SPEX / Dept. of Language and Speech
University of Nijmegen
Erasmusplein 1
NL-6525 HT Nijmegen
The Netherlands
E-mail: spex@spex.nl

SUBJECT:	Pre-validation Czech SPEECON corpus (Children)
AUTHORS:	Dorota Iskra
VERSION:	1.1
DATE:	17 March 2009

INTRODUCTION

The speech databases made within the SPEECON project were validated by SPEX, Nijmegen, the Netherlands, to assess their compliance with the SPEECON format and content specifications, as documented in Deliverables 2.1, and 4.1 of the project.

The validation results of the Czech SPEECON database (children) are contained in this document. This database is approved by the SPEECON consortium.

For validation SPEX receives all the documentation and label files as well as a small subset of signal files on which transcription validation is carried out. In the validation procedure we systematically check a list of validation criteria for a range of topics. In the following sections we will evaluate these criteria one by one. Validation results that call for attention because of deviations from the SPEECON specifications are marked by
⇒
so that they can be easily found.

CONTENTS

1 DOCUMENTATION

– File DESIGN.DOC is present

OK

– Language of doc file: English

OK

– Contact person: name, address, affiliation

OK

– Collectors and owners of the database

OK

– Number of disks

OK, section 2.4

– Contents of each disk

OK

– Formats of speech files

OK, section 3.1

– The directory structure of the disks
  - Database, block and session orderings
  - Directories DOC, INDEX, TABLE (and optionally HTML, PROMPT, SOURCE)

  ⇒ README.TXT should only be found on the documentation disk (section 2.2.1).

– File nomenclature
  - Root files
  - Names of speech files and label files

- Files in directories DOC, INDEX, TABLE (and optionally HTML, PROMPT, SOURCE)

OK

– Reference to the validation report made by SPEX (VALREP.DOC)

OK

– Contents and format of the label files
  - Clarification of attributes (three letter mnemonics)
  - Example of label file

OK, section 3.2
⇒ The labels MIP and MIT are missing from Table 12.

– Recording platform
  - Hardware set-up
  - Software set-up
  - Microphone types and positions

OK, section 6

– Speaker recruitment

OK, section 4.4

– Prompting
  - Presentation design
  - Prompting example for one recording session

OK, section 4

– Description of all the items in the database
  - Specification of the individual items in the database
  - Connection of prompted items to corpus codes in the database (in titles of subsections of individual corpus items)

OK, section 5
⇒ Not clear which corpus id is used for city and which for street names (section 5.3.13).
⇒ The symbols for rare phonemes in section 10.2 do not correspond to the ones used in the database.

⇒ *The phoneme /@/ is marked as rare in Table 41, but not listed as such in the paragraph above it.*

– A list of prompted digits in the language

OK, Table 17

– Tables with frequencies of the phones represented in the phonetically rich sentences (at transcription level)

OK, Table 41

– Transcription conventions
  - Procedure used
  - Quality assurance
  - Character set used for annotation (transcription) (ISO-8859 or other if needed)
  - Conventions used for transcription of spoken words
  - Annotation symbols for non-speech acoustic events: Filled Pause, Speaker Noise, Stationary Noise, Intermittent Noise, and additional, optional ones if used.
  - List of symbols used to denote word truncations, mispronunciations and not understandable speech
  - Case sensitivity of transcriptions
  - Use of punctuation

OK, section 9

– Lexicon information
  - Case sensitivity of transcriptions
  - Procedures to obtain phonemic forms from orthographic input
  - List of SAMPA phone symbols
  - List of Pinyins and Hepburn Romaji syllables (if applicable)
  - Statement whether or not the transcription and the lexicon are case sensitive
  - Information captured in the phone transcriptions (assimilation and reduction rules)
  - Statement whether multiple transcriptions are supported
  - Statement whether stress information is supplied
  - Statement whether there are any tags, and if so, the tagging conventions used, e.g., record (noun) vs. record (verb)
  - List of words that are from a foreign language
  - List of rare phonemes
  - Analysis of frequency of occurrence of the phonemes represented in the phonetically rich sentences, and in the full database (at transcription level); optional for statistics of diphones, triphones.
  - Any other language-dependent information or conventions

OK, section 10

– Speaker demographics
  - Which regions, how many of each
  - Motivation for selection of regions
  - Which age groups, how many of each
  - Sexes: boys, girls; how many of each.
  - Number of speakers
  - Number of sessions

  OK, section 8

– Recording environments
  - Description of the child environments of the sessions
  - Description of sub-environments per environment
  - Distribution of speakers over environments and sub-environments
  - Was a high pass filter (ME64 or similar) used or not

  OK, section 7. No filter was used.

2 DATABASE STRUCTURE, FORMATS AND FILE NAMES

– Directory / subdirectory conventions
Format of directory tree should be
\<database>\<block>\<session>
  - Database: defined as <dbName><#><language code>
    where <dbName> is ADULT for the adult speaker and CHILD for the children speaker database, <#> is 1 for SpeeCon, <language code> is the ISO 639 2-letter language code
  - Block: defined as BLOCK<nn> where <nn> is a progressive number from 00 to 99.
    Block numbers are unique over all disks.
    They correspond to the first two digits of <nnm> below.
  - Session: defined as SES<nnm> where <nnm> is the session code also appearing in file name

OK

– File naming conventions
  - All file names should obey the following pattern: DDNNMCCC.LLF
  - DD: database identification code
    For SPEECON: SA for adult and SC for child speakers
  - NNM: session code 000 to 999
  - CCC: item code;
  - LL: ISO-639 language code (with extensions)
  - F: speech file type
    0,1,2,3 is for signal files of the four channels
    O is for label file

OK

– NNM in filenames is not in conflict with BLOCK and SES numbers in pathname

OK

– Contents lowest level subdirectories should be of one recording session only

OK

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

OK

– A README.TXT file should be in the root of each database describing all (documentation) files. (This file is also allowed as README.HTM).

  ⇒ *In the text CDs are referred to whereas the list contains a DVD distribution.*

– 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: CHILD1EN_01.

  OK

– A copyright statement should be present in the file COPYRIGH.TXT (root)

  ⇒ *The copyright statement includes a reference to CD-ROM.*

– Documentation should be in \<database>\DOC
  - DESIGN.DOC
  - PLATFORM.DOC (optional)
  - TRANSCRP.DOC (optional)
  - SPELLALT.DOC (optional)
  - SAMPALEX.PS
  - ISO8859<n>.PS
  - SUMMAR0.TXT
  - SUMMAR{1|2|3}.TXT (optional, only needed if files are missing in other channels than 0)
  - VALREP.DOC

  OK, pdf versions of some files are provided too

– Tables should be in \<database>\TABLE
  - SPEAKER.TBL
  - LEXICON.TBL
  - REC_COND.TBL
  - SESSION.TBL

  OK

– Index files should be in \<database>\INDEX
  - CONTENT0.LST
  - CONTENT{1|2|3}.LST (optional, only needed for annotated channels)

  OK

– Prompt sheet files (optional) should be in \<database_name>\PROMPT

Not provided

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

OK

– 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.

OK

– The contents of the database as given in CONTENT{0|1|2|3}.LST should have the following order of attributes:
  - full pathname (DIR:)
  - speech file name (SRC:)
  - corpus code (CCD:)
  - speaker code (SCD:)
  - speaker sex (SEX:)
  - speaker age (AGE:)
  - speaker accent (ACC:)
  - scenario code (SCC:)
  - 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.

OK

– The contents of the SUMMAR{0|1|2|3}.TXT files should have the following order of attributes:
  - the full directory name where speech and label files are to be found (DIR)
  - the session number (SES)
  - two strings of typically N codes (CCD). Each item present in a string is represented by its code, separated by commas. The first string contains the item list as intended; the second string contains the item list as recorded. If the item is missing, a '---' should appear. The two strings are separated by a space.
  - recording date (RED)
  - recording time of first item (RET)
  - optional comment text
  - all these fields are separated by spaces
  - also the noise recordings and silent word recordings should be included

⇒ *In SUMMAR0.TXT spaces instead of tabs should be used to separate different fields.*

– All sessions indicated in the documentation SUMMAR{0|1|2|3}.TXT are present

OK

– Missing items per speaker should correspond to missing files reported in SUMMAR{0|1|2|3}.TXT)

OK

– The database should be free of viruses.

OK

3 CORPUS ITEMS: DESIGN AND COMPLETENESS

A. Check on mandatory corpus items

– Phonetically rich sentences (S01-S60)
- Read
- Max. 5 repetitions per sentence
- Min. of 600 different sentences
- At least 50 samples per phone at transcription level (except for rare phones)

OK, 600 different sentences were found. All except for rare phonemes were found with > 50 repetitions.

– Isolated digits (CI1-CI4)
- Prompted in words
- Min. occurrences per digit at transcription level: 120/D

OK, all the 16 digit forms were found with 10-15 repetitions each.

– Isolated digit string (CB1)
- Prompted in words
- Min. occurrences per digit at transcription level: 30

OK, all the 16 digit forms were found with 33-80 repetitions each.

– Connected digit strings (CC1-CC4)
- Prompted in words
- Min. occurrences per digit at transcription level: 600/D

OK, all the 16 digit forms were found with 47-88 repetitions each.

– Telephone number (CE1)
- Prompted as 9-13 digits number
- Including GSM numbers

OK

– Natural numbers (CN1-CN3)
- Prompted in words

OK

– Money amount (CM1)
  - Prompted in words

  OK

– Time phrases (CT1-CT2)
  - Prompted in words
  - One in analog format (CT1)
  - One in digital format (CT2)
  - CT1: Max. of 20 specific time words should be used

  ⇒ *The word "poledne" which was found in DESIGN.DOC was not found in the database.*

– Date phrases (CD1-CD3)
  - Prompted in words
  - One in analog format (CD1)
  - One in digital format (CD3)
  - CD2: max. 50 phrases

  OK, CD2: 39 different phrases were found

– Spelt words (CL1-CL3)
  - Prompted in letter sequences
  - One is artificial word
  - Min. occurrence per letter at transcription level: 525/L

  OK, all the 42 letters were found.
  ⇒ *Only 8 repetitions were found for the letter "dlouhé Y" (min is 12).*

– Person name (CP1)
  - Set of 50
  - Each name occurs max. 1 time

  OK

– City/street names (CO1-CO2)
  - Set of 50 cities
  - Set of 50 street names
  - Each name occurs max. 1 time

  OK

– Yes/No (CQ1, CQ2)
  - Min. of 45 yes/ 45 no at transcription level

  OK, 50 repetitions of yes/no were found

– E-mail and Web-address (CW1, CW2)
  - CW1: set of 50 Web-addresses
  - CW2: set of 50 E-mail addresses
  - Each name occurs max. 1 time

  OK

– Keyboard characters (CK1-CK2)
  - Max. 20 characters
  - Fixed set of 12 characters

  ⇒ 24 instead of 20 different keyboard characters were found.

– Application words (001-214)
  - Set of 122 words from 3 categories
  - Min. of 40 occurrences per word at transcription level
  - Each child speaks all the words

  OK, 177 different words were found at the prompt level with at least 42 repetitions at the transcription level. The words in the specifications match the ones in the DESIGN.DOC and these in turn match the ones in the database.

B. Checks on presence of corpus files (prompt level)

The following completeness checks are performed on obligatory SPEECON items only (i.e. all the items in the children's database). Results are presented only for the annotated channel 0. Results of other channels are reported only in case of serious deviations.

1. Files that are not there

The following is an overview of corpus items with a frequency lower than 50. Per item code the number of items not present in the database and empty files is given (merged together):

No files are missing

2. Effectively missing files

For the children's part 0 files with empty transcriptions were found (i.e. files with only silence, noise symbols or ** in their transcriptions). When merged with the missing files

listed above, the following frequencies of effectively missing items per item code are obtained:

No files are missing or empty

– For the children's part, SPEECON allows a maximum of 40% (=20) of the files for each mandatory corpus item as effectively missing.

  None

3. Corrupted speech files

For the children's part **5** corrupted files were found (utterances which have only truncated or mispronounced words). When merged with the effectively missing files listed above, the following frequencies per item code are obtained:

044: 1
101: 1
CK1: 1
CP1: 1
Y56: 1

– For the children's part SPEECON allows a maximum of 40% (=20) of the files for each mandatory corpus item as either effectively missing or corrupted.

  OK, max 2%

C. Automatic checks at transcription level

1. Match between prompt and transcription (only for isolated words; corpus codes: CI1-CI4, CP1, CO1-CO2, CQ1, CQ2, CK1-CK2, 001-214, Y01-Y..)

A mismatch between prompt and transcription is scored if the word in the prompt does not appear in the transcription, if there is no speech at all or only another word(s).

For the children's part, **137** files with a mismatch between prompt and transcription were found. When merged with the effectively missing or corrupted files under B.3, the following distribution is observed per item code (only items with frequency > 1 are included):

001 5
002 6
003 8
004 7
005 2
006 6
007 5

008 3
009 6
010 7
011 4
012 3
013 7
014 8
015 3
016 5
017 4
018 8
030 3
031 2
054 3
067 8
071 2
073 6
CP1 2

– For the children's part, SPEECON allows a mismatch between prompt and transcription text in a maximum of 40% (=20) of the files for mandatory isolated word items per item code (effectively missing and/or corrupted items included).

OK, max 8 files

2. Files containing truncation and mispronunciation marks

Truncations and mispronunciations (*, **, ~) are counted in the transcriptions of the individual items to obtain 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.

*For the children's part, **262** transcriptions containing a truncation or mispronunciation mark were found.*

4 SPEECH DATA FILES

– The speech files should be coded as PCM, 16 bit, 16 kHz, no compression

OK

– At least 90% of the sessions recorded in the *Children* environment must have a noise range between 30-70 dB(A) (DBA label)

OK, 100%

– For every new environment and position a set of room impulse responses are required:

Recording environment	**Number of measurements respective to positions**	
	***'medium distance'* positions**	***'far distance'* positions**
Children	3	3

OK

5 ANNOTATION FILES

– Each line must be delimited by <CR><LF>

OK

– Mandatory (SAM) mnemonics for label files:
LHD: SAM 6.1
DBN: SPEECON_<LL>
SES: <session number>
DIR: <with backslashes and no final backslash>
SRC: <filename of speech file>
CCD: <corpus code = item code>
REP: <PLC value of the SCC-label>
RED: <recording date, in format DD/Mmm/YYYY>
RET: <recording time, in format HH:MM:SS>
BEG: <begin sample, 0>
END: <end sample>
SAM: 16000 <sampling freq.>
SNB: 2, signed <number of bytes per sample>
SBF: lohi} <sample byte order, meaningless with single bytes>
SSB: 16 <number of significant bits per sample>
QNT: PCM <quantisation>
NCH: 4 <number of channels>
SCD: <speaker code>
SEX: {M|F|UNKNOWN}
AGE: <in years|unknown>
ACC: <regional accent, place of growing up>
SNQ: CHN0=, CHN1=, … <signal quality, SNR, per channel>
MIP: CHN0=CLOSE_HEADSET, CHN1=CLOSE_LAVALIER, CHN2=MEDIUM, CHN3={MEDIUM|FAR}
MIT: CHN0=SENNHEISER_ME104,CHN1=NOKIA, CHN2={ SENNHEISER_ME64| AKG},CHN3={MBF_HAUN|PEIKER}
SCC: ENV=CHILD,PLC=<env>_<nr>, POS={ CLOSE_WALL_nn | FAR_WALL_nn | NO_WALL_nn }, SIZ={SQM_00_10 | SQM_10_20 | SQM_20-30 | SQM_30+| SQM_100_200 | SQM_200+}, AUD={ON|OFF},DRV=
DBA: <dB (A) value>
LBD:
LBR: <start>, <end>, [gain], [minimum value], [maximum value], <orthographic prompt>
LBO: <start sample>, [centre sample], <end sample>, <transliteration>
ELF:

– Optional (SAM) mnemonics (i.e. may be omitted or left empty)
REG: <region of session>
TYP: orthographic

TXF: <name of the prompt sheet text file>
CMT: <comment>
ARC: <region or area code of session>
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>
EDU: <education level>
SOC: <Socio Economic Status>
HLT: <health>
TRD: <tiredness>
ASS: <assessment code>

– Only legal mnemonics (labels) are used

OK, an extra mnemonic EPI is used for phonetic transcription. Also the optional mnemonics SYS, EXP and DAT are used.

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

OK

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

OK

– Neither illegal attributes nor illegal values should appear

OK

– For MIP and MIT the following arrangements are respected:

SCENARIO	**CLOSE DISTANCE**		**MEDIUM DISTANCE**		**FAR DISTANCE**
Children	Sennheiser ME 104	Nokia Lavalier HDC-6D	-	Mikrofonbau Haun MBNM-550 E-L	Mikrofonbau Haun MBNM-550 E-L
Impulse	Mikrofonbau		Mikrofonbau		

response (_01-_03)	Haun MBNM-550 E-L		Haun MBNM-550 E-L		
Impulse response (_04-_06)	Mikrofonbau Haun MBNM-550 E-L				Mikrofonbau Haun MBNM-550 E-L

OK

6 LEXICON

A. Formal check

– Check lexicon existence (\<database>\TABLE\LEXICON.TBL)

OK

– The entries should be alphabetically ordered

OK

– Used SAMPA symbols are provided in \<database>\DOC\SAMPALEX.PS

OK

– In transcriptions only SAMPA symbols are allowed

OK

– All SAMPA phoneme symbols should be covered

OK

– Phoneme symbols must be separated by blanks

OK

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

OK

– Each line is delimited by <CR><LF>

OK

– All entries should have at least one phone transcription

OK

– 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)

  OK

– Orthographic entries are as a rule split by spaces only, not by apostrophes, and not by hyphens.

  OK

– Words with *, or ~ should not appear in the lexicon

  OK

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

  OK,

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

  8 entries were found which are not in the database.

– Lexicon contents should be taken from actual utterances, so the entries should exactly match the transcriptions.

  OK

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

  Not provided

B. Content check by expert phonetician (carried out at prevalidation)

The following is a description of a lexicon validation procedure which was carried out on a complete lexicon (both adults and children). This lexicon was delivered together with the prompt sheets for prevalidation. The numbers below, therefore, indicate errors for the combined adults' and children's lexicon.

1000 lexicon entries were checked for phonetic correctness by native speaker phoneticians that were not involved in the original transcription process

The validation of the phonemic correctness of the lexicon entries was organised as follows:

- 1000 entries were randomly extracted from the lexicon;
- Of phonemic transcriptions only the first one was kept;
- The check was carried out at the segmental level only (not on syllable boundaries or stress marks, if provided)
- The check was carried out by a phonetically educated person who is a native speaker of the language
- The given transcription received the benefit of the doubt
- The given transcription was correct if it represented a possible pronunciation of the word (which is not necessarily the most common)
- Each transcription was rated on a 3-point scale: OK; Minor error; Severe error
- A minor error occurred if only one symbol in the transcription was wrong
- A severe error occurred if more than one symbol was wrong

Criteria:

- A maximum of 10% minor errors is allowed.

  OK, 67 minor errors were found (6.7%)

- A maximum of 5% severe errors is allowed. Severe means more than one erroneous symbol in the transcription.

  OK, 4 severe errors were found (0.4%)

7 SPEAKERS

– Check existence speaker database files (SPEAKER.TBL)

OK

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

OK

– Each line is delimited by <CR><LF>

OK

– Each field is separated by [TAB] (ASCII 9)

OK

– A minimum of 50 children is recorded

OK, 50

B. Checks for children

– Ages: for children the following criteria apply:

Age interval:	Proportion of speakers	Requirement
08-10	≥ 30%	Mandatory
11-15	≥ 30%	Mandatory No boys with voice breaks may be included

OK, 22 speakers (44%) in the first and 28 in the second category were found

8 RECORDING CONDITIONS

– Check existence and format of recording conditions tables (\<database>\TABLE\REC_COND.TBL) Required attributes:
  - Session number SES
  - Microphone position(s) MIP
  - Microphone type(s) MIT
  - Scenario code SCC

  OK

– Check existence and format session tables (\<database>\TABLE\SESSION.TBL). Required attributes:
  - Session number SES
  - Speaker code SCD
  - Recording place REP
  - Recording date RED
  - Recording time (of first item) RET

  OK

– Optional attributes SESSION.TBL:
  - Prompt sheet text file TXF
  - Sheet number SHT

  Not used

– For the CHILDREN environment the following restrictions apply

Size (for documentation purpose only)	Place	Number of places	Position	Number of positions per place	Number of children per place and position	Number of children per position category
SQM_00_10, SQM_10_20, SQM_20_30, SQM_30+	CHILD_01, CHILD_02, ...	at least 1	CLOSE_WALL_01, ..., CLOSE_WALL_04	1-4	0- 25	20-30
			FAR_WALL_01,..., FAR_WALL_04	1-4	0- 25	20-30

OK, 10 places were found with max 2 positions per place, max 9 speakers per place and position, and 26 and 24 speakers per position category.

9 TRANSCRIPTION

A. Validation by software tools

– Transliteration 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.)

OK, case-insensitive, capital letters used for spelt sequences only

– Punctuation marks should not be used in the transliterations

OK

– Digits must appear in full orthographic form

OK

– In principle only the following symbols are allowed to indicate non-speech acoustic events:
[fil] [spk] [sta] [int]
Other symbols (and language equivalents) must be mentioned in the documentation

OK

– Asterisks should be used to indicate mispronunciations

OK

– Double asterisks should be used for not understandable parts

OK

– Tildes should be used to indicate truncations

OK

B. Validation by human experts

This validation involves 1000 short items and 1000 long items. 20% of the long items stemmed from the spontaneous speech. The items are proportionally selected from the

adults' database and the children's database. The results of the combined transcription validation for adult and child speakers are presented here.

The following corpus items are considered as short items: single word utterances (application words, single digits, Y/N questions, names, language names, keyboard characters). All other items are considered as long items.

A native speaker of the language performed the check on the speech part of each utterance. The transcription validation of the non-speech symbols (everything between squared brackets) was not necessarily done by a native speaker of the language, but by some-one experienced in listening to background noises and capable to decide which noises should be transcribed or not. The transcriptions in the label files were checked by listening to the corresponding speech files and by correcting the transcriptions if necessary. As a general rule, the delivered transcription should always have the benefit of the doubt; only overt errors should be corrected.

Three types of errors are distinguished:

1. Errors in the transcription of speech
2. Errors in the transcription of non-speech (background noises)
3. Channel mismatch

A channel mismatch means that recordings that are supposedly simultaneous do not contain the same utterance or contain only part of the same utterance. One file of the other channels is linked to each tested file of the close-talk channel in order to test this.

The following error criteria are used:

1. For speech a maximum of 5% of the validated utterances (=files) may contain a transcription error.

2. For non-speech a maximum of 20% of the validated utterances (=files) may contain a transcription error.

3. A maximum of 5% channel mismatches may be found

RESULTS

1. Long items

Transcription errors with respect to speech were found in **36** items. This amounts to **1.8%**, which is below the criterion of 5%.

Errors in the transcription of non-speech were found in **22** items. This amounts to **1.1%** of the items, which is below the criterion of 20%.

Channel mismatches were found in 0 items. This amounts to 0% of the items, which is below the criterion of 5%.

2. Short items

Transcription errors with respect to speech were found in **9** items. This amounts to **0.5%**, which is below the criterion of 5%.

Errors in the transcription of non-speech were found in **23** items. This amounts to **1.2%** of the items, which is below the criterion of 20%.

Channel mismatches were found in 0 items. This amounts to 0% of the items, which is below the criterion of 5%.

3. Overall result

*When long and short item sets were put together, errors were found with respect to the transcription of speech in **45** items. This amounts to **2.3%**, which is below the 5% criterion.*

*Errors in the transcription of non-speech were found in **45** items. This amounts to **2.3%** which is below the 20% criterion.*

Channel mismatches were found in 0 items. This amounts to 0% which is below the 5% criterion.

10 SUMMARY

This database is approved by the SPEECON consortium.

1. Documentation

- ⇒ *README.TXT should only be found on the documentation disk (section 2.2.1).*
- ⇒ *The labels MIP and MIT are missing from Table 12.*
- ⇒ *Not clear which corpus id is used for city and which for street names (section 5.3.13).*
- ⇒ *The symbols for rare phonemes in section 10.2 do not correspond to the ones used in the database.*
- ⇒ *The phoneme /@/ is marked as rare in Table 41, but not listed as such in the paragraph above it.*

2. Database structure, formats and file names

- ⇒ *In the text CDs are referred to whereas the list contains a DVD distribution.*
- ⇒ *The copyright statement includes a reference to CD-ROM.*
- ⇒ *In SUMMAR0.TXT spaces instead of tabs should be used to separate different fields.*

3. Corpus items: design and completeness

- ⇒ *The word “poledne” which was found in DESIGN.DOC was not found in the database.*
- ⇒ *Only 8 repetitions were found for the letter “dlouhé Y” (min is 12).*
- ⇒ *24 instead of 20 different keyboard characters were found.*

4. Speech data files

OK

5. Annotation files

OK

6. Lexicon

OK

7. Speakers

OK

8. Recording conditions

OK

9. Transcription

OK