

Tutorial Workshop "Speech Corpus Production and Validation"

LREC 2004, Lisbon

Monday, 24th May 2004
Time: 14:00 to 19:00

Schedule

14:00 - 14:15 Welcome, Introduction of Speakers and Tutorial Overview

14:15 - 14:45 Schiel: Speech Corpus Specification

14:45 - 15:45 Draxler: Speech Data Collection in Practice

15:45 - 16:15 coffee break

16:15 - 16:45 Schiel: Speech Corpus Annotation

16:45 - 17:45 Draxler: Speech Annotation in Practice

17:45 - 18:00 short break

18:00 - 18:45 van den Heuvel: Validation and Distribution of Speech Corpora

18:45 - 19:00 Discussion and Conclusion

Workshop Organisers

Christoph Draxler, BAS Bavarian Archive for Speech Signals
Ludwig-Maximilian-University Munich, Germany
draxler@bas.uni-muenchen.de

Henk van den Heuvel, SPEX Speech Processing Expertise Centre
Centre for Language and Speech Technology,
Radboud University Nijmegen
henk@spex.nl

Florian Schiel, Bavarian Archive for Speech Signals
Ludwig-Maximilian-University Munich, Germany
schiel@bas.uni-muenchen.de

Table of Contents

Speech Corpus Specification	4
Speech Data Collection in Practice	10
Recording Script DTD	16
SpeechRecorder Sample XML Recording Script	17
Speech Corpus Annotation	18
Speech Annotation in Practice	23
Validation and Distribution of Speech Corpora	28
Template Validation Report SALA II	36
Methodology for a Quick Quality Check of SLR and Phonetic Lexicons	65
A Bug Report Service for ELRA	96
SLR Validation: Current Trends and Developments	107

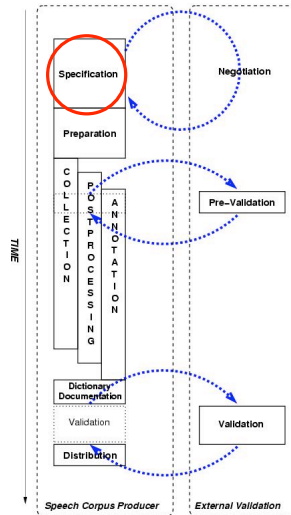
Author Index

Baum, Micha
Draxler, Christoph
Heuvel, Henk van den
Iskra, Dorota
Sanders, Eric
Schiel, Florian
Vriendt, Folkert de

Speech Corpus Specification

Florian Schiel

- First step of corpus production
- Defines all properties of desired speech corpus
- Basis for cost estimate
- Basis for production schedule



Before getting started: Some general rules

- All specifications should be fixed in a version numbered document called 'Specification X.Y'
- If you produce data as a contractor for a client let all versions be signed by your client.
- Allocate considerable time for the specification phase (10-25% of total project time); observe *Hofstetter's Law*
- Use the check list on page 65 of book
- Specify tolerance measures whenever applicable
- If not sure about feasibility, ask experts (LDC,ELDA,SPEX,BAS)

Overview: Important Parts of a Speech Corpus Specification

- **Speakers: number, profiles, distribution**
- **Spoken content**
- **Speaking style**
- **Recording procedure**
- **Annotation**
- **Meta data, documentation**
- **Technical specifications**
- **Final release: corpus structure, media, release plan**

In the following slides everything we deem to be absolutely necessary for a speech corpus specification will be underlined

Examples will appear in italic

Speaker number, profiles, distribution

- Distribution of sex: $m : f = 50 : 50$ 5% tolerance
- Distribution of age: 16-25 : 20%, 26-50 : 60%, >50 : 20%
5% tolerance
- L0 (mother tongue): German, max 3% non-native speakers
- Dialects: Distribution over classified dialects
North German 50%, South G. 50%
5% tolerance
- Education / Proficiency / Profession
Use a closed vocabulary!
School, College, University
Novice, Computer user, Expert
- Others: pathologies, foreign accent, speech rate, jewelry etc.

Spoken Content

Specify the spoken content by one of:

- **Vocabulary**
14 commands spoken 10 times by 5000 speakers
 - **Domain**
weather, fairy tales, last night's TV program
 - **Task**
travel planning, program the VCR, find a route on a map
 - **Phonological distribution**
distribution of phonemes, syllables, morphs
- or a combination (recommended):
14 commands spoken 10 times + 1 minute monologue about the weather

Eliciting Speaking Styles

Recommendation: use more than one speaking style!

Select from the following basic speaking styles:

- **Read speech**
prompt sheets, prompting from screen
Hints:
 - avoid words with ambiguous spellings (acronyms, numbers!)
 - avoid foreign names
 - define how punctuations are to be treated
 - avoid tongue twisters
 - for dictation task: define exact rules
 - avoid inappropriate, offensive language

Eliciting Speaking Styles

- **Answering Speech**
questions on prompt sheets or screen, acoustic prompting
Hints:
 - speakers very likely deviate from the intended closed vocabulary:
"Have you been to the cinema today?" (intended: 'No.')
 - "Of course not!"
 - avoid questions that are funny or intimate:
"Are you male?" (intended: 'Yes/No.')
 - "<laugh> What a revolting idea!"
 - questions should clearly indicate the length of the expected response:
Bad: "What did you have for breakfast?"
Good: "What is your phone number?"

Eliciting Speaking Styles

- **Command / Control Speech**

prompt sheets, prompting from screen, Wizard-of-Oz

Hints:

- read commands are not equal to real commands (prosody)
- real command speech can only be obtained with convincing Wizard-of-Oz settings or a real life system

- **Descriptive Speech**

show a picture or movie and ask for description

Hints:

- more spontaneous than read or answered speech
- easy way to get speech with restricted vocabulary

Eliciting Speaking Styles

- **Non-prompted Speech**

guided conversation, role models, task solving, Wizard-of-Oz

Hints:

- very similar to spontaneous speech
- restricted vocabulary
- requires speakers that can act convincingly

- **Spontaneous Speech**

stealth recording of a conversation

Hints:

- legally problematic
- technical quality often compromised

- **Emotional Speech**

two possibilities: acted or real

Recording Procedure

Specifies the recording situation (not the technical specs)

- **Acoustical environment**

echo canceled studio, studio, quiet office,

office with printer/telephone,

office with 1/2/5/10 employees,

quiet living room (furniture, open/closed windows)

etc.

- **The ,script‘**

Defines how the speaker acts:

speaker follows instructions while not changing position,

speaker drives a car, speaker moves in the living room,

speaker points to certain objects while speaking,

speaker uses a phone

etc.

Recording Procedure

- **Background noise**

none, natural, controlled: type and level (only in studio)

- **Type, number, position and distance of microphones**

Hint:

- Use a simple sketch in the specs to clarify the description
- Make some pictures (if recording site is accessible)

Annotation

(= all kinds of segmentation or labelling)

The specifications should contain:

Which?

Types, conventions/definitions/standards, coverage

How?

Procedures, manual/automatical, training of labelers

Quality?

Error tolerance, double checks, formal checks

Documentation and Meta Data

• **Specify Documentation**

Only necessary in the specification, if working with large group of partners:

text formats, documentation templates

• **Specify Meta data**

(= formal documentation of the corpus data)

Meta data are an essential part of each speech corpus.

Therefore their minimum contents should be defined in the specification:

speaker profiles, recording protocols

Hint:

Extensive formal (computer readable) meta data help with the later documentation!

Collect meta data from the very beginning!

Technical Specifications

(= the formal properties of signals and annotations, meta data, documentation)

• **Signals** (minimum requirements)

- Sample frequency
- Sample type and width
- Byte order (if applicable)
- Number of channels (in one file or separate files)
- File format (mostly WAV or NIST)

Multimodal corpora require adequate descriptions of all modalities other than speech.

Technical Specifications

• **Annotation format**

Recommendation: Use existing format that fits requirements

•SAM : ASCII, line structured, no hierarchy, no linking, not extendable

EAF : XML, extendable, no hierarchy, no linking, no points in time

BPF : ASCII, line structured, no hierarchy, linking on word level, extendable, overlapped speech

ESPS : ASCII, very simple, not supported any more

AGS : XML, very powerful, tool libraries

(see lecture ,Speech Corpus Annotation' for details)

Technical Specifications

- **Meta data format**

No widely accepted format yet.

Recommendation: IMDI (tools available, web-based)

www.mpi.nl/IMDI/

(see papers in parallel workshop)

Technical Specifications

- **Lexicon format**

(= list of all words together with additional information)

No widely accepted formats yet.

Hints:

- code orthography in unicode whenever possible
- code pronunciation in SAM-PA or X-SAM-PA
- use non-formatted, plain text or XML
- clearly define the type of pronunciation:
canonical, citation form, most-likely
- minimum content:
unambiguous orthography, word count, pronunciation

Final Release

- **Specify Corpus structure**

(= file system structure on media)

- only necessary in large projects with several partners
- separate signal data from symbolic data
- avoid directories with more than 1024 files

- **Specify Media**

Reliable, durable, platform independent media:

CD-R, DVD-R

Avoid unreliable media: *disk, tape, magneto-optical disk,*

CD-RW, DVD+RW, hard disk

- **Define Release Plan**

- mile stones for preparation, pre-validation, collection, postprocessing, annotation, validation phases
- define topics of pre-validation and validation

What is not part of the specification:

- **Logistics**
- **Recording tools, software**
- **Recruiting scheme**
- **Postprocessing procedures**
- **Annotation tools, software**
- **Distribution tools, software**

Considerations for the cost estimate

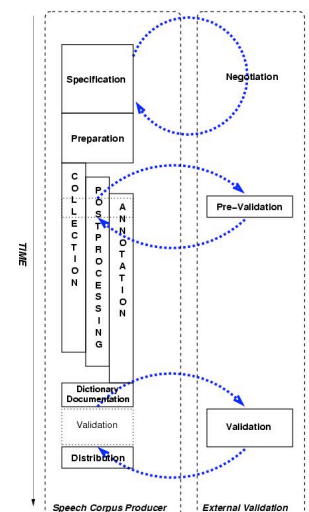
- Incentives for speakers
- Advertising for recruitment
- Special hardware
- Training = paid working time!
- Offline time for maintenance
- Costs for media and backup media
- Maintenance of web site / corpus after project termination

Speech Data Collection in Practice

Christoph Draxler
draxler@bas.uni-muenchen.de

Motivation

- You are here:
 - specification is done
- Now you have to
 - recruit speakers
 - allocate resources
 - prepare recordings
 - set up equipment
 - implement recording script
 - test, test, test



Overview

- Equipment
- Software: SpeechRecorder
 - introduction
 - demonstration
- Discussion

Types of speech corpora

- Speech technology
 - speech recognition
 - speech synthesis
 - speaker verification
 - language identification
- Speech research
 - phonetics
 - linguistics
 - medicine
 - ethnology
 - sociology

Equipment: Studio

- ≥ 2 microphones with pre-amplifiers
 - headset or clip microphone
 - table or room microphone
 - microphone array
 - conference microphones with push-to-talk buttons
- Sensor data
 - laryngograph, palatograph, etc.
- Professional digital audio mixer and audio card

Equipment: in the field

- 2 microphones with preamplifiers
 - headset and table microphone
- Digital recording device
 - laptop with external audio interface
 - portable hard disk recorder
 - tape devices
 - DAT recorder
 - DV camera with external microphones

Software: SpeechRecorder

- Multi-channel audio recording
- Multi-modal prompting
- Multiple configurable screens
- URL addressing
- Platform independent
- Localizable graphical user interface

Multi-channel audio recording

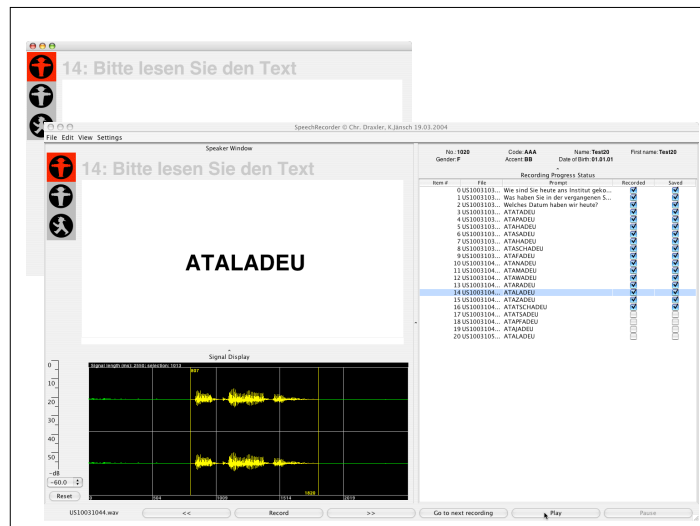
- `ipsk.audio` library
 - abstracts from Java Sound API details
 - wrapper classes for ASIO drivers
 - Digidesign, M-Audio, Emagic, etc.
- Alternative media APIs
 - QuickTime for Java (Mac and Windows)
 - Java Media Framework

Multi-modal prompting

- Prompt display
 - Unicode text
 - jpeg, gif image
 - wav audio
- Prompt sequence
 - sequential or random order
 - manual or automatic progress

Multiple configurable screens

- Speaker screen
 - recording indicator
 - instructions and prompt
- Experimenter screen
 - speaker screen
 - progress monitor
 - signal display, level meters
 - recording control buttons



URL addressing

- All resources are addressed via URLs
 - recording script
 - prompt data
 - signal files
- Perform recordings via the WWW
 - uncompressed audio

SpeechRecorder configuration

- Project: via configuration file
- Session: via parameters

```
SpeechRecorder
  recording_script
  [speaker_database
  [recording_directory]]
```

default values: anonymous speaker and user directory

Project configuration

- Input sources
- Signal parameters
- Audio library
- Recording sequence and mode
- Screen configuration

Sample project configurations

- | | |
|---|---|
| <ul style="list-style-type: none">• BITS synthesis corpus<ul style="list-style-type: none">– headset and table microphone, laryngograph– digital mixer, Digidesign audio card, standard PC– 48 KHz/16 bit– speaker and experimenter screen– text prompts– 5 speakers– > 2000 utterances each | <ul style="list-style-type: none">• Car command corpus<ul style="list-style-type: none">– 2 mouse microphones inside the car– USB audio device, laptop operated by co-driver– 44.1 KHz/16 bit– experimenter screen only– audio prompts– 25 speakers– ~ 90 utterances each |
|---|---|

Recording session configuration

- Recording script in XML format
- Tags for
 - metadata
 - media items
 - instructions to speakers
 - comments for experimenter
 - recording and nonrecording items

Recording script

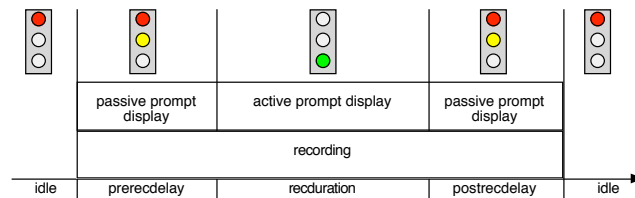
- Nonrecording items
 - speaker information and distraction
 - status feedback
 - any media type
- Recording items
 - elicit speech from speaker
 - provide hints to experimenter

Recording item

```
<!ELEMENT recording
  (recinstructions?, recprompt, recomment?)>
<!ELEMENT recinstructions mediaitem>
<!ELEMENT recprompt mediaitem>
<!ELEMENT recomment mediaitem>

<!ATTLIST recording file CDATA #REQUIRED
  recduration CDATA #REQUIRED
  prerecdelay CDATA #IMPLIED
  postrecdelay CDATA #IMPLIED
  finalsilence CDATA #IMPLIED
  beep CDATA #IMPLIED
  rectype CDATA #IMPLIED>
```

Recording phases



Sample recording script

```
<recording file="004.wav" prerecdelay="2000"
  recduration="60000" postrecdelay="500">
  <recinstructions mimetype="text/UTF-8">
    Please describe the picture
  </recinstructions>
  <recprompt>
    <mediaitem mimetype="image/jpeg"
      src="002.jpg"/>
  </recprompt>
</recording>
```

Demo

- SpeechRecorder installation
 - Edit recording script
 - Configure displays
 - Setup audio input
- Perform recording

Web recording: Access

- Project specific configuration
 - Java WebStart application
 - predefined recording script
 - bundled image, audio and video prompts
 - preset signal and display settings
- Optional login
 - check speaker identity

Web recording: Technology

- Record to client memory
 - perform signal quality checks
 - no cleanup on client hard disk necessary
 - potential danger of loss of data
- Automatic upload of recorded signals
 - background process
 - resume after connection failure

Projects using SpeechRecorder

- Bosch spoken commands in car
- BITS synthesis corpus
- IPA recordings in St. Petersburg
- Regional Variants of German - Junior
- Aphasia studies at Klinikum Bogenhausen
- your project here...

SpeechRecorder recording script DTD

This Document Type Description specifies the format for SpeechRecorder recording script files.

Version: 1.0 of April 2004

Author: Christoph Draxler, Klaus Jansch; Bayerisches Archiv für Sprachsignale, Universität München

```
<!ELEMENT session (metadata*, recordingscript)>
<!ATTLIST session id CDATA #REQUIRED>

<!ELEMENT metadata (key, value)+>
<!ELEMENT key (#PCDATA)>
<!ELEMENT value (#PCDATA)*>

<!ELEMENT recordingscript (nonrecording | recording)+>
<!ELEMENT nonrecording (mediaitem)>
<!ELEMENT recording (recinstructions?, recprompt, recomment?) >
<!ATTLIST recording
  file CDATA #REQUIRED
  recduration CDATA #REQUIRED
  prerecdelay CDATA #IMPLIED
  postrecdelay CDATA #IMPLIED
  finalsilence CDATA #IMPLIED
  beep CDATA #IMPLIED
  rectype CDATA #IMPLIED
>
<!ELEMENT recinstructions (#PCDATA) >
<!ATTLIST recinstructions
  mimetype CDATA #REQUIRED
  src CDATA #IMPLIED
>

<!ELEMENT recprompt (mediaitem)>
<!ELEMENT recomment (#PCDATA)>

<!ELEMENT mediaitem (#PCDATA)*>
<!ATTLIST mediaitem
  mimetype CDATA #REQUIRED
  src CDATA #IMPLIED
  alt CDATA #IMPLIED
  autoplay CDATA #IMPLIED
  modal CDATA #IMPLIED
  width CDATA #IMPLIED
  height CDATA #IMPLIED
  volume CDATA #IMPLIED
>
```


Sample XML recording script

This XML document is a sample recording script for the SpeechRecorder application.

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<!DOCTYPE session SYSTEM "SpeechRecPrompts.dtd">

<session id="LREC Demo recordings">
  <metadata>
    <key>
      Database name
    </key>
    <value>
      LREC Demo 2004
    </value>
  </metadata>

  <recordingscript>

    <recording prerecdelay="500" recduration="60000" postrecdelay="500" file="US10031030.wav">
      <recinstructions mimetype="text/UTF-8">
        Please answer
      </recinstructions>
      <recprompt>
        <mediaitem mimetype="text/UTF-8">
          How did you get here today?
        </mediaitem>
      </recprompt>
    </recording>

    <recording prerecdelay="500" recduration="60000" postrecdelay="500" file="US10031031.wav">
      <recinstructions mimetype="text/UTF-8">
        Please answer
      </recinstructions>
      <recprompt>
        <mediaitem mimetype="text/UTF-8">
          What did you do during the last hour?
        </mediaitem>
      </recprompt>
    </recording>

    <recording prerecdelay="500" recduration="60000" postrecdelay="500" file="US10031032.wav">
      <recinstructions mimetype="text/UTF-8">
        Please answer
      </recinstructions>
      <recprompt>
        <mediaitem mimetype="text/UTF-8">
          What day is it?
        </mediaitem>
      </recprompt>
    </recording>

    <recording prerecdelay="500" recduration="6000" postrecdelay="500" file="US10031034.wav">
      <recinstructions mimetype="text/UTF-8">
        Please read the text
      </recinstructions>
      <recprompt>
        <mediaitem mimetype="text/UTF-8">
          ATATADEU
        </mediaitem>
      </recprompt>
    </recording>

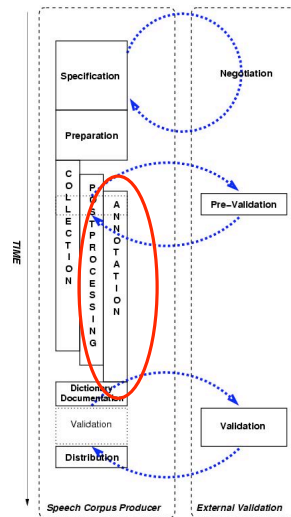
    <recording prerecdelay="500" recduration="6000" postrecdelay="500" file="US10031050.wav">
      <recinstructions mimetype="text/UTF-8">
        Please read the text
      </recinstructions>
      <recprompt>
        <mediaitem mimetype="text/UTF-8">
          ATALADEU
        </mediaitem>
      </recprompt>
    </recording>

  </recordingscript>
</session>
```

Speech Corpus Annotation

Florian Schiel

- All information related to signals
- Without annotation no corpus!
- Often the most costly part!
- Quality is everything!



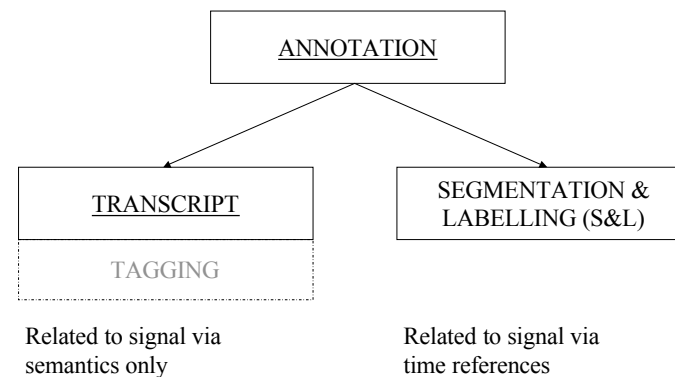
Before getting started: Some general rules

- Minimum required annotation is a basic transcript
- Use standards
- Use existing tools or libraries
- Allocate considerable time for the annotation phase (40-60% of total project time); observe *Hofstetter's Law*
- Use the check list on page 119
- If not sure about feasibility, ask experts (LDC, ELDA, SPEX, BAS)

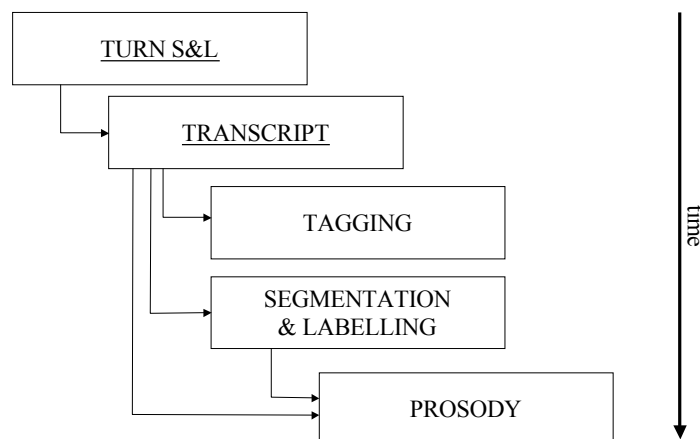
Overview: Annotation of a Speech Corpus

- General points: types, dependencies, hierarchy
- Transcription
- Tagging
- Segmentation and Labelling (S&L)
- Manual annotation tools
- Automatic annotation tools
- Formats
- Quality: double check, logging, inter-labeler agreement

Types of Annotation



Dependencies and Hierarchies



Transcription

General

- One transcription file (line) per signal file
- Minimum: *spoken words*
- Other: *noise, syntax, pronunciation, breaks, hesitations, accents, boundaries, overlaps, names, numbers, spellings etc.*
- Standard for spelling (Webster, German Duden, Oxford Dict.)
- No capital letters at sentence begin
- No punctuation (or separate them from words by blanks)
- Transcribe digits as words ('12.5' -> 'twelve point five')

Transcription

Format

- Use or provide transformation into standard format
- (Use 'readable' intermediate format for work)
- Common formats: *SAM, (SpeechDat), Verbmobil, MATE, EAF*

Transcription

Minimal software requirements

- text editor with 'hot keys', syntax parser (e.g. *Xemacs*)
- simple replay tool, markup and replay of parts
- use *WebTranscribe* (see demo)

Logistics

- train transcribers
- 'check-out' mechanism (data base) for parallel work
- two steps: labeling + correction (preferable one person!)
- formal checks: syntax + extract word list (typos)

Transcription

Example:

w253_hfd_001_AEW: hallo [PA] [B3 fall] . <#> <"ahm> [B2] ich wollt' fragen [NA] [B2] , was heute abend [NA] im Fernsehen [PA] kommt [B3 fall] .

w253_hfw_002_SMA: hallo . <P> <#> was kann ich f"ur Sie tun ?

w253_hfd_003_AEW: <"ah> [B2] ich w"urde ganz gern [NA] das Abendprogramm [PA] wissen [B3 fall] .

w253_hfw_004_SMA: wenn ich Ihnen einen Tip geben darf , <P> <#> heute kommt ~Der+Bulle+von+T"olz auf ~Sat-Eins um #zwanzig Uhr #f"unfzehn .

w253_hfd_005_AEW: -/und wa=-/ [B9] <"ah> [NA] [B2] gibt es heute [NA] abend eine *Sportshow [PA] [B3 cont] ? <P> zum Beispiel [NA] Fu"sball [PA] [B3 rise] ?

Tagging

- Markup of words or chunks based on categorical system
- Often based on an existing transcript

Examples:

- *Dialog acts, parts-of-speech, canonical pronunciation, prosody*

ORT: 0 good

ORT: 1 morning

ORT: 2 have

ORT: 3 we

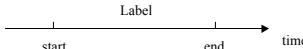
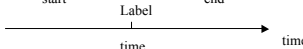
ORT: 4 met

ORT: 5 before

DIA: 0,1 GREETING_AB

DIA: 2,3,4,5 QUERY_AB

Segmentation & Labeling (S&L)

- *segment*: (start, end, label) 
- *point-in-time event*: (time, label) 
- S&L requires more knowledge than transcript (training budget!)
- time effort ~ 1 / length of units
- maximize inter- and intra-labeler agreement

Examples:

turns/sentences, dialog acts, phrases, words, syllables, phonemes, phonetic features, prosodic events

Segmentation & Labeling (S&L)

Software requirements: *Praat* (www.praat.org)

Logistics: same as transcription

Example: *SAP tier of BAS Partitur Format (BPF)*

SAP: 1232 167 0 Q%<
 SAP: 1399 2860 0 E:
 SAP: 4259 884 1 v
 SAP: 5143 832 1 a
 SAP: 5975 914 1 s%>
 SAP: 6889 599 2 h%<
 SAP: 7488 545 2 aq
 SAP: 8033 662 2 lq
 SAP: 8695 431 2 t
 SAP: 9126 480 2 -H
 SAP: 9606 0 2 @-
 SAP: 9606 429 2 n
 SAP: 10035 628 3 z
 SAP: 10663 733 3 i:-I

→ SAP: 10663 733 3 i:-I
 Tier marker
 Begin
 End
 Word number
 Phonemic substitution

Manual Annotation Tools

<i>Transcript:</i>	<i>WebTranscribe (see demo)</i>
	<i>ELAN (www.mpi.nl/tools/elan.html)</i>
<i>S&L:</i>	<i>Praat (www.praat.org)</i>
<i>Video:</i>	<i>ANVIL (www.dfki.de/~kipp/anvil)</i>
<i>Video + Transcript:</i>	<i>CLAN (childes.psy.cmu.edu/clan/)</i>

Automatic Annotation Tools

<i>Transcript:</i>	-
<i>Segmentation into words:</i>	Viterbi Alignment of HMMs e.g. <i>HTK, Aligner, MAUS</i>
<i>S&L:</i>	- Viterbi Alignment of HMMs - <i>MAUS</i> - <i>Elitist Approach</i> : segmentation of phonetic features
<i>Prosody:</i>	ToBi Light, e.g. IMS Stuttgart
<i>Parts-of-Speech:</i>	POS tagger, e.g. IMS Stuttgart
<i>Video + Transcript:</i>	-

Annotation File Formats

- SAM (www.phon.ucl.ac.uk/resource/eurom.html)
 - ASCII, good for simple, single speaker corpora
 - used in SpeechDat
- EAF (Eudico Annotation Format, MPI Nijmegen)
 - Unicode, XML
 - powerful, but not widely used in technical corpora
- BPF (BAS Partitur Format)
 - ASCII, simple, extendable
 - relates tiers over time and word order
- XWAVES
 - basic, but is used by EMU for hierarchical label database
- AGS (Annotation Graphs)
 - XML, extendable, C-library, Java API

Quality Control & Assessment

- Comprehensive, clear, constant guidelines
- Extensive, consistent, on-going training (forum, meetings)
- Second pass / correction pass, preferably by one person / trainer
Error logging: documents progress, may be used in training
- Formal checks (syntax, label inventory)
- Double/triple annotations of parts of the corpus:
 - inter labeler agreement
 - intra labeler agreement (over time)
 1. symmetric label accuracy (*Kipp, 1998*)
 2. histograms of boundary deviations

Examples

• WebCommand Transcription

CMT: *** Label file body ***
LBD:
LBR: 0,149216,,,,start my computer
LBO: 0,74608,149216,start my computer
ELF:

Examples

• Verbmobil POS Tagging

ORT: 0 also
ORT: 1 ausgerechnet
ORT: 2 habe
ORT: 3 ich
ORT: 4 am
ORT: 5 dritten
ORT: 6 Juli
POS: 0 ADV
POS: 1 ADJD
POS: 2 VAFIN
POS: 3 PPER
POS: 4 APPRART
POS: 5 ORD
POS: 6 NN

• Verbmobil Pronunciation

KAN: 0 Q'alzo+
KAN: 1 Q'aUsg@r"ECn@t
KAN: 2 h'a:b@+
KAN: 3 Q'IC+
KAN: 4 Q'am+
KAN: 5 dr'lt@n
KAN: 6 j'u:li:

• Verbmobil S&L

MAU: 0 799 -1 <p:>
MAU: 800 799 0 Q
MAU: 1600 799 0 a
MAU: 2400 799 0 z
MAU: 3200 799 0 o
MAU: 4000 1599 1 aU
MAU: 5600 479 1 s

URL

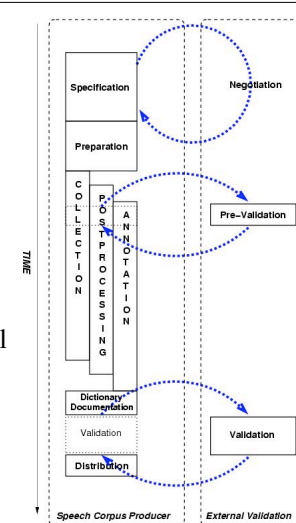
- BAS Home Page: <http://www.bas.uni-muenchen.de/Bas>
- SPEX Home Page: <http://www.spex.nl/>
- Steven Greenberg Elitist Approach: <http://www.icsi.berkeley.edu/~steveng/>
- MAUS: <ftp://ftp.bas.uni-muenchen.de/pub/BAS/SOFTW/MAUS>
- Praat: <http://www.praat.org>
- HTK: <http://htk.eng.cam.ac.uk/>
- ANVIL: <http://www.dfki.de/~kipp/anvil>
- WebTranscribe: <http://www.bas.uni-muenchen.de/Bas>
- CLAN: <http://childes.pry.cmu.edu/clan>
- SAM: <http://www.phon.ucl.ac.uk/resource/eurom.html>
- Eudico Annotation Format (EAF), ELAN: <http://www.mpi.nl/tools/elan.html>
- BAS Partitur Format: <http://www.bas.uni-muenchen.de/Bas/BasFormatseng.html>

Speech Annotation in Practice

Christoph Draxler
draxler@bas.uni-muenchen.de

Motivation

- You are here:
 - recordings have started
- Now you have to
 - annotate your data
 - implement quality control
 - submit data for (pre-)validation



Overview

- Annotation editors
- WebTranscribe
 - architecture
 - configuration
 - demonstration
- Discussion

Annotation editor requirements

- Tailored to the task
- Intuitive to use
- Extensible
- Scalable
- Platform independent

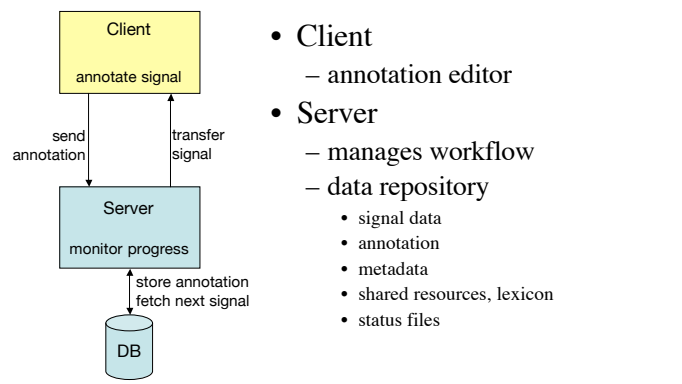
There is not one editor...

- Different annotation tiers...
 - phonetic segmentation, phonemic labelling
 - orthographic transcription
 - POS tagging, dialogue markup, syntax trees, etc.
- ...and formats...
 - free form text, implicitly structured text, text markup
- ...and different types of data
 - audio, video, sensor

...but the procedure is always the same

1. get signal data
2. enter or edit the annotation
3. assess the signal quality
4. perform formal checks on annotation
5. save annotation and meta-data
6. go back to square 1

Annotations in a client/server system



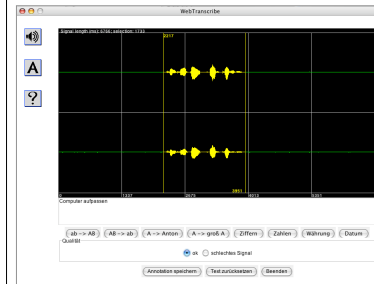
WebTranscribe

- Tailored to the task
- Intuitive to use
- Extensible
- Scalable
- Platform independent
- task-specific modules
- clean interface
- plug-in architecture
- any number of clients
- Java and any RDBMS
- zero client configuration
- localizable user interface

WebTranscribe server

- Simple web server with cgi-functionality
 - dynamic HTML pages
 - generated by scripts, e. g. perl
- Current implementation
 - Java servlets with Tomcat
 - annotations held in relational database
 - signals stored in file system

WebTranscribe client



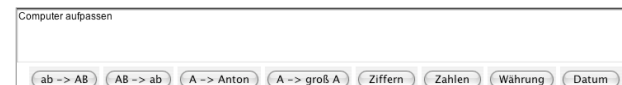
- Signal display
 - segmentation
 - audio output
- Annotation edit panel
 - annotation field
 - editing support
- Quality assessment
- Progress controls

Client implementation

- Applet
 - browser compatibility problems
 - limited access to local resources
- Java WebStart
 - easy software distribution
 - controlled access to local resources
 - secure environment
 - independent of browser

Annotation edit panel

- Text area for annotation text
- Editing support for often-needed tasks
 - digit and string conversion
 - marker insertion



Formal consistency checks

- Annotation accepted only if formal check succeeds
- Tokenizer, parser implemented in edit panel module
- Error handling on client
- Updates of central resources immediately available to client

Configuration

- Server contains project specific package
 - server URL and signal paths
 - database access
 - mapping of database table names to annotation variables
 - type of annotation
- No client configuration necessary

WebTranscribe demo


- Download software
- Configuration
- Sample annotations
 - RVG-J signal files
 - annotations according to SpeechDat
- Discussion

Final steps

- Signal files are stored in place
- Database stores all annotations
 - export database contents
 - export metadata
- Write documentation
- Submit speech database to validation

Extending WebTranscribe

- Additional annotation tiers
 - plug-in architecture
 - graphical annotations
- Enhanced signal display
 - zoom, scroll, multiple selections
 - additional display types, e.g. sonagram

SPEX 


Validation and Distribution of Speech Corpora

Henk van den Heuvel

SPEX: Speech Processing Expertise Centre

CLST: Centre for Language and Speech Technology

Radboud University Nijmegen, Netherlands

SPEX 


SPEX: Mission statement

The mission statement of SPEX is:

1. to provide and enrich spoken language resources and concomitant tools which meet high quality standards
2. to assess spoken language resources
3. and to create and maintain expertise in these fields

SPEX aims to operate:


- for both academic and commercial organisations
- as an independent academically embedded institution

SPEX 

SPEX: Organisation

Employees (in chronological order):

Lou Boves	(0.0 fte)
Henk van den Heuvel	(0.6 fte)
Eric Sanders	(0.5 fte)
Andrea Diersen	(0.7 fte)
Dorota Iskra	(1.0 fte)
Folkert de Vriend	(1.0 fte)
Micha Baum	(1.0 fte)

SPEX 

SPEX: Activities

SPEX's main activities at present are the creation, annotation and validation of spoken language resources.

- SPEX has been selected as the ELRA's primary Validation Centre for speech corpora. Further, SPEX acts as validation centre for several European projects in the SpeechDat framework.
- SPEX is also involved in the creation and/or annotation of SLR.
- SPEX fulfilled several tasks in the construction of the Dutch Spoken Corpus (CGN).
- Publication of results in proceedings, journals

SIPIEX **Overview of the presentation**

- **Validation**
 - What is SLR validation
 - Overview of validation checks
 - History of SLR validation
 - Aims of validation
 - Dimensions of validation
 - Validation flow and types
 - What can be checked automatically
 - Validation software
 - On the edge of SLR validation: phonetic lexica
 - SPEX and SLR validation
 - Validation at ELRA & LDC
- **Distribution**
 - Models of distribution
 - ELRA & LDC

SIPIEX **What is SLR Validation? (1)**


- **Basic question: What is a "good" SLR?**
 - "good" is what serves its purposes
 - Evaluation and Validation
- **Validation of SLRs:**
 1. Checking a SLR against a fixed set of requirements;
 2. Putting a quality stamp on a SLR as a result of the aforementioned check. If the database passes the check, then we say that it has been "validated"

SIPIEX **What is SLR Validation? (2)**

- **Validation criteria**
 - Specifications
 - Tolerance margins
- **Specs & Checks**
 - have a matrimony in validation
- **Validation and SLR repair are different things:**
 - Diagnosis and cure
- **Dangerous to combine !**


SIPIEX **Overview of checks**

- **Documentation**
- **Database format**
- **Design**
- **Speech files**
- **Label files**
- **Lexicon**
- **Speakers and recording environments**
- **Transcriptions**
 - Example: template report SALA II

SIPIEX 


History of SLR validation (1)

- Production of similar SLRs in (European) Consortia
 - SpeechDat family
- Principle of "Put in one, pull out many"
- "E-quality" (Equality in quality) of SLRs becomes of paramount importance
- Demand for independent validation institute

SIPIEX 


History of SLR validation (2)

- ELRA has a similar demand for quality control for the SLRs in the catalogue: customers value a quality stamp
- The same is true for the LDC

SIPIEX 

Aims of validation

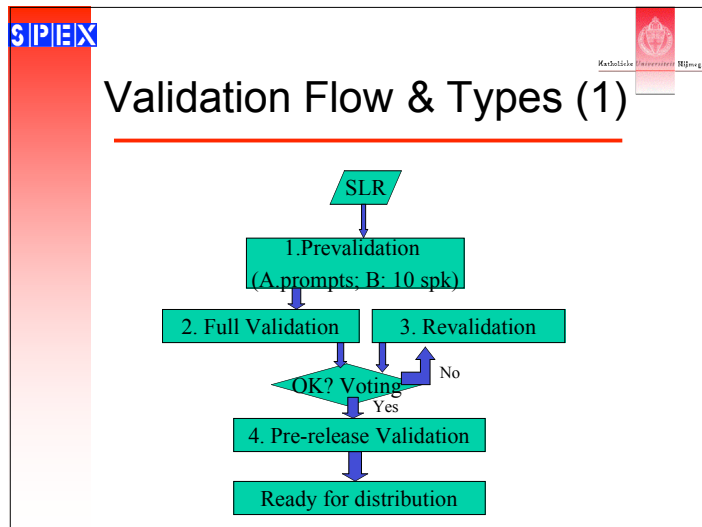
- Quality assurance
- Quality improvement


SIPIEX 


Dimensions of validation


- Two dimensions:
 - Dim. 1: checks vs specs
 - Dim. 2: subjective vs objective


Validator	Validation scheduling	
	During production	After production
Internal	(1)	(2)
External	(3)	(4)



- SIPIEX** 
- ## Validation Flow & Types (2)
- Objectives of prevalidation
 - Detect major shortcomings before recordings start
 - Develop software for:
 - Database formatting (producers)
 - Database validation (SPEX)
 - Types of prevalidation
 - Check of all prompt sheets and lexicon (before any recording): is db potentially OK?
 - Mini database of 10 speakers


- SIPIEX** 
- ## Validation Flow & Types (3)
- Full validation
 - On *complete* database
 - Preceded by a Quick Check on formats
 - All checks, incl. transcriptions/completeness checks
 - Voting procedure
 - Provider obtains validation report with request to comment to the report and to the list of irreparable shortcomings if any (design/transcription errors)
 - (Updated) report together with main shortcomings & reply provider is sent to consortium with request to vote
 - In case of rejection rectification of the corpus and revalidation is necessary

- SIPIEX** 
- ## Validation Flow & Types (4)
- Purpose of Pre-release-validation
 - Final check on master CD before distribution
 - Procedure
 - Check if all files are there
 - Check if most recent versions of files are there
 - One more run of validation software to preclude any hidden format defects
 - At remaining errors: rectification and revalidation necessary

SIPIEX 


Validation Flow & Types (5)

- **Evaluation:**
 - close involvement in the specification phase desired / recommended
 - How to avoid a full revalidation
 - resubmission of files "on the fly"
 - include minor corrections in the documentation file
 - Gap between validation and CD mastering should be kept minimal
 - Validation costs (paradox)

SIPIEX 


What can be checked automatically ?

Automatic	By hand/ear
	Documentation
Database format	
Design	
Speech files	Speech files
Label files	
Lexicon	Lexicon
Speakers and recording environments	
Transcriptions	Transcriptions
	Interpretation of output software
	Editing the validation report

SIPIEX 


Validation software

- Is it advantageous to distribute the validation software to database providers?
 - **Yes**
 - they can check in advance
 - **No**
 - No double check
 - Validation centre becomes helpdesk
 - Platforms, prog. languages, errors...
 - Delays in database delivery

SIPIEX 


On the edge of SLR validation: Phonetic Lexicons

- What is an SLR:
 - Speech database
 - Phonetic lexicon
- LC-STAR as example:
 - 12 lexicons with common/application words and names for ASR & TTS: Lemma, phon.transcriptions, POS tags
 - SPEX: XML-format, documentation, phon.transcriptions (see LREC 2004 paper)
 - CST: POS-tags
 - Bilingual lexicons?
 - Corpora?

SIPEX 


SPEX & SLR validation (1)

- Checks, specs & SPEX
 - Internal validation of data productions
 - External validation these data: by client or by another institute (CGN)
 - External validator in SpeechDat projects & successors and for ELRA

SIPEX 

SPEX & SLR validation (2)


Project	SLR	Period
SpeechDat(M)	8 FDB	1994-1996
SpeechDat(II)	20 FDB 5 MDB 3 SDB	1995-1998
Speechdat-Car	9 CDB	1998-2001
SpeechDat-East	5 FDB	1998-2000
SALA	4-5 FDB	1998-2000
SALA II	12 MDB	2001-2004
LILA	?? ?DB	2004-
SpeechDat-AT	1 FBD,1MDB	2000
SpeechDat-AU	1 FDB	2000
Speecon	18 HDB	1999-2003
NET-DC	1 BCNDB	2002
OrienTel	23 FDB	2001-2004
LC-STAR	12 LEX	2002-2005

SIPEX 

SPEX & SLR validation (3)


Principles:

- SPEX validates SLR (not WLR)
- SPEX aims at involvement in the specification phase of a project in order to avoid backward engineering and other infeasibilities afterwards
- SPEX never creates a database that it has to validate itself
- SPEX only checks databases, but does not modify them, to avoid that we check our own work

SIPEX 


Validation at ELRA

- Quality assessment of LR in catalogue
- VCOM with two validation centres
 - SPEX for SLR
 - CST (Copenhagen) for WLR
- Tasks
 - Validation manual
 - Bug report handling
 - Quick Quality Checks (QQCs)

SIPIEX 


ELRA's bug report service

- Accessible via <http://www.elra.info>
- Bug reports
- Formal error lists
 - Made by validation centre after verification
 - Accessible via web after approval provider
- Correction (by provider)
- Patches

SIPIEX 


ELRA's QQC procedure

- A QQC is a quick validation restricted to formal properties of a database and the documentation
- Done on LR in ELRA's catalogue or entering it
- Takes about 6 working hours
- Results in two reports:
 - For provider or end-user (about LR proper):
 - Based on check-list minimal requirements
 - Accessible via web after approval provider
 - For ELDA:
 - about information on description forms
- Updates of LR and/or description forms

SIPIEX 

Validation at LDC/BAS

- Self-produced corpora
- Internal validation
- External corpora are upgraded and reformatted to LDC's own quality standards / BAS has no external corpora
- There are no validation reports for LR available
- Bugs can be reported via website

SIPIEX 

So much for validation ...

SIPIEX

Kabulika Universitato HJmg

Distribution

- Do it yourself
- Do it via broker (ELRA or LDC)
- Advantages broker: a central place for
 - LR identification
 - Contracts/licenses
 - Marketing/pricing
 - Packaging/shipping
 - Quality maintenance

SIPIEX

Kabulika Universitato HJmg

Distribution at ELRA

- Steps:
 1. Description of LR (by description forms)
 2. Licensing
 - By tailoring generic contract models
 - Usage/pricing/royalties
 3. QQC (if not validated before)

```

graph TD
    Owners --> Providers
    Providers -- "Distribution agreement" --> ELRA
    ELRA -- "VAR agreement" --> VAR
    ELRA -- "End-user agreement" --> END_users1[END-users]
    VAR -- "End-user agreement" --> END_users2[END-users]
  
```

SIPIEX

Kabulika Universitato HJmg

Membership of ELRA/LDC

	ELRA	LDC
Fee	EUR 750 - 5,000	\$2000 - 20,000
LR price	Reduced for members	Free for membership year
Member binding	Fidelity program	On-line service

SIPIEX

Kabulika Universitato HJmg

ELRA Sales

Year	Speech	Written	Terminology	Total
2001	~250	~150	~100	408
2002	~350	~150	~100	481
2003	~250	~150	~100	362

Methodology for a Quick Quality Check of SLR and Phonetic Lexicons

<i>Project reference number</i>	ELRA/0201/VAL-1
<i>Project acronym</i>	ELRA VCom
<i>Project full title</i>	ELRA SLR Validation
<i>Project contact point</i>	Harald Höge
<i>Project web site</i>	http://www.elra.info
<i>EC project officer</i>	
<i>Document title</i>	Methodology for a Quick Quality Check of SLR and phonetic lexicons
<i>Deliverable ID</i>	D1.2
<i>Document type</i>	Report
<i>Dissemination level</i>	/ELRA Vcom / ELRA Board
<i>Contractual date of delivery</i>	
<i>Actual date of delivery</i>	31 March 2003
<i>Status & version</i>	V2.2: draft
<i>Work package & task ID</i>	
<i>Work package, task & deliverable responsible</i>	
<i>Number of pages</i>	32
<i>Author(s) & affiliation(s)</i>	Henk van den Heuvel, SPEX
<i>Additional contributor(s)</i>	
<i>Keywords</i>	Quality Control, Validation, SLR
<i>Abstract</i>	
<i>Additional notes & remarks</i>	

Document evolution

Version	Date	Status	Notes
1.0	25 October	First draft	
1.1	5 November	First Update after 30 October 2001 (meeting Nijmegen)	
1.2	9 November	Update after meeting Nijmegen, 30 October 2001	To be added: priority list of SLR for QQC !! Together with KC !!
1.3	10 April 2002	Pre-Final	Update after meeting in Pisa, 22 Jan. 2002; priority list added
1.4	16 Oct. 2002	Final	Update after meeting in Copenhagen, 9 Oct. 2002. Coversheet table modified.
2.0	31 May 2003	New draft	New approach: <ul style="list-style-type: none">- stars assessment represents minimal requirements instead of documentation match- part for phonetic lexicons added
2.1	14 Nov. 2003		<ul style="list-style-type: none">- Check of description forms included- Other comments at VCOM meeting of 19/9/03 included
2.2	31 March 2004	Draft	<ul style="list-style-type: none">- Comments after meeting Jan. 2004 (BRUSSELS) included- Split in QQC_DB (about the database quality) and QQC_DF (about correctness of description forms)

ELRA Contact :

ELRA - Distribution Agency (ELDA)

Dr. Khalid CHOUKRI

CEO

55-57, rue Brillat Savarin

F-75013, PARIS, FRANCE

Tel. +33 1 43 13 33 33

Fax. +33 1 43 13 33 30

Email: choukri@elda.fr

Table of contents

1	Introduction	4
2	Principles and realisation	4
2.1	QQC_DB	4
2.2	QQC_DF	6
3	Layout of the QQC report	7
4	A priority listing	8
5	References	8
6	Appendix A : Priority list of SLRs	9
7	Appendix B: QQC_DB templates	11
7.1	QQC_DB for SLR (ASR applications)	11
7.2	QQC_DB for SLR (phonetic lexicons)	17
7.3	QQC_DB for SLR (speech synthesis)	22
8	Appendix C: QQC_DF templates	23
8.1	QQC_DF for SLR (ASR applications)	23
8.2	QQC_DF for SLR (phonetic lexicons)	27
8.3	QQC_DF for SLR (speech synthesis)	31

1 Introduction

There are many Spoken Language Resources (SLRs) in ELRA's catalogue which have not been validated before. The same holds for many new databases that are offered to ELRA to be sold. A full validation protocol as outlined in [2] takes a lot of time (approx. 40 hrs). ELRA would like to have some first indications of the quality of a yet unvalidated SLR in the catalogue. To achieve this, two strategies will be followed.

1. Install a bug report service at ELRA web pages. The reports sent in by users of the SLRs give an indication of possibly deficient SLRs. A framework for this service is presented in [3].
2. A quick quality check should be construed, so that a first impression of the quality of an SLR can be obtained. This can help in establishing the priority list for the validation. A methodology for this quick quality check (abbreviated from now as QQC) is presented in this report.

In section 2 we describe the general aspects of the realisation of the QQCs. The general layout of QQC reports are presented in section 3. Section 4 presents the relevant parameters for a priority listing of SLR to be submitted for a QQC. Appendix A contains a priority listing. QQC templates are presented in Appendix B.

2 Principles and realisation

As points of departure for the QQC the following principles are adopted:

- A. The QQC mainly checks the database contents against a number of minimal requirements. These requirements are of a formal nature which enables a quick check. Content checks are included in other types of validations.
- B. Generally, a QQC should take about 6-7 hours work at maximum (for one person at SPEX)
- C. For each SLR two QQC reports are produced: One for the provider and users on the quality of the database proper (QQC_DB); one for ELDA on the quality of the information on the description forms (QQC_DF)

2.1 QQC_DB

The QQC report contains a quality assessment of the resource with respect to a number of minimum formal requirements. A star notation is used for this.

Meaning of the quality stars:

- * : The minimal criteria for this part of the lexicon are not fulfilled.

** : The minimal criteria for this part of the lexicon are reasonably well fulfilled.

*** : The minimal criteria for this part of the lexicon are all fulfilled.

Other values:

Not Applicable: This part is not applicable for this resource

Missing: This part is missing in the resource, but relevant

Also these values are given in the Quality value column (merging the three cells in the row).

The basic topics checked in a QQC are outlined in Table 1. Table 1 is adopted from section 3 of [2].

Database part	ELRA rating		
	*	**	***
1. Documentation			
2. Format			
3. Design & contents			
4. Speech signals			
5. Annotation files			
6. Speakers			
7. Environments			
8. Transcriptions			
9. Lexicon			

Table 1: Basic assessment sheet for a QQC on SLR in ELRA's catalogue

Depending on the type of resource the basic assessment sheet will contain the relevant elements.

Concluding remarks about the database will be added on the cover sheet juxtaposed to the assessment table. In each QQC report the individual checks and their results will be presented after the cover sheet. QQC_DB templates are presented in appendix B. There are different templates for different types

(application domains) of SLR.

The QQC_DB report is intended for ELRA's database users if the database is already in the catalogue and for the database providers if the database is new and not in the catalogue yet. ELDA will forward QQC reports to providers for comments. The resulting QQC report will be made available via ELRA's web pages (catalogue).

2.2 QQC_DF

Each database at ELRA is accompanied by one or two description forms: a general description form and/or a specific description form (depending on the type of resource). These description forms contain the basic information about a database according to ELRA. The description forms are filled out in cooperation with the LR provider. The form is used to inform potential customers about the database. The information provided on the description form should be correct. The correctness of this information is also a minimum requirement for a database and checked at the QQC.

The QQC_DF report contains a quality assessment of the correctness of the information on the description forms. A star notation is used for this as well.

Meaning of the quality stars:

* : The information provided is insufficient/incorrect

** : The information provided needs some improvement or extension

*** : The information provided is complete and correct

Other values:

Not Applicable: This part is not applicable for this resource

Missing: This part is missing in the resource, but relevant

Also these values are given in the Quality value column (merging the three cells in the row).

The basic topics checked in a QQC are outlined in Table 2.

Database part	ELRA rating		
	*	**	***
1. General description form			
2. Specific description form			

Table 2: Basic assessment sheet for a QQC on description forms of SLR in ELRA's catalogue

Concluding remarks about the database will be added on the cover sheet juxtaposed to the assessment table. In each QQC report the individual checks and their results will be presented after the cover sheet. QQC_DF templates are presented in appendix C.

The QQC report is intended for ELDA since ELDA is responsible for the contents of the description forms. ELDA will take care of the required modifications of the description forms according to the QQC_DF.

The resulting QQC report is not meant for publication in the internet, only for improving the content of the description forms; these forms, though, are available to the public.

3 Layout of the QQC report

Tables 1 and 2 serve as summary sheets for QQC reports for QQC_DB and QQC_DF, respectively. After the check of a part is completed, a star rating is given. Thus the complete table is filled out. It is followed by a more detailed account dealing with the individual checks. Finally, a brief conclusion is added to the summary sheet.

ELRA QQC REPORT FOR {SLR | LEXICA}
QQC IS CARRIED OUT BY SPEX, NIJMEGEN, THE NETHERLANDS

TITLE DATABASE:

VERSION OF DATABASE:

TYPE OF DATABASE:

DATABASE OWNER / PRODUCER:

ELRA CATALOGUE NUMBER:

AUTHORS:

DATE:

VERSION:

VERSION OF QQC TEMPLATE:

>> SUMMARY SHEET (TABLE 1) AND CONCLUDING COMMENTS

>> MORE DETAILED REPORT ON CHECKS IN SECTION 2

>> OTHER REMARKS, IF ANY

Templates of QQC reports according to this framework can be found in Appendices B and C.

4 A priority listing

A priority list of SLRs to be submitted to a QQC is set-up together by ELRA's CEO and SPEX. The order of SLRs in this list is determined by the following parameters:

1. Type of SLR. We will first concentrate on speech databases, and on pronunciation lexicons. For the time being other SLRs (such as multi-modal SLRs) are outside the scope of QQC.
2. Popularity. An SLR that sells well should have priority over an SLR that is not sold yet. On the other hand, SLRs that will not be sold do not need a QQC. Popularity is therefore a mix of copies sold and expected sell of copies.
3. Layout of the media. SLRs that have the documentation and label files separate from the speech files (e.g. SAM instead of NIST headers) can be quicker handled for a QQC and have therefore a (somewhat) higher priority.

The priority list is given in the Appendix A.

5 References

- [1] Henk van den Heuvel, Louis Boves, Eric Sanders (2000) *Validation of Content and Quality of SLR: Overview and Methodology*. ELRA/9901/VAL-1 Deliverable 1.1
- [2] Henk van den Heuvel (2000) *Procedures to Validate and Improve Existing SLR*. ELRA/9901/VAL-1 Deliverable 2.1
- [3] Henk van den Heuvel (2001) *A Bug Report Service for ELRA*. ELRA/2001/VAL-1 Deliverable 2.3

6 Appendix A : Priority list of SLRs

The following SLRs were sold more than 5 times (status as of 14 Jan. 2002).

Ref	resource name	# sold	valid
S0004	BDLEX 50000	31	
S0006	BREF-80	21	
S0021	M2VTS	15	
S0042	POLYCOST	15	
S0016	FRESCO - DB1	13	x
S0018	German SpeechDat(M) Database - DB1	13	x
S0035	PHONOLEX (BAS/DFKI)	13	
S0011	English SpeechDat(M) database - DB1	12	x
S0052	FIXED0IT - DB1	12	x
S0067	BREF-120 - A large corpus of French read speech	11	
S0005	BDSONS	10	
S0010	Dutch Polyphone Database	10	
S0023	PHONDAT 1 – PD1 (2nd edition)	9	
S0045	German Pronunciation Rules Set - PHONRUL 9.0	9	
S0065	Spanish SpeechDat(M) Database - DB1	9	x
S0009	COST232	8	
S0015	EUROM1i	8	
S0031	TED	8	
S0039	APASCI	8	
S0043	ONOMASTICA-COPERNICUS DATABASE	8	
S0060	MULTEXT Prosodic database	8	
S0025	SIEMENS 100 - SI100	7	

S0058	RVG1 (Regional Variants of German 1, Part 1)	7	
S0059	ILE: Italian Lexicon	7	
S0001	ACCOR – English	6	
S0032	TEDPhone	6	
S0061	French Speechdat(II) FDB-1000	6	x
S0074	British English SpeechDat(II) MDB-1000	6	x
S0007	BREF-POLYGLOT	5	
S0020	GRONINGEN	5	
S0051	German SpeechDat(II) FDB-1000	5	x
S0063	German SpeechDat-II FDB-4000	5	x

From this list a first priority list was compiled by ELRA:

Ref	resource name	# sold
S0004	BDLEX 50000	31
S0031	TED	8
S0067	BREF-120 - A large corpus of French read speech	11
S0025	SIEMENS 100 - S1100	7
S0042	POLYCOST	15
S0009	COST232	8
S0032	TEDPhone	6

The QQC's for this list were completed through 2002. The former approach was used to check a database against its own documentation (or lack of it), outlined in version 1.4 of this report.

7 Appendix B: QQC_DB templates

7.1 QQC_DB for SLR (ASR applications)

ELRA QQC REPORT FOR SLR (ASR APPLICATIONS)

QQC WAS CARRIED OUT BY SPEX, NIJMEGEN, THE NETHERLANDS

TITLE DATABASE:

VERSION OF DATABASE:

TYPE OF DATABASE:

DATABASE OWNER / PRODUCER:

ELRA CATALOGUE NUMBER:

AUTHORS OF QQC REPORT: Henk van den Heuvel

DATE:

VERSION OF REPORT:

VERSION OF QQC TEMPLATE 2.2

SUMMARY SHEET:

GENERAL REMARKS:

Database part	ELRA Rating		
	*	**	***
1. Documentation			
2. Format			
3. Design & contents			
4. Speech signals			
5. Annotation files			
6. Speakers			
7. Environments			
8. Transcriptions			
9. Lexicon			

The QQC report contains a quality assessment of the resource with respect to a number of minimum formal requirements that are outlined in this report.

Meaning of the quality stars:

- * The minimal criteria for this part of the SLR are not fulfilled.
- ** The minimal criteria for this part of the SLR are reasonably well fulfilled.
- *** The minimal criteria for this part of the SLR are all fulfilled.

MISSING: This part of the database is missing, but relevant

NOT APPLICABLE: This part is not applicable to this SLR

1 Quick Quality Check Report

1.1 Documentation

The most important topics should be covered and clearly described in the documentation:

- Owner and contact point
- db layout and media
- application potential for the SLR
- directory structure and file names
- recording equipment
- design and contents of the recordings
- coding and format of the speech files
- contents and format of the label files
- speakers
- recording environments distinguished
- transcription conventions and procedure
- lexicon: format and transcriptions included

1.2 Format

- The file names and directory structure should correspond to the documentation
- The resource is in a well-known standard

1.3 Design and contents

- All mandatory items according to the documentation should be included
- Number of missing files per corpus item should be less than 10%

1.4 Speech signals

- Empty speech files are not permitted
- Acoustic measurements on the speech files will be made, and the results reported. The acoustical measurements involved are:
 - o Clipping rate
 - o SNR
 - o Mean amplitude

1.5 Annotation files

- Empty label files are not permitted

- A random selection of the annotation/label files will be checked. They should be
 - o Readable
 - o Contain the information described in the documentation

1.6 Speakers

- Speaker distributions should be in agreement with documentation
- Proportion of each speaker sex should be appropriate for application
- Distribution of speaker accents should be in agreement with documentation

1.7 Environments

- Environment distributions should be in agreement with documentation

1.8 Transcription

- A max of 5% of the speech files may miss an orthographic transcription (no or empty transcription files)
- All non-speech markers should be described in documentation

1.9 Lexicon

- All documented phones should be used
- All used phones should be documented
- All words in the (orth.) transcriptions should be present in the lexicon
- All words in the lexicon must have at least one phon. transcription

1.10 Other remarks

7.2 QQC_DB for SLR (phonetic lexicons)

ELRA QUICK QUALITY CHECK (QQC) REPORT FOR SLR (LEXICA)

QQC WAS CARRIED OUT BY SPEX, NIJMEGEN, THE NETHERLANDS

TITLE DATABASE:

VERSION OF DATABASE:

TYPE OF DATABASE:

DATABASE OWNER / PRODUCER:

ELRA CATALOGUE NUMBER:

AUTHORS OF QQC REPORT: Henk van den Heuvel

DATE:

VERSION OF REPORT:

VERSION OF QQC TEMPLATE 2.2

SUMMARY SHEET:

GENERAL REMARKS:

Database part	ELRA Rating		
	*	**	***
1. Documentation			
2. Format			
3. Design & contents			
4. Transcriptions			

The QQC report contains a quality assessment of the resource with respect to a number of minimum formal requirements that are outlined in this report.

Meaning of the quality stars:

- * The minimal criteria for this part of the lexicon are not fulfilled.
- ** The minimal criteria for this part of the lexicon are reasonably well fulfilled.
- *** The minimal criteria for this part of the lexicon are all fulfilled.

MISSING: This part of the lexicon is missing, but relevant

NOT APPLICABLE: This part is not applicable to this lexicon

1 Quick Quality Check Report

1.1 Documentation

The most important topics should be covered and clearly described in the documentation:

- database layout and media
- application potential for the lexicon
- directory structure and file names
- origin of the entries
- format of the lexicon files
- transcription conventions for orthographic entries are described (spelling conventions, character set used, multiple word entries, treatment of abbreviations)
- transcription conventions for phonemic entries are described (phoneme set, phonological rules included, segmental information, supra-segmental information)
- morphological and syntactic information is described, if present (POS set used, plus attributes and values, multiple tagging possibilities)

- procedures for quality control are explained

1.2 Format

- The file names and directory structure should correspond to the documentation
 - The format of the lexicons should be in agreement with documentation
 - The format of other files in the database should be in agreement with documentation
 - Character set coding for orthographic entries should be commonly used (ISO, WINDOWS,...)
- The lexicon is in some well-known standard format

1.3 Design and contents

- All domains according to the documentation should be included

1.4 Transcription

- The character coding set for the orthographic entries as stated in the documentation should indeed be used in the lexicon
- All entries have a (phonemic or other) transcription
- The correct set of phone symbols should be used (according to

documentation). All phones documented are used, and all used phones are documented.

1.5 Other remarks

7.3 QQC_DB for SLR (speech synthesis)

!! To be made

8 Appendix C: QQC_DF templates

8.1 QQC_DF for SLR (ASR applications)

SLR QQC REPORT ON DESCRIPTION FORMS FOR SPEECH DATABASES
QQC WAS CARRIED OUT BY SPEX, NIJMEGEN, THE NETHERLANDS

TITLE DATABASE:

VERSION OF DATABASE:

TYPE OF DATABASE:

VERSION OF DESCRIPTION FORMS:

- General:
- Speech:

ELRA CATALOGUE NUMBER:

AUTHORS OF QQC REPORT: Henk van den Heuvel

DATE:

VERSION OF REPORT:

VERSION OF QQC TEMPLATE 2.2

SUMMARY SHEET:

GENERAL REMARKS:

Database part	ELRA Rating		
	*	**	***
1. General description form			
2. Specific description form			

The QQC report contains a quality assessment of the correctness of the information on the description forms. A star notation is used for this.

Meaning of the quality stars:

* : The information provided is insufficient/incorrect

** : The information provided needs some improvement or extension

*** : The information provided is complete and correct

Other values:

Not Applicable: This part is not applicable for this resource

Missing: This part is missing in the resource, but relevant

1 Quick Quality Check Report

1.1 General Description form

The most important topics should be covered and clearly described in the documentation:

- G1. General information
- G2. Producer/Provider
- G3. Prices
- G4. Availability
- G5. Additional information
- G6. Documentation
- G7. Validation
- G8. Distribution media
- G9. Sample of resource/demo
- G10. Short description of the resource

1.2 Special description form: Speech

- S1. General information

- S2. Speaker information

- S3. Lexicon

- S4. Linguistic information and segmentation

- S5. Technical information

- S6. Further comments

1.3 Other remarks

8.2 QQC_DF for SLR (phonetic lexicons)

SLR QQC REPORT ON DESCRIPTION FORMS FOR PHONETIC LEXICA
QQC WAS CARRIED OUT BY SPEX, NIJMEGEN, THE NETHERLANDS

TITLE DATABASE:

VERSION OF DATABASE:

TYPE OF DATABASE:

VERSION OF DESCRIPTION FORMS:

- General:
- Lexicon:

ELRA CATALOGUE NUMBER:

AUTHORS OF QQC REPORT: Henk van den Heuvel

DATE:

VERSION OF REPORT:

VERSION OF QQC TEMPLATE 2.2

SUMMARY SHEET:

GENERAL REMARKS:

Database part	ELRA Rating		
	*	**	***
1. General description form			
2. Specific description form			

The QQC report contains a quality assessment of the correctness of the information on the description forms. A star notation is used for this.

Meaning of the quality stars:

* : The information provided is insufficient/incorrect

** : The information provided needs some improvement or extension

*** : The information provided is complete and correct

Other values:

Not Applicable: This part is not applicable for this resource

Missing: This part is missing in the resource, but relevant

1 Quick Quality Check Report

1.1 General Description form

The most important topics should be covered and clearly described in the documentation:

- G1. General information
- G2. Producer/Provider
- G3. Prices
- G4. Availability
- G5. Additional information
- G6. Documentation
- G7. Validation
- G8. Distribution media
- G9. Sample of resource/demo
- G10. Short description of the resource

1.2 Special description form: Lexicon

- L1. General information

- L2. Description of the lexicon

- L3. Technical information

- L4. Additional information

- L5. Further comments

1.3 Other remarks

8.3 QQC_DF for SLR (speech synthesis)

Identical to the one presented in section 8.1.



SPEX / Dept. of Language and Speech
University of Nijmegen
Erasmusplein 1
NL-6525 HT Nijmegen
The Netherlands
e-mail: spex@spex.nl
Version validation criteria: 2.8

SUBJECT:	Validation Taal MDB SALA II corpus
AUTHORS:	Auteurs
VERSION:	RapportVersie
DATE:	Datum

INTRODUCTION

The speech databases made within the SALA II project were validated by SPEX, Nijmegen, the Netherlands, to assess their compliance with the SpeechDat format and content specifications, as documented in Technical Annex of the project.

The validation results of the Taal Mobile Network SALA II database (Aantal Sprekers speakers) are contained in this document.

This database was validated and ?? by the SALA II Consortium.

In the validation procedure we systematically check a list of validation criteria for a range of subjects. In the following sections we will evaluate these criteria one by one. Validation results that call for attention because of deviations from the SALA II specifications are marked by

⇒

so that they can be easily found.

The following subjects were validated:

1	DOCUMENTATION	3
2	DATABASE STRUCTURE, CONTENTS AND FILE NAMES	6
3	ITEMS.....	9
4	SAMPLED DATA FILES	16
5	ANNOTATION FILE.....	18
6	LEXICON	20
7	SPEAKERS	22
8	RECORDING CONDITIONS	25
9	TRANSCRIPTION	25

The document is concluded by

10	SUMMARY	27
----	---------------	----

1 DOCUMENTATION

- File DESIGN.DOC is present
??
- Language of doc file: English
??
- Contact person: name, address, affiliation
??
- Description of number of CDs and contents per CD
??

- The directory structure of the CDs
 - database, block and session orderings
 - directories DOC, INDEX, TABLE (and optionally PROMPT, SOURCE)
??

- The format of the speech files (A-law, Mu-law, 8 bit, 8 kHz, uncompressed)
??

- File nomenclature
 - root files
 - names of speech files and label files
 - files in directories DOC, INDEX, TABLE (and optionally PROMPT, SOURCE)
??

- Contents and format of the label files
 - clarification of attributes (three letter mnemonics)
 - example of labelfile
??

- Description of recording platform
??

- Explanation of speaker recruitment
??

- Prompting information
 - connection of sheet items to item numbers on CD
 - sheet example
 - items must be spread over the sheet to prevent list effects (e.g. three yes/no questions immediately after another are not allowed)
??

- Description of all recorded items
??

- Analysis of frequency of occurrence of the phones represented in the phonetically rich sentences and phon. rich words at transcription level (format: table) (!! New SALA II criterion)
??

- Analysis of frequency of occurrence of the phones in the full represented in the full database at transcription level (format: table) (!! New SALA II criterion)
??

- Transcription conventions
 - procedure
 - quality assurance
 - character set used for annotation (transcription) (ISO-8859)
 - annotations symbols for non-speech acoustic events must be mentioned Filled Pause, Speaker Noise, Stationary Noise, Intermittent Noise
 - list of symbols used to denote word truncations, mispronunciations, distortion due to the cellular network transmission, and not understandable speech
 - case sensitivity of transcriptions
 - use of punctuation
??

- Lexicon information
 - procedures to obtain phonemic forms from orthographic input (lexicon generation and lay out)
 - splitting of entries only at spaces
 - (Reference to) SAMPA symbols used
 - case sensitivity of entries (matching the transcriptions)
??

- Speaker demographics
 - which regions, how many of each
 - motivation for selection of regions
 - which age groups, how many of each
 - sexes: males, females, also children?; how many of each.
 - how many sessions by how many speakers
??

- Recording conditions:
 - description of recording environments
 - number of speakers per environment
??

- Information on test (set) specification
??

- The validation report made by SPEX (VALREP.DOC) is referred to
??

2 DATABASE STRUCTURE, CONTENTS AND FILE NAMES

- Directory/ subdirectory conventions
Format of directory tree should be
\`<database>\<block>\<session>`
 - database: defined as `<name><#><language code><name>` is MOBIL
`<#>` is 4 for SALA
`<language_code>` is the ISO two-letter code for the language
 - block: defined as `BLOCK<nn>` where `<nn>` is a progressive number from 00 to 99.
Block numbers are unique over all CD's.
They correspond to the first two digits of `<nnnn>` below.
 - session: defined as `SES<nnnn>` where `<nnnn>` is the session code also appearing in file name

??

- File naming conventions
All file names should obey the following pattern: `DDNNNNCC.LLF`
 - DD: database identification code
For SALA II: B4 = cellular net
 - NNNN: session code 0000 to 9999
 - CC: item code; first character is item type identifier,
second character is item number
 - LL: language code (as specified in Technical Annex)
 - F: speech file type
A is for A-law; U is for Mu-law;
O is for Orthographic label file

??

- NNNN in filenames is not in conflict with BLOCK and SES numbers in pathname
??

- Contents lowest level subdirectories should be of one call only
??

- All text files should be in MS-DOS format (`<CR><LF>`) at line ends
??

- A README.TXT file should be in the root describing all (documentation) files on the CD-ROM
??

- A file containing a shortened version of the volume name (11 chars max.) should be in the root directory. The name of this file is DISK.ID. This file supplies the volume label to UNIX systems that cannot read the physical volume label. Example of contents:
MOBIL4EV_01
??
- A copyright statement should be present in the file COPYRIGHT.TXT (root)
??
- Documentation should be in \<database_name>\DOC
 - DESIGN.DOC
 - TRANSCRIP.DOC (optional)
 - SPELLALT.DOC (optional)
 - SAMPALEX.PS
 - ISO8859<nr>.PS
 - SUMMARY.TXT
 - SAMPSTAT.TXT??
- The contents list (CONTENTS.LST) is in \<database_name>\INDEX
??
- Tables should be in \<database_name>\TABLE
 - SPEAKER.TBL
 - LEXICON.TBL
 - REC_COND.TBL (optional)
 - SESSION.TBL??
- Index files (optional) should be in \<database_name>\INDEX.
Mandatory are:
 - CONTENTS.LST
 - B4TST<language code>.SES??
- Prompt sheet files (optional) should be in \<database_name>\PROMPT
??
- All sessions indicated in the documentation SUMMARY.TXT are present on the CDs
??

- Empty (i.e. zero-length) files are not permitted
??

- File match: For each label file there must be one speech file and vice versa
??

- Part of the corpus is designed for training and a smaller part for testing:
 - For 1000 databases of 1000 sessions 200 test sessions are required, for databases with more than 2000 sessions 500 test sessions should be defined.

 - No overlap between train and test sessions is allowed.
??

- All table files, and index files should report the field names as the first row in the files using tabs as in the data records following.
??

- The contents of the database as given in CONTENTS.LST should comprise:
 - CD-ROM volume name (VOL:)
 - full pathname (DIR:)
 - speech file name (SRC:)
 - corpus code (CCD:)
 - speaker code (SCD:)
 - speaker sex (SEX:)
 - speaker age (AGE:)
 - speaker accent (ACC:)
 - orthographic transcription of uttered item (LBO:)
 - The first line should be a header specifying the information in each record.
 - This file must be supplied as an ASCII TAB delimited file.??

- The contents of the SUMMARY.TXT files should comprise:
 - The full directory name where speech and label files are to be found (DIR:)
 - the session number (SES:)
 - a string of typically N codes. Each item present is represented by its code. If the item is missing, a '--' should appear.
 - recording date (RED:)
 - recording time of first item (RET:)
 - optional comment text
 - all these fields are separated by spacesNote: The contents of the SUMMARY.TXT file are not CD-dependent.
??

- Missing items per session
Check with documentation (SUMMARY.TXT)
??

- The database should be free of viruses
??

3 ITEMS

A. Check on mandatory corpus items

- 6 common application words (code A1-6)
 - read
 - set of 25-30 should be used, 25 of which are fixed for all
 - minimum number of examples of each word = $\#sessions^1 / 8$ (at transcription level) (!! New SALA II criterion)
 ??

- 2 isolated digits (code I1-2)
 - read or prompted
 ??

- 1 sequence of 10 isolated digits (code B1)
 - each sequence must include all digits
 - optional are hash and star
 ??

- 4 connected digits (code C1-4)
 - 5+ digit number to identify the prompt sheet (optional) (C1)
 - read
 ??

- 9-11 digit telephone number (C2)
 - read
 - local numbers
 - inclusion of at least 50% cellular telephone numbers mandatory (!! New SALA II criterion)
 ??

- 16 digit credit card number (C3)
 - read
 - set of 150
 - if there is a checksum then formula must be provided
 ??

¹ #sessions refers to the agreed sizes of the databases. It can have only two values: 1000 or 4000. This also holds if more recordings than 1000 or 4000 are included in the database.

- 6 digit PIN code (C4)
 - read
 - set of 150
 ??

- ~30 digits per session are required
 - read
 - set of 150
 ??

- digits must appear numerically on the sheet, not as words
 - read
 - set of 150
 ??

- 1 date (code D1)
 - spontaneous
 ??

- 1 date (code D2)
 - read, wordstyle
 - analogue form
 - covering all weekdays and months, ordinals and year expressions (also exceeding 2000)
 ??

- 1 general or relative date (code D3)
 - read
 - analogue
 - should include forms such as TODAY, TOMORROW, THE DAY AFTER TOMORROW, THE NEXT DAY, THE DAY AFTER THAT, NEXT WEEK, GOOD FRIDAY, EASTER MONDAY, etc.
 ??

- 1 application word phrase (code E1)
 - application word is embedded in phrase
 - read or spontaneous
 - At least 5 different phrases are required for each application word
 - a length of minimal 3 words per sentence is required (!! New SALA II criterion)
 ??

- 3 spelled words (code L1-3)
 - L1 is spontaneous name spelling linked to O1 (or to another item explicitly documented)
 - others are read
 - equal balance of all vocabulary letters
artificial words can be used to enforce this balance
 - average length at least 7 letters
 - may include names, cities and other frequently spelled items
 - should primarily include equivalents of:
A-Z, accent words, DOUBLE, APOSTROPHE, HYPHEN

??

- 1 money amounts (code M1)
 - read
 - currency words should be included
 - mixture of small amounts including decimals
and large amounts not including decimals

??

- 1 natural number (code N1)
 - read
 - provided as numbers (numerically)
 - decimal numbers are only allowed for additional natural numbers
 - numbers should all be smaller than 1,000,000

??

- 6 directory assistance names (code O1-7)
 - 1 spontaneous name (e.g. forename) (O1)
 - 1 spontaneous city name (O2)
 - 1 read city name (list of at least 500 most frequent) (O3)
 - 1 read company/ agency name (list of at least 500 most frequent) (O5)
 - 1 read proper name, fore- and surname (O7)
(list of 150 names: both male and female names)

??

- 2 yes/ no questions (code Q1-2)
 - spontaneous, not prompted
 - one question should elicit (predominantly) 'no' answers;
the other (predominantly) 'yes' answers
 - also fuzzy answers should be envisaged

??

- 9 phonetically rich sentences (code S1-9)
 - read
 - minimum number of phone examples = #sessions/10
 - at transcription level (!! New SALA II criterion)
 - Exception: rare phonemes:
 - these appear mainly in loan words AND
 - a max. of 10% of all phonemes in the language may be rare
 - each sentence may appear a max. of 10 times at prompt level (!!New SALA II criterion)

??

- 1 time of day (code T1)
 - spontaneous

??

- 1 time phrase (code T2)
 - read
 - analogue form
 - equal balance of all words
 - should include equivalents of:
AM/ PM, HALF/ QUARTER, PAST/ TO, NOON, MIDNIGHT, MORNING,
AFTERNOON, EVENING, NIGHT, TODAY, YESTERDAY, TOMORROW

??

- 4 phonetically rich words (code W1-4)
 - read
 - minimum number of phone examples = #sessions/10
 - at transcription level (!! New SALA II criterion)
 - Exception: rare phonemes:
 - these appear mainly in loan words AND
 - a max. of 10% of all phonemes in the language may be rare
 - each word may appear a max. of 5 times at prompt level (!! New SALA II criterion)

??

- Any additional, optional material:
??

B. Checks on presence of corpus files

The following completeness checks are performed :

1. Structurally missing corpus items

- Which items are not recorded at all?

??

2. Incidentally missing files

a. files that are not there

We found ?? missing files, according to the following distribution over the corpus items:
??

b. files with empty transcriptions in the LBO label field (effectively missing files)

We found ?? files with empty transcriptions (only noise symbols and/ or **). If we merge these files with the missing files (being ??) given above then we get the following distribution:

??

c. corrupted speech files

If we regard utterances which have only truncated or mispronounced words as corrupted files, and merge these with the effectively missing files under b, then the following distribution emerges:

??

d. files containing truncation and mispronunciation marks

We found ?? transcriptions with at least one *, or %, or **, or ~, according to the following distribution over the items:

??

(* , % , ** , ~ are counted in the transcriptions of the individual items to get an idea of distorted speech data. This will not be used to reject or approve a database but it will be supplied as supplementary information.)

3. Overall conclusion

SALA II has the following criteria for missing items:

- A maximum of 5% of the files of each mandatory item (corpus code) may be effectively missing.
- As missing files are counted: absent files, and files containing non-speech events only.
- For the phon. rich sentences a maximum of 10% of the files may be effectively missing or corrupted
- There will be no further comparison of prompt and transcription text in order to decide if a file is effectively missing.
As a consequence: If there is some speech in the transcription, then the file will NOT be considered missing, even if it is in fact useless.

??

4 SAMPLED DATA FILES

1. Coding

- A-law or Mu-law, 8 bit, 8 kHz, no compression
??

2. Sample distribution

Several sample statistics are generated: File length, clipping rate, mean sample value, Signal-to-Noise Ratio (SNR). Statistics were generated on file level by the producer of the database, using SPEX software. The results were delivered to SPEX. SPEX compiled histograms on the basis these results. These histograms are presented below, both on file level and on directory (call) level. The histograms are presented as they are and not further interpreted by SPEX. On the basis of these data the user of the database should be able to decide which acoustic quality is still acceptable for the application at hand. Statistics on the acoustics of individual speech files can be retrieved from file \DOC\SAMPSTAT.TXT.

The SAMPSTAT.TXT file contains on each line 12 fields separated by a colon. The values of the fields are:

- name of the speech file
- number of the channel (here 0 always)
- cut/nocut value whether dtmf signal was cut out of the signal
- maximum sample value
- minimum sample value
- total number of samples
- global clip rate (computed with max. possible sample value)
- local clip rate (computed with max. occurring sample value)
- mean sample value
- snr quick
- snr bins
- noise percentage

2.1 File length

We calculated the length of the files in seconds in order to trace spurious recordings if files were of extraordinary length.

Duration distribution over calls/ directories:

Length (s) #Occurrences
??

2.2 min-max samples

We provide a histogram with clipping ratios. The clipping ratio is defined as the proportion of samples in a file that is equal to the maximum/ minimum value, divided by all samples in the file.

The histogram, then, is an overview of how many files were found in a set of clipping rate intervals.

Clip distribution over calls/ directories:

Clipping rate (in %)	Occurrences
??	

2.3 Mean values

We computed the mean sample value of each item in each call. We provide a histogram with mean values below. The histogram, then, is an overview of how many files were found in a set of mean sample value intervals. This overview can be used to trace files with large DC-offsets.

Mean distribution over calls/ directories:

Mean	Occurrences
??	

2.4 Signal to Noise Ratio

We split each signal file into contiguous windows of 10 ms and computed the Mean Square (energy) in each window. The mean sample value over the complete file was subtracted from each individual sample value before MS was computed. 30% of the windows that contained the lowest energy were assumed to contain line noise. In this way the signal to noise ratio could be calculated for each file by dividing the mean energy over all windows by the mean energy of the 30% sample mentioned above. The result was multiplied by $10 \cdot \log$ for scaling.

SNR distribution over calls/ directories:

SNR	occurrences
??	

5 ANNOTATION FILE

- Each line must be delimited by <CR><LF>
??

- Mandatory (SAM) mnemonics:
 - LHD: SAM, 6.00
 - DBN: SALA_II_<country>_<language>_Mobile_Network
 - VOL: MOBIL4<LL>_<nr>
 - SES: <session number>
 - DIR: <with backslashes and no final backslash>
 - SRC: <filename of speech file>
 - CCD: <corpus code = item code>
 - REP: <location of recording equipment>
 - RED: <recording date, in format DD/Mmm/YYYY>
 - RET: <recording time, in format HH:MM:SS>
 - BEG: <begin sample, usually 0>
 - END: <end sample>
 - SAM: 8000 < = sampling freq.>
 - SNB: 1 < = number of bytes per sample>
 - SBF: < = sample byte order, meaningless with single bytes>
 - SSB: 8 < = number of significant bits per sample>
 - QNT: A-LAW | MU-LAW < = quantisation>
 - SCD: <speaker code>
 - SEX: M/ F/ UNKNOWN
 - AGE: <in years/ unknown>
 - ACC: <regional accent, place of growing up>
 - REG: <region of call>
 - ENV: <environment of call>
{HOME_OFFICE|PUBLIC_PLACE|STREET|VEHICLE|CAR_KIT}
 - PHM: <telephone model> {CELLULAR|CELLULAR/HF|CELLULAR/HH} (!!New SALA II criterion !!)
 - NET: {TACS| AMPS| GSM| TDMA|CDMA| DECT|OTHER}
 - LBD:
 - LBR: <start>, <end>, [gain], [minimum value], [maximum value],
<orthographic prompt>
 - LBO: <start sample>, [centre sample], <end sample>, <transliteration>
 - ELF: <end label file>

- Optional (SAM) mnemonics (may be omitted or left empty)
 - TYP: orthographic
 - TXF: <name of the prompt sheet text file>
 - CMT: <comment>
 - NCH: 1 <= number of channels recorded>
 - ARC: <region or area code of call>
 - SHT: <sheet number for prompts>
 - CMP: <compression, should be empty if used>
 - EXP: <labelling expert>
 - SYS: <labelling system>
 - DAT: <date of completion of labelling>
 - SPA: <SAMPA version>
 - DSC: <= discontinuity marker>
 - EDU: <education level>
 - SOC: <Socio Economic Status>
 - HLT: <health>
 - TRD: <tiredness>
 - RCC: <recording conditions code>
 - CRP: <= corpus repetition, empty>
 - ASS: <assessment code>

- Order restrictions:
 - LHD and TYP are first
 - LBR and LBO come after LBD
 - ELF is end of file keyword
 - ??

- No illegal mnemonics used
 - ??

- There are no mnemonics missing
 - ??

- All files must contain the same mnemonics. This holds as well for the optional mnemonics.
 - ??

- No illegal field values should appear
 - ??

- Each lowest subdirectory does not refer to multiple sheet ids.
 - ??

- For spontaneous speech LBR should contain a mnemonic word.
 - D1 : <date>
 - L1 : <forename_spelled>
 - O1 : <forename>
 - O2 : <city>
 - Q1 : <yes_question> or <no_question>
 - Q2 : <yes_question> or <no_question>
 - T1 : <time>
 - ??

- Assessment of speech items in terms of SNR, presence of additional noise, adherence to prompting text is provided (optional)
??

6 LEXICON

- Check lexicon existence (LEXICON.TBL)
??

- The entries should be alphabetically ordered
??

- Used SAMPA symbols are provided in SAMPALLEX.PS
??

- In transcriptions only SAMPA symbols are allowed
??

- All SAMPA phoneme symbols should be covered
??

- Phoneme symbols must be separated by blanks
??

- A line in the lexicon should have the following format
<grapheme form> <TAB> [<frequency> <TAB>] <phoneme transcription> [<altern.>]
[TAB] is ASCII 9.
??

- Each line is delimited by <CR><LF>
??

- All entries should have at least one phone transcription
??

- Alternative transcriptions are optional.
They may follow the first transcription, separated by [TAB] or have a separate entry
(only in case also frequency information is supplied)
??

- Orthographic entries are taken from the LBO-transcriptions from the label files. These
LBO-transcriptions are as a rule split by spaces only, not by apostrophes, and not by
hyphens.
??

- Words appearing only with * or ~ or % should not appear in the lexicon
??

- The lexicon should be complete
 - Check for undercompleteness (are all words in lexicon)

 - Check for overcompleteness
(Undercompleteness is worse than overcompleteness. Overcompleteness cannot be a reason for rejection)
??

- Lexicon contents should be taken from actual utterances (from LBO), so the entries should exactly match the transcriptions.
??

- Optional information: stress, word / morphological / syllabic boundaries.
But, if provided, then it should follow the SpeechDat conventions.
??

7 SPEAKERS

- Obligatory information in SPEAKER.TBL:
 - unique number (speaker/ caller) SCD
 - sex SEX
 - age AGE
 - accent ACC
 - ??

- Optional information:
 - height HET
 - weight WET
 - native language NLN
 - ethnic group ETH
 - education level EDL
 - smoking habits SMK
 - pathologies PTH
 - socio-economic status SOC
 - health HLT
 - tiredness TRD
 - ??

- Each speaker only calls once. There is a tolerance of 5% of the speakers who may call twice. (!! New SALA II criterion !!)
 - ??

- Balance of sexes
 - How many males, how many females, should match specification in documentation file
 - Misbalance may not exceed 5% (Each sex must be represented between 45-55% of the sessions)
 - ??

- Balance of dialect regions
 - which dialect regions and how many of each should match specification in documentation file
 - ACC is used to check dialect balance, according to motivation in DESIGN.DOC
 - At least #sessions/20 speakers per dialect should be included (!! New SALA II criterion)
 - For the US English database the minimum number of speakers per dialect is 223 and the maximum is 670.
 - ??

- Balance of ages
 - which age groups and how many of each should match specification in documentation file
 - Criteria
 - < 16 : \geq 1% of speakers strongly recommended
 - 16-30 : \geq 20% of speakers mandatory
 - 31-45 : \geq 20% of speakers mandatory
 - 46-60 : \geq 15% of speakers mandatory
- (The age criteria are meant for the whole database; they are not to be applied for male and female speakers separately)
- ??

8 RECORDING CONDITIONS

- Obligatory attributes of the (optional) REC_COND.TBL file:
 - recording conditions code RCC
 - region of call REG
 - environment ENV
 - ??

- Obligatory attributes of the SESSION.TBL
 - Session code SES
 - Recording date RED
 - Recording time (of first item) RET
 - Speaker code SCD
 - Speaker age AGE
 - Speaker sex SEX
 - Speaker dialect region ACC
 - Calling region REG
 - Calling environment ENV
 - Phone model PHM
 - Telephone network (if included) NET
 - ??

- The recordings are distributed as follows (check ENV) (!! New SALA II criterion):

Environment	Full database distribution	Each dialect region distribution
1. Car, train, bus	20 % \pm 5%	\geq 20%
2. Public place	25 % \pm 5%	
3. Street	25 % \pm 5%	
4. Home/Office	25 % \pm 5%	\geq 20%
5. Car kit (hand free mode)	5 % \pm 1%	No restriction

- In each dialect at least 20% of the speakers are recorded in environments 1-3
- In each dialect at least 20% of the speakers are recorded in the home/office environment
- ??

- Recordings from the fixed net are not included
- ??

9 TRANSCRIPTION

A. Validation by software tools

- Transliterations is case-sensitive unless specified otherwise.
(In general lower case is used also at sentence beginning Only exception: proper names and spelled words, ZIP codes, acronyms and abbreviations.
In the latter case blanks should be used in between the letters.)
??
- Punctuation marks should not be used in the transliterations
??
- Digits must appear in full orthographic form
??
- In principle only the following symbols are allowed to indicate non-speech acoustic events:
[fil] [spk] [sta] [int] [dit]
Other symbols (and language equivalents) must be mentioned in the documentation
??
- Asterisks should be used to indicate mispronunciations
??
- Double asterisks should be used for not understandable parts
??
- Tildes should be used to indicate truncations
??
- Percent signs should be used to indicate speech distortions due to transmission characteristics of the cellular network
??

B. Validation by a native speaker of the language

This validation was carried out by taking 1000 short items and 1000 long items. The transcriptions in the label files for these samples were checked by listening to the corresponding speech files and correcting the transcription if necessary. In case of doubt nothing was corrected.

This check was performed by a native speaker of the language. The background noise markers were checked by a trained (non-native) validator.

Short items are:

- isolated digit
- time phrases
- date phrases
- yes/no questions
- names
- application words
- phonetically rich words

Long items are:

- isolated digit string
- connected digits
- natural numbers
- money amounts
- spelled words
- application phrases
- phonetically rich sentences

- - The evaluation comprised the following guidelines:
 - Two types of errors were distinguished: speech and non-speech transcription errors
 - Non-speech refers to [fil] [spk] [sta] [int] only
 - For non-speech all symbols were mapped to one during validation. i.e. If a non-speech symbol was at the proper location then it was validated as correct (regardless if it was the correct non-speech symbol or not). The only exception is [sta] which should be properly marked in the transcriptions.
 - Only noise deletions in the transcription were counted as wrong, not noise insertions.
 - The given transcription is given the benefit of the doubt; only obvious errors are corrected.
 - Errors were only determined on item level, not on word level
 - For speech a maximum of 5% of the validated items (=files) may contain a transcription error
 - For non-speech a maximum of 20% of the validated items (=files) may contain a transcription error.

C. Results

1. Long items

Transcription errors with respect to speech were found in ?? items. This amounts to ??%, which is below the criterion of 5%.

Errors in the transcription of non-speech were found in ?? items. This amounts to ??% of the items, which is below the criterion of 20%.

2. Short items

Errors with respect to the transcription of speech were found in ?? items. This amounts to ??%, which is below the criterion of 5%.

Errors in the transcription of non-speech were found in ?? items. This amounts to ??% of the items, which is below the criterion of 20%.

3. Overall result

When we take the long and short item sets together, we find errors with respect to the transcription of speech in ?? items. This amounts to ??%, which is below the 5% criterion. Errors in the transcription of non-speech were found in ?? items. This amounts to ??%, which is below the 20% criterion.

4. Further remarks

??

10 SUMMARY

The Taal database was validated and ?? accepted by the SALA II consortium.

Below we give a brief overview of our findings for this database. The subsections follow the order of the various topics in the previous sections of the report.

1. Documentation

??

2. Database structure and file names

??

3. Items

??

4. Sampled data files

The speech data files are in the correct format (A-law, Mu-law). A file with acoustic characteristics of each file is delivered (SAMPSTAT.TXT). Histograms of a number of acoustic characteristics of the files (duration, mean sample value, clipping rate, SNR) were generated and included in section 4 of this report. Acoustical details of individual files can be looked up in the SAMPSTAT.TXT file.

??

5. Label files

??

6. Lexicon

??

7. Speakers

??

8. Recording conditions

??

9. Transcription

??

A Bug Report Service for ELRA

<i>Project reference number</i>	ELRA/0201/VAL-1
<i>Project acronym</i>	VCom
<i>Project full title</i>	ELRA SLR Validation
<i>Project contact point</i>	Harald Höge
<i>Project web site</i>	
<i>EC project officer</i>	
<i>Document title</i>	A Bug Report Service for ELRA
<i>Deliverable ID</i>	D2.3
<i>Document type</i>	Report
<i>Dissemination level</i>	ELRA Vcom/ELRA Board
<i>Contractual date of delivery</i>	
<i>Actual date of delivery</i>	11 December 2002
<i>Status & version</i>	V1.8
<i>Work package & task ID</i>	
<i>Work package, task & deliverable responsible</i>	
<i>Number of pages</i>	11
<i>Author(s) & affiliation(s)</i>	Henk van den Heuvel, SPEX
<i>Additional contributor(s)</i>	
<i>Keywords</i>	Quality Control, Validation, SLR
<i>Abstract</i>	
<i>Additional notes & remarks</i>	

Document evolution

Version	Date	Status	Notes
1.0	18 May 2001	First draft	
1.1	28 Aug. 2001	Draft	Update after meeting on 2 nd of July
1.2	21 Sep. 2001	Draft	Specification of patch files added
1.3	17 Oct. 2001	Draft	Version before meeting 30 th of Oct.
1.4	1 Nov. 2001	Intermediary	
1.5	16 Nov. 2001	Pre-final	Outcomes of meeting 30 th Oct. included;
1.6	10 April 2002		Outcomes of meeting 22 nd Jan. 2002 in Pisa included
1.7	4 October 2002	Updated	Outcomes of Paris meeting (Apr. 2002) and later events included
1.8	16 October 2002	Final	Outcomes Copenhagen (9 Oct. 2002) included)

ELRA Contact :

ELRA - Distribution Agency (ELDA)

Dr. Khalid CHOUKRI

CEO

55-57, rue Brillat Savarin

F-75013, PARIS, FRANCE

Tel. +33 1 43 13 33 33

Fax. +33 1 43 13 33 30

Email: choukri@elda.fr

Table of contents

<u>1.</u>	<u>INTRODUCTION</u>	4
<u>2.</u>	<u>A framework for a bug report service</u>	5
<u>2.1</u>	<u>Type of errors</u>	5
<u>2.2</u>	<u>Appropriate actions to bug reports</u>	5
<u>2.3</u>	<u>Ownership issues</u>	7
<u>3.</u>	<u>Implementation</u>	7
<u>3.1</u>	<u>Bug reports</u>	7
<u>3.2</u>	<u>Formal error lists</u>	8
<u>3.3</u>	<u>Archiving</u>	9
<u>3.4</u>	<u>Rewards for bug reporters</u>	9
<u>4.</u>	<u>REFERENCES</u>	9
<u>5.</u>	<u>Appendix: Screen Shot of Bug Report Sheet</u>	10

1. INTRODUCTION

A glance at the catalogues of database distribution agencies such as ELRA (the European Language Resources Association) and LDC (the Linguistic Data Consortium) shows that language resources (LRs) in general and spoken language resources (SLRs) in particular have grown rapidly in number and in size over the last ten years. Such developments pose a growing demand on LR maintenance, quality control and improvement.

Nowadays, a quality check (also termed 'validation') is integrated in the production of many European SLRs. Validation entails that, during production and immediately after completion, the SLRs created in a project are checked against a set of criteria based on the original specifications and accompanying tolerance margins; a SLR can only be released if it passes the validation. Typical examples of such validated SLRs are the databases in the SpeechDat family.

However, this type of validation can only be one slice in the cake of a comprehensive LR quality control procedure. Firstly, many existing SLRs were produced in a project that did not have a validation component. Secondly, bugs may also be found when a validated LR is actually used, e.g. if a SLR is used for training an automatic speech recognizer. An adequate way of reporting the bugs gives way to a wealth of possible LR improvements that otherwise remain unaccomplished.

It is fairly easy to introduce a bug report service at ELRA's web pages. It will, however, run counterproductive if it is not embedded in a decent framework of bug administration, communication with the reporter, error listing, database validation and correction, and issuing new LR releases. In other words, appropriate actions on short notice should be taken when a bug report comes in. On the other hand, immediate correction of any error reported, whatever its seriousness, puts too strong strains firstly on ELRA's possibilities to rectify errors, and secondly on LR users to cope with a horrendous diversity of version numbers (for example if they would like to compare recognition results obtained on a speech database with other labs). A satisfactory midway has to be found.

In this report we propose:

1. a framework of error handling by means of a bug report service for ELRA
2. a possible implementation of such a bug report service.

2. A framework for a bug report service

2.1 *Type of errors*

A proper framework for a bug report service should provide a list of satisfactory actions (both to ELRA and the customer) to various types of errors. This framework is devised for Spoken LR (SLR), but can be tailored to other types of LR where appropriate.

The first distinction to be made is that between small and severe errors that are reported. Severe errors refer to substantial deficiencies in elementary properties of the database [1]:

1. the quality of the speech files
2. the quality of (orthographic) transcriptions
3. the lexicon (with phone transcriptions)

(Relatively) Small errors typically refer to errors in:

4. file names and directories
5. annotation/label files
6. metadata (e.g. speaker table)

We note that “small” errors may be treated as severe if they show up in huge quantities. Conversely, a “severe” error may be treated as small if there is only very few of them.

2.2 *Appropriate actions to bug reports*

In principle, only errors in text files will be repaired. Speech files will not be touched. The following procedure for the processing of bug reports will be used:

1. Bug reports are sent to SPEX via the public validation page of SPEX; SPEX acknowledges the receipt of the report.
2. The bug report is verified by SPEX and, if accepted by SPEX, added to the formal error list (FEL) maintained by SPEX (for each SLR a separate FEL exists). The updated list is sent to ELDA. The FEL is sent to the provider for feedback (action ELDA).
3. The formal error list is linked to each SLR in the catalogue (the list may be empty) provided the provider of the SLR allows to do so (action ELDA).
4. The access to the FEL is free of charge and allows bug reporting users to see the status of the bugs of an SLR.
5. Based on an update of the FEL the provider of that SLR is asked by ELDA to correct that part of the SLR which was reported to be faulty. ELDA sends the corrected part to SPEX.
6. If the provider refuses to correct the incorrect files, ELDA or other institutions selected by ELDA produce the corrected part.

7. ELDA sends the corrected part to SPEX. SPEX produces a patch from the corrected part. This patch produces a new version of the SLR from the old version. The version of the patch and the version of the SLR have to be consistent. SPEX checks that the patch integrates properly the corrected part of the SLR into the latest version of the SLR. SPEX sends the patch to ELDA. ELDA puts the patch into the catalogue.
8. If the provider of the SLR agrees, ELDA produces a new version of the SLR with this patch. This new version of the SLR is put in the catalogue.
9. The patches can be ordered through ELDA. The corresponding information (cost, version) has to be integrated into the catalogue.

The verification of a reported error will be performed by SPEX if the error is not language-specific. If the error is language-specific (e.g. errors in the orthographical transcriptions), then SPEX will consult a qualified institution in its validation network to check the errors. Such an external check will typically be done if a series of such language-specific errors are collected for the SLR (and not when just one error is reported). ELRA will pay a reasonable remuneration to the external validator if so required. ELDA will encourage the provider of a SLR to comment to the FEL of the SLR.

ELDA inquires if providers agree to the publication of the FELs on ELRA's website.

As long as there is only a relatively small number of errors reported and verified, the users should consult the formal error list of an SLR and use this information as pleases him.

The formal error lists made and maintained by SPEX.

When severe (or many small) errors are reported, then rectification of the erroneous files becomes necessary.

If severe errors are found in more than one elementary property (see section 2.1) of a LR, then a full validation of the database can be considered. If a (partly or full) validation is deemed necessary by ELRA's SLR VCom, SPEX will include the database in its general validation queue. This makes it difficult to predict when the validation will take place, but it can be monitored via the publicly accessible validation status tables that SPEX maintains.

Obviously, SPEX should not carry out any rectifications of SLR, since a conflict of interests emerges when the corrections need to be validated. In essence, this implies that correction and validation should be iterated until a satisfactory result is achieved.

The rectification of erroneous files is therefore coordinated by ELDA. Minor changes can be performed by ELDA itself. If major changes are needed, ELDA will contact one of its (language-specific) production institutes to fix the files. The provider of the SLR will be asked first. Alternatively, if customers (e.g. the reporting one) made the necessary corrections already, then these could be purchased by ELDA (and validated by SPEX). The reporting customer could also be subcontracted to carry out the work. Once corrected the files are sent to SPEX. SPEX compares the updated files with the formal bug report, and makes a corresponding patch file.

The following notes are in place with respect to the correction patch for an SLR:

- The patch is a tar file containing all the files that need to be replaced.
- The patch adds/substitutes text files; it leaves the signal files unchanged;
- If several patches are made for a specific version of an SLR; then they are made additive, not cumulative;
- The patch is owned by ELRA and maintained by SPEX;
- The patch files may only be used for internal use by the receivers and not be distributed further;
- A patch is associated only with a specific version of the SLR, not with another version. It should not be supplied with any other version than for which it was made.

If validation shows that the errors observed render the database below minimum quality standards (see section 3 of [1]), then this information is added to the error list of the database as well. ELRA should in that case decide what to do with the SLR until the errors are corrected.

If the time between bug reporting and appropriate action is short, then this will encourage SLR users to use the service and feel positive about it. Error verification time will be short, presumably about two weeks; however a validation may take longer depending on the length of the general validation queue at SPEX. The progress can then be monitored via the publicly accessible validation status tables that SPEX maintains.

2.3 Ownership issues

In principle the reporter of the bug is the owner of this information. Therefore, it has to be made explicit in all information about the bug report service that the bug reporter transfers all (non-exclusive) exploitation rights on this information to ELRA.

The original SLR and the patches remain strictly separated.

- The SLR is owned by the provider; the patch is owned by ELRA
- When the patch is run by a user, the original version can be restored by just copying the original CDs

3. Implementation

3.1 Bug reports

1. The bug report sheet is a slot-based html-page, with slots for the following information (slots):

- Database name
- Code in ELRA's catalogue
- Coordinates (name, affiliation, e-mail address) of the reporter
- Errors to report
- Desired prize

The bug report sheet will have a section in which the bug handling procedure is described. The bug report sheet will also make explicit that the bug reporter transfers all rights on the reported information to ELRA (on a non-exclusive basis).

The bug report page also contains a brief explanation of the procedure for bug report handling as presented in section 2.2. See the appendix for a screen shot.

2. After completion the bug report sheet is (automatically) sent to

- the validation centre (SPEX)
- ELDA staff

The reporter of the errors should be stimulated to be as precise as possible in his bug reports; s/he should report file names, errors, and suggested corrections.

SPEX will make the html page and maintain it at its own validation portal. A link to the page will be established from ELRA's pages (validation page and the SLR catalogue pages).

The bug report service will be tested with SpeechDat partners first before it is made available for a wider public. In the testing phase it will reside at SPEX's validation portal. After the test it will be installed in the formal ELRA webpages.

3.2 Formal error lists

After verification of a reported error, SPEX will update the formal error list for an SLR and send notification to ELDA. Formal error lists for all SLR in ELRA's catalogue will be maintained by SPEX. ELDA will install an FTP server which can be accessed by SPEX (via password). SPEX can put updates of FELs at this location.

There will be a standard FEL for each SLR with the text: "No imperfections reported so far." In case errors are found, the text is changed into: "No database is perfect. The quality of this resource can be further improved if the following modifications are carried out.", followed by the verified error list.

3.3 Archiving

Each formal error list and each patch should be administered as belonging to a specific version of an SLR. This is reflected in the file name of a formal error list and of a patch file. They are not valid for any other versions. Especially, if the provider of the SLR releases a new version, this becomes relevant. SPEX can be given instructions to update the formal error list for the new version, and ELDA can make a new patch file based on SPEX's findings.

3.4 Rewards for bug reporters

To stimulate the submission of bug reports, two prizes (PDA's in the range of 600 - 800 Euro's) will be given once a year. One goes to the best contributor, i.e. the person who reports the most, serious, true bugs in a clear manner. The other goes to one of the other contributors by means of a random draw. They will be presented to the winners at the major conferences, e.g. LREC.

SPEX proposes the best bug reporter to the Vcom. The Vcom decides.

ELDA contacts the winner for the details of the desired prize; ELDA purchases the prize and takes care of correct delivery to the winner.

4. REFERENCES

- [1] Henk van den Heuvel (2000) *Procedures to Validate and Improve Existing SLR*.
ELRA/9901/VAL-1 Deliverable 2.1

Appendix: Screen Shot of Bug Report Sheet



ELRA's SLR Catalogue: Bug Reporting

<input type="text"/>	Referen	<input type="text"/> (opti
<input type="text"/>	Yc	<input type="text"/>
PDA		<input type="text"/>

Bug description:
(be as precise as possible; report per file name: found errors and suggested corrections. [click here for some examples.](#))

NOTE: By submitting this report you transfer all exploitation rights to ELRA (on a non-exclusive basis)

Examples:

- File B10003S1.ITO should have following orthographic transcription: 'e pericoloso sporgersi [spk]'
- SPEAKER.TBL has wrong speaker gender codes for 005, 066, 888
- File B10003S1.ITO contains illegal characters at file end; so do files B10003T1.ITO, B10103T1.ITO and all files in BLOCK05
- README.TXT is completely wrong; from another database?

- LEXICON.TBL uses SAMPA symbol A: everywhere, whereas o: is correct
- I have a list of wrong transcriptions and other errors per annotation file for this database. Please contact me to obtain this (big) file

In case of questions contact [Henk van den Heuvel](#) at SPEX.

SLR Validation: Current Trends and Developments

Henk van den Heuvel, Dorota Iskra, Eric Sanders, Folkert de Vriend

SPEX (Speech Processing Expertise Centre), Department of Language and Speech, Nijmegen, the Netherlands

e-mail: {henk,dorota,eric,folkert}@spex.nl

Abstract

This paper deals with the quality evaluation (validation) of Spoken Language Resources (SLR). The current situation in terms of relevant validation criteria and procedures is briefly presented. Next, a number of validation issues related to new data formats (XML-based annotations, UTF-16 encoding) are discussed. Further, new validation cycles that were introduced in a series of new projects like SpeeCon and OrienTel are addressed: prompt sheet validation, lexicon validation and pre-release validation. Finally, SPEX's current and future activities as ELRA's validation centre for SLR are outlined.

1. Introduction

Validation, as we will use the term here, refers to the quality evaluation of a database against a checklist of relevant criteria (Van den Heuvel et al., 2003; Fersøe, 2003; Schiel and Draxler, 2003). For the validation of language resources (LR) in general, and spoken language resources (SLR) in particular the relevant criteria are dependent on the application domain targeted with the SLR at hand and the setting in which the criteria are developed. Basically, two settings should be distinguished. The first setting is the situation in which SLR are completed in a framework where, validation by an external (i.e. non-producing partner) is an integral part of the SLR production process and the validation centre is involved from the beginning of the specifications of the databases. Therefore, the validation criteria are closely linked to the specifications. Examples of such validation scenarios are SpeechDat (II) (Van den Heuvel, 1996), SpeechDat-Car (Van den Heuvel, 1999), SALA (Van den Heuvel, 1997), SpeeCon (Van den Heuvel et al., 2001), OrienTel (Iskra et al., 2002), and more recently LC-STAR (Shammas and Van den Heuvel, 2003).

The other setting is that in which validation is not an integral part of the SLR production and should be done post-hoc. The European Language Resources Association (ELRA) faces this situation for part of the LR in its catalogue. ELRA regards quality assessment as an important element for the LR that it distributes. For this reason, ELRA is developing a set of minimum requirements which the various types of resources in its catalogue should fulfill. Obviously, these minimum requirements do not simply coincide with the specifications of the database proper (Van den Heuvel et al., 2003).

In this paper relevant issues as experienced by SPEX in both validation settings are presented. We start with an overview of the current situation and the new challenges encountered and then deal with new developments in more detail.

2. Current Situation and New Challenges

For a SLR the validation criteria typically comprise the following elements:

1. Documentation. It is checked if all relevant aspects of an SLR (see 2-8 below) are properly described in

terms of the three C's: clarity, completeness and correctness.

2. Database format. It is checked if all relevant files (documentation, speech files, label files, lexicon) are present in the appropriate directory structure and with the correct format.
3. Design. The appropriateness and the completeness of the recorded items are addressed for the purpose of the envisaged application(s).
4. Speech files. The acoustical quality of the speech files is measured in terms of, e.g., (average) duration, clipping rate, SNR, mean sample value. Also auditory inspection of signal quality belongs to this category.
5. Label files. It is checked if the label files obey the correct format, and if they can be automatically parsed without yielding erroneous information or even system crashes.
6. Phonemic lexicon. The lexicon should contain appropriate phonemic (or allophonic) transcriptions of all words in the orthographic transcriptions of a SLR.
7. Speaker & environment distributions. The recorded speakers should present a fair sample of the population of interest in terms of (typically) gender, age and dialectal background. Also the recording environments should be representative for the targeted applications.
8. Orthographic transcriptions. A native speaker of the language checks a sufficiently large sample of the orthographic transcriptions by comparing these to the speech in the signal files and the transcription protocol.

Formats and formal criteria can be tested automatically. The content of a database such as the correctness of the orthographic or phonemic transcriptions, but also the contents of the documentation require manual labour.

The associated criteria can be found in detail in the validation deliverables given in the reference section for individual projects as mentioned in section 1 above.

The annotation of SLR in the SpeechDat-family is rather flat and is captured by label files following the SAM-standards (SAM, 1992). However, for SLR with more complex annotation layers the SAM concept is not well suited. More appropriate annotation formalisms for such SLR are ATLAS (Laprun et al., 2002), and IMDI (Broeder et al., 2001). Examples of hierarchically structured annotation layers recently validated by SPEX

are broadcast news databases developed for TC-STAR_P (<http://www.tc-star.org>) and the phonetic lexicons produced in the LC-STAR project (Hartikainen et al, 2003). Annotation of these databases is in XML-based encodings. The new challenges that such new formats pose for validation are discussed in section 3.

Traditionally, the validation scenario in a SpeechDat approach consisted of two phases: 1) prevalidation, 2) full validation. During prevalidation, the recordings of the first 10 speakers are evaluated in order to find systematic errors at an early stage of the speech collection. For these 10 speakers identical checks are carried out as will be the case later for the complete database. These checks are executed on the speech files, label files, and documentation files and refer to the aspects mentioned above. For a full validation, all the checks which were executed in the prevalidation phase are carried out again, this time, however, on a complete database. Furthermore, orthographic transcriptions are evaluated by native speakers and the database is checked against a number of distribution criteria, such as gender or environment distributions, which is only possible when all the database recordings are available.

This scenario was followed in projects such as SpeechDat (II), SpeechDat-East, SpeechDat-Car and SALA (I & II). (<http://www.speechdat.com>). The experience of both producers and the validation centre was that the two validation stages were not sufficient to detect certain errors both in the early design phase and in the very final phase when, after full validation, some final corrections were made by the producing parties without these corrections being re-checked. Therefore, in more recent projects like SpeeCon and Orientel, new validation stages were introduced in order to minimize such risks. These new stages are presented in section 4 of the paper.

As mentioned in section 1, there is also the setting in which validation has to be done post-hoc. In section 5 we provide a concise update of the latest validation activities at ELRA, where SPEX acts as the validation centre for SLR.

3. New Data Formats

As mentioned above, annotations in other than SAM oriented formats require new validation approaches. Such annotations are found in, e.g., broadcast news databases (BCN) from the TC-STAR_P project and phonemic lexicons, from the LC-STAR project. These databases differ from the SpeechDat format in a number of ways.

	SpeechDat-family	BCN (TC-STAR_P)	LC-STAR
Speech	Many short utterances	Very long items with complete broadcast	No speech
Database Structure	Many files in relatively complex directory structure	Few files in simple directory structure	Very few files in very simple directory structure

	Meta files in SAM or tab separated	Meta files in XML	Files in XML
Character coding	ANSI/other	ANSI/other	Unicode UTF-16

Table 1: Differences between LR types validated by SPEX

Table 1 gives the most important differences between SpeechDat, TC-STAR_P and LC-STAR databases. In the following we will discuss how these differences influence the validation procedure.

Speech

Where SpeechDat-like databases have many items containing short phrases like numbers, names or dates, typically lasting between two and ten seconds, the BCN databases from TC-STAR_P have huge speech files up to half an hour or longer. LC-STAR contains only lexicons and no speech files at all.

Because of the length of the speech files in BCN databases, it is impossible to make a straightforward random selection of speech files for the validation of orthographic transcriptions. Therefore, a semi-random selection (accounting for all sorts of distributions, like gender and accent of speaker) of the transcriptions is made, and checked against the corresponding speech segments. In order to do this, the time stamps of the selected parts are searched automatically in the XML label files and used to cut the speech segments out of the large speech files. Speech segments of the same speaker are grouped together in order to allow the validator to assess if subsequent segments come from the same speaker as indicated by the producer. For this more sophisticated procedure new software tools were developed.

The quality of the speech is checked by computing a number of statistics of the signal, like clipping rate and signal-to-noise ratio (see section 2). These statistics, however, are only meaningful in relatively short speech clips up to a few minutes at most. To compute meaningful signal statistics on very large files, these files have to be divided in smaller segments first, so that portions with bad signal quality can be detected and are not averaged out.

Database Structure

SLR in the SpeechDat-family have a relatively complex directory structure accompanied by simply structured information files. These label files are encoded in SAM, a scheme that was standardised by the EAGLES group (Gibbon et al., 1997). Other metafiles in the SpeechDat-approach usually contain just tab-separated fields. The TC-STAR_P BCN databases and LC-STAR lexicons have a simple directory structure but more complex structured information files. The multi-layered annotations in the BCN and the lexicons in LC-STAR are in XML format.

For validation this means that the relevant information has to be extracted by parsing XML-format files. That implies that the validation software for automatic checks either has to be adapted or, alternatively, already existing off-the-shelf tools can be used. These tools are typically freely or commercially available parsers, like XML-Spy, which

can, for instance, check the XML against a Document Type Definition (DTD). This means that instead of writing new software only a set of proper DTD rules have to be stated. The definition of these rules forms part of the specifications of the database and are not directly developed for validation. They have to be used, however, by the validation centre to carry out the check against the DTD.

Nonetheless, additional smart parsing procedures were needed to check for instance sufficient coverage of certain POS-tags in the LC-STAR XML-based lexicons (De Vriend et al. 2004).

Character coding

For European languages plain ANSI character encoding was sufficient, but with databases in all kinds of other languages appearing, a lot of other character encodings are needed. In Oriental and SpeeCon languages like Mandarin, Arabic, Hebrew and Korean are recorded, to name a few. For transcription validation of more 'exotic' character codings tools are required that are able to handle codings other than those in the ISO-8859 series.

Unicode is becoming a new standard and is also used in LC-STAR. In this case the software has to be able to cope with UTF-16, in which characters are coded in two bytes. This poses special challenges for comparing strings, inserting characters in strings, and generating validation output.

4. New Procedures

Since its specification in the early nineties, the validation procedure as described in section 2 has undergone a number of changes. These are due to, on the one hand the experience of the validation centre, but on the other hand the needs of the producers. The current procedure which has been applied in the more recent projects such as SpeeCon and OrientTel comprises the following validation stages:

- 1) *prompt sheet validation*
- 2) *lexicon validation by an external expert*
- 3) pre-validation of the first 10 recorded speakers
- 4) validation of a complete database
- 5) *pre-release validation*

The stages 1, 2 and 5 are new and were not applied in the first SpeechDat projects. In the following section these new stages are presented in more detail together with a motivation for their introduction.

1) Prompt sheet validation

Before embarking on recording speakers, the producers design reading scripts. These scripts should be an ideal reflection of the specifications with regard to the content of the corpus items and the number of repetitions for each item. Since things are bound to go wrong during the recordings due to problems with the recording platform, of speakers omitting certain items altogether, not reading them correctly, stuttering or speaking in an environment

with high background noise, the reading scripts have to meet the upper bounds of what is achievable in a database.

The validation of the prompt sheets comprises checks with regard to the presence of the corpus items, adherence of their design to the specifications as well as the number of repetitions at word or sentence level calculated for the complete database.

If at this stage the prompt sheets do not fulfil the validation criteria (the absolute minimum which is required in the end), measures can still be easily taken to repair the errors since no recordings have been made yet. Database producers indicate to highly appreciate this part of validation which allows them to spot and repair errors in an early design stage.

The prompt sheet validation is also a test for the specifications as it pinpoints parts which are underspecified and need further clarification.

2) Lexicon validation by an external expert

A formal check of the lexicon with regard to the format and the use of legal phoneme symbols is part of all the validation stages and can be carried out by the validation centre itself. From experience in the SpeechDat projects, however, a need to check the quality of the phonemic transcriptions has arisen. Since this work needs to be done by phoneticians of each language, the validation centre delegates this task to external experts. There are two conditions for the selection of these experts: they have to be native speakers of their language and must have a phonetic training. These experts check manually a relevant sample of the lexicon. They are instructed to give the provided pronunciation the benefit of the doubt and only to correct transcriptions that reflect an overtly wrong pronunciation. This is in order to prevent marking as errors differences which are due to different phonetic theories or different ideas about what the 'most common' or 'best' pronunciation is.

5) Pre-release validation

The validation of a complete database results in a report containing a list of errors which were found in the database. Some of them are irreparable and related to flaws in the design of the database or the recordings themselves. However, a large number are usually minor and refer to the documentation, label files or other text files which are produced during post-processing. These errors can easily be repaired and the producers are willing to do that. The danger, however, is the introduction of new errors or format inconsistencies during the rectification. Therefore, a pre-release validation has been introduced so that the envisaged master disks can be checked again by the validation centre. The purpose of this validation is to make sure that the minor errors which were found during complete validation are repaired and that no new errors have been introduced. If the pre-release validation is finished with a positive result, the database is ready for distribution and the producers are not allowed to make any more changes, however minor.

It may seem that with these new validation stages the procedure becomes more complex. Our experience, however, is that it also becomes more structured and more efficient with as a result a higher quality of the final product. It should be stressed that the extra stages 1 & 2 are of importance for data collections of which the contents are predictable in advance, whereas the pre-release validation is of relevance for all SLR that need an update after validation.

5. SLR validation and ELRA

SPEX is ELRA's official validation centre for SLR. This work is typical for a setting in which quality assessment and LR repair is performed on a post-hoc basis. SPEX maintains a bug report service for SLR and conducts Quick Quality Checks (QQC) for SLR that are in ELRA's catalogue or are about to enter it.

For the bug report service we refer to <http://www.elra.info/> (Services around LRs > Validation > Bug report service). Attractive prizes are offered at a regular basis for the best bugs reported.

A QQC is a shortened version of a full validation still addressing all 8 relevant aspects listed in the introduction section, but only at a formal level for which mainly automatic format checks can be defined and applied. Exception is the documentation which is always manually checked. A QQC can be carried out in say 6 hours whereas a normal full validation takes at least 25 hours.

Depending on the type of application domain of the SLR a set of minimal requirements is formulated (Van den Heuvel et al., 2003). Different sets have now been defined for SLR for Automatic Speech Recognition and phonetic lexicons. Sets of minimal requirements are currently under development for speech synthesis SLR. The QQC will consist of two parts in the future. The first report will present validation results on the SLR proper and will contain comments to the provider; the second report will present validation results on the description forms that ELRA provides with the SLR, and will be directed to ELRA.

ELRA has a validation centre for written language resources (WLR) as well, being CST in Copenhagen. Also CST is developing templates for QQCs and a bug report service for WLR (Fersøe, Monachini, 2004).

6. Conclusion and prospects

Validation of SLR is not static, neither in content nor in procedure. Validation criteria are dynamically adapted to the application domains of the SLR at hand and to the settings in which validation is required.

Apart from that, new data formats require new checks or new implementations of existing checks. This was illustrated on the basis of recent validation work in the TC-STAR_P and LC-STAR project.

The procedures for validation have not reached an end-point either. The introduction of new validation stages at the very beginning and at the very end of database production allows us to more closely assist SLR producers in making a better product.

For existing databases for ELRA's catalogue, new quick check templates are under development to allow for rapid and efficient validation of a SLR at the formal level.

7. References

- Broeder, D., Offenga, F., Willems, D., Wittenburg, P. (2001) The IMDI meta-data set, its tools and accessible linguistic databases. Proceedings IRCS Workshop on linguistic databases, 11-13 December 2001. Philadelphia USA. http://www ldc.upenn.edu/annotation/database/papers/Broeder_etal/32.3.broeder.pdf.
- De Vriend, F., Castell, N., Giménez, J., Maltese, G. (2004). LC-STAR: XML-coded Phonetic Lexica and Bilingual Corpora for Speech-to-Speech Translation Proceedings LREC'2004 Workshop on XML-based richly annotated corpora, 29th May 2004.
- Fersøe, H. (2003). Validation Manual for lexical. ELRA/0209/VAL-1 Deliverable D1.1A.
- Fersøe, H., Monachini, M. (2004). ELRA Validation Methodology and Standard Promotion for Linguistic Resources. In: Proceedings LREC 2004, Lisbon.
- Gibbon, D., Moore, R., Winski, R., (Eds) 1997. *Handbook of standards and resources for spoken language systems*. Mouton, de Gruyter. Berlin, New York.
- Hartikainen, E., Maltese, G., Moreno, A., Shammass, S., Ziegenhain, U. (2003). Large lexica for Speech-to-Speech Translation: from specification to creation. Proceedings Eurospeech 2003, Geneva, Switzerland, September 2003.
- Iskra, D., Van den Heuvel, H., Gedge, O., Shammass, S. (2002). Specification of Validation Criteria. OrientTel Technical Report D6.2. <http://www.orientel.org>.
- Laprun, C., Fiscus, J.G., Garofolo, J., Pajot, S. (2002) A practical introduction to ATLAS. Proceedings LREC 2002, Las Palmas.
- SAM (1992). User guide to ETR tools. SAM: Multi-lingual speech Input/Output Assessment, Methodology and Standardisation. Ref: SAM-UCL-G007.
- Shammass, S., van den Heuvel (2004). Specification of validation criteria for lexicons for recognition and synthesis. LC-Star Technical Report D6.1. <http://www.lc-star.com>.
- Schiel, F., Draxler (2003). Production and validation of speech corpora. Bastard Verlag München. <http://www.phonetik.uni-muenchen.de/Bas/>.
- Van den Heuvel, H. , Choukri, K., Höge, H., Maegaard, B., Odijk, J., Mapelli, V. (2003). Quality Control of Language Resources at ELRA. Proceedings Eurospeech 2003, Geneva, Switzerland, pp. 1541-1544.
- Van den Heuvel, H. (1996): *Validation criteria*. SpeechDat Technical Report SD1.3.3. <http://www.speechdat.com>
- Van den Heuvel, H. (1999): *Validation criteria*. SpeechDat Car Technical Report CD1.3.1, 1999. <http://www.speechdat.com>
- Van den Heuvel, H., Shammass, S., Moyal, A. (2001): Definition of validation criteria. SpeeCon Technical Report D4.1. <http://www.speecon.com/>

