BMBF

Verbmobil
Verbundvorhaben

Standardisierung orthographischer Transkriptionen: Ein SGML/TEI-basierter Vorschlag für VERBMOBIL

Andreas Witt, Harald Lüngen, Dafydd Gibbon

Universität Bielefeld

Vm Memo 117
Januar 1997

Januar 1997

Andreas Witt, Harald Lüngen, Dafydd Gibbon

Universität Bielefeld (UBI)
Fakultät für Linguistik und Literaturwissenschaft
Universitätsstr. 25
Postfach 10 01 31
33501 Bielefeld

Tel.: (0521) 106 - 3510
Fax: (0521) 106 - 6008
e-mail: {luengen|gibbon}@Spectrum.Uni-Bielefeld.DE

Gehört zum Antragsabschnitt: 6.4.1 Kommunikation,
WWW-Dokumentation, Sprecherdatenbasis

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 B 2 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

Inhaltsverzeichnis

1 Einleitung

Die Standard Generalized Markup Language ist seit 1986 der internationaler Standard (ISO 8879:1986) für die Kodierung und Annotation von Texten. In ihr werden Text– und Strukturinformationen kodiert.

Die *Text Encoding Inititive* (TEI) hat einen Standard zur Annotation nichttechnischer Texte im Paradigma der SGML veröffentlicht.

Im folgenden wird kurz SGML vorgestellt und auf die Nutzbarkeit für VERBMOBIL eingegangen. Darauf wird auf das *base tag set* "TEI.spoken" eingegangen. In den Anhängen befindet sich eine tabellenartige Gegenüberstellung von TRL-Elementen mit TEI-SGML-Elementen, und als Beispiel eine TEI-konforme Kodierung des Dialogs n010k von der CDROM 1. In Anhang D befindet sich eine Liste internationaler wissenschaftlicher Projekte, in denen Texte nach dem TEI-Standard formatiert werden, um die bereits enorme Verbreitung dieses Standards zu unterstreichen.

2 Die *Standard Generalized Markup Language* (SGML)

Die *Standard Generalized Markup Language* (SGML) ist eine Metasprache für spezialisierte Markup-Sprachen, d.h. eine Sprache, mit der verschiedene anwendungsortientierte Markup-Sprachen beschrieben werden (z.B. HTML).

In Abbildung 1 ist der prinzipielle Ablauf der Erstellung eines SGML-Dokuments zu sehen. In SGML wird eine Strukturbeschreibung einer Klasse von Dokumenten definiert. Die bekannteste dieser so definierten Klassen von Dokumenten ist die *Hyper Text Markup Language* (HTML).

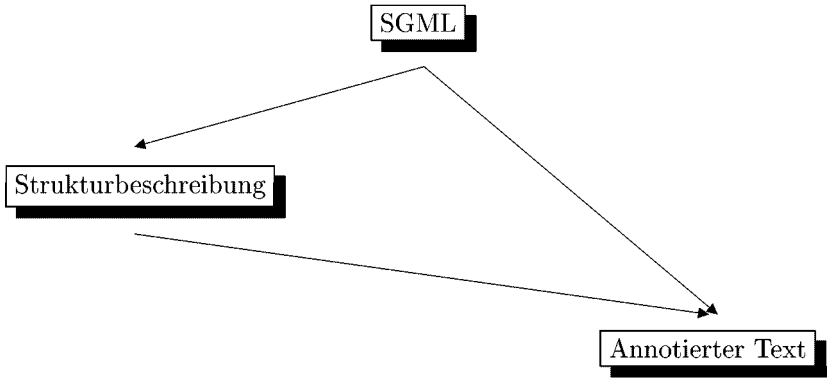

Abbildung 1: Definition eines SGML-Dokuments

Die Strukturbeschreibung heißt *Document Type Definition* (DTD). In einer DTD werden alle Elemente definiert, die in der Klasse der Dokumente vorkommen. Darüberhinaus wird ausgesagt, in welcher Beziehung zueinander diese Elemente stehen und wie oft sie autreten können.

Beispiel: DTD für ein Vorlesungsverzeichnis

```
<!ELEMENT vv              - -  (veranst+)  >
<!ELEMENT veranstaltung   - -  ((titel, sws, termin+, anbieter,
                                 kommentar?))>
<!ELEMENT titel           - -  (#PCDATA)>
<!ELEMENT anbieter        - -  (#PCDATA)>
<!ELEMENT sws             - -  (#PCDATA)>
<!ELEMENT kommentar       - -  (p+)>
<!ELEMENT p               - -  (#PCDATA)>
<!ELEMENT termin          - -  ((tag, beginn, ende))  >
<!ELEMENT tag             - -  (#PCDATA)>
<!ELEMENT beginn          - -  (#PCDATA)>
<!ELEMENT ende            - -  (#PCDATA)>
```

In dieser einfachen DTD für ein Vorlesungsverzeichnis wird z.B. zum Ausdruck gebracht, daß eine Veranstaltung (`veranst`) genau einen `titel` und mindestens einen `termin` besitzt. Der `titel` besteht einfach aus Zeichen `#PCDATA`, wohingegen im jeder der Termin weiter strukturiert werden muß. Neben der Definition der Elemente wird in der DTD festgelegt, welche Attribute diese Elemente besitzen können.

Beispiel: Attributdefinition

```
<!ATTLIST veranstaltung
        bnummer number              #required
        art     (vorlesung|seminar) seminar   >
```

Diese Attributspezifizierung sagt aus, daß das Element `veranstaltung` die Attribute `bnummer` und `art` enthält. Der Wert von `veranstaltung` muß ein Zahl sein. Der Wert des Attributs `art` ist entweder `vorlesung` oder `seminar`, wobei der Defaultwert `seminar` lautet.

Vorteile:
- Trennung von Inhalt und Form
- Strukturinformationen ohne Ambiguitäten
- Einfache Wiederverwendbarkeit von Textteilen
- Vorhandensein leistungsfähiger Software (z.B. Parse– , Formatier– und Transformierprogramme)

3 Die Richtlinien der Text Encoding Inititive (TEI)

Die *Text Encoding Initiative* (TEI) ist ein internationales Projekt, welches Richtlinien zur Kodierung geisteswissenschaftlicher Texte ausgearbeitet hat. Die TEI wird von drei Organisationen organisiert und finanziert:

- Association for Computational Linguistics (ACL)

- Association for Literary and Linguistic Computing (ALLC)
- the Association for Computing and the Humanities (ACH)

Die erste vollständige Version der *TEI Guidelines for electronic text encoding and interchange* wurden im Mai 1994 nach einer Entwicklungszeit von 6 Jahren veröffentlicht. Diese Richtlinien enthalten eine komplexe DTD. In dieser DTD werden Elemente für verschiedene wissenschaftliche Gebiete definiert. Nun ist es jedoch nicht so, daß alle Gebiete dieselben Strukturinformationen verwenden, da sich z.B. die Struktur von Lexika substantiell von der Struktur textkritischer Ausgaben literarischer Werke unterscheidet. Dies wird in der TEI-DTD dadurch berücksichtigt, daß verschiedene *Base Tag Sets* zusätzlich zu einem festen *Core Tag Set* Verwendung finden. Die Möglichkeit der Ergänzung oder Veränderung besteht durch sog. *User Defined Tag Sets.*
Eines der *Base Tag Sets* dient der Definition der Struktur von orthographischer Transkription (Transliterieration) gesprochener Sprache.

4 Kodierung gesprochener Sprache und Anwendbarkeit für die VERBMOBIL-Transliterationen

Zur Transliteration und Transkription gesprochener Sprache wird das *Base Tag Set* "TEI.spoken" verwendet, vgl. Sperberg-McQueen/ Burnard (1994), Kapitel 11, "Transcription of Speech". Diese Menge stellt Elemente zur Verfügung um sprachliche und nichtsprachliche Informationen formal zu notieren. Nahezu alle in den VERBMOBIL-Transliterationen verwendeten Annotierungen sind inhaltlich auch in 'TEI.spoken" vorgesehen, z.B. Elemente für artikulatorische und nicht-artikulatorische non-verbale Phänomene, für Pausen, Abbrüche, Wiederaufnahmen, für Überlappungen etc. In Anhang A befindet sich eine tabellenartige Gegenüberstellung der derzeitigen VERBMOBIL-Konventionen (vgl. Kohler et al. (1994)) und ihrer "TEI.spoken"-Entsprechungen. Da die TEI-Konvention auch gewisse Freiheiten zur Kodierung spezieller Phänomene erlaubt, müssen einige wenige zusätzliche Entscheidungen getroffen werden, etwa bei der Einführung von *user-defined tags* für technische Abbrüche und Neologismen.
Da die formalsprachliche Definition des VERBMOBIL-TRL-Formats weitestgehend regulär ist, kann ein mit Unix-Tools geschriebener Konverter diese in das SGML-Format überführen. Wo das VERBMOBIL-TRL-Format nicht regulär ist, etwa bei Überlappungen oder bei der Kodierung von Aussprachevarianten mittels eines Indexes ('haben wir <!2 hammer>') wird eine mächtigere Grammatik benötigt. Wenn eine solche vorliegt, wie im Zuge der Erweiterung des TRL-Formats für VERBMOBIL geplant, kann mit relativ geringem Aufwand ein vollständiger Konverter geschrieben werden.

Literatur

Burger, Susanne: *Konventionslexikon zur Transliteration von Spontansprache.* Institut für Phonetik und sprachliche Kommunikation München, 1996.

http://www.phonetik.uni-muenchen.de/VMTraLex.html

Kohler et al.: *Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL - 3.0.* IDPS Kiel, 1994. VERBMOBIL Technisches Dokument Nr. 11.

Sperberg-McQueen, C.M. and Lou Burnard: *Guidelines for Electronic Text Encoding and Interchange (TEI P3).* Volumes I and II. Chicago 1994.

A Tabellarische Gegenüberstellung von TRL-Elementen und TEI-SGML-Elementen

TRL-vm1	**TEI-SGML**
Umlaute	
“a	ä
“o	ö
“u	ü
“A	Ä
“O	Ö
“U	Ü
“s	ß
Satzzeichen	
[STRING]	[STRING]
.	.
?	?
,	,
Nonverbale Produktionen	
<”ah>	<vocal descr=’äh’>
<”ahm>	<vocal descr=’ähm’>
<hm>	<vocal descr=’hm’>
<h”as>	<vocal descr=’häs’>
<Husten>	<vocal descr=’Husten’>
<Lachen>	<vocal descr=’Lachen’>
<R”auspern>	<vocal descr=’Räuspern’>
<Schlucken>	<vocal descr=’Schlucken>
<Schmatzen>	<vocal descr=’Schmatzen>
<Geräusch>	<vocal descr=’Geräusch’>
<Z”ogern>	<vocal descr=’Zögern’>
<Atmen>	<vocal descr=’Atmen’>
<Z>	<vocal descr=’Z’>
<A>	<vocal descr=’A’>
<P>	<pause>
[STRING2]<Z>[STRING2]	[STRING1]<vocal descr=’Z’>[STRING2]
[STRING2]<A>[STRING2]	[STRING1]<vocal descr=’A’>[STRING2]
Nicht-artikulatorisch	
<#Klicken>	<event descr=’Klicken’>
<#Klingeln>	<event descr=’Klingeln’>
<#Klopfen>	<event descr=’Klopfen’>
<#Mikrobe>	<event descr=’Mikrobe’>
<#Mikrowind>	<event descr=’Mikrowind’>
<#Quietschen>	<event descr=’Quietschen’>
<#Rascheln>	<event descr=’Rascheln’>
<#>	<event>
Kommentare	
;[COMMENT]	<!– – [COMMENT] – –>
<;[COMMENT]>	<!– – [COMMENT] – –>
Aussprachevarianten	
[STRING1] <!1 [STRING2]>	<reg orig=[STRING2]> [STRING1] </reg>
[STRING1] [STRING2] <!2 [STRING3]>	<reg orig=[STRING3]> [STRING1] [STRING2] </reg>

TRL-vm1	**TEI-SGML**
Gleichzeitigkeit von Schallereignissen	
<:<#[STRING1]> [STRING2]:>	<event descr=[STRING1] start=P1 end=P2> <anchor id=P1> [STRING2] <anchor id=P2>
<:<[STRING1]> [STRING2]:>	<vocal descr=[STRING1] start=P1 end=P2> <anchor id=P1> [STRING2] <anchor id=P2>
Technische Abbrüche	
[STRING]<;T>	[STRING]<tech>
<;T>[STRING]	<tech>[STRING]
<;T>	<tech>
Abbrüche	
[STRING]/–	<del type=break_off> [STRING] </del>
Wiederaufnahme	
+/[STRING]/+	<del type=resumption> [STRING] </del>
+/[STRING1]/+ [STRING2]/+	<del type=resumption> [STRING1] </del> <del type=resumption> [STRING2] </del>
+/[STRING1]/+ [STRING2]/+ [STRING3]/+	<del type=resumption> [STRING1] </del> <del type=resumption> [STRING2] </del> <del type=resumption> [STRING3] </del>
Wortfragmente	
[STRING]=	<del type=truncation> [STRING] </del>
Wortunterbrechungen	
[STRING1]_ [STRING2] _[STRING3]	[STRING1]<del type=word_interruption> [STRING2] </del> [STRING3]
Nichtwörter	
*[STRING]	<neol> [STRING] </neol>
Schwerverständliches, Unverständliches	
%[STRING]	<unclear> [STRING] </unclear>
<%>	<unclear>
SAMPA–Aussprachekommentar	
[STRING] <! [SAMPA]>	<reg orig=[SAMPA]> [STRING] </reg>
Überlappung	
(@ [STRING])	<anchor id=P[N]> [STRING] <anchor id=P[N+1]>)
([STRING] @)	<overlap start=P[N] end=P[N+1]> [STRING] </overlap>

B Konverter für die regulären Anteile an der TRL-Definition

```
#!/bin/sh
# trl2sgml
# D. Gibbon
# 16 Jan 1997
#-----------------------------------------------------------------
# Convert VM-1 orthographic transcription to TEI/SGML
# by UNIX rexexp (character regular expression) filter cascade
#-----------------------------------------------------------------

# Preprocessor for text format normalisation: one turn per line
#
# - Remove comment lines (this will be changed to convert to SGML comments)
# - Convert line-feed to blank
# - Reduce blank sequences to single blank
# - Restore line-feed before TURN-ID
# - Remove leading and trailing blanks
# - Retain only non-empty lines
# - Store preprocessed transcription
cat $1 |
sed "s/^;.*$//g" |
tr "\012" " " |
sed "s/  */ /g
     s/......: /§&/g" |
tr "§" "\012" |
sed "s/^ //g
     s/ $//g" |
grep . > $1.tmp

#-----------------------------------------------------------------
# Converter: UNIX rexexp emulation of finite state transducer
#
# Note on formal language type:
# - A filter does not necessarily check for well-formedness
# - Nested bracketings can also be translated by an fst (this may not be
#   immediately obvious)
# - Numbered comments on pronunciation variants require a stack of unknown
    depth and therefore are not regular sequences (they resemble context-free
#   constructions where the stack pointer is output instead of a string)
# - Sequences of overlaps require two stacks in an indexed (context-sensitive)
#   grammar

gawk '
```

```
# File ID extraction (requires modification)

BEGIN{
FILE=ARGV[1]
gsub(/\.tmp/,"",FILE)
CDROM=FILE
gsub(/-.*/,"",CDROM)
gsub(/.*-/,"",FILE)
print "<!-- SGML Orthographic transcription file " FILE " -->"
print "<!-- Generated by trl2sgml V.0.00 16 January 1997 D. Gibbon -->"
print "\n<body CDROM=" CDROM " file=" FILE ">\n"
}
{

#-----------------------------------------------------------------

# Turn ID extraction (requires further modification)

ID=$1
gsub(/:/,"",ID)
ID="<u file=" FILE " id=" ID ">\n"
gsub(/......:/,ID)
gsub(/$/,"\n</u>\n")

# Hesitations

gsub(/<\"ah>/," <vocal descr=\047&auml;h\047> ")
gsub(/<\"ahm>/," <vocal descr=\047&auml;hm\047> ")
gsub(/<hm>/," <vocal descr=\047hm\047> ")
gsub(/<h\"as>/," <vocal descr=\047h&auml;s\047> ")

# Noises

gsub(/<Husten>/,"\n   <vocal descr=\047Z\047> ")
gsub(/<Lachen>/,"\n   <vocal descr=\047Z\047> ")
gsub(/<R\"auspern>/,"\n   <vocal descr=\047Z\047> ")
gsub(/<Schlucken>/,"\n   <vocal descr=\047Z\047> ")
gsub(/<Schmatzen>/,"\n   <vocal descr=\047Z\047> ")
gsub(/<Ger\"ausch>/,"\n   <vocal descr=\047Z\047> ")

# Vocalisations

gsub(/<Z>/,"<vocal descr=\047Z\047>")
gsub(/<A>/,"\n   <vocal descr=\047A\047> ")
gsub(/<P>/,"\n   <pause> ")

# Nonhuman noises
```

```
gsub(/<#Klicken>/,"\n <event descr=\047Klicken\047> ")
gsub(/<#Klingeln>/,"\n <event descr=\047Klingeln\047> ")
gsub(/<#Klopfen>/,"\n <event descr=\047Klopfen\047> ")
gsub(/<#Mikrobe>/,"\n <event descr=\047Mokrobe\047> ")
gsub(/<#Mikrowind>/,"\n <event descr=\047Mikrowind\047> ")
gsub(/<#Quietschen>/,"\n <event descr=\047Quietschen\047> ")
gsub(/<#Rascheln>/,"\n <event descr=\047Rascheln\047> ")
gsub(/<#>/,"\n<event>")

# Pronunciation variants
# THIS IS ONLY A PSEUDO-TRANSLATION: A STACK MACHINE IS NEEDED

gsub(/<![0-9]* [^>]*>/,"\n <reg number=&")
gsub(/<!/,"")
gsub(/number=[0-9]* /,"&orig=")
gsub(/orig=[^>]*/,"&\"")
gsub(/orig=/,"&\"")

# Heterophonous homographs
# NOT TESTED

gsub(/<! [^>]*>/,"<sampa>&</sampa>")
gsub(/<!/,"")
gsub(/><\/sampa>/,"</sampa>")

# Special comment types

gsub(/<;T>/,"<tech>")
gsub(/<;[^>]*>/,"\n <!-- & -->\n")
gsub(/<;/,"")
gsub(/> --/," --")

# NOTE: Line comments of this type were removed in the preprocessor ...
# gsub(/^;.*$/,"\n<!-- & -->\n")
# NOT TESTED
# Interruption
gsub(/ [A-Za-z"]*\/-/ ,"\n <del type=break_off>&</del>")

# Resumption
# THE DOUBLE RESUMPTION TYPE IS NOT IMPLEMENTED
# AN ADDITIONAL CONVENTION IS NEEDED

gsub(/\+\//,"\n <del type=resumption> ")
gsub(/\/\+/," </del>")

# Truncation
# NOT TESTED
```

```
gsub(/ [A-Za-z"]*= /,"\n <del type=truncation>&</del>")

# Non-words
# EMBEDDED NEOLOGISMS NOT IMPLEMENTED; CONVENTION NEEDED

gsub(/ \*[A-Za-z"]* /,"<neol>&</neol>")

# Hardly comprehensible

gsub(/ \%[A-Za-z"]* /,"\n <unclear>&</unclear")

# Incomprehensible
gsub(/<\%>/,"<unclear>")


# Broken words

gsub(/ [A-Za-z"]*_ /,"\n <del type=word_interruption>&")
gsub(/ _[A-Za-z"]* /,"&</del> ")
gsub(/_ /,"")
gsub(/ _/,"")

# Simultaneity
# NOT IMPLEMENTED -- INDEXED GRAMMAR NEEDED!

# Punctuation

# gsub(/[.,?]/,"&\n")

# Umlaut

gsub(/\"a/,"\\\&auml;")
gsub(/\"o/,"\\\&ouml;")
gsub(/\"u/,"\\\&uuml;")
gsub(/\"A/,"\\\&Auml;")
gsub(/\"O/,"\\\&Ouml;")
gsub(/\"U/,"\\\&Uuml;")
gsub(/\"s/,"\\\&szlig;")

print
}

END {
print "</body>"
}
 ' $1.tmp |

sed "s/  */ /g" |
```

```
grep -v "^[ ]*$"
```

C Beispieldialog n010k von CDROM 1

```
;N010K
; Vorsicht!
; Originaldateinamen und Originalturnnamen (nach jedem Turn als
;Kommentar-
; zeile) beziehen sich auf die urspr"unglichen Namen aus Karlsruhe, die
;nicht
; den Namenskonventionen entsprachen.
; Die Turnnummerierung wurde in der alten Karlsruher Version gelassen,
; weil auch die entsprechenden Sprachfiles so nummeriert sind.
; Das hei"st, jede Terminabsprache f"angt mit 001 an.
;
; Originaldatei: mhs2_mjg1_tsponti1.trans
; Terminabsprache a:

HS2001: <A> <Schmatzen> <A> oh , h<Z>allo , Herr Gramlich . gut <!1
        gud> , da"s ich Sie hier treffe . ich<Z> <!1 'ch> h"att' gern
        einen <!1 'en> Termin mit Ihnen ausgemacht f"ur<Z> eine <!1
        'n'> Projektleiter_ <A> _sitzung . <A> <#Klicken>
;mhs2_1_01

JG1002: <Ger"ausch> <A> ja , ich<Z>/- <"ah> <P> von mir aus k"onnen wir
        das <!3 k"omma des> machen . am besten <!1 beschden> w"ar' der
        <A> Monat Juli . <P> <A> un<Z>d <!1 un'> <P> +/wann/+ welche
        Wochentage w"urden Ihnen am<Z> ehesten liegen ? <#Klicken>
;mjg1_1_02

HS2003: <;T>a geht 's nicht <!1 ned> , da bin ich in Urlaub , Dienstag
        <!1 Dienschtag> auch<Z> , Mittwoch w"ar' in Ordnung <!1 Odnung>
        . ab zehn Uhr . da komm' ich vom Arzt <A> . <#Klicken> <A>
;mhs2_1_03

JG1004: <A> Mittwoch w"ar' bei mir auch<Z> in Ordnung . allerdings erst
        ab<Z> <A> <P> achtzehn Uhr , nach der Vorlesung . <#Klicken>
;mjg1_1_04

HS2005: <;T>e , da bin ich auf einer <!1 ner> Sitzung <A> . wie
        w"ar's<Z> Freitag , ab<Z> sechzehn Uhr <A> ? <#Klicken>
;mhs2_1_05

JG1006: <Ger"ausch> <"ah> freitags nach sechzehn Uhr . <A> <P>
        <Ger"ausch> besser w"ar's <"ah> sechzehn Uhr drei"sig .
        <#Klicken>
;mjg1_1_06

HS2007: <Ger"ausch> Freitag , sechzehn Uhr drei"sig <!1 dreisich> <A> .
        aber zwanzig Uhr mu"s ich wieder weg <#Mikrowind> . <A>
```

```
;mhs2_1_07

JG1008: gut , dann <P> machen wir das <!3 mach ma des> Freitag . <P>
        zwanzig Uhr war 's <Lachen> <;leise> ? <#Klicken>
;mjg1_1_08

HS2009: ich geb' dann noch dem <!1 'm> Herrn<Z> <A> Sch"afer Bescheid ,
        da"s der dann <;verschliffene Artikulation> auch da ist . <A>
        tsch"u"s <A> . <#Klicken>
;mhs2_1_09

JG1010: gut , ade . <P> <#Klicken>
;mjg1_1_10
;
; Originaldatei: mhs2_mjg1_tsponti2.trans
; Terminabsprache b:

JG1001: <#> <A> guten <!1 gu'n> Tag , Herr Schulz . ich h"att' gern mit
        Ihnen einen <!1 'en> Termin ausgemacht , und zwar<Z> irgendwann
        im Juli . <A> da Sie mich jetzt schon dreimal versetzt haben ,
        <"ah> w"urd' ich das <!1 des> jetzt gern z"ugig hinter <!1
        hinda> mich bringen , und zwar am Dienstag , um<Z> <A>
        f"unfzehn Uhr f"unfundzwanzig . <#Klicken>
;mjg1_2_01

HS2002: <A> also es tut mir <!1 ma> leid , <#Mikrowind> da"s ich Sie
        dreimal versetzt hab' , aber es ging wirklich nicht <!1 net>
        anders <!1 anerschd> , ich mu"ste <!1 mu"st> dringend <#> auf
        'ne Tagung in die $U $S $A . <A> aber Dienstag <!1 Dienschtag>
        ist ganz schlecht . wie w"ar's mit einem anderen <!3 mim an'n>
        <;verschliffene Artikulation> Wochentag ? <#Klicken>
        <#Mikrowind>
;mhs2_2_02

JG1003: <A> <P> <"ahm> ich h"att' Ihnen als einzige Alternative noch
        anzubieten den Mittwoch , <A> irgendwann zwischen sieben und
        zw"olf Uhr . <#Klicken>
;mjg1_2_03

HS2004: <;T>is neun bin ich weg . <Ger"ausch> <A> es <!1 's> geht nicht
        <!1 net> zwischen sieben und neun <!3 siebne'neun>
        <;verschliffene Artikulation> , keine Chance <A> . h"ochstens
        <!1 hechschstns> von elf bis zw"olf , das <!1 des> wird aber
        dann ziemlich knapp . oder wir machen 's am Freitag . w"urd's
        da wirklich nicht <!1 net> gehen ? <A> <#Klicken>
;mhs2_2_04

JG1005: <A> ja , dann m"ussen wir <!3 m"u"s' ma's> halt<Z> Mittwoch ,
```

```
        zwischen elf und zw"olf machen . <P> <A> <"ah> das <!1 des>
        reicht ja auch . ungef"ahr eine <!1 'ne> halbe Stunde . also
        w"urd' ich mal sagen , <A> elf Uhr . <#Klicken>
;mjg1_2_05

HS2006: <A> <Schmatzen> okay , in Ordnung <!1 Odnung> , elf Uhr in
        Ihrem B"uro . sagen <!1 sang> Sie den andern Bescheid <A> ?
        <#Klicken> <A>
;mhs2_2_06

JG1007: <A> ja , das<Z> <!1 des> lass' ich meine Sekret"arin
        "ubernehmen . <#Klicken>
;mjg1_2_07

HS2008: <A> <#Klicken> okay . tsch"u"s , bis Mittwoch . <#Mikrowind>
        <#Klicken>
;mhs2_2_08

JG1009: auf Wiedersehen , Herr Schulz . <#Klicken>
;mjg1_2_09
```

D TEI-konforme Aufbereitung des Beispieldialogs

```
<!DOCTYPE TEI.2 system 'tei2.dtd'
                        [<!ENTITY % TEI.spoken 'INCLUDE' >
                         <!ENTITY % TEI.corpus 'INCLUDE' >
                         <!ENTITY % TEI.extensions.ent SYSTEM "ISOlat1.ent">
                        ]>

<tei.2>
  <teiheader>
    <filedesc>
      <titlestmt>
        <title>n010k</title>
      </titlestmt>
      <publicationstmt>
        <authority>vm</authority>
      </publicationstmt>
      <sourcedesc>
        <recordingstmt>
          <recording type="audio">
            <date>tt.mm.jj</date>
          </recording>
        </recordingstmt>
      </sourcedesc>
    </filedesc>
    <profiledesc>
      <particdesc>
        <person id="hs">
          <p>Person 1</p>
        </person >
        <person id="jg">
        </person>
      </particdesc>
    </profiledesc>
  </teiheader>
  <text>
    <body>
<u who=hs n=2001>
 <vocal desc='A'>
 <vocal desc='Z'>
 <vocal desc='A'> oh , h<vocal desc='Z'>allo , Herr Gramlich .
 <reg n=1 orig="gud"> gut  </reg> , da&szlig; ich Sie hier treffe .
 <reg n=1 orig="'ch"> ich</reg><vocal desc='Z'>
 h&auml;tt' gern  <reg n=1 orig="'en"> einen </reg>
Termin mit Ihnen ausgemacht f&uuml;r<vocal desc='Z'>
 <reg n=1 orig="'n'">eine  </reg>
 <del type=word-interruption> Projektleiter</del>
 <vocal desc='A'> sitzung .
```

```
 <vocal desc='A'>
 <event desc='Klicken'>
</u>
<u  who=jg n=1002>
 <vocal desc='Z'>
 <vocal desc='A'> ja , ich<vocal desc='Z'>/- <vocal desc='&auml;h'>
 <pause> von mir aus k&ouml;nnen wir
 <reg n=3 orig="k&ouml;mma des"> das  </reg>
 machen . am
 <reg n=1 orig="beschden"> besten  </reg>
w&auml;r' der
 <vocal desc='A'> Monat Juli .
 <pause>
 <vocal desc='A'> un<vocal desc='Z'>d
 <reg n=1 orig="&uuml;n"> </reg>
 <pause>
 <del type=resumption> wann </del> welche Wochentage w&uuml;rden Ihnen am<vocal desc='Z'> ehe
 <event desc='Klicken'>
</u>
<u who=hs n=2003>
 <tech>a geht 's
 <reg n=1 orig="ned">nicht </reg>
, da bin ich in Urlaub ,
 <reg n=1 orig="Dienschtag"> Dienstag </reg>
auch<vocal desc='Z'> , Mittwoch w&auml;r' in
 <reg n=1 orig="&Ouml;dnung"> Ordnung </reg>
. ab zehn Uhr . da komm' ich vom Arzt
 <vocal desc='A'> .
 <event desc='Klicken'>
 <vocal desc='A'>
</u>
<u  who=jg n=1004>
 <vocal desc='A'> Mittwoch w&auml;r' bei mir auch<vocal desc='Z'> in Ordnung . allerdings ers
 <vocal desc='A'>
 <pause> achtzehn Uhr , nach der Vorlesung .
 <event desc='Klicken'>
</u>
<u who=hs n=2005>
 <tech>e , da bin ich auf
 <reg n=1 orig="ner"> einer </reg>
Sitzung
 <vocal desc='A'> . wie w&auml;r's<vocal desc='Z'> Freitag , ab<vocal desc='Z'> sechzehn Uhr
 <vocal desc='A'> ?
 <event desc='Klicken'>
</u>
<u  who=jg n=1006>
 <vocal desc='Z'> <vocal desc='&auml;h'> freitags nach sechzehn Uhr .
 <vocal desc='A'>
```

```
 <pause>
 <vocal desc='Z'> besser w&auml;r's <vocal desc='&auml;h'> sechzehn Uhr drei&szlig;ig .
 <event desc='Klicken'>
</u>
<u who=hs n=2007>
 <vocal desc='Z'> Freitag , sechzehn Uhr
 <reg n=1 orig="dreisich"> drei&szlig;ig </reg>
 <vocal desc='A'> . aber zwanzig Uhr mu&szlig; ich wieder weg
 <event desc='Mikrowind'> .
 <vocal desc='A'>
</u>
<u  who=jg n=1008>
 gut , dann
 <pause>
 <reg n=3 orig="mach ma des"> machen wir das </reg>
 Freitag .
 <pause> zwanzig Uhr war 's
 <vocal desc='Z'>
 <!-- leise -->
 <event desc='Klicken'>
</u>
<u who=hs n=2009>
 ich geb' dann noch
 <reg n=1 orig="'m"> dem </reg>
Herrn<vocal desc='Z'>
 <vocal desc='A'> Sch&auml;fer Bescheid , da&szlig; der dann
 <!-- verschliffene Artikulation -->
 auch da ist .
 <vocal desc='A'> tsch&uuml;&szlig;
 <vocal desc='A'> .
 <event desc='Klicken'>
</u>
<u  who=jg n=1010>
 gut , ade .
 <pause>
 <event desc='Klicken'>
</u>
<u  who=jg n=1001>
<event>
 <vocal desc='A'>
 <reg n=1 orig="gu'n"> guten  </reg>
Tag , Herr Schulz . ich h&auml;tt' gern mit Ihnen
 <reg n=1 orig="'en"> einen  </reg>
Termin ausgemacht , und zwar<vocal desc='Z'> irgendwann im Juli .
 <vocal desc='A'> da Sie mich jetzt schon dreimal versetzt haben , <vocal desc='&auml;h'>
w&uuml;rd' ich
 <reg n=1 orig="des"> das </reg>
jetzt gern z&uuml;gig
```

```
 <reg n=1 orig="hinda"> hinter </reg>
mich bringen , und zwar am Dienstag , um<vocal desc='Z'>
 <vocal desc='A'> f&uuml;nfzehn Uhr f&uuml;nfundzwanzig .
 <event desc='Klicken'>
</u>
<u  who=hs n=2002>
 <vocal desc='A'> also es tut
 <reg n=1 orig="ma"> mir </reg>
leid ,
 <event desc='Mikrowind'> da&szlig; ich Sie dreimal versetzt hab' , aber es ging wirklich
 <reg n=1 orig="net"> nicht </reg>
 <reg n=1 orig="&auml;nerschd"> anders</reg>
, ich
 <reg n=1 orig="mu&szlig;t"> mu&szlig;te </reg>
dringend
<event> auf 'ne Tagung in die $U $S $A .
 <vocal desc='A'> aber
 <reg n=1 orig="Dienschtag"> Dienstag </reg>
ist ganz schlecht . wie w&auml;r's
 <reg n=3 orig="mim an'n"> mit einem anderen </reg>
 <!-- verschliffene Artikulation -->
 Wochentag ?
 <event desc='Klicken'>
 <event desc='Mikrowind'>
</u>
<u  who=jg n=1003>
 <vocal desc='A'>
 <pause> <vocal desc='&auml;hm'> ich h&auml;tt' Ihnen als einzige Alternative noch
anzubieten den Mittwoch ,
 <vocal desc='A'> irgendwann zwischen sieben und zw&ouml;lf Uhr .
 <event desc='Klicken'>
</u>
<u  who=hs n=2004>
 <tech>is neun bin ich weg .
 <vocal desc='Z'>
 <vocal desc='A'>
 <reg n=1 orig="'s"> es </reg>
geht
 <reg n=1 orig="net"> nicht </reg>
zwischen
 <reg n=3 orig="&szlig;iebne'neun">sieben und neun </reg>
 <!-- verschliffene Artikulation -->
 , keine Chance
 <vocal desc='A'> .
 <reg n=1 orig="hechschstns"> h&ouml;chstens </reg>
von elf bis zw&ouml;lf ,
 <reg n=1 orig="des"> das  </reg>
wird aber dann ziemlich knapp . oder wir machen 's am Freitag .
```

```
w&uuml;rd's da wirklich
 <reg n=1 orig="net"> nicht  </reg>
gehen ?
 <vocal desc='A'>
 <event desc='Klicken'>
</u>
<u  who=jg n=1005>
 <vocal desc='A'> ja , dann
 <reg n=3 orig="m&uuml;&szlig;' ma's"> m&uuml;ssen wir </reg>
halt<vocal desc='Z'> Mittwoch , zwischen elf und zw&ouml;lf machen .
 <pause>
 <vocal desc='A'> <vocal desc='&uml;h'>
 <reg n=1 orig="des"> das </reg>
reicht ja auch . ungef&auml;hr
 <reg n=1 orig="'ne"> eine </reg>
halbe Stunde . also w&uuml;rd' ich mal sagen ,
 <vocal desc='A'> elf Uhr .
 <event desc='Klicken'>
</u>
<u  who=hs n=2006>
 <vocal desc='A'>
 <vocal desc='Z'> okay , in
 <reg n=1 orig="&Ouml;dnung">Ordnung </reg>
 , elf Uhr in Ihrem B&uuml;ro .
 <reg n=1 orig="&szlig;ang"> sagen </reg>
Sie den andern Bescheid
 <vocal desc='A'> ?
 <event desc='Klicken'>
 <vocal desc='A'>
</u>
<u  who=jg n=1007>
 <vocal desc='A'> ja , das<vocal desc='Z'>
 <reg n=1 orig="des"> </reg>
lass' ich meine Sekret&auml;rin &uuml;bernehmen .
 <event desc='Klicken'>
</u>
<u  who=hs n=2008>
 <vocal desc='A'>
 <event desc='Klicken'> okay . tsch&uuml;&szlig; , bis Mittwoch .
 <event desc='Mikrowind'>
 <event desc='Klicken'>
</u>
<u  who=jg n=1009>
 auf Wiedersehen , Herr Schulz .
 <event desc='Klicken'>
</u>
    </body>
  </text>
```

```
</tei.2>
```

E Liste von Projekten

Die folgenden Projekte nutzen zur Annotation die DTD der Text Encoding Initiative. Diese Liste wurde vom Server der TEI übernommen (http://www-tei.uic.edu/orgs/tei/app/alpha.html).

1. American Memory Project (Library of Congress)
   at http://lcweb2.loc.gov/ammem/ammemhome.html
2. American Verse Project
   at http://www.hti.umich.edu/english/amverse
3. Bodleian Library: Toyota City Imaging Project
   at http://www.bodley.ox.ac.uk/toyota
4. British National Corpus Project (BNC)
   at http://info.ox.ac.uk/bnc
5. Brown University Scholarly Technology Group
   at http://www.stg.brown.edu
6. Brown University Women Writer's Project
   at http://www.wwp.brown.edu
7. Cambridge University Press
   at http://www.cup.cam.ac.uk/Reviews&blurbs/CDROMtop.html
8. Canterbury Tales Project
   at http://www.shef.ac.uk/uni/projects/ctp/index.html
9. Center for Electronic Text in the Law
   at http://www.law.uc.edu/CENTER
10. Center for Electronic Texts in the Humanities
    at http://www.ceth.rutgers.edu
11. Chadwyck-Healey English Poetry Full-Text Database
    at http://etext.virginia.edu/epd.html
12. Charrette Project
    at http://www.princeton.edu/~lancelot
13. Chiba Corpus of Map Task Dialogues in Japanese
    at http://cogsci.L.chiba-u.ac.jp/MapTask
14. Chronicon: An On-Line Journal of History
    at http://www.ucc.ie/chronicon
15. Consortium for Interchange of Museum Information (CIMI)
    at http://www.cimi.org/cimi
16. CURIA Project: The Thesaurus Linguarum Hiberniae
    at http://curia.ucc.ie/curia
17. Danish Spoken Language Dialogue Systems Project
    at http://www.cog.ruc.dk/projects/Dialogue/user-95
18. Digital Literature (Strindberg & Other Swedish Texts)
    at http://www.nada.kth.se/~broady/diglit.html
19. Documenting the American South, or the Southern Experience in 19th-Century America
    at http://www.unc.edu/~nsmith/
20. EAGLES (Expert Advisory Group on Language Engineering Standards)
    at http://www.ilc.pi.cnr.it/EAGLES/home.html
21. Edinburgh Map Task Corpus

at http://www.cogsci.ed.ac.uk/elsnet/Resources/Map-Task/ mt_corpus.html
22. Electronic New Testament Manuscript Project
at http://www.entmp.org
23. Electronic Thesaurus Linguae Latinae (TLL)
at http://www.cs.usask.ca/faculty/devito/e-TLL
24. The English-Norwegian Parallel Corpus
at http://www.hd.uib.no/enpc.html
25. European Corpus Initiative
at http://www.cogsci.ed.ac.uk/elsnet/eci.html
26. Grammars for Parsing of Dictionaries
at http://www.cs.umd.edu/users/ericp
    - Example of MITRE Dictionary Encoding
27. HyperGrammar
at http://www.uottawa.ca/academic/arts/writcent
28. Indiana University Library Electronic Text Resources (LETRS)
at http://www.indiana.edu/~letrs
29. Japanese/English Bilingual Corpus
30. Japanese Text Initiative
at http://etext.virginia.edu/japanese
31. Kolb-Proust Archive for Research
at http://www.grainger.uiuc.edu/kolbp
32. Leiden Armenian Database
33. LE (Language Engineering)-PAROLE
at http://www.ilc.pi.cnr.it/parole/parole.html (Under Construction)
34. Lingua Project
at http://www.loria.fr/exterieur/equipe/dialogue/lingua
35. Mexican Bibliography
36. Model Editions Partnership: Historical Editions in the Digital Age at
http://mep.cla.sc.edu
37. Multilingual Corpora for Cooperation (MLCC)
at http://issco-www.unige.ch/projects/MLCC.html
38. Multilingual (Basque/Spanish/French/English) Dictionaries
at http://fims-www.massey.ac.nz/~JPatrick
39. Multilingual Text Tools and Corpora (MULTEXT)
at http://www.lpl.univ.aix.fr/projects/multext
40. Multilingual Text Tools and Corpora for Central & Eastern European
Languages (MULTEXT-EAST)
at http://nl.ijs.si/ME/
41. Network of Literary Archives
at http://gonzo.hd.uib.no/Nola/Nola.html
42. Old Finnish Texts Corpus
at http://syy.oulu.fi/kurssit/epubl/home.html
43. Orlando Project: An Integrated History of Womens Writing In The British
Isles
at http://www.ualberta.ca/ORLANDO
44. Oxford Text Archive
at http://info.ox.ac.uk/ota/
45. Perseus Project

at http://www.perseus.tufts.edu
46. Piers Plowman Electronic Archive
at http://jefferson.village.virginia.edu/piers/archive.goals.html
47. Pirque Rabbi Eliezer Electronic Text Editing Project
at http://www.usc.edu/dept/huc-la/pre-project
48. Serbian Proverb Project
at http://www.matf.bg.ac.yu/proverb/home.html
(may not be operational)
49. Silfide (Serveur Interactif pour la Langue Française, son Identité, sa Diffusion, son Étude)
at http://www.loria.fr/Projet/Silfide(Frames)
at http://www.loria.fr/Projet/Silfide/silfide.html(No Frames)
50. Spanish Dramatic Texts
51. Spanish Language of New Mexico: Legal Documents and Official Correspondence from Pre-Statehood New Mexico, 1684-1894
52. Thesaurus Musicarum Italicarum (TMI)
at http://www.let.ruu.nl/C+L/wiering/tmi_home.htm
53. University of Michigan Humanities Text Initiative (HTI)
at http://www.hti.umich.edu
54. University of Virginia Electronic Text Center
at http://www.lib.virginia.edu/../etext/ETC.html
    - Electronic Archive of Early American Fiction (at University of Virginia)(rare books)
      at http://etext.lib.virginia.edu/eaf
55. Victorian Women Writers' Project
at http://www.indiana.edu/~letrs/vwwp/index.html
    - Willett, Perry. "The Victorian Women Writers Project: The Library as a Creator and Publisher of Electronic Texts."The Public-Access Computer Systems Review, Vol. 7, NO. 6 (1996). (Refereed Article)
        + HTML version: http://info.lib.uh.edu/pr/v7/n6/will7n6.html
        + ASCII version: http://info.lib.uh.edu/pr/v7/n6/willett.7n6
        + Listserv: Send e-mail message GET WILLETT PRV7N6 F=MAIL to listserv@uhupvm1.uh.edu.
56. Voltaire Foundation
at http://www.voltaire.ox.ac.uk
57. Waterloo Lutheran Seminary Historic Text Archive
at http://info.wlu.ca/~wwwsem/wlstext.html