

SPEX QQC REPORT

TITLE DATABASE: COST 232 MULTI-ENGLISH database
 DATABASE OWNER / PRODUCER: COST 232 consortium
 ELRA CATALOGUE NUMBER: S0009

AUTHORS OF QQC REPORT: Henk van den Heuvel
 DATE: 1 July 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:

There are three text files that serve as documentation: datafile.txt, datadesc.txt, badandmi.txt. They are all on disk 1 and only pertain to that disk. For disk 2 there is no documentation at all.

- db layout and media
 - ⇒ *Unclear. There is no proper overview of the contents of the individual CDs and the contents of various directories and files. The contents of the tools-directory are not explained; the contents of files with extensions .dat and .tel are not explained.*
 - ⇒ *The information in the db232/tools directory (on disk 2) is not explained.*
- application potential for the SLR
 - ⇒ *Not explained. Surprisingly, the language of the recordings is not documented! From the vocabulary in section 3 we infer that it is English.*
- directory structure and file names
 - ⇒ *Not explained. Only some of the filenames are explained in datadesc.txt.*
- recording equipment
 - ⇒ *Not explained.*
- design and contents of the recordings
 - ⇒ *The vocabulary of the recordings is in section 3 of datadesc.txt. But it is not a full description of the design. It only pertains to disk1 (db232eng); The W-codes for disk 2 refer to other prompt information than those for disk 1.*
- coding and format of the speech files
 - ⇒ *Unclear. There are .dat files. They contain a header and speech data. The information in the header is not explained. The signals are in A-law (not explicated either).*
 - ⇒ *Also the .tel files (disk 2) are speech files (A-law). They do not have a header and there is no information on their coding.*

- contents and format of the label files
 - ⇒ *There are no label files. The .dat files on disk 1 have a header with some information. The .tel files on disk2 do not have a header.*

- speakers
 - ⇒ *Unclear. The file db232/tools/forms.txt seems to contain speaker information, but it is unclear what information is contained in each column.*
 - ⇒ *For disk 1, there are 469 speakers, 246 males and 223 females, but this is not documented either. There is one call per speaker as far as we can judge.*
 - ⇒ *For disk 2, the situation is not clear either. If we count the lines in the db232/tools/db*.txt files, then we find 268 participants. If we count the lines in the db232/tools/forms.txt file then we find 797 participants/calls. If we count the numbers of files in one directory then we find 977 to 1033 calls per directory, probably reflecting about 4 calls per speaker.*

- recording environments distinguished
 - ⇒ *Unclear. The filenames of .dat files contain information about the telephone used. (But also U is used for unknown cases; not in the documentation).*

- transcription conventions
 - ⇒ *There are no transcriptions. Most of the time the prompts were spoken. For the (two) spontaneous items, there is no transcription.*

- lexicon: format and transcriptions included
 - ⇒ *There is no lexicon.*

1.2 Format

- The file names and directory structure should correspond to the documentation
 - ⇒ *This holds for disk 1. There is neither documentation nor help for the files on disk 2.*
 - ⇒ *Remark: The chosen file names are impractical because they do not contain the item code (word number) in the file names. Due to this there are many files with the same filenames, which may get lost during any kind of database processing.*

1.3 Design and contents

- All mandatory items according to the documentation should be included
 - ⇒ *For disk 2 there is no item w001.*

- Number of effectively missing files per corpus item should be appropriate
 - ⇒ *Disk 1:*
 - All w-directories contain speech for 469 speakers. Exceptions: w0008 (463), w0011 (465), w0018 (468), w0021 (467), w0033 (462). This is not found in the documentation. The directory bad_utte contains different information than explained in db232eng/badandmi.txt
 - ⇒ *Disk 2:*
 - All w-directories contain 977 to 1033 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

Clipping rates (based on extreme (min./max.) sample values in the files) were computed for all signal files. The results were grouped per corpus item (w-item). This was the result:

Clip val.	Number of corpus items
0.0 - 0.2 :	20
0.4 - 0.6 :	3
1.8 - 2.0 :	1

For SNR the following distribution over the items was found

SNR range	Number of items
5 - 10 :	1
30 - 35 :	12
35 - 40 :	11

For mean amplitude there was no group with an alarming average sample value.

The group with the deviating values for clipping rate and SNR was W0033. Since this is the NOISE-item the result is not unexpected.

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
 - o Readable
 - o Contain the information described in the documentation
- ⇒ *There are no label files. The files on disk 1 have a header, e.g.*
SAMPLE_RATE=8000.0|CODING_METHOD=ALAW|CONDITIONS=ISDT|
DATABASE=COST232|DATA_TYPE=INTEGER|DATA_SIZE=8|FRAME_S
IZE=1|RECORDING_DATE=25-AUG-1993|13:44:52|TIME_STAMP=25-
AUG-1993
13:44:52|TALKER_NUMBER=5531|FILE_STATUS=NOT_CHECKED|STA
RT_POINT=0|ENDPOINTER_TYPE=COLLECTION_WINDOW|SESSION_
NUMBER=1|CLASS=1|REPETITION=1
- ⇒ *The files on disk 2 do not have a header.*

1.6 Speakers

- Speaker distributions should be in agreement with documentation

This cannot be checked since the documentation is unclear about the speaker distribution. For disk 1 the distribution of speakers per country is about:

Country	Male speakers (calls)	Female speakers (calls)
BE	16	16
PO	18	17
CZ	8	7
SL	32	32
DE	7	9
SP	23	15
GE	19	19
SU	16	17
IT	25	15
SW	29	28
NO	13	20
UK	38	27

For disk 2 we can infer a distribution from file db232/tools/forms.txt.

If we sort these figures to speaker gender, then we obtain:

Country	Male speakers (calls)	Female speakers (calls)
BE	35	28
PO	32	32
CZ	15	17
SL	28	32
DE	16	16
SP	40	32
GE	38	36
SU	44	44
IT	32	37
SW	44	44
NO	24	24
UK	56	52

1.7 Environments

- Environment distributions should be in agreement with documentation

Not applicable

1.8 Transcription

- how many speech files miss an orthographic transcription?

There is no transcription.

- All non-speech markers should be described in documentation

NA

1.9 Lexicon

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

There is no lexicon

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

NA

1.10 Other remarks

CD2 is unreadable from a UNIX platform. Due to missing x-attributes it can only be read by the root user.

2. Recommendations

The quality of the database could be enhanced if:

- A proper documentation file is added
- Clarity on the number of speakers, speaker distribution of age, region and gender, and number of calls per speakers is added (as replacement of form.txt file)
- Orthographic transcriptions of the signals are added