OrienTel Deliverable D6.2

Specification of Validation Criteria

Project reference number	IST-2001-28373
Project acronym	OrienTel
Project full title	Multilingual Access to Interactive Communication Services for the Mediterranean and the Middle East
Project contact point	Philips Speech Processing Aachen Rainer Siemund, project co-ordinator Kackertstr. 10 D-52074 Aachen, Germany Phone: +49 241 8871 392, Fax: +49 241 8871 149 Email: rainer.siemund@philips.com
Project web site	http://www.orientel.org
EC project officer	Domenico Perotta
Document title	Specification of Validation Criteria
Deliverable ID	D6.2
Document type	Report
Dissemination level	Public
Contractual date of delivery	April 2002
Actual date of delivery	25 August 2002
Status & version	Fnal, v1.2
Work package & task ID	WP6, T6.2
Work package, task & deliverable responsible	NSC
Number of pages	23
Author(s) & affiliation(s)	Dorota Iskra (SPEX) Henk van den Heuvel (SPEX) Oren Gedge (NSC) Sherrie Shammass (NSC)
Additional contibutor(s)	
Keywords	speech, databases, validation, telephony

Abstract	This document contains the specifications of the validation criteria that all the databases within the OrienTel project should fulfil. In this way it constitutes an important framework for WP6 on assessment and evaluation. This document gives an overview of the aspects of the databases which are validated, such as, e.g. documentation, completeness of database items or acoustic quality of the speech signal. It presents a set of tolerance margins, otherwise called validation criteria, which are employed to accept or reject a database. It also outlines the validation procedure which is done in a number of stages.
Additional notes & remarks	

Document Evolution

Version	**Date**	**Status**	**Notes**
1.0	18.04.2002	First draft	Distributed among all partners
1.1	14.08.2002	Pre-final	Updated according to D2.1 v1.6
1.2	25.08.2002	Final	Incorporating comments from partners

Table of Contents

1 Executive summary

This document contains the specifications of the validation criteria that all the databases within the OrienTel project should fulfil. In this way it constitutes an important framework for WP6 on assessment and evaluation. This document gives an overview of the aspects of the databases which are validated, such as, e.g. documentation, completeness of database items or acoustic quality of the speech signal. It presents a set of tolerance margins, otherwise called validation criteria, which are employed to accept or reject a database. It also outlines the validation procedure which is done in a number of stages.

2 Introduction

In order to safeguard the quality of the databases and check their compliance with the specifications, all the databases in the OrienTel project are subject to validation. During validation all the databases are checked against a number of validation criteria. These criteria are derived from the specifications. First of all a number of important and quantifiable specifications are selected. Then tolerance margins are set up which must be fulfilled by all the databases.

The aim of this document is to specify validation criteria that OrienTel databases should fulfil and give an overview of the aspects of the databases that are checked in the process of validation. This document lists the criteria against which the databases are checked and which are employed to accept or reject a database.

The principles of validation and the criteria listed here have evolved over a number of previous SpeechDat family projects. Therefore, a number of documents have been used as a reference here ([1], [2], [3]). This means that the principles and the validation criteria have been extensively tested before. Arabic languages we are dealing with in OrienTel, however, pose new challenges that have to be accounted for during validation as well.

Apart from very specific validation criteria (like the allowed number of missing files) the databases should also fulfil a lot of other requirements that immediately follow from the specifications of the databases. These specifications are related to the database format and structure, to the transcription conventions, speaker demographics, environmental conditions, and the lexicon contents. A summary of or a reference to these specifications is contained in the present document, as their fulfilment is of immediate importance for the acceptability of a database. Details will have to be looked up in the deliverable on database specifications: D2.1 [4]. In case of conflicts specifications as outlined in D2.1 should be complied with firstThe following aspects of the validation criteria are addressed in the sections below:

1. Documentation.
2. Database structure, file names and contents.
3. Database items and completeness.
4. Acoustic quality of the speech files.
5. Annotation files.
6. Lexicon.
7. Speaker information and distribution.

8. Recording conditions.
9. Transcription quality.

The criteria outlined in this report will be transformed into a precise checklist which will serve as the basis for the database validation reports and distributed among the consortium partners.

3 Documentation

Each database must be accompanied by a DESIGN.DOC which is written in English and includes the following information:

- contact person: name, address, affiliation;
- distribution media
  - number of disks;
  - contents of each disk;
  - layout of the disk file system;
- formats of speech and label files;
- file nomenclature and directory structure
- reference to the validation report VALREP.DOC
- speaker recruitment strategies employed
- prompting
  - presentation design (e.g. which items were spread over a recording session to prevent list effects);
  - prompting example for one recording session;
- database design
  - number of items in the prompting material;
  - specification of the individual items of the prompting material including lists of vocabulary for each section;
  - for spontaneous items the texts prompted to the speakers should be included together with an English translation
  - a list of digit forms in the language;
  - connection of prompt items to item numbers in the database (to be provided in the subheadings of individual corpus items); recording platform description;
  - hardware
  - software
  - telephone network
  - telephone handsets

- description of the different recording environments and their distribution in the database
- speaker demographics information:

  - a complete and comprehensive rationale of the regional pronunciation variants that are distinguished;
  - a description of accent regions, how many speakers in each region;
  - which age groups, how many speakers in each group;
  - gender: male, female, children, how many speakers in each group;
- annotation information:

  - procedure used;
  - quality assurance;
  - a list of non-standard and alternative spellings (or reference to file SPELLALT.DOC);
  - standard character set used for transcription (ISO-8859<n> or other if needed)
  - other language-dependent information such as abbreviations, proper name conventions, contractions (July or july, isn't, cannot or can not, etc.);
  - annotation symbols for non-speech acoustic events including the standard defined (i.e., [fil], [spk], [sta], [int]) and other language-specific symbols;
  - markers for mispronunciations, recording truncations, unintelligible speech, GSM distortion;
- lexicon information:

  - procedures to obtain phonemic forms from orthographic input;
  - list of SAMPA phone symbols;
  - list of orthographic symbols for languages not using Latin alphabet;
  - whether or not the transcription and the lexicon are case-sensitive;
  - whether multiple transcriptions are supported;
  - whether frequency information is provided;
  - whether stress information is supplied;
  - 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;
  - tables with frequency of occurrence of phones represented in the phonetically rich material (combined words and sentences) and in the full database (at transcription level);
  - list of rare phonemes;
  - any other language-dependent information or conventions;

- indication of quality assurance: estimated percentage of files that were double-checked by the producer;
- any other information useful to characterise the database.

A template file with section headers and default information to be included into each (sub)section will be distributed among partners by the validation centre (SPEX).

4 Database Structure, File Names and Contents

4.1 File names for label files and speech files and directory names

The databases should comply with the following directory structure:

\<database>\< block>\<session>

Where:

<database>	defined as: <DBname><#><language code>. where <DBname> is ORNTL <#> is 5 for OrienTel <language code> is a 2-letter language code, see [4]
<block>	defined as: BLOCK<NN> where <NN> is a progressive number from 00 to 99. Block numbers are unique on all disks and the same as the first 2 digits used in <NNMM> described below.
<session >	defined as: SES<NNMM> where <NN> is a block number and MM is a session number from 00 to 99.

Table 1: Directory structure

Both signal files and label files have to be put in the terminal node subdirectories.

In addition to the previous structures the following directories are used to store the other (non-speech data) files:

\ *(root)*	README.TXT file containing a short description of the database and the files, DISK.ID file and COPYRIGH.TXT
\<database>\DOC	documentation
\<database>\TABLE	speaker, session, recording condition, overlap and lexicon tables
\<database>\INDEX	index files, i.e. contents file
\<database>\PROMPT	prompt sheets (optional)
\<database>\SOURCE	any source code supplied (optional)

Table 2: Directory structure for non-speech data files

The filenames should correspond to the following template:

```
<dbID><NNMM><CC>.<LL><F>
```

where:

<dbID>	database identification code (00-ZZ) For OrienTel: A5
<NNMM>	recording session progressive number (0000-9999)
<CC>	corpus code (A0-Z9) obtained by collating the corpus and the item identifiers
<LL>	2-character language code (e.g. A1 for Standard UAE Arabic)
<F>	file type code: O=Orthographic label file, A=A-law speech file

Table 3: OrienTel file naming conventions

NNMM in file names may not be in conflict with BLOCK and SES numbers in path name.

4.2 The DOC directory

The following files should be in \<database_name>\DOC:

- DESIGN.DOC
- PLATFORM.DOC
- TRANSCRP.DOC (optional)
- SPELLALT.DOC (optional)
- SAMPALEX.PS
- ISO8859<n>.PS
- SUMMARY.TXT
- SAMPSTAT.TXT
- VALREP.DOC

The validation of the DESIGN.DOC main documentation file is described in section 3. PLATFORM.DOC contains platform specifications. TRANSCRP.DOC contains transcription instructions to the transcribers (in the native language and/or in English). ISO8859<n>.PS is a postscript file containing the ISO-8859-<n> character table used for orthographic transcription. The SAMPALEX file lists the SAMPA symbols used for the phonemic transcriptions in the lexicon together with an example. SUMMARY.TXT contains an overview of all items recorded for each session. SAMPSTAT.TXT is the output of the acoustical check on the speech files performed by each partner. The file VALREP.DOC which contains the validation report is created by the validation centre.

4.3 The TABLE directory

Tables should be in \<database>\TABLE

- LEXICON.TBL

- SPEAKER.TBL
- SESSION.TBL
- REC_COND.TBL
- OVERLAP.TBL

The validation of LEXICON.TBL is dealt with in section 8; the validation criteria for the SPEAKER.TBL are given in section 9, and for SESSION.TBL, REC_COND.TBL and OVERLAP.TBL files in section 10.

4.4 Other directories

The root directory should contain the files:

- README.TXT: ASCII text file containing a description of the files in the database
- README.HTM: with browser access to all documentation directories (optional)
- COPYRIGH.TXT: copyright statement in ASCII
- DISK.ID: 11-character string with volume name

Index files should be in \<database>\INDEX. The obligatory files are CONTENTS.LST, A5TRN<LL>.SES and A5TST<LL>.SES which should have the format specified in [4].

Prompt sheet files (optional) should be in \<database>\PROMPT.

Any delivered program code (optional) should be stored in \<database>\SOURCE.

4.5 Other requirements

All text files should have <CR><LF> at line ends. This concerns all label files, all table (.TBL) files, all index (.LST) files, and all (.TXT) files.

All table files and index files (but *not* SUMMARY.TXT) should report the field names collected in each record as the first row (header) of the file. In this header tabs should be used to separate the fields just like in the rest of the file.

Correct item codes should be used as defined in [4].

Empty files are illegal. This is of special relevance for speech and label files.

For each label file there must be one corresponding speech file and vice versa.

Obviously the database should not be infected by any viruses.

5 Database Items and Completeness

5.1 Mandatory items specifications

It will be checked if all mandatory items are recorded. The mandatory items for each database are listed in the tables in [4]. Each individual item should meet the specifications in [4].

The checks are carried out at two levels: the prompt level and the transcription level, the latter is marked in italics in the table below.

corpus ID	# of items FL	# of items MSA	# of items MCA	Utterance description	Validation criteria / Min coverage
	2	**3**	**2**	**isolated digit items:**	
I1	1	1	1	• single isolated digit	*Per digit: (80% of # sessions) / # digit forms*
B1	1	0	1	• sequence of 10 isolated digits in one utterance (written in digits)	*Per digit: (80% of # sessions) * 10 / # digit forms*
B1-B2	0	2	0	• sequence of 5 isolated digits in one utterance (10 different digits)	*Per digit: ((80% of # sessions) * 2 * 5) / # digit forms*
	4	**7**	**4**	**digit/number strings:**	*Per digit: ((80% of # sessions) * average # digits in all strings (30)) / # digit forms*
C1	1	1	1	• prompt sheet number (6-digits including any check digit))	String length: 6 digits
C2	1	0	1	• telephone number (usual way, 6-15 digits) including local, fixed and GSM numbers	String length: 6-15 digits; GSM numbers included
C3	1	0	1	• spontaneous telephone number	
C4	1	0	1	• credit-card-like number (14-16 digits, including any check digit)	Length: 14-16 digits
C5	1	0	1	• 6-digit PIN code	Length: 6 digits
C2-C7	0	6	0	• strings of 4 digits in written format (24 digits in all)	Length: 4 digits
	2	**2**	**1**	**natural number**	
N1-N2	2	1	1	• natural number in written format for MSA	
N2	0	1	0	• string of natural numbers in written format	
	1	**2**	**2**	**money amounts: (local, Euro, $, £, whatever appropriate)**	
M1	1	0	0	• currency amounts, mixed sizes, in local and foreign currencies	40-60% local currency (main currency word)
M1	0	1	1	• currency amount, mixed sizes and units in local currency	

M2	0	1	1	• currency amount, mixed sizes and units in foreign currency (Euro, $)	
	2	**2**	**2**	**yes/no questions:**	
Q1	1	0	1	• predominantly *yes* including ‘fuzzy’ yes/no (spontaneous)	*60 % * # sessions*
Q2	1	0	1	• predominantly *no* including ‘fuzzy’ yes/no (spontaneous)	*60% * # sessions*
Q1	0	1	0	• the word "yes"	*80% * # sessions*
Q2	0	1	0	• the word "no"	*80% * # sessions*
	3	**3**	**4**	**dates:**	
D1	1	0	1	• birth date (spontaneous)	
D2	1	1	1	• prompted date phrase, word not digital format (Western calendar), **note:** be careful about local names for months in Levantine area	*Per month name: 80% # sessions / # months* *Per day name: 80% # sessions / # days*
D4	0	1	1	• prompted date phrase, in word not digital format (Islamic calendar)	*Per month name: 80% # sessions / # months* *Per day name: 80% # sessions / # days*
D3	1	1	1	• relative and general date expression	From a set of 50
	2	**1**	**2**	**times:**	
T1	1	0	1	• time of day (spontaneous)	
T2	1	1	1	• prompted time phrase in analogue form	From a set of 20
A1-A6	**6**	**6**	**6**	**application keywords/keyphrases**	From a set of 25-100 words (25 function words + up to 3 synonyms =100) *Per word: 80% (# sessions * 6) / # target words*

E1	**1**	**1**	**1**	**word spotting phrase using embedded application words**	*Per word: 80% (# sessions) / # target words* Max phrase length 5 words Minimum 2 phrases per word for DB<=500 speakers Minimum 4 phrases per word for DB>500 speakers
	5	**3**	**5**	**directory assistance names:**	
O2	1	0	1	• city of birth/growing up (spontaneous)	
O3	1	1	1	• most frequent cities both local/foreign	From a set of 150
O5	1	1	1	• most frequent companies/agencies	From a set of 250
O7	1	1	1	• personal name (first name and family name	From a set of 150
O1	1	0	1	• personal first name (spontaneous)	
	3	**2**	**3**	**spellings (letter/alphabet strings, for this prompt sheet written as characters *a, b, c* …)**	
L1-L2	0	2	0	• string of 4-letter sequences (written out as Aleph, Bae, Jim, etc., standard pronunciation)	*Per letter: 80% of # sessions * 4 * 2 / # letters*
L1	1	0	1	• real/artificial words to maximise letter coverage	*Per letter: 80% of # sessions * 15 (average # letters in all spelt items) / # letters*
L2	1	0	1	• spelling e.g. of directory assistance city name	
L3	1	0	1	• spelling of personal first name (from same set as personal first names + family names)	
W1-W4	**4+**	**4+**	**4+**	**phonetically rich words**	*Minimum 100 repetitions per phoneme* Maximum 5 repetitions of each word
S1-S9	**9**	**9**	**9**	**phonetically rich sentences**	*Minimum 500 repetitions per phoneme* Maximum 10 repetitions of each sentence
	2	**4**	**2**	**spontaneous (for control)**	

X1	0	1	0	• city of childhood/growing up (spontaneous)	
X2	0	1	0	• caller age	
X3	1	1	1	• place of call (environment)	
X4	1	1	1	• phone type (mobile, fixed, handset, handsfree)	
	0	**0**	**1**	**free spontaneous speech**	
F1	0	0	1	1 out of a list of 3 topics	
	47+	**49+**	**49+**	**TOTAL utterances**	

Table 4: Validation criteria for foreign language (FL), Modern Standard Arabic (MSA) and Modern Colloquial Arabic (MCA)

5.2 *Validation of missing items*

For each database it will be checked if all mandatory items are present in sufficient quantities.

- A maximum of 5% of the files of each mandatory item (corpus code) may be *effectively missing*;
- A maximum of 7% of the files of each mandatory item (corpus code) may be *effectively missing* or contain *corrupted* speech only;
- A maximum of 10% of the files of mandatory isolated word items may show a *mismatch* between prompt and transcription text; this percentage includes the effectively missing and corrupted files.

Effectively missing files are: absent files, and files containing only non-speech (i.e., noise symbols between square brackets and/or items marked as being unintelligible) according to the transcriptions. Files with only *corrupted speech* are files for which *each* word is mispronounced or truncated according to the transcription.

For the isolated word items a further comparison of prompt and transcription is made. In case the word in the prompt does not appear in the transcription (no speech at all or only another or other word(s) instead), then this should be considered as a mismatch. A maximum of 10% of the files may be *mismatching* in this way. It is obvious that effectively missing and corrupted files contribute to this count as well. If the word is present but is transcribed as mispronounced, truncated or unintelligible, then it is *not* considered as a mismatch.

The following corpus items are involved in this mismatch check:

- application words (A1-6)

- isolated digit (I1)

- city name (O3)

- company name (O5)

- person name (O7)

- yes/no in MSA (Q1-2)

- phonetically rich words (W1-4)

A count of isolated word items that are mispronounced, truncated or unintelligible will be done in order to obtain an idea of probably useless data. This will not be used to reject or approve a database, but it will be supplied as supplementary information in the validation report.

The checks on item completeness rely on a correct transcription of the speech. To verify the transcriptions themselves, a manual check on 2000 of the transcriptions will also be carried out (see section 11).

Similar items may compensate for each other in order to meet the completeness criteria. Items with the same corpus identifier can compensate for each other. Exceptions are:

- yes/no questions
- different types of names
- I1, B1-2, and C1-7 can compensate for each other although they do not have the same corpus identifier.

5.3 Validation of missing words/digits

The check on the completeness of each corpus code described in section 5.2 is accompanied by a detailed completeness check of individual word items within the corpus.

These checks are carried out ***at transcription level***. A word is counted as present if it is in the transcription, even if it is truncated or mispronounced. Only if the word is not present in the transcription, is it considered as missing.

As a general rule it is stated that at least 80% of the maximum achievable word tokens should have been recorded. The checks which are carried out at the transcription level are indicated in italics in Table 4.

For the phonetically rich words and phonetically rich sentences the following criteria apply:

- there are at least 100 repetitions of each phoneme in phonetically rich words at transcription level
- there are at least 500 repetitions of each phoneme in phonetically rich sentences at transcription level
- rare phonemes are an exception to this since they appear mainly in loan words
- a max. of 10% of all phonemes in the language may be rare.

6 Acoustic Quality of the Speech Files

The following acoustic measurements are performed on each speech file of a database: file length, mean sample value, clipping rate, and SNR value. These measurements are carried out by each individual partner using SPEX software. The results are passed on to SPEX (as file <database>\DOC\SAMPSTAT.TXT), together with the database to be validated. SPEX summarises the results of these acoustic measurements in the validation report by means of histograms. These histograms are generated at the level of both files and directories (calls).

The histograms are presented in the validation report just 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 a given application.

7 Annotation Files

Checks will be performed to make sure that:

- Correct labels and correct accompanying values are used
- There are no empty label files
- Each line is delimited by <CR><LF> (DOS format)

The correct labels and field values are described in [4].

8 Lexicon

8.1 Format checks

For the lexicon table the following checks are carried out:

- Format check
- All and only SAMPA phoneme symbols are used
- The lexicon contains all words in the transcriptions except distorted words (i.e., mispronounced or truncated words)
- If tagging is supplied, check that all tag symbols are defined and only those symbols are used

The format of the lexicon is described in [4].

The lexicon should be complete. The completeness check is carried out on orthographic transcriptions in the label files in order to find out if all the transcribed words are in the lexicon. Undercompleteness is not permitted, overcompleteness is.

8.2 Validation of phonemic transcriptions

1000 lexicon entries will be checked for phonetic correctness by native speaker phoneticians that were not involved in the original transcription process, or by comparing with other available pronunciation lexicons.

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

- 1000 entries are randomly extracted from the lexicon;
- Only the first phonemic transcriptions is kept in case of multiple transcriptions;
- The check is carried out at the segmental level only (not at syllable boundaries or stress marks, if provided)
- The check is carried out by a phonetically trained person who is a native speaker of the language
- The given transcription receives the benefit of the doubt

- The given transcription is correct if it represents a possible pronunciation of the word (which is not necessarily the most common)
- Each transcription is rated on a 3-point scale: OK; Minor error; Major error
- A max. of 10% minor errors are allowed; and a max. of 5% major errors are allowed
- A minor error occurs if only one symbol in the transcription is wrong
- A major error occurs if more than one symbol is wrong

Since only a sample of 1000 entries are evaluated, the detected errors give the following confidence intervals when extrapolated to the entire DB.

Error percentage	Confidence interval
5%	3.6% - 6.4%
10%	8.1% – 11.9%

Table 5: Confidence intervals

9 Speaker Information and Distribution

The speaker table file SPEAKER.TBL should have the format specified in [4]. The format requirements can be summarised as follows:

- Each line should end with <CR><LF>
- [TAB]s are used between field values.

The obligatory fields which SPEAKER.TBL must contain are also specified in [4]. It should be stressed here, however, that the speaker code must be unique and independent of the recording session.

The minimum number of speakers to be recorded is database-dependent and can be found in [4]. Speaker distribution criteria are summarised below.

9.1 Speaker gender

- Gender misbalance may be 5% at maximum for the *total database*. This means that the proportion of sessions from male and female speakers must be in the interval 45-55% for both genders for all the age categories together.
- Gender misbalance may be 5% at maximum for *each network*. This means that the proportion of sessions from male and female speakers must be in the interval 45-55% for both genders for all the age categories together.
- For each *recording environment* the proportion of each gender should be between 30% and 70%.

9.2 Speaker age

The following validation criteria are applied for speaker age distribution:

Age interval	Proportion of speakers
16-30	≥ 30%

31-45	≥ 20%
46-60	≥ 10%

Table 6: Speaker age distribution

9.3 *Speaker accent distribution*

Dialect requirements concern only *colloquial* varieties of the language and apply to the whole database.

- Each accent region is represented by at least 20 sessions.
- Furthermore, each accent region is represented by (with a max deviation of 50%):

  #dialect speakers = #speakers / #dialect regions

  e.g., if #dialect speakers=20, between 10 and 30 speakers from a given dialect must be recorded.

Speaker balances are validated by checking the label files and counting how many speakers were recorded in each category. The result is then compared to the information in the database documentation (DESIGN.DOC) and the speaker and session tables.

10 Recording Conditions

The session table (SESSION.TBL), the recording condition table (REC_COND.TBL) and the overlapping speakers table (OVERLAP.TBL) files should have the format specified in [4]. Moreover, they should contain the fields which are specified in [4].

OVERLAP.TBL is a table of 3 (or 2) columns for FL, MCA and MSA containing speaker codes. A maximum of 15% (263 speakers) overlap is allowed (except for Saudi Arabia: 17%. i.e. 298 speakers), except between MSA and MCA where no overlap is allowed. In numbers:

- 163 speakers between MCA and FL
- 100 speakers between MSA and FL
- 300 speakers where only two language varieties are collected.

The environment requirements are as follows:

Network	Environment	Speakers
Fixed 30% ± 5%	Home / Office	>= 75%
	Public place / Booth	optional
Mobile 70% ± 5%	Home / Office	>= 20%
	Public place / Street	>= 20%
	Vehicle	>= 15%
	Hands free car kit	>= 5% (optional)

Table 7: Environment distribution

11 Transcription

For transcription validation 2000 utterances are selected randomly and their transcription is checked manually. Transcription validation of speech is carried out by a trained native speaker of the language concerned, who did not participate in the original transcription process. The transcription validation of the non-speech symbols is not necessarily done by a native speaker of the language, but by someone experienced in listening to background noises and capable of deciding which noises should be transcribed or not. The transcriptions in the label files are checked by listening to the corresponding speech files and by correcting the transcriptions if necessary. As a general rule it is maintained that the delivered transcription should always receive the benefit of the doubt and that only overt errors should be corrected.

11.1 Type of errors

Two types of errors are distinguished:

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

Errors in the transcription of truncations, mispronunciations, word fragments and unintelligible fragments are counted as errors in the transcription of speech. Only errors in the transcription of non-speech acoustic events (i.e., in [fil], [spk], [sta], and [int]) are counted as non-speech transcription errors.

11.2 Transliteration

The following validation criteria are applied to orthographic transcriptions:

- Transliterations are case-sensitive unless specified otherwise in the documentation
- Punctuation marks should not be used in the transliterations
- Digits and numbers must appear in full orthographic form
- In principle only the following symbols are allowed to indicate non-speech acoustic events: [fil], [spk], [sta] and [int].
- Asterisks should be used to indicate mispronunciations
- Double asterisks should be used to indicate unintelligible arts of utterances.
- Tildes should be used to indicate recording truncations (and can therefore only appear at the beginning and/or at the end of the utterance, unless there is a drop-out)
- The percent sign (%) should be used for typical GSM distortions. The full description of the relevant transcription conventions can be found in [4].

These criteria are checked both automatically on the *full* database, and by the native speaker on the *subset* for transcription validation.

11.3 Criteria for validation

The main criteria for the validation of the transcriptions by the expert are:

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

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

Only erroneous omissions of noise symbols are considered errors in non-speech.

All non-speech symbols are mapped onto one during validation, i.e. if a non-speech symbol was at the proper location then it is validated as correct, regardless if it is the *correct* non-speech symbol or not. Only stationary noise may not be confused with another type of noise.

The error percentage is determined at item level, not at word level.

11.4 Statistical reliability

A random sample of 1000 utterances from long items and 1000 utterances from short items are checked for each complete database. The only exception is that the free spontaneous item in MCA (F1) is included in the validation set by default.

The following corpus items are considered short items: single word utterances (application words, single digits, yes/no items, names, phonetically rich words). All other items are considered long items.

For each set of 1000 items the (95%) confidence intervals for varying error percentages are:

Error percentage	Confidence interval
5%	3.6% - 6.4%
10%	8.1% – 11.9%
50%	46.9% - 53.1%
95%	93.6% - 96.4%

Table 8: Confidence intervals for 1000 items

And for the full set of 2000 items the confidence intervals are:

Error percentage	Confidence interval
5%	4.0% - 6.0%
10%	8.7% - 11.3%
50%	47.8% - 52.2%
95%	94.0% - 96.0%

Table 9: Confidence intervals for 2000 items

11.5 Spelling check

A formal spelling check of the orthographic transcriptions will not be carried out by the validation centre. It is recommended that partners report the results of a spelling check that they carried out themselves in the documentation of the database.

12 Validation procedures

A database is validated in at least three stages: prevalidation, validation and pre-release validation.

12.1 Prevalidation

The delivery for prevalidation contains two parts:

A. The prompt files as designed for the complete database. Also the lexicon table file for all read items should be included.

B. A complete minidatabase of 10 sessions, uniformly distributed over all the environments (2-3 sessions per environment). This minidatabase contains all speech and label files and all other files that are required for a normal validation, but, of course, tailored to only the included speakers .

The goals of the prevalidation are:

- To detect errors in the database design before the main series of recordings start;
- To stimulate partners to write their database formatting software at an early stage of the project;
- To stimulate the validation centre to write the validation software at an early stage of the project;

The following checks are carried out for part A and B, respectively:

A.	**B.**
The completeness checks as described in section 5, as far as possible for read material.	All checks described in sections 3-10 as far as possible, typically not including completeness checks for corpus items, speaker and rec. environment distributions
The lexicon checks as described in sections 8.	The automatic check on transcription symbols
	Quick check on the use of non-speech transcription symbols by a non-native speaker

12.2 Validation

For validation the procedure is as follows:

1. The producing partner sends the database to the validation centre
2. Immediately after receiving the database, a quick check verifies if all needed files are present and if the label files have the correct format
3. After a successful quick check, all checks described in the previous sections 3-10 are carried out when the database reaches its turn in the queue;
4. The result of the validation, the validation report VALREP.DOC is sent to the producing partner.

5. SPEX asks the producer for clarifications of deviations observed during validation and communicate both deviations and clarifications to the consortium. The consortium decides about the approval of the database. In case SPEX does not find any serious deviations, the database is accepted without voting.

The prevalidation and validation costs are estimated as follows:

Infrastructural costs (specifications, implementation, management, travel/subsistence)

	44 k€ /8 partners =		5,5 k€
Contribution per DB in case of	2 DBs =	2, 75 k€, 3 DBs =	1,83 k€
Prevalidation:			
prompts+lexicon(1,5 kEur) + miniDB			2,5 k€
Validation			3 k€
Pre-release val.			0,5 k€
Total per DB:		8,75 kEur	7,83 k€

The total costs per database are:

- 8,75 k€ for partners submitting two DBs
- 7,83 k€ for partners submitting three DBs.

The total costs per partner are:

- 17,5 k€ for partners submitting two DBs (Siemens, Knowledge and NSC)
- 23,5 k€ for partners submitting three DBs.

All these prices are exclusive of VAT.

12.3 Revalidation

In case a database is not approved or only conditionally approved by the consortium, an extra validation is needed at the costs of the producing partner/owner.

This may boil down to a full revalidation or to the revalidation of parts of the database. The costs of such a revalidation depend on the items to be revalidated. A table for partial revalidation is shown below:

0 data transport from disk and other preparation	0.1 k€
1 completeness of recordings	0.8 k€
2 transcription quality (orthographic)	0.8 k€
4 lexicon (format, phoneme symbols, completeness)	0.1 k€
5 acoustical quality of the speech files	0.3 k€
6 database format/structure and file names	0.1 k€
7 label files (mnemonics and values used)	0.3 k€
8 speaker information	0.1 k€
9 environmental information	0.2 k€
10 documentation files	0.2 k€

As a consequence, the price for a full revalidation is 3 k€ (exclusive of VAT).

If the revalidation work is minor and can be completed in less than one working day, SPEX will not charge the database owner, since then the overhead incurred by the billing process might be higher than the total amount of the invoice.

Full validations of other databases always have priority over revalidations in our planning. Revalidations come at the end of the queue, and are only scheduled when there is an empty time slot.

In case of minor modifications, the validation centre can agree to accept them if an extra section in DESIGN.DOC listing the modifications made after the validation report is written. This should, however, always be discussed with the validation centre first.

12.4 Pre-release validation

When a database is approved, the final master disks must be made. Prior to multiplication, SPEX carries out an additional check on the non-speech data disk. This validation includes the following checks:

- structure of disks;
- version of DESIGN.DOC;
- version of all re-submitted files;
- version of the validation report.

Once this disk is approved, multiplication and distribution of the database can commence.

13 References

[1] van den Heuvel, H.*: Validation criteria*. SpeechDat(II) Technical Report SD1.3.3. Version 1.9, 1997.

[2] van den Heuvel, H.*: Validation criteria*. SpeechDat-Car Technical Report D1.3.1. Version 3.4, 1999.

[3] van den Heuvel, H.: *Definition of Validation Criteria*. SpeeCon Technical Report D41. Version 2.0, 2002

[4] Gedge, O., Shammass, S. et al.: *Speech Database Design*. OrienTel Technical Report D21. Version 1.6, 2002