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

Labeling von User–States im Mensch–Maschine Dialog – User–State–Kodierkonventionen SmartKom Version 2

Silke Steininger

Olga Dioubina

Rolf Siepmann

Cibran Beiras–Cunqueiro

Angelika Glesner

Ludwig Maximilians Universität München

Technisches Dokument Nr. 17
September 2001

September 2001

Silke Steininger, Olga Dioubina, Rolf Siepmann, Angelika Glesner, Cibran Beiras–Cunqueiro

Ludwig Maximilans Universität München
Schellingstr. 3
80799 München

Tel.: (089) 2180/5751 oder 2180/2760
FAX: (089) 2800362

E–Mail: [kstein] [olga]@phonetik.uni–muenchen.de

Dieses Technische Dokument gehört zu Teilprojekt 1: Modalitätsspezifische Analysatoren und Projektfeld II: Anwendungsszenarien

Das diesem Technischen Dokument zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01 IL 905 gefördert. Die Verantwortung für den Inhalt liegt bei den Autoren.

Inhaltsverzeichnis

1 Einführung

1.1 Überblick

Dieses Dokument beschreibt die Vorgehensweise beim Labeln der Benutzerzustände (User−States) im Projekt SmartKom.

Für das Projekt SmartKom werden Versuchspersonen bei der Benutzung eines Dialogsystems aufgenommen, das (Spontan−)Sprache und gestische Eingaben interpretieren kann. Außerdem soll der emotionale Zustand (und zusätzlich auch nicht rein emotionale Zustände wie Nachdenken) des Benutzers (in der Mimik und der Sprache) erkannt werden.

Bei den Aufnahmen der Mensch−Maschine−Dialoge sollen die Versuchspersonen mit Hilfe des Systems Aufgaben aus verschiedenen Anwendungsgebieten (Planung eines Kinobesuchs, einer Stadt Besichtigung, eines Fernsehabends u.ä.) lösen.
Das Ergebnis sind einerseits Videodaten, andererseits spontansprachliche Daten, die als Grundlage zur Forschung und Entwicklung im Bereich der Spracherkennung dienen.

Die vorliegenden Kodierkonventionen wurden anhand einer Teilmenge der WOZ−Aufnahmen festgelegt (die Kino−Aufnahmen). Ausgangspunkt der Label waren die Vereinbarungen auf dem Datenworkshop bei der LMU in München (11.10.−13.10.99) und einige Änderungen dazu, die in zwei Arbeitstreffen zwischen LMU und FAU Erlangen (4.5.01, 23.5.01) besprochen wurden.

Bei der Festlegung der Label−Kategorien standen folgende Ziele im Vordergrund:

- möglichst große Aussagekraft in Bezug auf den Mensch−Maschine−Dialog
- Gute Eignung als Trainingsmaterial für die Erkenner
- gute Beobachtbarkeit/Unterscheidbarkeit
- einfaches, schnelles System

Weitere Informationen über das User−State Labeln in [5], [6].

Die drei Arbeitsschritte beim Labeling der Benutzerzustände

Das Labeln der User−States umfaßt drei getrennte Arbeitsschritte:

1. **Das holistische Labeln** der Mimikspur mit Audio (User−State holistisch, USH)

Beim holistischen Labeln werden die Videosequenzen von einem Labeler mit einem Tool [1] anhand der hier beschriebenen Kodierkonventionen bearbeitet. Jedem identifizierten Benutzerzustand wird eines von sieben USER−STATE LABELN zugeordnet, jeweils mit Start− und Endzeitpunkt. Der Dialog wird vollständig gelabelt, d.h. es gibt keine ungelabelten Abschnitte. Das Label wird nach dem Gesamteindruck vergeben, der sich aus der Mimik, aber auch aus Informationen wie Stimme, Wortwahl und situativem Kontext zusammensetzt. Wichtig bei der Beachtung des Kontextes ist, daß tatsächlich der *Eindruck* einer Emotion besteht. Legt lediglich das Verhalten der Vp eine bestimmte Emotion nahe, fehlt der emotionale Eindruck jedoch, wird das Labeln *nicht* vergeben!

Das Label wird außerdem mit einem RATING versehen, der sich auf die Intensität/Stärke des Zustandes (schwach oder stark) bezieht. Dieses Rating wird nach dem subjektiven Eindruck vergeben.

Neben den User−State Labeln gibt es NON−USER−STATE LABEL, die vergeben werden, wenn sich die Hand, Teile der Hand, ein Stift oder ein Objekt im Gesicht befinden. Sie können parallel zu den User−State Labeln vorkommen.

Das holistische Labeln dient dazu, möglichst alle Episoden zu erfassen, in denen der Benutzer einen interessanten Zustand zeigt. Daher werden alle Kontextinformationen berücksichtigt.

Die Klassifikation eines Benutzerzustandes beim holistischen Labeln umfaßt also:

- User−State Label
- Non−User−State Label
- Rating

2. Das Labeln der Mimikspur ohne Audio (User−State Mimik, USM)

Ein Labeler (der nicht am USH−Labeln beteiligt ist) bearbeitet die Videospur ohne Audiospur. Als Vorlage bekommt er ein gefiltertes USH−File, in dem alle User−State Label (außer Neutral) gelöscht sind, alle Segmentgrenzen und alle Non−User−State Label jedoch noch enthalten sind. Beim USM Labeln werden alle neutralen Segmente und alle Non−User−State Segmente ignoriert und nur bei den User State Segmenten neue Label und neue Ratings vergeben, sowie u.U. neue Grenzen.

Das Mimik−Labeln dient überwiegend dem Training des Mimikerkenners. Der Erkenner−Algorithmus kann nur die Mimikdaten verarbeiten, die Kontextinformationen sind für ihn irrelevant. Ein Ärger−Zustand, der ohne Kontext wie Freude erscheint, ist für ihn praktisch unmöglich zu erkennen. Daher haben wir mit dem Labelschritt ohne Audio versucht, die Voraussetzungen für Mimikerkenner und Mensch anzugleichen. Für einen Labeler ist die Situation ohne Audio ähnlich − aber nicht 100%. Wenn er einen "stummen" User−State sieht, dann fehlt ihm lediglich der Audio−Kontext. Andere Kontextinformationen (wie z.B. Weltwissen) kann er nutzen.

Man könnte vermuten, daß ein menschlicher Beobachter viele Episoden verpassen würde, wenn er sie nur aus der Mimik erschließen muß (da er gewohnt ist, emotionale Zustände aus dem Gesamtkontext zu erschließen, nicht nur aus der Mimik). Daher haben wir uns entschlossen, zuerst den holistischen Labelschritt durchzuführen und anschließend das mimische Labeln.

Hier könnte man einwenden, daß es durchaus sein kann, daß die Labeler beim Labeln ohne Audio sogar genauer sind als beim holistischen Labeln, da sie ev. nicht durch die Sprache abgelenkt werden. Dem möchten wir entgegensetzen, daß die Fehler, die beim holistischen Labeln gemacht werden, valider sind: Sie entsprechen eher den Fehlern bei der Emotionserkennung, die ein Mensch in einer natürlichen Kommunikationssituation macht, in der er sein Gegenüber sehen und hören kann. Auch aus diesem Grund sollte der holistische Schritt in unseren Augen vor dem Schritt ohne Audio erfolgen − nicht möglichst genaues Labeln ist das Ziel, sondern ein Labeln das Kategorien erzeugt, die möglichst genau den Kategorien entsprechen, die ein Zuhörer in einem Gespräch seinem Gegenüber zuordnet.

Die beschriebene "Filtermethode" wurde als Mittelweg zwischen rascher Verarbeitung und möglichst unbeeinflußtem Labeln gewählt.

3. Das Annotieren der Audio Spur (User−State Prosodie, TRP)

Hier handelt es sich um einen getrennten Schritt mit einer anderen Methodik. Ein gefiltertes Transliterationsfile wird noch einmal abgehört und mit neun formalen Labeln versehen. Die Label beziehen sich auf prosodische Eigenschaften der Sprache, von denen angenommen wird, daß sie mit den Benutzerzuständen der Sprecher korrelieren.

Ziel ist, im Vergleich mit den USH Labeln für die User−State Erkennung in der Stimme prosodische Merkmale zu finden, die sich für die automatische User−State Erkennung nutzen lassen.

1.2 Eingrenzung des Labelvorgangs

Das Labeln teilt den Mensch−Maschine Dialog in Episoden auf, in denen jeweils ein bestimmter Benutzer−Zustand vorherrscht. Die gewählten Label sind Sammel−Kategorien, d.h. das Label Ärger/Mißerfolg umfaßt Ärger, Irritation, Verärgerung, Wut, Unmut etc. Die einzelnen Kategorien wurden nach ihrer Aussagekraft für den Mensch−Maschine Dialog und nach ihrer gegenseitigen Abgrenzbarkeit gewählt.

Das prosodische Labeln (TRP) erfaßt den Benutzerzustand in der Stimme nach formalen Kriterien, von denen vermutet wird, daß sie ein Ausdruck des zugrunde liegenden emotionalen Zustandes sind. Es hat den Vorteil, daß die Kriterien gut objektivierbar sind und vermutlich eine hohe Inter−Labeler Reliabilität ergeben. Der Nachteil liegt in einer wahrscheinlich eingeschränkten Validität: Es ist nicht sicher, daß die gewählten formalen Kriterien wirklich die sind, die den zugrundeliegenden Zustand gut beschreiben.

Das holistische Labeln hat sicher eine höhere Validität. Allerdings definieren sich die Label explizit nach dem subjektiven Eindruck des Labelers. Die Inter−Labeler Übereinstimmung muß daher niedriger ausfallen.

Wir hoffen, durch eine Kombination einer formalen und einer subjektiven Herangehensweise, sowie durch die Berücksichtigung bzw. Nicht−Berücksichtigung von Kontextinformationen ein möglichst gutes Abbild der für den Mensch−Maschine Dialog wichtigen Benutzerzustände zu erfassen.

2 Definitionen User−States USH/USM

2.1 Die User−State Label

USM und USH unterscheiden sich nur in der Vorgehensweise. Die Label/die Kategorien beim USH− und beim USM−Labeln sind identisch.

Jede Episode, in der der Benutzer einen einheitlichen Zustand zeigt, wird einem der folgenden Label zugeordnet:

1. Freude/Erfolg
2. Ärger/Mißerfolg
3. Ratlosigkeit
4. Überlegen/Nachdenken
5. Überraschung
6. Neutral
7. Restklasse

Das Kriterium für diese Unterteilung ist der subjektive Eindruck des Labelers. Ziel des Labelvorgangs ist es also nicht, den "wahren" Benutzer−Zustand zu erfassen. Wir betrachten es als völlig ausreichend, den Zustand zu erfassen, der bei einem Kommunikationspartner (in diesem Fall das System) ankommt. Die zu beantwortende Frage ist also "In welchem Zustand ist der Benutzer offensichtlich?" und NICHT "Was empfindet der Benutzer tatsächlich?"
Da Vp ihre Emotionen sehr individuell zeigen und auch individuell unterschiedlich stark, orientiert sich die Vergabe der Label und der Ratings an der Vp. Ob "Freude/Erfolg" als "stark" gelabelt wird, hängt entsprechend davon ab, wie stark die Vp insgesamt "Freund/Erfolg" zeigt, bzw. ihre User−States allgemein.

Die Grenzen einer Episode werden jeweils dort gesetzt, wo eine Änderung im Benutzerzustand ersichtlich wird.
Wenn sich die Intensität innerhalb eines User−States bzw. einer Sequenz ändert, so wird dies auch festgehalten, z.B.:

Überlegen/Nachdenken	stark
Überlegen/Nachdenken	schwach

2.2 Die Non−User−State Label

Wenn das Gesicht zum Teil von einer oder beiden Händen, Teilen der Hand oder einem Stift verdeckt sind oder wenn das Gesicht teilweise nicht in der Kamera zu sehen ist, wird eines der folgenden Labels vergeben:

1. Hand im Gesicht
2. Stift im Gesicht
3. Teilweise nicht im Bild
4. Hand im Gesicht/Mund
5. Hand im Gesicht/Augen
6. Hand im Gesicht/Nase
7. Stift im Gesicht/Mund
8. Stift im Gesicht/Augen
9. Stift im Gesicht/Nase
10. Objekt im Gesicht

"Teilweise nicht im Bild" wird vergeben, wenn das Gesicht der Vp von der Kamera nicht vollständig erfaßt wird (weil sie sich z.B. bewegt hat). Wenn nur einzelne Haarpartien nicht in der Kamera zu sehen sind, dann wird das Label *nicht* vergeben.
Label 4−9 beschreiben Verdeckungen von Mund, Augen und Nase mit der Hand bzw. einem Stift. Befinden sich Hand oder Stift woanders im Gesicht oder ist die Verdeckung nicht klar abgrenzbar bezüglich Mund, Augen oder Nase (z.B. wenn zwei dieser Regionen überdeckt werden), dann wird Label 1 oder 2 vergeben.

Die Non−User−State Label werden parallel zu den User−State Labeln vergeben, d.h. beide Kategorien können gleichzeitig vorkommen.

2.3 Die User−State Label im Einzelnen

2.3.1 FREUDE/ERFOLG:

Hat der Labeler den Eindruck, daß die Vp irgendeine positive Emotion, wie Freude, Erfolg, Glück, Zustimmung o.ä. ausdrückt, dann wird der Event als Freude/Erfolg gelabelt. Formale Hinweise können auf mimischer Ebene sein: Lächeln, Hochgezogene Augenbrauen, leichtes Zurückschlagen des Kopfes, weiter geöffnete Augen, sichtbare Zähne. Auf stimmlicher Ebene sind höhere und/oder lautere Stimme, freundlicher Tonfall und Lachen hinweisgebend.

2.3.2 ÄRGER/MISSERFOLG:

Als Ärger/Misserfolg werden alle negativen Äußerungen der Vp vermerkt, wie z.B.: Gereiztheit, Genervtheit, Enttäuschung, Ärger, Irritation, Unmut. Hinweisgebend sind Merkmale wie zusammengezogene Augenbrauen, ein breitgemachter Mund, der Frustration ausdrückt, nach oben verdrehte oder geschlossene Augen, gerümpfte Nase, aufgeklappter Mund. In der Stimme sind folgende Kriterien Hinweise für diese Kategorie: lautere und oft tiefere Stimme, Vp spricht langsamer und deutlicher, ermahnender Ton, eventuell auch abgehakt.

2.3.3 ÜBERRASCHUNG/VERWUNDERUNG:

Wird immer vergeben, wenn die Vp den Eindruck macht, sie sei überrascht oder verwundert. Diese Kategorie kommt häufig am Anfang einer Aufnahme vor und ist fast immer eine Reaktion auf eine Informationsausgabe des Systems. Formale Kriterien können auf der Mimikebene sein: Schnell hochgezogene Augenbrauen, Mund und/oder Augen öffnen sich, Kopf geht ein wenig zurück. In der Stimme kann folgendes für die Emotion "Überraschung" hinweisgebend sein: Ausrufe wie "huch", "ach" oder
Fragesätze, z.B. "Was war denn das?", lautere Stimme und höhere Stimme.

2.3.4 ÜBERLEGEN/NACHDENKEN:

Als Überlegen/Nachdenken werden alle Episoden gelabelt, in denen die Vpn nachdenklich, überlegend o.ä. wirkt. Die Aufgabe der Vpn führt dazu, daß in einem Großteil der Sessions konzentriert geschaut, gelesen, gesucht u.ä. wird. Dieses *Verhalten* wird nicht als der User–State "Überlegen/Nachdenken" gelabelt. Erst bei einem deutlichen Eindruck, daß die Vpn nachdenkt, wird das Label vergeben. D.h. wenn das Gesicht nicht mehr entspannt ist (Stirnrunzeln, auf die Lippe beißen o.ä.) und/oder die Stimme überlegend wirkt. Formale Hinweise in der Mimik können auch sein: Augen konzentriert nach oben richten, Unterlippe nach innen ziehen, zusammengekniffene Augen, angespannte Augenbrauen, offener Mund, Luft anhalten, Oberkörper anspannen, z. B. Schultern, Kopf hin und her werfen. In der Stimme: Äußerungen wie "mmh" oder "ääh", gedehnte Sprache, Flüstern, mit sich selbst sprechen, Brabbeln, "Sing–Sang" in der Stimme, Knarrstimme.

2.3.5 RATLOSIGKEIT:

Wird vergeben, wenn die Vp den Eindruck macht, sie ist ratlos, verwirrt, hilflos; wenn sie nicht weiß, wie es weitergeht, wenn sie fragend aussieht (wobei nicht tatsächlich eine Frage gestellt werden muß). Wichtig bei der Vergabe des Labels ist es, daß die Vp tatsächlich von ihrer Emotion ratlos wirkt. Legt ihr Verhalten Ratlosigkeit nahe, fehlt aber der emotionale Eindruck, dann wird das Label nicht vergeben! Wichtige Hinweise im Gesicht sind: Verdutzter Ausdruck, hilfloser Blick mit hochgezogenen Augenbrauen, vorgestülpter Mund, gerümpfte Nase, Mund klappt auf, Kopf schräg halten, nach Hilfe suchender Blick. Stimmliche Merkmale: Hilfesuchende Fragen, z.B. "Was soll ich jetzt
tun"?, Frageintonation.

2.3.6 NEUTRAL:

Diese Kategorie wird immer dann vergeben, wenn der Benutzer in keinem speziellen emotionalen oder anderweitigen Zustand ist. Er ist ruhig, das Gesicht ist entspannt, die Stimme ist gleichmäßig oder er zeigt eine der obigen Zustände in einem so geringen Ausmaß, daß die Zuweisung zu einem der obigen Label nicht gerechtfertigt erscheint.

Die Dauer ist üblicherweise sehr lang, da es sich (meist) um die Grundhaltung der Vp handelt.

"Neutral" wird aus offensichtlichen Gründen nicht als stark oder schwach gerated.

2.3.7 RESTKLASSE:

Alle Zustände, die sich nicht in eine der bisherigen Kategorien einteilen lassen, werden der "Restklasse" zugeordnet, also alle nicht–neutralen, nicht identifizierbaren emotionalen Zustände. Auch gemischte Emotionen fallen unter diese Kategorie. Bei diesem Label wird wie bei Neutral kein Rating vergeben.

Ein Sonderfall, der ebenfalls als "Restklasse" gelabelt wird, ist ein "schwarzes Bild", das entstehen kann, wenn die Mimikspur kürzer ist als die anderen Spuren. Betrifft die Sessions w403, w405, w406, w412 und w418.

3 Labeln der Prosodie (TRP)

Überblick

Die prosodische Annotation soll die jeweilige Einstellung des Sprechers zum System innerhalb der Dialoge anhand von formalen Kriterien erfassen. Sie beruht auf den Kriterien von Fischer [2].

Ein Labeler bekommt dazu ein gefiltertes Transliterationsfile und fügt mit einem Labeltool [1] die unten aufgeführten Label ein. Es wird vorerst mit gefilterten Basistransliterationen [3] gearbeitet und später geprüft, inwieweit die prosodischen UserState−Label mit Annotationen der Basistransliteration konvergieren. Die gefilterten Basistransliterationen enthalten keine −linguistischen− prosodischen Label für Pausen, Grenzen, Akzente und für finale Phrasenverläufe. Außerdem sind Zögerungen eliminiert. Im einzelnen gibt es die folgenden prosodischen Label zur Annotation der UserStates:

1.[PAUSE_PHRASE]
2.[PAUSE_WORD]
3.[PAUSE_SYLL]
4.[LENGTH_SYLL]
5.[EMPHASIS]
6.[STRONG_EMPH]
7.[CLEAR_ART]
8.[HYPER_ART]
9.[LAUGHTER]

Bei der Vergabe der Label gelten die folgenden Regeln:

– Jeder Dialog wird vor der prosodischen Annotation einmal angehört, um einschätzen zu können, wann der Sprecher für seine Verhältnisse "normal spricht"; dies ist für die Einschätzung von [EMPHASIS], [STRONG_EMPH], [CLEAR_ART] und [HYPER_ART] wichtig.
– An dieser Stelle können bereits globale Kommentare wie z.B. "Sprecher ist sehr aufgeregt" in den Header eingetragen werden.
– Ein Wort kann mehrere prosodische Label haben, die hinter dem Wort aufgelistet sind. Bei mehreren Pausen zwischen verschiedenen Silben eines Wortes wird nur ein [PAUSE_SYLL] vergeben.
– Nach dem letzten Label folgt das Zeitbullet, das die Länge des gesamten gelabelten Wortes beinhaltet.
– Im Fall von [PAUSE_PHRASE] und [PAUSE_WORD] umfaßt das Zeitbullet das vorhergehende Wort plus die Länge der Pause.
– Werden zwei Label bei einem Wort vergeben, von denen nur eines eine Pause erfaßt, dann muß für jedes Label ein Zeitbullet gesetzt werden.
– Bei Ausdrücken, die in der Basistransliteration mit einem Pluszeichen verschriftet sind, kommt das Label und das Zeitbullet ¥ direkt hinter den entsprechenden Teilausdruck, z.B. "der Film Aimee [PAUSE_SYLL] ¥ +und+Jaguar."
– Es werden auch Gliederungspartikel wie "hm", "äh" usw. prosodisch annotiert.
– Bei auftretenden Aussprachekommentaren stehen die prosodischen Label hinter dem Kommentar. Ein an das Wort angehängte Label für Off−Talk wird nicht abgetrennt.

– Die Prosodie–Label (PL) stehen grundsätzlich mit einem Leerzeichen hinter dem Wort; anschließend folgt das ebenfalls mit einem Leerzeichen verbundene Zeitbullet.
– Bei Offtalk <OOT>, <ROT> stehen die PL dahinter
– Wenn Aussprachekommentare (= von der Verschriftung abweichende Aussprachen) auftreten, dann stehen PL nach diesen Kommentaren.
– Wenn PL ein Wort eines Aussprachekommentars betrifft, dann kommt das Label direkt hinter das Wort, z.B.: Jacob+der+Lügner <!1 Jakob [EMPHASIS] {%snd ...}und der Lügner>.
– Wenn PL z.B. in Jacob [EMPHASIS] {%snd ...}+der+Lügner, dann PL immer VOR dem Pluszeichen.
– Bei agrammatischen Phänomenen, Wiederholungen, Korrekturen +/ ... /+ sowie bei false starts –/ ... / – stehen PL und Zeitinformation innerhalb der Transliterationslabel dem entsprechenden Wort.
– Bei mehreren PLs hinter einem Wort besteht keine vorgegebene Reihenfolge. In diesem Fall wird genau ein Zeitbullet gesetzt; Ausnahme sind Pausen, die ein eigenes Zeitbullet erhalten.
– Wort + PL + Satzzeichen, also PL vor den Satzzeichen.
– Wort + PL + Turnabbruch, also PL vor dem Zeichen für Turnabbruch.
– Falls es sich um eine asyndetische Reihung handelt, z.B. "Ich würde gerne nach Berlin [Pause], München [Pause], Wien fahren", werden extrem lange Pausen als [Pause_Phrase] gelabelt. Das Label plus Zeitbullet stehen stets vor dem Komma, d.h. direkt nach dem betroffenen Wort.
– Wenn mehrere Wörter innerhalb einer Phrase als [Clear_Art] gelabelt werden müssen (oder auch anderes), sollten sie auch jeweils einzeln gelabelt werden. Das Prinzip ist die Wörter zu labeln, nicht die Phrasen.
– Wenn es Schwierigkeiten gibt, beim allein stehenden Wort zwischen [EMPHASIS] und [STRONG_EMPHASIS] zu unterscheiden, sollte man den vorgehenden Satz oder auch den ganzen Dialog nachmal anhören, um die kontextuelle und sprecherabhängige Gesprächsdynamik einschätzen zu können.

3.1 Die TRP Label im Einzelnen

zu 1. [PAUSE_PHRASE]: Pausen zwischen sinntragenden Einheiten. Nicht gemeint sind Pausen zwischen einzelnen Sätzen oder Pausen zwischen Haupt– und Nebensatz, es sei denn, sie ist besonders lang (Anmerkung: Sind in der Basistransliteration mit PP gekennzeichnet). Gemeint ist vielmehr eine nicht übliche Pause, wie z.B. in "der Film [PAUSE_PHRASE] im Kino Europa."

zu 2. [PAUSE_WORD]: Pausen zwischen Worten
Bsp.: der Film im [PAUSE_WORD] Kino Europa.

zu 3. [PAUSE_SYLL]: Pausen zwischen den Silben einen Wortes
Bsp.: Wochen<P>end<P>termin (Anmerkung: In der Basistransliteration werden an dieser Stelle vermutlich Zögerungen <Z> gekennzeichnet).

zu 4. [LENGTH_SYLL]: Silbendehnung
Kann beliebige Silbe(n) eines Wortes erfassen.
Bsp. der Filmmmm [LENGTH_SYLL] im Kiiiino [LENGTH_SYLL] Europa.
(Anmerkung: Kann in der Basistransliteration einer Zögerung <Z> entsprechen)

zu 5. [EMPHASIS]: starke Betonung eines Wortes bzw. einer Silbe eines Wortes
Bsp. MONtag (Anmerkung: Kann in der Basistransliteration dem Prominenzakzent entsprechen).

zu 6. [STRONG_EMPH]: sehr starke Betonung eines Wortes bzw. einer Silbe eines Wortes. Bsp. MOONtag (Anmerkung: Kann in der Basistransliteration dem Emphaseakzent entsprechen).

zu 7. [CLEAR_ART]: deutliche Aussprache
Die deutliche Aussprache kann als schwache Form von [HYPER_ART] gelten. In diesem Fall wird versucht, eher hochdeutsch zu sprechen, also nicht umgangssprachlich und nicht dialektal, etwa wie Nachrichtensprecher.

zu 8. [HYPER_ART]: Hyperartikulierte Aussprache
(Übermäßige) Steigerung der deutlichen Aussprache.

zu 9 [LAUGHTER]: von Lachen oder Seufzen verzerrte Sprache
Dies betrifft Wörter, die von Lachen oder Seufzen überlagert sind, z.B. Termihihin. auch: Nervosität; irrelevant sind durch technische Geräusche u.ä. verzerrte Sprache.

4 Ablauf des Label−Vorgangs

USH

1. Erstlabeling mit dem Programm Clan − Segmentierung und Labeln der User−States

2. Korrektur − Überprüfung der gelabelten User−States auf inhaltliche und formale Korrektheit.

3. End−Korrektur − Abermalige Überprüfung der gelabelten User−States auf inhaltliche und formale Korrektheit. Erstlabeler, Korrektor und End−Korrektor sind natürlich unterschiedliche Personen.

4. Umwandlung des Clan−Files in das USH−Format (siehe Abschnitt 5) mit dem Skript [skchatoush *cha]. Ev. Verbesserung von formalen Fehlern (das Umwandlungsskript entdeckt einige Fehler).

5. Ausführung des Skriptes [synchrocha *cha]. Clan setzt eine Segmentgrenze nicht immer an der Framegrenze. Das Skript korrigiert dies.

6. Überprüfen der formalen Korrektheit mit einem Checker: [test_ush *ush]. Der Checker überprüft die Schreibweise aller Labels, den formalen Aufbau des Files und die Bündigkeit der Label.

7. Ev. Korrektur des Labelfiles und erneute Umwandlung und erneuter Check.

8. Ausfüllen des Headers.

9. Auslieferung.

USM

1. Umwandlung des fertigen *ush Files in das Clanformat mit dem Skript [ush2cha *ush]. Dabei werden alle User−State Label (außer "Neutral") gelöscht. Die Non−User−State Label bleiben erhalten.

2. Erstlabeling mit dem Programm Clan − Segmentierung und Labeln der User−States.

3. Korrektur − Überprüfung der gelabelten User−States auf inhaltliche und formale Korrektheit.

4. Umwandlung des Clan−Files in das USM−Format (siehe Abschnitt 5) mit dem Skript [skchatousm *cha]. Ev. Verbesserung von formalen Fehlern (das Umwandlungsskript entdeckt einige Fehler).

5. Ausführung des Skriptes [synchrocha *cha]. Clan setzt eine Segmentgrenze nicht immer an der Framegrenze. Das Skript korrigiert dies.

6. Überprüfen der formalen Korrektheit mit einem Checker: [test_usm *usm]. Der Checker überprüft die Schreibweise aller Labels, den formalen Aufbau des Files und die Bündigkeit der Label.

7. Ev. Korrektur des Labelfiles und erneute Umwandlung und erneuter Check.

8. Ausfüllen des Headers.

9. Überprüfung, ob USH und USM gleich lang sind mit dem Skript [cmpusmush]. Ev. Korrektur der unterschiedlichen Länge. Falls sich beide Files nur um einige wenige Frames unterscheiden, kann das längere gekürzt werden mit dem Skript [cutusmush].

10. Überprüfung, ob die Non−User−State Label bei USH und USM gleich lang sind mit dem Skript [chk_nonemotional_synch]. Ev. Korrektur von Fehlern und erneuter Check.

11. Auslieferung.

4.1 TRP

1. Konvertierung
Die TRL−Files wurden mit zwei Scripts i) und ii) konvertiert.
i) ~**smart/bin/filter_trl2xusp.pl** erzeugt aus den trl−Files (Eingabe) xusp−Files (Ausgabe), die die folgenden Label der trl−Files nicht mehr enthalten:
– prosodische Label: B2 cont/ fall/ rise; B2, B9, EK, NA, PA
– Geräusche
– Pausen
–!Key
– ~ (Namenstilden)
– Überlagerungszeichen @x, x ist eine Zahl, und deren Klammerung
ii) ~**smart/bin/xusp2clan.pl** konvertiert die xusp−Files (Eingabe) in cha−Files (Ausgabe), die von Clan gelesen werden können.

2. Prosodische Annotation mit Clan.

Clan ist eine Windows/Mac−Applikation – http://childes.psy.cmu.edu/html/clan.html – mit der die cha−Files prosodisch annotiert werden.

Arbeiten mit Clan:

a) cha−file öffnen

b) "In sonic mode" im Menue Mode wählen und korrespondierenden Sound der transliterierten Aufnahme angeben; es sollten die wav−Dateien, die mit dem Headset (*h*.wav) – oder Richtmikrophon (*d*.wav) aufgenommen wurden, verwendet werden; Idealerweise sollten cha− und wav−File lokal auf dem Rechner in einem Verzeichnis liegen (Clan merkt sich keine Verzeichnisnamen).

c) **STRG−W** ruft die möglichen Label auf, die dann in den TRL−Text eingefügt werden können; das Zeitbullet des markierten Signalabschnitts wird durch den Button "S" (unten links beim Signal) oder mit **STRG−I** eingefügt.

d) Damit mit STRG−W die Label aufgerufen werden können, müssen die Label in die Datei CED.Prefs eingetragen sein; diese Datei wird bei der Installation von Clan in dem Windowssystemverzeichnis unter .../Clan/ erzeugt und kann editiert werden.

3. Umwandlung des Clan−Files in das TRP−Format und Vergleich mit dem entsprechenden TRL−File mit dem Skript [chkchawithparse *cha *trl]. Ev. Verbesserung von allen möglichen Fehlern (das Umwandlungsskript entdeckt einige Fehler) findet in *cha (!) nicht *trp File statt.

4. Erneute Umwandlung und erneuter Check mit dem gleichen Skript.

5. Auslieferung der *trp Files.

5 Struktur der Label−Files

Beschreibung Label−File USH

Ein USH−Label−File ist ähnlich zu einem BAS Partitur File (BPF) aufgebaut und orientiert sich an den Standards des BAS [4].

Der Name ergibt sich aus der Session−Nr. und der Endung "ush", also z.B. w059_pk.ush oder w091_mt.ush.

Am Anfang steht ein Header, der enthält:
- DVD: Nummer der DVD auf der die Session distribuiert wird
- Dialog: Nummer der gelabelten Session
- Version: Bei jeder Änderung im USH wird die Versionsnr. hochgezählt.
- zuletzt bearbeitet am: Datum der Endkorrektur
- Aufnahme−Qualität: ok oder Angabe von ev. Problemen
- Anmerkungen: sonstige Anmerkungen
- Erst−Labeling, Korrektur, End−Korrektur: Namen der Labeler
- VKP: Sprecherkürzel

Jede Headerzeile ist mit einem Strichpunkt als Kommentar gekennzeichnet.
Das File endet mit einer Kommentarzeile (; EOF = End of File).

Jedes User−State Segment beginnt mit USH:
Anschließend sind − jeweils durch einen Tabulator getrennt − aufgeführt:
- Onset User−State Segment
- Dauer User−State Segment
- Label (User−State Label oder Non−User−State Label)
- Rating (außer bei Non−User−State Label, Neutral und Restklasse)
- Anmerkung

Gibt es kein Rating oder keine Anmerkung, steht an der Stelle ein Tabulator.

Alle Zeiten sind in Samples angegeben, um konform zu den BPFs zu sein. Um eine Zeitangabe in sec zu erhalten, muß der Wert durch 16 000 geteilt werden (da die Sample Rate 16 kHz ist).

Alle Label sind unter den entsprechenden Abschnitten oben erklärt.

Beispiel Label−File USH

```
; DVD: 5
; Dialog: w037_pk
; Version: 1.0
; zuletzt bearbeitet am: 7.9.01
```

```
; Aufnahme-Qualität: Video verschwommen
; Anmerkungen: --
; Erst-Labeling USH: Roberto
; Korrektur USH: Angelika
; End-Korrektur USH: Silke
; VPK: AAS
;
USH:    0       199680  Neutral
USH:    199680  14720 Restklasse
USH:    214400  72960  Neutral
USH:    287360  111360 Überlegen/Nachdenken   schwach
USH:    398720  696960  Neutral

[snip]

USH:    3616640         87040   Neutral
USH:    3703680         65280   Überlegen/Nachdenken   schwach
USH:    3768960         53760   Überlegen/Nachdenken   stark
USH:    3810560         16640   Hand im Gesicht
USH:    3822720         225920  Neutral
; EOF
```

Beschreibung Label−File USM

Der Name eines USM Labelfiles ergibt sich aus der Session−Nr. und der Endung "usm", also z.B. w059_pk.usm.

Ein USM−Label−File ist genauso aufgebaut wie ein USH−File mit folgenden Unterschieden:

– zuletzt bearbeitet am: Datum der Endkorrektur USM
– Labeling USH: Erst−Labeler, Korrektor, End−Korrektor des USH−Durchgangs
– Labeling USM: Erst−Labeler, Korrektor USM
– Jedes User−State Segment beginnt mit "USM:"

Bei USM gibt es nur einen Korrektur−Durchgang.

Alle Label sind unter den entsprechenden Abschnitten oben erklärt.

Beispiel Label−File USM

```
; DVD: 133
; Dialog: w153_mn.usm
; Version: 1.2
; zuletzt bearbeitet am: 03.04.03
; Aufnahme-Qualität: Das Bild ruckelt
; Anmerkungen: --
; Labeling USH: Roberto, Angelika, Silke
; Labeling USM: Nadja, Hana
; VPK: ACY
;
USM:  0    298880 Neutral
USM:  149760 156160 Teilweise nicht im Bild
USM:  298880 58880  Überlegen/Nachdenken  schwach
USM:  357760 281600 Neutral
USM:  368640 34560  Teilweise nicht im Bild

[snip]

USM:    4105600         524160  Neutral
USM:    4629760         10240   Überlegen/Nachdenken   schwach
;EOF
```

Beschreibung Label−File TRP

Der Name eines TRP Labelfiles ergibt sich aus der Session−Nr. und der Endung "trp", also z.B. w059_pk.trp.

Die Label−Files der UserStates Prosodie (TRP) unterscheiden sich in drei Punkten von den üblichen Trl−Files (vgl. [3]):
– Der Header weist am Anfang vier zusätzliche Zeilen auf, die mit Prosodie−UserStates beginnen, s.u. 4.1
– Version: Bei jeder Änderung im TRP wird die Versionsnr. hochgezählt. Gestartet wird mit 1.0 (nicht mit der Versionsnr./TRV des TRLs).

– Der Trl–Text ist um die oben beschriebenen Symbole in eckigen Klammern angereichert, sofern die entsprechenden Phänomene auftreten.
– Die prosodischen Label sind mit Zeitinformationen der folgenden Art verbunden (vgl. 3.1): {%dateityp:"DATEINAME"_Beginn(ms)_Ende(ms)}

5.1 Beispiel Beschreibung Label–File TRP

```
;Prosodie-UserStates, Erstverschriftung: olga
;Prosodie-UserStates, Korrektur: rolf
;Prosodie-UserStates, Datum: 3.07.01
;Prosodie-UserStates, Kommentar:
; DVD: 8.0
; Version: 1.0
; Dialog: p002_pk
; zuletzt bearbeitet am: 21.11.00
; Tonqualität: SON teilweise "ubersteuert
; ATMO: Demovideo
; die "sprachlichen Besonderheiten" (Prosodie, <L>) des Wizards wurden
; nicht berücksichtigt
; Erst: sonjan Pros: sonjab, Korr: dani
; Offtalk: wenig
; VPK: SON
;
p002_pkd_001_SON: also , Aladdin [LAUGHTER] {%snd:"W047_PKD_AAX.WAV"_38590_39070} , wach auf .

p002_pkw_002_SMA: herzlich willkommen bei SmartKom . wie kann ich Ihnen helfen ?

[snip]

;EOF
```

6 Literatur

[1] Label–Tool: Clan, siehe http://childes.psy.cmu.edu/

[2] Fischer, K. (1999). Annotating Emotional Language Data. Verbmobil Report 236.

[3] Beringer, N., Oppermann, D., Burger, S. (00): Transliteration spontanprachlicher Daten–Lexikon der Transliterationskonventionen–SmartKom (Version1). SmartKom Technisches Dokument Nr. 2, Februar 2000.

[4] Bayerisches Archiv für Sprachsignale – BAS, http://www.phonetik.uni–muenchen.de/Bas/

§ Danksagung

Unser Dank geht an alle USH–, USM– und TRP–Labeler: Roberto, Stefanie, Iris, Karin, Andrea, Anke, Christa, Angelika, Christelle, Nadja, Anke, Sonja, Hana, Irene, Ulrich.