezogen werden.

Für Komposita gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

Anmerkungen:

Bei Wortunterbrechung zwischen Kompositateilen steht der Bindestrich vor dem wiederaufnehmenden Wortteil (nach der Unterbrechung).

Sonderfall:

Kompositumsergänzung:

Bei zusammengesetzten oder abgeleiteten Wörtern, bei denen ein gemeinsamer Bestandteil nur einmal genannt wurde, steht **doppelter Bindestrich (--)**:

Geld-- und andere Sorgen,

Hin-- und Rückfahrt,

ein-- bis zweimal,

Lederherstellung und --vertrieb

Das mit Bindestrichen **endende** Wort kann **nicht artikulatorisch oder technisch abgebrochen** sein,

wird aber ansonsten wie eine lexikalische Einheit behandelt.

Das mit Bindestrichen **beginnende** Wort kann **nicht zusätzlich klassifiziert** sein, wird aber ansonsten auch nach den Konventionen lexikalischer Einheiten verschriftet.

****deutsch****

1.5. Klassifizierte lexikalische Einheiten

1.5.1. Buchstabierung und Abkürzungen

Symbol:

\$ (bei ausgesprochenen Buchstaben)

Beispiele:

... mein Name ist ~J<Z>"ansch , \$J \$"A \$N \$S \$C \$H ...

... in die \$U-\$S-\$A fahren ...

... die \$A-\$B-\$R-Filiale ...

... \$H \$A doppel-\$M \$E \$R ...

... gestern war die OPEC-Konferenz ...

Definition:

Als Buchstabierung gelten ausgesprochene Buchstaben, etwa zur Verdeutlichung der Schreibweise eines Namens oder bei Abkürzungen, bei denen die Buchstaben einzeln ausgesprochen werden (z.B. USA).

Abkürzungen, die als Wort ausgesprochen werden, werden nicht als Buchstabierung behandelt (z.B. OPEC-Länder, CeBIT-Messe).

Transliterationskonvention:

Jedem ausgesprochenen Buchstaben (Buchstabiereinheit) wird \$ vorangestellt. Als reine Buchstabiereinheiten gelten Buchstabierungen von Namen oder Wörtern, deren korrekte Schreibweise dem Gesprächspartner verdeutlicht werden soll. Die Buchstabiereinheiten sind hier

durch Leerstellen getrennt und werden jeweils wie lexikalische Einheiten behandelt.

Buchstabiersequenz:

Werden Buchstaben in Abkürzungen oder Komposita ausgesprochen, so wird dem jeweiligen Buchstaben ebenfalls \$ vorangestellt. Die zusammengehörigen Buchstabiereinheiten oder Komposita-Teile werden mit Bindestrich aneinandergehängt und wie Komposita behandelt.

\$U-\$S-\$A

\$U-Bahn

scharf-\$S

\$S-\$Z

\$A-\$B-\$C-Filiale

Abkürzung:

Als Wörter ausgesprochene Abkürzungen werden nach üblicher Schreibweise festgehalten und wie lexikalische Einheiten behandelt.

DIN

Benelux

OPEC

AStA

CeBIT

Für Buchstabierungen und Abkürzungen gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

Anmerkungen:

ACHTUNG:

\$F-\$A-\$Z für gesprochenes "eff a zet"

aber: FAZ für gesprochenes "Faz"

1.5.2. Namen

Symbol:

~

Beispiele:

... mein Name ist ~Hans ~Ableitner . wir sollen zu ~\$I-\$B-\$M fahren
und zwar an ~Bu"s-und-Betttag

... Herr ~Huber , wir treffen uns im Wirtshaus ~Zur-blauen-Traube
in ~Ettlingen ...

Definition:

Markiert werden alle Namen, die nicht in eine andere Sprache übersetzt werden sollen.

- Personennamen (auch Vornamen)
- Straßennamen

- Hotel- und Restaurantnamen
- Instituts- und Firmennamen
- Ortsnamen (aber nicht Ortsnamen, die übersetzt werden können, wie München, Wien, ...)
- Feiertagsnamen (aber nicht Feiertagsnamen, die übersetzt werden können, wie Ostern, Weihnachten)

Transliterationskonvention:

Namen werden mit vorgestelltem ~ markiert.

Bei Namen, die aus mehreren Wörtern bestehen, werden die Namensbestandteile mit Bindestrichen aneinandergehängt (z.B. ~Zur-blauen-Traube).

Ist ein Name nur Teil eines ansonsten aus anderen Wörtern bestehenden Kompositums, wird der Name nicht markiert (z.B. Berlin-Video, Elisabeth-Kirche).

Für Namen gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

1.5.3. Zahlen

Symbol:

#

Beispiele:

... am #f"unf-zehnten oder am #zwei-und-zwanzigsten h"atte ich
Zeit ..
... in #vier-zehn Tagen , am #siebzehnten Mai ginge es ...

Definition:

Als Zahlen gelten reine Zahlwörter oder Zahlenkombinationen, auch Ordnungszahlen.

Transliterationskonvention:

Zahlen und zusammengesetzte Zahlen werden mit vorangestelltem # markiert.

Zusammengesetzte Zahlen

Die Zahlen von 13 bis einschließlich 99 (zweistellige Zahlen) und die von "ein" bis "neunzehn" gezählten Hunderter (d.h. einschließlich Jahreszahlen wie neun-zehn-hundert)

werden mit Bindestrichen zusammengeschrieben.

#zwei-und-zwanzig

#drei-zehnter

#ein-hundert #f"unf-und-zwanzig

Alle übrigen Zahlenkombinationen werden mit Leerzeichen (oder Zeilenumbruch-Leerzeichen) getrennt (dreistellige Zahlen und größer):

#neun-zehn-hundert #drei-und-neunzig

#drei #Millionen #neun-und-vierzig #tausend #sechs-hundert #vier-zehn

#drei-hundert und #neun-und-vierzig

Ist eine Zahl nur Teil eines ansonsten aus anderen Wörtern bestehendem Kompositums, wird die Zahl **nicht** markiert.

Acht-Uhr-Maschine

Fünf-Tages-Seminar

Für Zahlen gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

Anmerkungen:

ACHTUNG:

#sechzehn

#siebzehn

("sech" und "sieb" sind keine eigenständigen Zahlwörter!)

1.5.4. Neologismen

Symbol:

*

Beispiele:

... *haarknapp <:<#Rascheln> um:> einen <!1 ein'> Tag verfehlt ...

... ich *verschraubel das jetzt mal

... was *exkursieren Sie denn ? ...

... *Diaabend-Weintrink-Revisionstreffen ...

Definition:

Als Nichtwort gelten:

- Neologismen (Wortneuschöpfungen)
- unsinnige Wortkombinationen
- Versprecher, die zu Neologismen führen

Transliterationskonvention:

Einem Neologismus wird * ohne Leerstelle vorangestellt.

Unwörter können **nicht** artikulatorisch abgebrochen sein.

Für Neologismen gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

Anmerkungen:

Kleinere Versprecher sollten in der richtigen Version mit Aussprachekommentar, völlig unsinnige Lautkombinationen als unverständliche Sprachproduktion verschriftet werden.

1.5.5. fremdsprachige Wörter

Symbol:

<*XXX>

wobei XXX:

ENG = englisch

JAP = japanisch

GER = deutsch (bei nicht-deutschen Dialogen oder mehrsprachigen Dialogen)

auch

FRA =französisch

ITA = italienisch

SPA = spanisch

etc.

Beispiele:

... das finde ich jetzt <*ENG>strange ...
... <*JAP>sayonara , Herr ~Fujisaki
... <*FRA>bien . <*FRA>c'est <*FRA>la <*FRA>vie würde ich sagen
...
... tschau , <*ITA>bella ...

Definition:

Fremdsprachige Wörter sind Wörter aus einer anderen Sprache als der im Dialogbeitrag mehrheitlich verwendeten.

Transliterationskonvention:

Fremdsprachige Wörter werden mit vorangestelltem <*ENG> (<*JAP>, <*ITA>, <*FRA>,<*GER>, ...) markiert.

Für fremdsprachige Wörter gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

Anmerkungen:

ABER: Manche fremdsprachigen Wörter stehen z.B. im deutschen Duden und brauchen im Fall deutscher Dialoge dann nicht als fremdsprachig markiert werden.

tschau

Meeting

****deutsch****

ABER: fremdsprachige Wörter, die mit Apostroph geschrieben werden, sollten, auch wenn sie im deutschen Duden zu finden sind, als fremdsprachig markiert werden, weil sonst Konflikte mit Text-Filtern für deutsche Wort-Reduktionen auftreten können.

<*FRA>d´accord

deutsch

1.6. Lexikalische Einheiten mit artikulatorischen Besonderheiten oder Defekten

1.6.1. Zögerung

Symbol:

<Z>

Beispiele:

... ich h"atte Zeit<Z> +/am<Z>/+ <A> <Schmatzen> ab Dienstag ...
... ich dacht<Z>e ger<Z>ade ...

Definition:

Werden Laute innerhalb einer lexikalischen Einheit verhältnismäßig lang gedehnt, z.B. gedehnte Laute vor Phrasengrenzen ("Prefinal Lengthening") oder im Sinne einer Häsitiation, wird dies als Zögerung eines Lautes protokolliert.

Das gilt auch für Plosive mit überlanger Verschußphase, starker oder andauernder Aspiration.

Transliterationskonvention:

<Z> wird ohne Leerstelle direkt an den gedehnten Laut angehängt. Bei Zögerung innerhalb eines Wortes erfolgt die Verschriftung des Wortendes direkt im Anschluß an <Z>.

Für Wörter mit gezögerten Lauten gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

1.6.2. schwerverständliche Wörter

Symbol:

%

Beispiele:

... #eins% , #zwei% , #drei% , #vier% ...
... wann h"atten Sie da% bitte Zeit ...

Definition:

Schwerverständlich sind alle Wörter, die nicht eindeutig verstanden werden können, weil sie durch Geräusche oder Artikulation akustisch schlecht gehört werden, oder bedingt durch dialektale oder artikulatorische Varianten, bei denen nur schwer auf die hochsprachige Version rückgeschlossen werden kann.

Transliterationskonvention:

Verschriftet wird ein Wort, das der schwerverständlichen Äußerung ähnlich klingt und auch im Kontext plausibel erscheint. Dem in der üblichen Orthographie transliterierten schwerverständlichen Wort wird % ohne Leerstelle angehängt.

Für schwerverständliche Wörter gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

Anmerkungen:

Schwerverständliche Wörter können nicht am Wortende abgebrochen sein.

Werden Wörter wegen dialektaler oder aussprachebedingter Variation schlecht verstanden, werden sie nach den Konventionen für schwerverständlich verschriftet. Ein Aussprachekommentar ist hier nicht nötig.

Wörter, die keinerlei Rückschlüsse auf hochsprachliche Formen zulassen und nicht als einem Dialekt zugehörig identifiziert werden können, werden nach der Konvention für unverständliche Sprachproduktionen verschriftet.

1.6.3. Abgebrochene lexikalische Einheiten

Definition:

Lexikalische Einheiten können artikulatorisch oder technisch abgebrochen oder unterbrochen sein und gehen als markierte Wortfragmente in die Transliteration mit ein.

Transliteration:

Die Abbruchstelle wird markiert, das verbleibende Wortfragment wird, solange sein Ursprung klar ist, in üblicher Orthographie nach den Konventionen für lexikalische Einheiten verschriftet. Klassifizierungssymbole bleiben erhalten (z.B. bei abgebrochene Namen oder Zahlen).

Anmerkung:

Abgebrochene lexikalische Einheiten können nicht zusätzlich als schwerverständlich markiert werden.

1.6.3.1. artikulatorischer Abbruch

Symbol:

=

Beispiele:

```
... +/#sieb=/+ #siebzehnter ...  
... -/im Ja=/- also ich sag' Ihnen jetzt ...
```

Definition:

Artikulatorisch abgebrochen heißt, der Sprecher selbst bricht eine lexikalische Einheit ab, meist um sich dann zu korrigieren.

Transliterationskonvention:

An ein artikulatorisch abgebrochenes Wort wird = ohne Leerstelle angehängt. Das Wortfragment wird bis zur Abbruchsstelle orthographisch dem wahrscheinlich zugrunde liegenden Wort entsprechend verschriftet.

Bei artikulatorischem Abbruch am Turnende wird statt des in diesem Fall unsinnigen Interpunktionszeichen ein technischer Turnabbruch protokolliert, auch wenn der Turn nicht aus technischen Gründen abgebrochen wurde.

z.B. .. -/gr"u"s Gott , Herr<Z>/- wie war der Na= <*T>t

1.6.3.2. Artikulatorische Unterbrechung lexikalischer Einheiten

Symbol:

_ (Subline)

Beispiele:

```
... <:<#> Ver_:> <A> <:<#> _pflichtungen:> ...  
... statt_ +/f=/+ <h"as> _findet ...  
... Acht-Uhr_ <"ah> _-Maschine ..  
... ~Zur-blauen_ +/Treb=/+ _-Traube ...
```

Definition:

Ein Wort kann durch Versprecher, eine Pause, Atmen oder durch Häsitationen unterbrochen sein. Nach den unterbrechenden Elementen wird das unterbrochene Wort fortgesetzt.

Transliterationskonvention:

An der Unterbrechungsstelle wird _ (Subline) ohne Leerstelle an das Wortfragment angehängt. Danach werden nach einer Leerstelle (oder Zeilenumbruch-Leerstelle) die unterbrechenden Elemente transliteriert.

Nach dem letzten dieser Elemente folgt wieder Leerstelle, nach einem weiteren Subline wird ohne Leerstelle die Fortsetzung des unterbrochenen Wortes verschriftet.

Reihenfolge der Markierungen:

Wortklassensymbole bleiben vor dem ersten Wortteil erhalten.

Beide Wortteile ("vorn_" und "_hinten") können jeweils sprecherüberlagert und/oder geräuschüberlagert sein. Die Überlagerungsmarkierungen stehen gegebenenfalls nach (vorderer Teil) bzw. vor (hinterer Teil) dem Subline.

Anmerkungen:

Falls unbedingt erforderlich, könnte im Falle unterbrochener Wörter der letzte Wortteil (die Fortsetzung des unterbrochenen Wortes) als schwerverständlich oder abgebrochen markiert werden.

Ist ein Kompositum zwischen den Kompositumsteilen unterbrochen, steht ein möglicher Bindestrich nach dem zweiten Subline vor dem zweiten Wortteil.

z.B. Wort_ <"ah> _-Unterbrechung

1.6.3.3. Technischer Abbruch

Symbol:

<*T> (technische Turnunterbrechung)

<*T>t (technischer Turnabbruch)

<T_>.. (Wortanfang fehlt)

..<_T> (Wortende fehlt)

Beispiele:

```
... auf Wie<_T> <*T>t
... ich h"atte am #vier<_T> <*T> <T_>wanzigsten Zeit ...
... danke , wi<_T> <*T> <T_>ffen uns ...
... k"onnten wir uns <*T> oder #acht-zehnten M"arz treffen ? ...
... dann bis <*T>t
... <T_>"u"s Gott <*T> <T_>ber , wir <T_>ssen noch einen Termin
ausmachen ...
... ich hab nur am <T_>zehnten Zeit ....
... danke , wi<_T> <*T>t
```

Definition:

Wackelkontakte im Aufnahmeequipment, Bedienungsfehler oder verspätetes/verfrühtes Knopfdrücken bei Knopfdruckaufnahmen führen zu technischen Unterbrechungen oder Abbrüchen der Dialogaufnahme.

Im Signalfile ist an diesen Stellen kein Amplitudenausschlag mehr zu sehen.

Aufnahmestörungen treten am Turnanfang, während des Turns und am Turnende auf.

Lexikalische Einheiten können davon direkt betroffen sein, wenn Anfang oder Ende abgeschnitten wurden.

Transliterationskonvention:

Turnanfang:

Setzt die Aufnahme erst während eines Dialogbeitrags ein, so wird <T_> ohne Leerstelle vor die

lexikalische Einheit oder abgeschnittene lexikalische Einheit gesetzt.

Während des Turns:

Bricht die Aufnahme während einer lexikalischen Einheit ab, dann wird an das Fragment <_T> ohne Leerstelle gehängt.

Fehlen durch die Aufnahmeunterbrechung Teile des Dialogbeitrags, steht anstelle dieser Elemente <*T> zwischen Leerzeichen für Turnunterbrechung.

Setzt die Aufnahme während einer lexikalischen Einheit wieder ein, wird für den fehlenden Wortanfang <T_> ohne Leerstelle vor das Fragment gesetzt.

Turnende:

Ist die Aufnahme am Turnende während einer lexikalischen Einheit abgebrochen, so wird an das Wortfragment <_T> ohne Leerstelle angehängt. Nach einer Leerstelle wird mit <*T>t der Turnabbruch signalisiert.

Wenn keine lexikalische Einheit direkt vom Abbruch betroffen ist, aber eine Äußerungsphrase eindeutig abgebrochen wurde, steht nach dem letzten Wort und einer Leerstelle <*T>t für Turnabbruch.

Bei einem Turnabbruch steht kein Interpunktionszeichen am Ende.

Anmerkungen:

Technischer Turnabbruch wird auch dann protokolliert, wenn der Sprecher seinen Dialogbeitrag selbst abbricht.

Beim Zusammentreffen von technischer Turnunterbrechung oder technischem Turnabbruch und Interpunktionszeichen, fällt das Interpunktionszeichen immer weg.

1.7. Aussprachekommentare

Symbol:

<!n ..>

wobei n = Anzahl der betroffenen lexikalischen Einheiten

Beispiele:

```
... Donnerstag <!1 Donnaschag> ...
... nat"urlich <!1 ´t"urlich> ...
... irgendwie <!1 irgen´wie> ...
... und dann <!2 un´a´> ...
... #sieben-und-zwanzig <!1 sienzanzesch> ...
... k"onnen wir <!2 k"omma> ...
... kannst du <!2 kannst> ....
... damit w"ar´ das <!1 des> eigentlich klar ...
... dann kommen !1 komm´> Sie <!1 Se> doch ...
... wenn wir <!2 wemma> ´s die Woche noch machen ...
```

Definition:

Dialektaussprachen, andere Stilformen, Versprecher oder sonstige Abweichungen von der üblichen Aussprache werden in korrekter Form nach Duden verschriftet. Im Aussprachekommentar wird versucht, mittels Orthographie und Apostrophregelung die Abweichungen so zu protokollieren, daß schon aus der Transliteration erste Informationen über Aussprachevarianten gezogen werden können und für tiefere Analysen die interessanten Stellen markiert sind.

Transliterationskonvention:

Nach einer Leerstelle (oder Zeilenumbruch-Leerstelle) folgt der betroffenen / den betroffenen lexikalischen Einheit/en der Aussprachekommentar.

Der Aussprachekommentar steht nach <! , einer Zahl, die die Anzahl der betroffenen lexikalischen Einheiten bezeichnet und einer Leerstelle. Der Aussprachekommentar schließt mit >.

Innerhalb der Kommentarklammern gelten in Zweifelsfällen die orthographischen Regeln. (Groß- und Kleinschreibung wird beibehalten, ebenso doppel-s oder ie-Schreibung, vorausgesetzt, die Aussprache legt nichts anderes nahe.

Ansonsten kann die Orthographie zur Verdeutlichung bestimmter Varianten benutzt werden (besonders langes /i/ statt kurzem /i/ wird dann z.B. mit "ie" transliteriert).

Lautelisionspositionen werden mit **einem** Apostroph gekennzeichnet.

Entfällt bei einem Wort der Endlaut/Endlaute , beim folgenden der Anfangslaut /Anfangslaute, so steht auch hier für die Elisionsposition **nur ein** Apostroph für die Elisionsstelle und keine Leerstelle zwischen den Wörtern.

Werden in der Aussprachevariante Laute durch andere Laute ersetzt oder Laute hinzugefügt, dann wird versucht, mittels geeigneter Buchstaben die veränderten Laute zu beschreiben. Fehlen in diesen Wörtern zusätzlich Laute, steht in diesem Fall **kein** Apostroph. Das veränderte Wort wird so transliteriert, als ob es sich um ein neues Wort handeln würde.

Zwischen enklitisierten Varianten steht kein Apostroph. Diese Wörter werden zusammengeschrieben.

Anmerkungen:

Die Zusammenfassung mehrerer variiert lexikalischer Einheiten in einen Aussprachekommentar sollte eher vermieden werden. Nur bei stark zusammengezogenen oder enklitisierten Einheiten bezieht sich der Kommentar notwendigerweise auf mehrere Einheiten.

****deutsch****

Aussprachevarianten wie /könig/ vs /könich/ brauchen nicht kommentiert werden, da beide Versionen im Deutschen üblich sind.

****deutsch****

Hinweis: **Genauer definierte Regeln** zur konsistenteren Verschriftung von Aussprachekommentaren finden sich im Verbmobil-Memo 111: Aussprachevarianten in der VERBMOBIL-Transliteration - Regeln zur konsistenteren Verschriftung (Burger, Kachelrieß, München, August 1996) [4].

2. syntaktisch-semantische Gliederung

Definition der Elementklasse:

Syntaktisch-semantische Gliederungen sind nicht-zeitintensive Markierungen zur Strukturierung des Satzflusses.

Soweit in der Spontansprache möglich, wird versucht, reguläre Satz und Nebensatzstrukturen mittels eines Subsets an Interpunktionszeichen zu markieren.

Grammatisch irreguläre Phänomene wie Korrekturen oder Satzabbrüche werden so markiert, daß sie mit Hilfe eines geeigneten Filters so aus dem Text genommen werden können, daß sinnvolle syntaktisch-semantische Strukturen entstehen.

2.1. Interpunktion

Symbol:

•
?
,

Definition:

Punkt, Fragezeichen und Komma stehen als Interpunktions-Subset zur syntaktischen Markierung regulärer Satzteile zur Verfügung. Generell ist aber die Zeichensetzung bei Spontansprache mit Schwierigkeiten verbunden, da grammatikalisch "richtige" Sätze häufig fehlen.

Transliterationskonvention:

Die Interpunktionszeichen stehen wie alle Turnelemente zwischen Leerzeichen (oder Zeilenumbruch-Leerzeichen).

Nach Punkt und Fragezeichen wird klein weiter geschrieben, sofern es sich nicht um ein Substantiv handelt. Vor und nach dem Interpunktionszeichen steht jeweils ein Leerzeichen.

Die Zeichensetzung erfolgt nach den Regeln der für die verwendete Sprache gültigen Grammatik, soweit möglich.

Punkt:

In Zweifelsfällen entscheidet, ob ein " ." gesetzt wird oder nicht:

- Grammatik
- Intonation
- Pause, Atmen
- Beginn eines neuen Gedankens

Beispiel:

... so , guten Tag . <A> <"ah> mein Name ist ~J"ansch . gestern

hatte ich schon mal ...

Fragezeichen:

In Zweifelsfällen entscheidet, ob ein " ? " gesetzt wird oder nicht:

- Fragewörter
- Satzbau
- Intonation
- Kontext

Beispiel:

... <A> wie schaut 's denn aus , den darauffolgenden Sonntag ,
den <:#Mikrobe> #neun-und-zwanzigsten:> bei Ihnen ?
geht 's da ? ...

Komma:

In Zweifelsfällen entscheidet, ob ein " , " gesetzt wird oder nicht:

- Grammatik
- einen Nebensatz einleitende Partikel
- Intonation
- zusammengehörende Gedanken
- Phrasenmarkierungen wie Atmen oder Pause

z.B. :

a, ich könnte schon , mittwochs hab´ ich immer Zeit .

b, ich könnte schon . <A> <P> mittwochs hab´ ich immer Zeit .

Beispiel:

... <"ahm> morgen , Freitag , <h"as> wie ich seh´ , <"ah> mu"s
ich feststellen , da"s ich <"ah> "überhaupt keine Zeit hab´

Anmerkungen:

Vor den Interpunktionszeichen finden sich nur lexikalische Einheiten oder Kommentare. Alle anderen Ereignisse, auch Atmen, Häsitationen oder komplett Unverständliches werden immer nach den Interpunktionszeichen verschriftet.

Komma kann nicht am Turnende stehen.

Nach Abbruchmarkierungen steht keine Interpunktion (Wortabbruch genauso wie Satzabbruch oder technische Unterbrechung).

2.2. agrammatische Phänomene

Definition:

Agrammatische Phänomene treten dann auf, wenn ein Sprecher sich innerhalb einer Äußerungsphrase unterbricht. Nach der Abbruchstelle werden Teile der Äußerungsphrasen wiederholt oder korrigiert (Wiederholung/Korrektur) oder aber es erfolgt keine Wiederaufnahme

und der Sprecher beginnt einen neuen Gedankengang (False Start). Markiert werden der Anfang der agrammatischen Phrase und die Abbruchstelle.

Siehe auch [3].

Transliteration:

Anfangs- und Endklammer der agrammatischen Phrase werden immer ohne Leerstelle vor (Anfang) oder nach (Ende) der ersten bzw. letzten lexikalischen Einheit verschriftet.

Klassifizierungssymbole und Abbruchmarkierungen gehören zur lexikalischen Einheit.

+/~Huber/+

+/Kart=/+

Ist eine der zu markierenden lexikalischen Einheiten geräusch- oder sprecherüberlagert, dann steht die Phrasenmarkierung vor bzw. nach der Überlagerungsmarkierung.

+/@1Kartoffel/+

+/Kartoffel1 @/+

+/<:<#> Kartoffel:>/+

Kommentare zu den betroffenen lexikalischen Einheiten stehen nach einer Leerstelle und der Phrasenmarkierung.

-/haben wir/- <!2 hamma>

2.2.1. Wiederholung / Korrektur

Symbol:

+/. .. /+

Beispiele:

.... +/<:<#> am:>/+ <:<#Mikrobe> am:> Donnerstag kann ich erst ...

... +/im Sep=/+ im September ...

... die Woche +/von/+ <"ah> mit Freitag ...

... also +/das/+ +/das/+ das #zweite ...

Definition:

Bei Wiederholung/Korrektur werden Teile von Äußerungsphrasen wiederholt oder korrigiert. Markiert werden

- der Anfang des im folgenden Wiederholten/Korrigierten (Reparandum)

- die Abbruchstelle, nach der dann wiederholt oder korrigiert wird (Reparatum).

Das "Reparandum" wird so geklammert, daß nach dessen Herausnahme zusammen mit der verbleibenden Wiederholung/Korrektur ein nahezu korrektes Satzgefüge entsteht.

Transliterationskonvention:

Direkt vor die erste lexikalische Einheit des Reparandums wird ohne Leerstelle die Anfangsklammer (+/) gesetzt.

Direkt an die letzte lexikalische Einheit des Reparandums, an der Satzgefüge-Abbruchstelle, folgt ohne Leerstelle die Endklammer (/+).

Wird die Korrektur/Wiederholung noch einmal korrigiert/wiederholt, dann wird wieder Anfang und Ende des neuen Reparandums geklammert.

Anmerkungen:

Nach einem Reparandum muß immer eine Korrektur/Wiederholung erfolgen.

2.2.2. False Starts (oder Neustart)

Symbol:

-/..../-

Beispiele:

...-/ab dem #dritten August <A> bis zum/- <P> Moment , ich ...
... -/ja , ich hab' da eigentlich/- also , ich bin vom
#neun-zehnten bis ...

Definition:

Der Sprecher beginnt eine Äußerung und bricht diese ab (False Start). Nach der Abbruchstelle erfolgt keine Wiederaufnahme und der Sprecher beginnt einen neuen Gedankengang (Neustart). Markiert wird der gesamte False Start, also der Anfang des abgebrochenen Satzgefüges, meistens nach einem Interpunktionszeichen, und die Abbruchstelle.

Transliterationskonvention:

Direkt vor die erste lexikalische Einheit des abgebrochenen Satzgefüges wird ohne Leerstelle die Anfangsklammer(-/) gesetzt.

Direkt an die letzte lexikalische Einheit des abgebrochenen Satzgefüges, an der Satzgefüge-Abbruchstelle, folgt ohne Leerstelle die Endklammer (/).

Anmerkungen:

False Starts können nicht am Turnende stehen, da ja dann kein neuer Gedanke mehr folgt. In diesem Falle wird ein Turnabbruch (<*T>t) protokolliert.

3. nonverbale artikulatorische Produktionen

Definition der Elementklasse:

Artikulatorische Produktionen des Sprechers ohne erkennbaren semantischen Gehalt heißen nonverbale artikulatorische Produktionen.

Hierunter fallen:

1. Atmen
2. Häsitationen
3. vollkommen unverständliche Äußerungen
4. artikulatorische Geräusche wie Lachen und Husten

Verschriftungsregeln:

Alle nonverbalen artikulatorischen Produktionen stehen zwischen spitzen Klammern (<>). Sie können als zeitintensive Turnelemente sprecherüberlagert sein, Atmen, Häsitationen oder Unverständliches können auch aktiv Beiträge des anderen Sprechers überlagern. Überlagernde artikulatorischen Geräusche werden nach den Konventionen für Geräuschüberlagerung protokolliert.

Elementumgebung:

Diese Turnelemente können **nicht vor** Interpunktionszeichen stehen.

3.1. Atmen

Symbol:

<A>

Beispiele:

... ist G<Z>"urtner , <A> <"ahm> \$G \$"U \$R \$T \$N \$E <A> \$R

Definition:

Atmen als einzelnes Turnelement wird bei deutlich hörbarem Ein- oder Ausatmen protokolliert.

Transliterationskonvention:

Atmen wird mit <A> transliteriert.

Anmerkungen:

Ausatmen nach einem Plosiv wird nicht als Atmen, sondern als gezögerte Plosion markiert (<Z>). Treffen Atmen und Interpunktionszeichen zusammen, dann steht <A> immer nach dem Interpunktionszeichen.

3.2. Häsitationen

Symbol:

<"ah>
<"ahm>
<hm>
<h"as>

Beispiele:

... <"ah> wir hatten bereits telefoniert<Z> ...
... gu<Z>t , <"ahm> wie w"ar' es bei Ihnen am<Z> #neun-zehnten
Juli ...
... -/ich denke/- <hm> also bei mir ginge es sehr gut ...
... w"urd' ich sagen , <A> <h"as> wenn wir ...

Definition:

Häsitationen (auch gefüllte Pausen genannt) sind Zögerungen, die als Einzelereignis zwischen Äußerungen zu finden sind (im Gegensatz zu den gezögerten Lauten innerhalb eines Wortes <Z>).

Transliterationskonvention:

Häsitationen stehen zwischen spitzen Klammern.

Die Vielzahl der möglichen Häsitationen, die mittels der Orthographie nicht entsprechend repräsentiert werden kann, wird mit einer der vier Häsitationsklassen transliteriert:

- <"ah>: rein vokalische Artikulation, unabhängig von der Vokalqualität
- <"ahm>: vokalische Artikulation + nasale Artikulation
- <hm>: rein nasale Artikulation
- <h"as>: seltenere Artikulationen, die nicht in eine der ersten drei Klassen einzuordnen sind. Hierunter fallen z.B. /brrt/ /pf/ /puh/ etc.

Anmerkungen:

Häsitationen werden gegebenenfalls immer nach Interpunktionszeichen transliteriert.

3.3. Unverständliche Sprachproduktionen

Symbol:

<%>

Beispiele:

... aber <A> <"ah> <%> <"ahm> wie w"ar' 's denn ...
... <%> fr"uher geht 's leider nicht ...
... ich k"onnte am <%> geht% das bei Ihnen ?

Definition:

Wird eine Äußerung eines Sprechers überhaupt nicht verstanden, wegen akustischer Überlagerung, oder unverständlicher Artikulation, kann also auch durch den Kontext nicht identifiziert werden (im Gegensatz zu schwerverständlichen Äußerungen), dann wird an Stelle dieser Äußerung das Symbol für vollkommen unverständliche Sprachproduktion transliteriert.

Transliterationskonvention:

Für Unverständliches steht <%> .

Anmerkungen:

<%> steht gegebenenfalls immer nach einem Interpunktionszeichen. Im Zweifelsfall muß auf das Interpunktionszeichen verzichtet werden.

3.4. nonverbale artikulatorische Geräusche

Symbol:

<Schmatzen>

<Schlucken>

<R"auspern>

<Husten>

<Lachen>

<Ger"ausch> (= sonstiges artikulatorisches oder nicht-identifizierbares artikulatorisches Geräusch)

Beispiele:

... Mittagessen <P> <Schmatzen> und<Z> , na ja ...
... ordentlich planen mu"s und <"ah> <Schlucken> wann w"urde Ihnen
...
... einverstanden . <P> <R"auspern> auf Wiedersehen ...
... +/je/+ <Husten> je schneller wir das machen ...
... bei <:<Lachen> mir:> terminlich sehr ung"unstig ...
... dann/- <"ah> <A> <Ger"ausch> das sind #sechs Termine ...
... k"onnten Sie am #dritten ? <Husten> Entschuldigung

Definition:

Nonverbale artikulatorische Geräusche bezeichnen artikulatorische Produktionen des Sprechers, die den Sprachfluß unterbrechen oder Äußerungen überlagern.

Transliterationskonvention:

Alle nonverbalen artikulatorischen Geräusche stehen zwischen spitzen Klammern.
Um die Vielzahl der artikulatorischen Geräusche einzuschränken, stehen nur noch die fünf

häufigsten artikulatorischen Geräusche zur Verfügung. Alle anderen artikulatorischen Geräusche werden mit der Allgemeinklasse <Ger"ausch> verschriftet.

Anmerkungen:

Falls andere Personen im Hintergrund ein artikulatorisches Geräusch verursachen, wird dies auch als artikulatorisches Geräusch, unter Umständen als Geräuschüberlagerung, in die Transliteration mit übernommen.

Hintergrundreden wird mit <Ger"ausch> verschriftet.

In diesem Fall wird aber nicht Sprecherüberlagerung, sondern Geräuschüberlagerung protokolliert.

Artikulatorische Geräusche stehen gegebenenfalls nach einem Interpunktionszeichen.

Artikulatorische Geräusche können nur passiv sprecherüberlagert sein (Das heißt z.B., Sprecher A hustet und Sprecher B spricht, während Sprecher A hustet). Bei aktiver Überlagerung gelten die Konventionen zur Geräuschüberlagerung lexikalischer Einheiten.

4. Geräusche und technische Artefakte

Symbol:

<#Klicken>
<#Klingeln>
<#Klopfen>
<#Mikrobe>
<#Mikrowind>
<#Rascheln>
<#Quietschen>
<#>

Beispiele:

```
... de<Z>m <:<#Klicken> #neun-und-zwanzigsten:> August ...  
... <:<#Klingeln> am:> <:<#Klingeln> Mittwoch:> <:<#Klingeln>  
st"anden:> ...  
... <:<#Klopfen> gern:> <:<#Klopfen> so:> <:<#Klopfen> machen:>  
...  
... <#Mikrobe> das tut mir leid ...  
... ich k"onnte Ihnen <:<#Mikrowind> vorschlagen:> ...  
... nach Berlin <P> ausmachen . <#Rascheln> <"ahm> ...  
... der gesamte <:<#Quietschen> Rest:> des Mais ...  
... k"ame mir gelegen , <A> <#> allerdings ...
```

Definition:

Bei Geräuschen und technischen Artefakten handelt es sich um diejenigen hörbaren Turnelemente, die nicht durch die Artikulation eines Sprechers produziert wurden.

Diese Ereignisse treten zum Teil im Zusammenhang mit der Bandaufnahme des Dialogs auf (Mikrofonberührungen, Klicken bei Knopfdruckaufnahmen) oder werden durch Gegenstände im

Hintergrund verursacht (Rascheln, Klopfen, Klingeln). Sie sind entweder während Sprechpausen zu hören oder überlagern lexikalische Einheiten.

Transliterationskonvention:

Technische Geräusche werden in spitzen Klammern und mit vorangestelltem # verschriftet. Sie können als einzelne Ereignisse auftreten oder lexikalische Einheiten überlagern.

Um die Vielzahl der technischen Geräusche einzuschränken, stehen nur noch die sieben häufigsten Geräusche zur Verfügung. Alle anderen technischen Geräusche werden mit der Allgemeinklasse <#> verschriftet:

<#Klicken> für z.B. Klicken des Knopfdrucks

<#Klingeln> Telefonklingeln

<#Klopfen> z.B. Tischberührung

<#Mikrobe> Mikrofonberührung

<#Mikrowind> ins Mikrofon blasen oder ins Mikrofon atmen

<#Rascheln> z.B. Papierrascheln

<#Quietschen> z.B. Stuhlquietschen

<#> Geräusche, die in keine der anderen Kategorien passen oder nicht identifiziert werden können

Anmerkungen:

Technische Geräusche stehen gegebenenfalls hinter Interpunktionszeichen.

Technische Geräusche können nur passiv sprecherüberlagert sein (Das heißt z.B., während einer Sprechpause von Sprecher A ist Rascheln zu hören, Sprecher B spricht während dieser Sprechpause in den Dialogbeitrag von Sprecher A hinein). Bei aktiver Überlagerung gelten die Konventionen zur Geräuschüberlagerung lexikalischer Einheiten.

5. Sprechpausen

Symbol:

<P>

Beispiele:

... gr"u"s Gott , Herr <P> ~Huber . <P> wir m"ussen noch ...
... am Dienstag habe ich um<Z> <P> <Husten> <P> <A> #vier-zehn
Uhr Zeit

Definition:

Wenn ein Sprecher seinen Sprachfluß kurz stoppt, um beispielsweise Grenzen zu markieren oder kurz nachzudenken, wird eine Pause protokolliert. Während einer Pause ist im Signalfile noch ein kleiner Amplituden-Ausschlag, bedingt durch normales Hintergrundrauschen oder Aufnahmeauschlag zu sehen (im Gegensatz zur technischen Aufnahmeunterbrechung).

Transliterationskonvention:

Pausen werden mit <P> transliteriert.

Ereignisse, die während einer Sprechpause stattfinden, wie Atmen oder Geräusche, überlagern nicht die Pause, sondern werden als eigene Elemente transliteriert.

Anmerkungen:

Sprechpausen am Anfang oder Ende eines Turns werden **nicht** transliteriert.

Pausen, die mit Interpunktionszeichen zusammentreffen, stehen nach dem Interpunktionszeichen.

Pausen können als zeitintensives Turnelement passiv sprecherüberlagert sein. Das heißt, der Dialogpartner kann in eine Sprechpause des anderen Partners hineinsprechen, überlagert aber damit den Beitrag des Partners, weil dieser noch nicht zu Ende ist.

6. akustische Überlagerung

6.1. Sprecherüberlagerung

Symbol:

..n@ (passive Sprecherüberlagerung lexikalischer Einheiten)

@n.. (aktive Sprecherüberlagerung lexikalischer Einheiten)

..n@> (passive Sprecherüberlagerung sonstiger Ereignisse)

<@n.. (aktive Sprecherüberlagerung sonstiger Ereignisse)

Beispiel:

Spr A: hallo1@ , <"ah">1@> ich bin der Herr Huber2@.

Spr B: @1hallo . <P> @2ah @2ja . wie geht es <:<#>Ihnen3@:> denn3@ ?

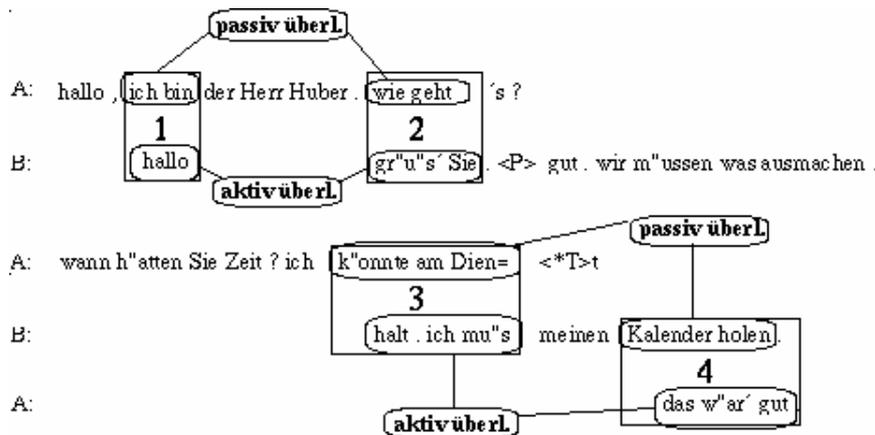
Spr A: @3gu<Z>t . <@3<A> wir müssen noch einen Termin aus_ <"ah"> _machen4@ . wann4@ können Sie ?

Spr B: @4ah @4ja . ich könnte am ...

Definition:

Bei Dialogen, bei denen die Sprecher nicht durch technische Gegebenheiten dazu gezwungen sind, nur dann zu sprechen, wenn der Gesprächspartner gerade nicht spricht (z.B. bei Knopfdruckdialogen), kann es zur gegenseitigen Überlagerung von Gesprächsbeitrags-elementen kommen (bestätigende Interjektionen, "dazwischenreden", unterbrechen usw.).

schematische Darstellung:



Sprecher B überlagert das "ich bin" im ersten Beitrag von Sprecher A mit "hallo" und das "wie geht 's" mit "grüß Sie".

Sprecher B unterbricht den zweiten Beitrag von Sprecher A, wobei sich die Teile "könnte am Dien=" (Sprecher A) und "halt, ich muß" (Sprecher B) überlappen. Sprecher A bricht hier seinen Dialogbeitrag ab.

Sprecher B wird dann in der Fortführung seiner Äußerung bei "Kalender holen" von "das wäre gut" (Sprecher A) überlagert.

Aus der Darstellung wird klar,

- daß jeder Sprecherbeitrag auch bei zeitlicher Überlappung in einem eigenen Turn steht
- daß passive (das Überlagerte) und aktive (das Überlagernde) Überlagerung durch gleichzeitiges Auftreten ein Paar bilden.

Transliteration:

Bei einer Sprecherüberlagerung werden zeitintensive Turnelemente, die gleichzeitig stattfinden, markiert.

sprecherüberlagerte /überlagernde Turnelemente können sein:

Passiv überlagerte Elemente, die als sprecherüberlagert markiert werden:

- lexikalische Einheiten
- Atmen
- alle nonverbalen artikulatorischen Produktionen
- länger anhaltende Geräusche, falls sie als Einzelelement auftreten
- Sprechpausen

Aktiv überlagernde Elemente, die als sprecherüberlagernd markiert werden:

- lexikalische Einheiten
- Atmen
- Häsitationen
- unverständliche artikulatorische Produktionen

ABER: artikulatorische und technische **Geräusche, die aktiv Sprachproduktionen überlagern,** werden als **Geräuschüberlagerung** im überlagerten Turn transliteriert!

Kennzeichnung:

Bei jedem passiv überlagerten Element wird die Überlagerungsmarkierung ohne Leerstelle an das Element hinten angehängt (unabhängig von davorstehenden Markierungen wie Abbruch oder schließende spitze Klammer bei z.B. Atmen).

z.B.

Kartoffel1@

Kart=1@

<A>1@>

Jedem aktiv überlagernden Element wird die Überlagerungsmarkierung ohne Leerstelle vorangestellt (unabhängig von eventuell weiteren Symbolen wie Wortklassenmarkierungen oder öffnende spitze Klammer bei z.B. Häsitiation)

z.B.

@1Kartoffel

@1#drei

<@1<"ah>

Um die Zusammengehörigkeit einer passiven mit einer aktiven Überlagerung auszudrücken, steht in den Überlagerungsmarkierungen eine Zahl, die für das Überlagerungspaar gleich bleibt (auch bei mehreren betroffenen Elementen).

Diese Zahl wird während des gesamten Dialogs hochgezählt.

(erstes Überlagerungspaar, zweites Überlagerungspaar, drittes ..)

z.B.

A: hallo , ich1@ bin1@ der Herr Huber . wie2@ geht2@ ´s ?

B: @1hallo . <P> @2gr"u"s´ @2Sie .

Bei Multiparty-Aufnahmen, bei denen theoretisch mehrere Sprecher gleichzeitig sprechen können, wird der überlagerte Beitrag eines Sprechers A als **passiv** überlagert gekennzeichnet. Die Überlagerungen der anderen Sprecher, auch wenn diese sich wieder gegenseitig überlagern, werden als **aktiv** überlagernd gekennzeichnet, die Zahl in den Markierungen bleibt **immer die gleiche wie die der passiven Überlagerung**, um die Zusammengehörigkeit der Überlagerungen zu kennzeichnen.

z.B.

A: hallo , ich1@ bin1@ der Herr Huber . wie2@ geht ´s ?

B: @1hallo . <P> @2ach . <P> ja , gut3@ .

C: @1hallo . <P> @3gut .

Anmerkungen:

Für sprecherüberlagerte lexikalische Einheiten gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten,

für sprecherüberlagerte sonstige Ereignisse gelten die entsprechenden Konventionen zur Transliteration dieser Ereignisse.

Agrammatische Phrasenklammern und Geräuschüberlagerungsklammern stehen nach der passiven, bzw. vor der aktiven Überlagerungsmarkierung.

Bei unterbrochenen lexikalischen Einheiten wird jeder Wortteil einzeln markiert. Die Markierungen für Sprecherüberlagerung stehen dann nach dem Subline der Unterbrechungsstelle bzw. vor dem Subline der Wiederaufnahme.

z.B.

A: Hutschachtel . wiederholen1@ Sie1@ das1@ .

B: @1*Hund_ <@1<"ah> @1_schachtel . oder Hut2@_ +/sche=2@/+ _schachtel .
A: @2genau .

Siehe auch C.1. Reihenfolge der Markierungssymbole .

Überlagert ein Sprecher immer wieder den Beitrag des anderen mit kleinen Einwüfen, so steht zwischen den überlagerten Elementen Sprechpause <P> . Falls während der Pausen Geräusche zu hören sind, werden diese als Geräuschüberlagerungen im **Turn des anderen Sprechers** transliteriert.

Um die Konventionen parsbar zu halten, müssen sich die Symbole für überlagernde/überlagerte lexikalische Einheiten (..n@ / @n..) und denen der sonstigen überlagernden Elemente wie Atmen, Pausen oder Häsitationen (..n@> / <@n..) etwas unterscheiden. Das zu beschreibende Phänomen ist davon jedoch nicht betroffen.

Die Zählung der überlagerten Stellen im Dialog erfolgt **unabhängig von den verschiedenen Markierungssymbolen**, also bei Sprecherüberlagerung lexikalischer Einheiten mit lexikalischen Einheiten, anderer Elemente mit anderen Elementen oder einer Mischung der beiden Markierungsmöglichkeiten (..1@ / @1.. , ..2@> / <2@.. , ..3@ / <@3.. , ..4@> / @4.. , ...).

6.2. Geräuschüberlagerung

Symbol:

<:<..> ..:>

<..> steht für alle möglichen artikulatorischen und technischen Geräuschkategorien (z.B. <:<#> ..:>).

Beispiele:

```
... de<Z>m <:<#Klicken> #neun-und-zwanzigsten:> August ...  
... bei <:<Lachen> mir:> terminlich sehr ung"unstig ...  
... <:<#Klopfen> <Lachen> k"amen:> <:<#Klopfen> <Lachen> mir:>  
<:<#Klopfen> <Lachen> die:> <:<#Klopfen> <Lachen> Monate:> April  
...  
... <:<Lachen> ich:> <:<#Klopfen> <Lachen> habe:> <:<Lachen>  
heute:> keine Zeit ...
```

Definition:

Lexikalische Ereignisse können von **einem oder mehreren** Geräuschen überlagert sein,
- die entweder von einem Sprecher produziert werden (z.B. Husten) oder verursacht werden (z.B. in das Mikrofon blasen) oder
- die im Hintergrund zu hören sind wie Rascheln, Klopfen oder Hintergrundgemurmel.

Transliterationskonvention:

Eine geräuschüberlagerte lexikalische Einheit wird zusammen mit der entsprechenden Geräuschkategorie-Bezeichnung mit <: und :> geklammert.

Nach der öffnenden Klammer (<:) folgt das überlagernde Geräusch, eine Leerstelle und eventuell weitere überlagernde Geräusche plus jeweils folgender Leerstelle, danach wird die überlagerte lexikalische Einheit transliteriert (gegebenenfalls mit entsprechenden Anfangssymbolen wie bei Namen oder Sprecherüberlagerung).

Direkt im Anschluß an die lexikalische Einheit folgt ohne Leerstelle die schließende Klammer (:>), unabhängig mit welchem Symbol die lexikalische Einheit abschließt (Abbruch, schwerverständlich, Wortunterbrechung oder Sprecherüberlagerung).

Anmerkungen:

Nur bei lexikalische Einheiten wird Geräuschüberlagerung transliteriert.

Für jede überlagerte lexikalische Einheit wird die Geräuschüberlagerung **extra** markiert, **eine Klammerung mehrerer lexikalischer Einheiten**, die von einem durchgehenden Geräusch überlagert werden, **ist nicht mehr möglich!**

Auch bei Wortunterbrechung wird jeder Wortteil **extra** mit Geräuschüberlagerung markiert. Falls ein ganzer Turn mit einem durchgehenden Geräusch überlagert ist, kann auf die Einzelmarkierung verzichtet werden und stattdessen in einem globalen Kommentar darauf hingewiesen werden.

Für geräuschüberlagerte lexikalische Einheiten gelten ansonsten die allgemeinen Konventionen für lexikalische Einheiten.

Die öffnende Geräuschüberlagerungsklammer einschließlich der überlagernden Geräusche steht

- nach der öffnenden Klammer der agrammatischen Phrase ohne Leerstelle
- vor allen anderen Symbolen wie Sprecherüberlagerung, technisch vorne abgeschnittenem Wort oder Wortklassifizierungssymbolen oder dem Subline einer Wortfortführung bei Wortunterbrechung.

Die schließende Überlagerungsklammer steht

- nach eventuellen Symbolen, die artikulatorische Besonderheiten der lexikalischen Einheit bezeichnen
- nach dem Symbol für passive Sprecherüberlagerung,
- nach dem Subline einer Wortunterbrechung,
- vor der Endmarkierung bei agrammatischen Phrasen.

Siehe auch C.1. Reihenfolge der Markierungssymbole.

7. lokale Kommentare

Symbol:

<;..>

Beispiele:

```
... zum Beispiel <:"übersteuert> ...
... einen Termin vereinbaren <;heiser> ...
... am #eins% , #zwei% , #drei% <;Zahlen gefl"ustert> #dritten
M"arz ...
```

Definition:

Hinter jedem Turnelement kann im Prinzip ein lokaler Kommentar stehen. In einem lokalen Kommentar werden Bemerkungen zu Besonderheiten einzelner Turnelemente direkt nach den betreffenden Turnelementen vermerkt (im Gegensatz zum globalen Kommentar zum gesamten Turn). Das können bestimmte Sprechstile wie "emphatisch" oder "geflüstert" sein oder Bemerkungen bezüglich falsch verwendeter Grammatik.

Transliteration:

Lokale Kommentare werden im Text hinter der relevanten Stelle und einer Leerstelle eingefügt, durch ein Semikolon eingeleitet und in spitze Klammern eingeschlossen.

Innerhalb der Kommentare gelten die normalen orthographischen Regeln, bei Umlauten wird TeX-Schreibweise verwendet.

8. Sonderkommentare

8.1. Code-Wörter

falls zur Steuerung des VERBMOBIL ein Codewort benötigt wird:

Symbol:

!KEY!

Beispiele:

```
.. ich h"atte am Dienstag Zeit !KEY!Verbmobil-Ende ...
```

Definition:

Um das VERBMOBIL per Stimme zu steuern, sei es, daß z.B. Äußerungsanfang und Äußerungsende akustisch gekennzeichnet werden sollen, eine Teiläußerung übersetzt werden oder VERBMOBIL ein Gesprächsprotokoll anfertigen soll, kann ein sogenanntes Codewort benutzt werden, das nicht in die Übersetzung mit eingeht, sondern eine Funktion des VERBMOBILs aufruft.

Transliterationskonvention:

Wörter, die nur als Systemsteuerelement dienen, werden mit vorangestelltem !KEY! markiert.

Code-Wörter werden wie lokale Kommentare behandelt und nicht wie lexikalische Einheiten. Sie können auch ohne Interpunktionszeichen am Turnende stehen.

8.2. szenario-bedingte Pause

Symbol:

<PP>

Definition:

Falls es das Aufnahme-Szenario erfordert, können lange Pausen, etwa während der Sprecher auf die Übersetzung des VERBMOBILs wartet, extra protokolliert werden.

Diese Pausen unterscheiden sich von den normalen Sprechpausen, weil sie durch die Aufnahmesituation bedingt künstlich erzeugt werden und nicht wie Sprechpausen syntaktische oder semantische Funktionen im Sprachfluß wahrnehmen.

Transliterationskonvention:

Künstliche Pausen werden mit <PP> transliteriert.

Diese Pausen werden wie lokale Kommentare behandelt und können auch am Anfang oder Ende eines Turns transliteriert werden.

C. Tabellen

1. Reihenfolge der Markierungssymbole bei lexikalischen Einheiten

agram.	Geräusch	Spr.üb.	Anf.Symb.	Lex.	defekt.	Spr.üb.	Geräusch	agram.
+/ -/	<:<...>	@1	\$ ~ # * <*ENG> _[-] <T_> (-)	Wort	= <_T> % (-)	1@	:>	/+ /-

2. Liste aller verwendbarer Symbole

<*tENG>	fremdsprachiger Turn (JAP, GER, ..)
;..	globaler Kommentar
..'..	Apostroph (Wort-Reduktion)

..-.. (--)	Bindestrich (Komposita)
\$.	Buchstabierung und Buchstabiersequenzen
≈	Namen
#..	Zahlen
*..	Neologismus
<*XXX>..	fremdsprachiges Wort (FRA,ITA, ..)
..<Z>..	Zögerung
..%	schwerverständliches Wort
..=	artikulatorischer Wortabbruch
.._	Wortunterbrechung, linkes Fragment
_..	Wortunterbrechung, rechtes Fragment
<T_>..	technischer Wortabbruch, vorne
..<_T>	technischer Wortabbruch, hinten
<*T>	technische Turn-Unterbrechung
<*T>t	technischer Turn-Abbruch
<!n ..>	Aussprachekommentar
. / ? / ,	Interpunktion
+/.	Beginn Wiederholung/Korrektur
../+	Ende Wiederholung/Korrektur
-/..	Beginn False Start
../-	Ende False Start
<A>	Atmen
<"ah>	Häsitation/gefüllte Pause
<"ahm>	Häsitation/gefüllte Pause
<hm>	Häsitation/gefüllte Pause
<h"as>	Häsitation/gefüllte Pause
<%>	unverständliche Sprachproduktion
<Schmatzen>	nonverbale artikulatorische Geräusche
<Schlucken>	nonverbale artikulatorische Geräusche
<R"auspern>	nonverbale artikulatorische Geräusche
<Husten>	nonverbale artikulatorische Geräusche
<Lachen>	nonverbale artikulatorische Geräusche
<Ger"ausch>	nonverbale artikulatorische Geräusche

<#Klicken>	technische Geräusche
<#Klingeln>	technische Geräusche
<#Klopfen>	technische Geräusche
<#Mikrobe>	technische Geräusche
<#Mikrowind>	technische Geräusche
<#Rascheln>	technische Geräusche
<#Quietschen>	technische Geräusche
<#>	technische Geräusche
<P>	Sprechpause
@n..	aktive Sprecherüberlagerung
..n@	passive Sprecherüberlagerung
<@n..	aktive Sprecherüberlagerung von Ereignissen
..n@>	passiver Sprecherüberlagerung von Ereignissen
<:<..> ..	Beginn Geräuschüberlagerung
..:>	Ende Geräuschüberlagerung
<;,..>	lokaler Kommentar
!KEY!..	Codewort
<PP>	szenario-bedingte Pause

D. Literatur:

[1] K.Kohler , G.Lex , M.Pätzold , M.Scheffers , A.Simpson, W.Thon: Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL -3.0 .Verbmobil-Technischer Report, Nr. 11 . Kiel. September 1994.

[2] S. Burger: Transliterationslexikon. Verbmobil Technisches Dokument, Nr. 36. München. Oktober 1995

[3] A. Batliner, S. Burger, A. Kießling: Außergrammatische Phänomene in der Spontansprache: Gegenstandsbereich, Beschreibung, Merkmalinventar. Verbmobil-Report, Nr. 5. München, Erlangen. Februar 1994.

[4] S.Burger, E.Kachelrieß: Aussprachevarianten in der VERBMOBIL-Transliteration - Regeln zur konsistenteren Verschriftung. Verbmobil-Memo, Nr. 111. August 1996.

[5] A. Batliner, A. Kiessling, S. Burger, E. Noeth: Filled Pauses in Spontaneous Speech. Verbmobil-Report, Nr. 88. München, Erlangen. Juli 1995.