

SPEX QQC REPORT

TITLE DATABASE: TED (Translanguage English Database) & TED laryngo
 DATABASE OWNER / PRODUCER: University of Munich, LIMSI-CNRS
 ELRA CATALOGUE NUMBER: S0031

AUTHORS OF QQC REPORT: Henk van den Heuvel
 DATE: 28 June 2002
 VERSION: 1.0

SUMMARY SHEET:

Database part	Applicability (y/n)	Quality value		
		*	**	***
1. Documentation	Y		**	
2. Format	Y			***
3. Design & contents	Y			***
4. Speech signals	Y			***
5. Annotation files	Y	*		
6. Speakers	Y		**	
7. Environments	N			
8. Transcriptions	N			
9. Lexicon	N			

For each applicable part a star assessment is given.

1. * This value is given if there is not a proper and reliable fit of the contents of the SLR and the information about this part as presented in the documentation.
2. ** This value is given if the documentation well accounts for the contents of the SLR. Some small deviations are permitted.
3. *** This value is given if there is no mismatch between the documentation and the contents of the SLR.

1 Quick Quality Check Report

1.1 Documentation

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

The documentation is in the following files in the DOC directory: DATA.DOC, FILE.DOC, RECORD.DOC, SHORTEN.DOC, TED.DOC. The files are on all CD-ROMS.

- db layout and media

OK, files: README, FILE.DOC

- application potential for the SLR

OK, TED.DOC

- directory structure and file names

OK, files: README, FILE.DOC

- recording equipment

OK, TED.DOC, RECORD.DOC

- design and contents of the recordings

OK, DATA.DOC

- coding and format of the speech files

OK, FILE.DOC

- contents and format of the label files

Not well described.

- speakers

OK, see DATA.DOC.

CD5 contains somewhat different data/speakers than documented in DATA.DOC:

rs61s300 is not on the disk; au00s100 and wh00s300 are on the disk but are not documented; al20s200 is documented as ao20s200.

In total, the 5 CDs contain 190 speeches of 184 different speakers.

- recording environments distinguished

N/A

- transcription conventions

There are no transcriptions of all speeches!

(Transcripts for 39 of the sessions are provided on an extra CD issued by the LDC (S0120 in ELRA's catalogue). Transcription conventions for these transcripts are documented in the UTF_1_0_V2.{PDF|PS} files in the DOC directory of the CD.)

- lexicon: format and transcriptions included

N/A

The TED laryngo CD-ROM contains the speeches and the corresponding laryngograph signals for 11 presentations of 11 speakers. The speech files are also provided on the 5 TED CD-ROMS (the laryngograph signals are not).

The encoding of the laryngograph signal files is not explained.

1.2 Format

- The file names and directory structure should correspond to the documentation

OK

1.3 Design and contents

- All mandatory items according to the documentation should be included

OK

- Number of effectively missing files per corpus item should be appropriate

OK. There are no missing files.

1.4 Speech signals

- For 2 CDs of the SLR acoustic measurements on the speech files will be

made, and the results reported. The acoustical measurements involved are:

- Clipping rate
- SNR
- Mean amplitude

The acoustic measurements were carried out on the recorded speeches, but not on the laryngograph signals. The measurements were not further grouped; they were computed per file, i.e. per speaker.

All files had a clipping rate below 0.2%.

The SNR-histogram over the files looks as follows:

SNR range	Number of speakers
10 - 15 :	2
15 - 20 :	55
20 - 25 :	34
25 - 30 :	50
30 - 35 :	43
35 - 40 :	5
45 - 50 :	1

The mean sample values for all files are between -300 and 300.

Thus, none of the files has alarming average acoustic characteristics.

1.5 Annotation files

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

The information in the SAM files (with extension SEO is not described in the documentation. An example of a SAM label file is:

LHD: V1.0

FIL: speech

DBN: TED

DIR: .
SRC: ad43s400.pes
SAM: 16000
BEG: 0
END: 15733556
RED: 21/9/93
REP: Berlin_Eurospeech93
SNB: 2
SBF: 01
SSB: 16
RCC: 1
NCH: 1
LBD:
ELF:

All SEO files contain the same SAM-labels. All recorded sessions are claimed to be recorded on 21 Sep. 1993.

1.6 Speakers

- Speaker distributions should be in agreement with documentation

OK

All speakers are represented with one presentation. Five speakers are contained with more presentations:

3 we
2 ms
2 mr
2 cl
2 al

1.7 Environments

- Environment distributions should be in agreement with documentation

N/A

1.8 Transcription

- how many speech files miss an orthographic transcription?

For 39 speech files a transcript is provided on an additional CD issued by the LDC.

- All non-speech markers should be described in documentation

N/A

1.9 Lexicon

- The correct set of phone symbols should be used (according to documentation)

N/A. There is no lexicon!

- All words in the (orth.) transcriptions should be present in the lexicon

N/A

1.10 Other remarks

The associated text materials that are mentioned (consisting of ASCII versions of over 400 proceedings papers and oral preparations that were supplied by the authors, as well as, 250 speaker questionnaires) are delivered in the TEDTEXT directory of the TED laryngo CD-ROM. This is not clear from the documentation. Further there is no link in file names between the texts and the speakers; and not all speakers on the CD-ROM released their oral text (there are oral texts for 60 papers). There are questionnaires for 254 speakers, but again there is no clear link from the names of these files to the recorded data.

2 Recommendations

- Transcripts of all speech files should be created. Otherwise the speech data cannot really be used.
- A lexicon with phonemic transcriptions of each word would be desirable.
- Link tables from speakers to data (recordings, presentation texts, questionnaires) should be created