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

DELIVERABLE 1.1

Validation of Content and Quality of Existing SLR: Overview and Methodology

Deliverable Identification:	**ELRA/9901/VAL-1 Deliverable 1.1**
Title:	**Validation of Content and Quality of SLR: Overview and Methodology**
Release:	2.1
Issued:	21 January 2000
Origin:	ELDA/ELRA
Author:	Henk van den Heuvel, Louis Boves, Eric Sanders (Speech Processing Expertise Center, SPEX,The Netherlands)
Version:	☐ Internal draft ☐ Circulated draft ☐ Final
Status:	☐ Public ☐ Restricted ☐ Confidential
Revision dates:	
Print date:	
Number of pages:	
Distribution:	CEO /ELRA Board members

TABLE OF CONTENTS

1 INTRODUCTION

This report describes several aspects of the validation of SLR (Spoken Language Resources) of existing databases. Throughout this document 'validation' is understood as the quality evaluation of a database against a checklist of relevant criteria. An earlier document [1], known as the Validation Manual for SLR made explicit the criteria that spoken language resources to be distributed via ELRA/ELDA should fulfil, as well as the procedures that ELRA will install to verify that resources offered for distribution do indeed meet these criteria. This report tailors, updates and extends these criteria for the validation of existing databases.

The ISO 9000-3 guidelines for computer software requirements [2] stress the need to develop a quality system and a manual that describes it, together with the need to develop quality plans which show how one intends to fulfil quality system requirements (section 4.2). Furthermore, in section 4.10 of the guidelines it is insisted that procedures should be developed to inspect, test and verify that final products meet all specified requirements. According to section 4.16 the quality records should be filed and well protected. According to section 4.13 of the guidelines procedures should be developed to correct or prevent non-conformities. This deliverable aims to provide some first substantial directives to follow these guidelines for SLR validation.

SLR validation can be performed in two fundamentally different ways: (a) Quality assessment issues are already addressed in the specification phase of the SLR. That is, during the definition of the specifications the feasibility of its evaluation and the criteria to be employed for such an evaluation are already taken into account. (b) A SLR is created, and the validation criteria and procedure are defined afterwards. In this way the risk is increased that the validation of some parts of the specification may become infeasible.

Furthermore, validation can be done in house (internal validation) or by another organisation (external validation). The two dimensions thus identified are shown in the following scheme.

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

Table 1: Four types of validation strategies

(1) in this table is in fact essential for proper database production. Each database producer should safeguard the database quality during the collection and processing of the data in order to ascertain that the specifications are met. A final check (2) should be an obvious, be it ideally superfluous, part of this procedure. In principle, this is the way in which the Linguistic Data Consortium (LDC) operates. Alternatively, or in addition, an external organisation can be contracted to carry out the validation of an SLR. In that case the best approach is that this organisation is closely involved in the definition of the specifications, and performs quality checks for all phases of the production process (3), followed by a final check after database completion (4).

The optimal strategy is to have all (1), (2), (3), (4) done. In fact, this strategy was adopted by the SpeechDat projects, where all producers performed internal quality checks, whilst SPEX served as an independent external validation centre, being closely involved in the specifications and performing intermediate and final quality assessments.

For a limited validation approach the numbers in Table 1 above reflect the order of importance. The internal quality control during production is the most important quality safeguard. In contrast, to have only an external validation after the database is produced is the least preferable option.

Unfortunately, this last case may be typical for the validation of many of the SLR of the present ELRA catalogue. The databases are created (and even sold), but the validation has to be carried out yet. Of course, one may have some faith that internal quality checks in the spirit of (1) and (2) took place for individual databases. This will, hopefully, facilitate the job of the *posthoc* validation to be carried out by the validation centre. This deliverable addresses the formulation of the validation criteria for the SLR in the ELRA catalogue.

The contents of this report can now be summarised as follows.

1. It gives an overview of SLR validation guidelines published or used by other keyplayers in the field (section 2);
2. It updates the validation guidelines given in [1] (section 3);
3. It gives a priority scheme for the validation criteria (section 4);

Deliverable 2.1 [8] in this series describes the procedures for validating and improving the SLR in the ELRA catalogue, and is thus the logical sequel to this report.

2 OVERVIEW OF OTHER SLR VALIDATION ACTIVITIES

The market for large SLR has been strongly growing over the last years. At several places and in various projects large speech databases are being produced. A good example of such activities are the databases produced in the SpeechDat projects. The reason for this growth is that large collections of speech material are needed to build reliably working automatic speech recognisers, even for 'simple' applications like digit recognition. As a consequence, one would expect that quality assessment of such databases is a quite 'hot topic' in the area of SLR production. However, as it appears from our inquiries, the topic is not considered 'hot' or too 'hot' to risk burnt fingers.

Many organisations who are active in disseminating information on SLR and guidelines to produce them are considerably less active in (their reporting about) the validation of such SLR.

The WWW pages of COCOSDA at http://www.itl.atr.cp.jp/cocosda/ do not contain any information about SLR validation. Explicit questions as to validation activities addressed to N. Cambell did not result in further information on the topic.

The Expert Advisory Group on Language Engineering Standards (EAGLES) is not active in SLR validation, as appears from a search of their Web site at http://www.ilc.pi.cnr.it/EAGLES96/guide/guide.html and explicit questions to their representatives Christoph Draxler and Dafydd Gibbon.

Also a query for SLR validation activities at CSLU (Center for Spoken Language

Understanding), see http://cslu.cse.ogi.edu/, rendered unsuccessful. Validation of SLR was not part of the Survey of the State of the Art in Human Language Technology [3], perhaps because in 1996 validation of SLR was not really state of the art.

One of the main actors in the field is the Linguistic Data Consortium, LDC. Stefanie Strassel of LDC provided us with the following information. LDC does not validate SLR produced by others (external speech databases). In the exceptional case, when a corpus from an outside source is published, most of LDC's Quality Control effort goes into normalizing the formatting of the corpus. That might include things like:

- checking that all components of the corpus mentioned in the documentation are present;
- checking that all components are formatted as stated in corpus documentation;
- checking all supporting components (like tables of speaker attributes) are consistently formatted.

When a problem exists with the formatting, LDC takes steps to remedy the problem before distributing the corpus. That might entail converting the current format into something more user-friendly, or it might entail creating some supporting tabulation to give easier access to the corpus (for instance, a table linking Call Ids with Speaker IDs where that table doesn't already exist).

In the normal case, however, LDC produces SLR itself and quality assessment is integrated in the production protocol of the SLR (indicated by (1) and (2) in Table 1). For instance, every transcription of a speech utterance is checked, and afterwards another 5% of the data is "spot checked" by the team leader; the performance of individual annotators is monitored daily; the annotators receive regular personalised feedback; there are weekly meetings and e-mail lists for the annotators. With respect to the production of the pronunciation lexicon each word is reviewed to check for transcription errors; phonemic transcriptions are generated by rule, and hand checked, or entirely entered by hand, if necessary. After the production cycle, but prior to publication, sanity checks are carried out, on e.g. speech and text file headers, illegal characters, symbols, words, missing attributes, file sizes, plausible word/second rates. For each database produced by LDC, users can report bugs via LDC Online (at http://www.ldc.upenn.edu/). The report is submitted to the responsible technician for checking and, if needed, for rectification.

The SpeechDat projects are a typical example where database validation was an integral part of the project (http://www.speechdat.org/). All databases are validated by an independent organisation, which was actively involved during the specification cycle of the project. To this end, rather extensive validation criteria and protocols were developed [4,5,6]. Those in [4] served as background for ELRA's validation manual for SLR [1].

3 VALIDATION CRITERIA

Below the minimum content and quality requirements for SLR, as we see it, are listed. The criteria outlined in [1] are taken as the basis. However, the criteria in [1] were typically set up for new SLR, whereas we deal here with the case of existing databases which should undergo a *posthoc* validation. This calls for a somewhat other approach. First of all, the basic SLR design, the recorded materials, the speakers participating and the speaker settings/environments are *faits accomplis,* which can therefore not be changed anymore, whereas, in contrast, many additional information, documentation, and formats can still be adapted. The orthographic annotation presents a specific case in this respect. In principle, all annotated text can be modified, however, practical reasons, such as cost and time considerations, preclude such an operation in most cases.

Validation criteria come in three distinct categories:

1. Documentation, addressing the written design specification that should accompany every database; this includes issues related to the number of speakers, selection of speakers, recording conditions, etc. The documentation should include an explanation and motivation of the decision that were made in designing and building the database. For obvious reasons, the documentation cannot be evaluated in an automatic way.
2. Formal and technical criteria, addressing issues like the medium on which a database is delivered, the structure of the directory trees, the format of the speech files and of the annotation files, contents lists, speaker tables, etc. Formal and technical criteria are by definition language independent and most are amenable to (semi-)automatic checks
3. Content related criteria, addressing among others the prompting material, the orthoghraphic transcription conventions, the phonemic transcriptions included in the lexicon, etc. Content related criteria are to a large extent language dependent, and are not amenable to (semi-)automatic checks.

Documentation is of a paramount value for a proper validation. The documentation accompanying a database should clearly describe the standards to which the database was collected. By doing so, it also defines the standards by which the database should be evaluated and validated. Thus, it is explicitly acknowledged that there is, as yet, no fixed set of standards that every speech database must adhere to (except the very basic ones detailed in this document). Rather, a specific spoken database will always have been built with some specific goal in mind, that determines the decisions made in the design to a very large extent. It is up to the user (buyer) of a database to decide whether a specific database is appropriate for the research or development project to be undertaken.

In all cases the major starting point for the validation will be the documentation provided with the database itself, and validation will reflect the degree to which the database meets the goals set forth in the documentation. For each database that cannot meet the basic criteria explained in this document, it must be clearly motivated why conformance was not possible or not necessary. At the same time it must be clearly motivated why, for whom and for what types of research, development or applications the database is of value.

The first source of information about a database is given by the provider when he fills out the SLR description form developed by ELRA. This contains a concise description of the database contents and formats. This form should be viewed as an important first aid to asses the quality of a database. SPEX created an updated version of this form to be used for future providers. A copy of this updated form is contained in appendix A of this deliverable.

In the specific case where existing databases are validated, ELRA can request the producer or owner to add missing information which is valuable for the database, especially if it can be easily generated. For this reason, the requirements for the documentation and other meta-information can be quite considerable, since the addition of this type of information is generally not time-consuming or expensive.

The list of validation criteria listed below, is taken from [1] and modified where considered appropriate. It was tuned to the validation of existing databases, for which less detailed and restrictive requirements can be set. As a result, the explicit SpeechDat formats and standards were removed and replaced by more general criteria. Still, the remaining criteria can be considered as the minimal requirements that a database should meet. It should be evident from this account that standards and validation criteria for *new* databases should be retrieved in [1], and not in the present document.

Despite this, it is very well possible that extremely valuable databases continue to be distributed through ELRA that do not meet some of the basic criteria detailed below. This may be because it is the only database available for a given language, or for a very specific task. ELRA will include these databases in its catalogue, properly flagged to be recognisable as 'non-conformant'.

3.1 Documentation

Each database offered for distribution via ELRA must come with a documentation in English. Although it is appreciated that in many cases English is not the native language of the authors, the text should meet high standards of quality, because the documentation is the starting point for validation (and usage!) and therefore must avoid any unnecessary ambiguity. For databases relating to other languages than English, complete documentation in the source language is recommended, in addition to the documentation in English.

The documentation must be provided in printed form; at least one clearly readable copy must be provided to ELRA. This is necessary to verify the correct processing of the electronic copies of the documentation from the media on which the database is delivered.

In addition, a readme file should be in the root directory of each medium, describing all files (including the documentation files) contained in the database.

Documentation in electronic form must be either in flat ASCII or in the form of a Word document. Other formats, specifically page description formats like PostScript, are not desirable, because the documentation text is used to derive essential validation information and criteria. Anyway, the readme file should specify the version of the text editor that has been used to create the documentation files.

For each database the documentation must contain at least the following information:

3.1.1 Administrative information

A database should be accompanied by suitable administrative and technical information, including [1]

- contact person: name, address, affiliation, and position in the organisation;
  address information must be complete, including telephone and fax numbers and e-mail address.
  Information about the position of the contact person in the organisation is essential, to ensure that requests regarding the database will be routed to the appropriate department in a large company, also after the original contact person has left, or has changed position.
- the number of CDs or other physical media (e.g. DAT tapes);
  this information is obviously needed to enable ELRA to check the package for completeness
- if the database is delivered on another medium than CD-ROM or DVD the written documentation should detail all specifications of the alternative medium, e.g. type of tape,

[1] Whenever a bullet list is given in section 3, the order of enumeration tries to reflect the order of importance, although, in principle, each listed item should be considered as an essential requirement.

type/level of tape drive used to write the tapes, etc.

- the contents of each CD, DVD or tape;
  this information is needed to allow ELRA to check the completeness of the information on each individual medium
- copyright statement
  A copyright statement should be provided in the root directory of each media comprising the database;
- information on IPR (Intellectual Property Rights)
  the documentation should clearly state the rights and restrictions on the use of the database for ELRA and for ELRA's customers. Distinction may be made between use for research and commercial use, or between distribution in and outside of Europe

3.1.2 Technical Information

The minimum documentation must comprise sufficient technical information about the contents and structure of the database to allow ELRA and its customers efficient and effective access to the data. This part of the documentation must include:

- the layout of the CD-ROMs, DVDs or tapes;
  the directory structures must be clearly described.
- file nomenclature and directory structure;
  both classifying and identifying codes for directory names and file names may be used. If classifying file names are used, unambiguous explanations of the classification scheme must be provided. If identifying codes are used, full and complete specifications of the mapping of the codes onto the contents of the directories and files must be provided.
  File names should adhere to the limitations imposed by the most widely used operating systems. In no case may filenames contain blanks, even if that is allowed in some modern Windows based systems.
- formats of the signal files and of the label files
  file formats should adhere to published standards. ELRA prefers the use of the standards defined by the SAM project and amended by SpeechDat, both for signal files and annotation files. Other standards, like the NIST standard and industry standard formats for audio files, are acceptable as well. Database providers who use other formats may be requested to provide tools for converting their formats to SAM format.
- coding;
  databases offered for distribution via ELRA should not use proprietary coding algorithms, since these would restrict access to the data. If coding methods adhering to published standards are used (e.g., Mu-Law or A-law) the coding method should clearly be stated.
- compression;
  the use of generally available and well-supported compression techniques are allowed. The code implementing the compression-decompression algorithm must be made available, except when standard compression tools like `ZIP` are used.
- sampling frequency;
  for each of the signal types in the database the sampling frequency must be specified. For databases comprising several parallel channels the sampling frequencies may differ between the channels. The sampling frequency should be adequate for the type of signal recorded. The use of 'exotic' sampling frequencies is not recommended.
- number of bits per sample;
  the documentation must state the number of bits per sample as well as the byte order (for codes comprising more than 8 bits/sample) after decompression. For databases comprising multiple parallel channels this information must be provided for each channel.

- Multiplexed signals; in case the SLR contains multiplexed signals the exact (de)multiplexing algorithm should be given. Preferably, the software needed to obtain the demultiplexed signals should accompany the database.

3.1.3 Database Contents

The documentation should clearly describe the purpose with which the resource was collected and the types of speech material recorded (e.g. multi-party conversations, human-human dialogues, human-machine dialogues, read sentences, connected and/or isolated digits, isolated words, etc. In addition to the types of speech the linguistic contents of the speech and the selection of the speakers must be described.

3.1.3.1 Linguistic contents

For speech material that is expressly prompted, the following information should be provided:

- a specification of the individual items of the prompting material;
- specification (and motivation) for the sheet design (e.g. how items were spread over the sheet to prevent list effects);
- in the case of text prompting, an example prompting sheet should be provided; in the case of speech prompting the text version of the prompting speech must be provided.

For speech material that is not expressly prompted minimum requirements for the specification of the contents of the material are more difficult to give. Below a number of suggestions is given regarding information that is typically necessary for several types of SLR:

- For multi-party conversations, the number of speakers participating, the topic(s) discussed, the type of setting (formal or informal).
- For human-human dialogues the type of dialogue (problem solving, information seeking, chat, etc.), the relation between the speakers, the topic(s) under discussion, the degree of formality, and the use of scenarios (if any).
- For human-machine dialogues the domain(s) and topic(s) under discussion, the dialogue strategy followed by the machine (system driven, mixed initiative), the type of system (test, operational service), and the instruction of the speakers (if any).

3.1.3.2 Speaker information

The information about the speakers that is available may differ between databases. It depends on the way in which speech has been recorded. For instance, if speakers calling to an operational service were recorded, their identity nor any other relevant characteristics may be known. However, if the speakers were expressly recruited for inclusion in the database, the following information should be considered as minimally required:

- speaker recruitment strategies
- number of speakers
- distribution of speakers over the categories of
  - sex

  - age
  - dialect regions, including a reasoned description of the regional pronunciation variants that are distinguished. A description of the criteria used to define and distinguish regional variants must be included. If these criteria have been described and documented in the open literature, a reference to the publication(s) is sufficient.

Speaker information is especially important in databases that have been recorded for the purpose of speaker recognition (identification or verification) research.

3.1.3.3 Recording platforms and recording conditions

The information about the recording procedure that can reasonably be expected to be of importance and that therefore must be provided includes for telephone databases:

- recording platform and telephone link description (analogue, digital);
- network from which the call originated;
- environment in which the caller was speaking (quiet office, pay phone in public location, etc.)
- handsets

For recordings in automobile environments information must be included on

- the type of car or public transport vehicle the caller was in during the recording
- the (average) speed of the vehicle
- the status of the windows (open/closed)
- for cars the type of pavement of the roads.

Other relevant topics are

- the position of the speaker (driver or co-driver)
- audio equipment playing during the recordings,
- the positioning and types of microphones.

It is appreciated that part or all of this information may be uncertain or simply completely absent if the database is recorded in an operational service. Some recording protocols may provide calling line identification (CLI) information, that might be of use. However, the use of CLI information may be restricted by privacy protection laws.

For databases collected without a telephone transmission link between speaker and recording platform the following minimal information should be specified (cf. the EAGLES Handbook on Spoken Language):

- recording platform (analogue or digital).
  - In the case of an analogue platform the effective bandwidth of the recordings must be specified.
  - For digital recordings the effective bandwidth is supposed to be from zero to half the sampling frequency; deviations from this default should be explicitly documented.
  - Number of channels recorded and channel separation.
- microphone information should include

quality control with short delay feedback is essential to guarantee high quality transcriptions (see below);

- selection of annotators
- the criteria for determining the suitability of the annotators must be specified and explained.
- training of annotators. A concise description should be provided of the procedures used to train the annotators for their task.
- transcription tools used. The description of the tools should at least contain an indication of the ways in which spell checkers were used. It should also contain an indication of the extent to which tools were used to 'predict' the transcription, for instance the prompting text in read material, or an automatic speech recogniser for extemporaneous speech.

The transcription quality of the databases must be assessed by means of an additional manual check of the (orthographic) transcriptions. The following guidelines are highly recommended and should be reported in the documentation.

- At least 5% of the transcriptions should be checked.
- The second annotator is a different person than the one who made the first transcriptions. The check of the transcriptions must be carried out by a person who is fluent in the language of the database. If it is not possible to find a native speaker of the language, the person who does the check should at least have near-native command of the language.
- Experience shows that quality of transcription and checking is enhanced if the persons who do the work are highly educated.

3.1.5 Lexicon

Each database should be accompanied by a lexicon comprising all words occurring in the annotations. The description in the documentation should include:

- The format of the lexicon
- Procedures used to obtain phonemic forms from the orthographic input
- Symbols in the transcriptions used as delimiter to obtain the lexicon entries
- An explanation of or reference to the phoneme set used
- Phonological or higher order phenomena accounted for in the phonemic transcriptions
- Case sensitivity of entries (matching the transcriptions)

3.1.6 Statistical information

Minimal statistical information that should be included in the documentation comprises:

- analysis of frequency of occurrence of the sub-word units; frequency information on phones is mandatory; information on the frequency of diphones, triphones, syllables, etc. is optional;
- word frequency tables per item type

Alternatively, the software could be provided that enable the customer to generate these statistics afterwards (See appendix B).

3.1.7 Additional information

The documentation may contain information about

- any other language-dependent information or conventions
- indication of how many of the files were double checked by the producer together with percentage of detected errors;
- any other information useful to characterise the database.

3.2 Formal and technical criteria

3.2.1 Directory names and file names

Both signal files and label files have to be put in the terminal node subdirectories. The directory location of a file should preferably be recoverable from the filename.

The directories listed in Table 2 will be typically available to store the other (non-speech data) files. Obviously the names of the directories need not be the same; this is indicated by the triangular brackets. It is the concept that counts.

\ *(root)*	The readme file, the copyright file
\…\<DOC>	Documentation files
\…\<INDEX>	Index files, e.g. contents file, corpus contents files, corpus list files, ...
\…\<TABLE>	Speaker, session, recording condition and lexicon tables
\…\<SOURCE>	Any source code supplied
\…\<PROMPT>	Prompt sheet if present (with appropriate sub-directory structure if needed);

Table 2: Directory structure for non-speech data files

All these support files should be preferably duplicated on each CD-ROM.

- Empty (i.e. zero-length) speech and annotation files are not permitted.
-

F5

- a summary file which typically contains records which list for each recording trial/session the items recorded and included in the database.
- a (postscript) file containing the character table used for the (orthographic) transcriptions.
- a (postscript) file listing the phoneme symbols used for the phonemic transcriptions in the lexicon
- a list of alternative spellings used for specific words. This must be present, even if no spelling alternatives are allowed in the transcriptions. In that case, the contents of the file may be empty, or just consist of the message "No alternative spellings allowed". Alternatively, a similar remark can be included in the main documentation file.

There is a substantial number of databases that have been collected using other paradigms than used in SpeechDat/Polyphone. For these databases other conventions for the specification of items and item categories must be followed. In databases collected 'in-service' the SpeechDat type of item information may not exist at all, or may be subject to different coding schemes. For instance, recordings made in a menu driven information system might be indicated and classified by means of the menu node in which they were produced. In mixed-initiative dialogue systems an unambiguous concept of menu node does not exist. In these cases a simple scheme for the classification of items may not be meaningful.

- *An additional file, VALREP.TXT, containing the validation report will be created and added by ELRA's validation centre.*

3.2.3 Other directories

The <INDEX> directory should contain a contents list file which lists all information contained in the label files in one big file. This file can then be easily used for queries in the database, without the need to have access to all individual label files. The contents list file contains one record per line, one record corresponding to one recording unit (session). The fields in each record contain the following information:

- full path name
- speech file name
- session number
- item code
- speaker information:
  - speaker code
  - speaker sex
  - speaker age
  - speaker accent
- recording conditions:
  - setting or environment
  - telephone network
  - telephone model
  - microphone type
  - microphone position
  - recording date
  - recording time
- orthographic transcription of the uttered item

- prompted text of the uttered item
- item repetition number (if relevant)

The fields are delimited by [TAB]s. An example of a contents list file can be found in SpeechDat deliverable SD1.3.1, section 7.1 [7].

Tables should be in \…\<TABLE>. Relevant are:

- A lexicon
- A speaker table and/or a session table
- A recording condition table (highly recommended for all databases where recording conditions differed substantially between sessions. No condition information within a session is expected. However, if such information does exist, it should be included here, in the form of identifiable fields in the record describing a session.)

Apart from the lexicon, all table files contain information that can in principle be derived from the contents list file. Therefore, the only obligatory files are the lexicon table and the contents list file. The speaker, session, and recording condition tables are recommended files. It is recommended to include the tool to generate contents list from the label files in the <SOURCE> direcory. This allows the customer to regenerate or customise the contents list. Also the software that generates the table files (if provided) should be included in the <SOURCE> directory.

This approach warrants that corrections in the annotation only need to be hand placed in the label files. By applying the software that generate the content list (and the table files) it is safeguarded that all changes automatically percolate to this meta-level.

The validation of the lexicon is dealt with in section 3.6. Examples of speaker tables, session tables and recording condition tables are presented in [7] and will not be further dealt with here. Prompt sheet files (optional) should be in \…\<PROMPT>. However, if printed prompt sheets have been used to elicit the speech, at least a template of the prompt sheets must be included in the main documentation file. It is highly recommended that all different prompt sheets used for the elicitation are included as well.

All delivered program code should be stored in \…\<SOURCE>.

3.3 Database items and completeness

Below, the completeness and acceptability criteria agreed in the SpeechDat project are given. It must be emphasized that these criteria can only be used as a sensible example of the type of acceptance and validation criteria that must apply. The details of the SpeechDat database design and the attendant acceptance criteria were only established when a fair number of POLYPHONE style databases (including the US English and US Spanish, the Dutch and Swiss French) had already been recorded. Consequently, these early POLYPHONE style databases do not conform to the SpeechDat specifications. However, that does not mean that these early databases would therefore be less valuable.

3.3.1 Completeness

In SpeechDat the following criteria for 'completeness' were agreed:

At least 95% of the files of each mandatory item (corpus code) must be present for a database to be considered complete. As missing files are counted: absent files, and files containing only non-speech or corrupted speech according to the transcriptions. There will be no further automatic comparison of prompt and transcription text in order to decide if a file is effectively missing.

For the mobile database recorded in SpeechDat an additional completeness criterion was specified, that derived from the fact that for this database each of the speakers had to call from four different locations (quiet office or living room; noisy public environment; moving vehicle; standing near a busy road). A recording could only be included in the database if all four recordings of the speaker were available. In this way an exact match was enforced for the four recording conditions.

A comparison of the acceptance criteria should make it clear that different assessment criteria will apply to different types of database. For each specific type the assessment criteria are derived from the purpose with which the database is collected. For a customer the acceptance criteria will be determined by the specific R&D project for which the database is acquired. For the posthoc validations the 95% completeness criterion for each item will be maintained as a sensible measure of completeness. Deviations from this criterion will be reported, but will not render a database invaluable on a priori basis. Obviously, deviations may be compensated for by similar items in the database.

In addition to the formal completeness checks, the following content-related issues need to be checked:

- are all intended words present and in sufficient quantities?
- are all phones present in phonetically rich sentences (if recorded) and in sufficient quantities (the minimum number of occurrences for each phone being typically #sessions/10) ?
- are all digits and numbers in the <u>prompt</u> text of digit and number items in <u>numerical</u> format? This is done to make sure that the way in which the digits and numbers were pronounced was informative for the conventional ways for completing this task in the given language. This could not have been accomplished if the digits and numbers would have been spelled out.
- Proper distribution of digits in number items.
- are formats of numbers correct ?

- is the distribution of letters in spelt items uniform? (i.e. reflecting the distribution in the language with sufficient samples for training for each common letter of the language)

3.4 Acoustic quality of the speech files

ELRA provides software (developed by SPEX) for calculating the following acoustic measurements on each speech file of a database: file length, mean sample value, clipping rate, and SNR value. These measurements should be carried out by each database provider. The results should be included in the database before it is submitted for validation (as file <database>\<DOC>\SAMPSTAT.TXT). The validation office (for instance SPEX) can summarise the results of these acoustic measurements in the validation report by means of histograms. These histograms are generated both on file level and on directory (call) level. The availability of the report generation software can be negotiated with ELRA.

The histograms are presented in the validation report just as they are and not further interpreted by SPEX (unless this is explicitly required according to the specification of the database). In the normal case, the statistics just help the user of the database in deciding which sessions have an acceptable acoustic quality for the application at hand.

3.5 Annotation files

3.5.1 General criteria

The general criteria are meant to guarantee usability of the database on the widest possible range of computer platforms and operating systems.

The specifications in this section are heavily biased by the SpeechDat decision to follow the SAM standards for annotation. Therefore, they should be interpreted as specifications of the type, contents and quality of the information that should be provided by a database supplier. The exact format of the information may be different. Ideally, any other format should allow automatic conversion to the SAM-style format described below. Actually, successful automatic conversion to the SAM style format would prove the conformance of the information with the specifications given below.

- Empty label files should not occur
- All files must contain the same mnemonics

3.5.2 Mandatory information in the label files

The following labels are considered as the minimum mandatory set (the three-letter label mnemonics are just examples taken from SpeechDat, and are obviously not literally required):

DBN: <language>_<database type>
VOL: <language code>_<nr>
SES: <session number>
DIR: <full pathname with backslashes and without 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>
SAM: < =sampling freq.>
BEG: <begin sample, usually 0>
END: <end sample>
SNB: < =number of bytes per sample>
SBF: <sample byte order, meaningless with single bytes>
SSB: < =number of significant bits per sample>
QNT: <quantisation>; in other words: speech coding standard used
SCD: <speaker code>
SEX: <speaker sex>
AGE: <in years/unknown>
ACC: <regional accent, place of growing up>
REG: <region of call>
ENV: <environment of call>
LBR: <orthographic prompt>

LBO: <transliteration>

The following mnemonics are optional:

TYP: orthographic
TXF: <name of the prompt sheet text file>
CMT: <comment>
NCH: < =number of channels contained in the speech file>
ARC: <region or area code of call>
SHT: <sheet number for prompts>
CRP: <recording repetition number; repetition number for multiple pronunciations of the same utterance by the same speaker>
CMP: <compression software used; field should be empty if this mnemonic is used!>
EXP: <labelling expert>
SYS: <labelling system>
DAT: <date of completion of labelling>
SPA: <SAMPA version>
PHM: <telephone model>; alternatively, mnemonics to identify microphone type & positioning and other recording equipment
NET: <network>
EDU: <education level>
SOC: <Socio Economic Status>
HLT: <health>
TRD: <tiredness>
STR: <subjective stress level>
RCC: <recording conditions code>
SNL: <subjective noise level>
ASS: <assessment code>

But they may be considered mandatory for specific types of databases.

In case SAM labels are used in the database, they should adhere to the SAM conventions.

3.6 Lexicon

For the lexicon (in <database>\TABLE\LEXICON.TBL) the following checks are carried out:

- The entries should be taken from the orthographic transcriptions;
- A list of delimiters used to generate the orthographic entries in the lexicon must be provided. Preferably, words are split by spaces only, not by apostrophes, and not by hyphens;
- Each entry should have at least one phonemic transcription;
- The lexicon should be complete. A check is carried out on the orthographic transcriptions in the label files in order to find out if the lexicon is undercomplete or overcomplete. Undercompleteness of the lexicon is not acceptable, whereas overcompleteness is not problematic;
- Words which only occur with a distortion marker may not appear in the lexicon;
- The orthographic lexicon entries should exactly match the transcriptions;
- Frequency information is optional. Also alternative transcriptions are optional, unless the design specifications of the SLR say otherwise;
- The entries should be alphabetically ordered;
- Optional information that may be present in the phonemic transcriptions include: stress, word/morphological/syllabic boundaries.

The lexicon validation is focused on the format of the lexicon table only; the lexicon contents (i.e. the correctness of the phonemic transcriptions) are not validated.

Obviously, changes in the orthographic transcriptions directly affect the lexicon. It is however impossible for the user to regenerate the lexicon after (new) words are added in the transcriptions. A software tool for this makes no sense since it is impossible for the owner to include a database containing all words and their phone transcriptions in the database. Therefore, this regeneration can only be done by the database producer, and quick and efficient ways should be developed to update deficiencies in the lexicon. (This is different from the updating of contents list, which can be directly carried out by the database user if the appropriate software is available).

3.7 Speaker information and distribution

As a general rule the speaker distribution should be in accordance with the speaker specifications laid down in the documentation file.

If such specifications are missing, then SpeechDat rules can be used as a sensible alternative.

In SpeechDat a misbalance of sexes of 5% at maximum has been allowed. This means that the proportion of calls from male and female speakers must be in the interval 45-55% for both sexes. This seems a reasonable criterion to employ for the validation of other SLR as well.

Unless otherwise documented, for speaker ages the following criteria (taken from SpeechDat) will be employed and deviations reported.

Age interval:	Proportion:	Requirement:
<16	>= 1%	Recommended
16-30	>= 20%	Mandatory
31-45	>= 20%	Mandatory
46-60	>= 15%	Mandatory

Table 3: Required age distribution

The age criteria are meant for the whole database; they need not apply, in a more strict sense, for male and female speakers separately.

The balance of regions is validated by checking the speaker file and counting how many speakers called from which region. The result is then compared to the information in the database documentation. As a general rule the region distribution in the database should be the same as the distribution of the population, with a deviation of max. 5% per region, and a minimum of 5% of the speakers in the database per region.

Obviously, specific purpose SLR (e.g. speaker verification databases) require specific criteria.

3.8 Recording conditions

For the recording conditions the same general rule applies as set for the speaker characteristics: The recording conditions should comply with the specifications for the individual database.

In cases where such specifications are not made explicit, the following guidelines from SpeechDat are followed during validation:

At least the following global recording conditions could be distinguished:

1. Home/office
2. Public place (background talking)

and additionally for recordings made via a cellular network:

3. Pedestrian by road side (traffic emission noise) (ENV-value: STREET)

4. Passenger in moving car, bus (traffic immision noise) (ENV-value: VEHICLE)

An even distribution of recordings over the environments should be envisaged. A deviation of 10% from an even distribution is acceptable. (e.g. for a SLR of 250 recordings per environment, a deviation of plus or minus 25 calls per environment is tolerable).

Still, these can only be rough guidelines. Depending on the objectives of the databases the criteria to be fulfilled may be much more stringent. But this needs to decided on a case by case basis.

3.9 Orthographic transcription

3.9.1 Requirements

The following criteria are considered essential for the orthographic transcriptions:

- The conventions should be transparant
- The annotation symbols should be parsable by automatic procedures
- Digits and numbers must appear in full orthographic form
- Background noises should be annotated. These should be distinguished in noises from the speaker (filled pauses and other speaker noises) and non-speaker noises. For the non-speaker noises two categories at least should be distinguished: transient and stationary noises.
- Further, mispronunciations, non-understandable parts, recording truncations and (in recordings over the cellular network) transmission distortions should be indicated, using different markers
- The transliterations are case-sensitive unless specified otherwise in the documentation
- Punctuation marks should preferably not be used in the transliterations, and certainly not if files are cut into single utterance portions

In case the SLR does not comply with these requirements, the validation centre should make an estimate of the efforts involved in transforming the transcriptions to these standards

3.9.2 Procedure

Two types of transcription errors are distinguished:

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

Errors in the transcription of truncations, mispronunciations, word fragments and not-understandable fragments are counted as errors in the transcription of speech. Only errors in the transcription of non-speech acoustic events (e.g, in filled pauses, speaker noises, stationary noises and transient noises) are counted as non-speech errors.

The transcription validation is carried out by a trained native speaker of the language concerned. The transcriptions in the label files are checked by listening to the corresponding speech files and correcting the transcriptions if necessary. As a general rule it is maintained that the delivered transcription should always have the benefit of the doubt and that only overt errors should be corrected. A subdivision is made in long utterances and short utterances, if needed.

Short utterances are typically:

- isolated digit
- time phrases
- date phrases
- yes/no questions
- names

- application words
- phonetically rich words
- spontaneous words

Long utterances are typically:

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

3.9.3 Criteria for validation

The main criteria for the validation of the transcriptions are:

- For speech a maximum of 5% of the validated utterances may contain a transcription error.
- For non-speech a maximum of 20% of the validated utterances may contain a transcription error.

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.

Further, both noise *deletions* in the transcription and noise *insertions* in the transcriptions are counted.

3.9.4 Statistical reliability

When transcription accuracy is reported in terms of percentage agreement, confidence intervals for the measurements must also be reported. Of course, these confidence intervals depend on the number of items that was checked. Table 4 gives confidence measures for the SpeechDat databases, where 1000 short items and 1000 long items are checked for all databases. Since these confident intervals are only dependent on the sample size, and not on the full database size, it is reasonable to use the same sample sizes for other SLR.

Error percentage		
	Long items	**short items**
5%	3.6% - 6.4%	3.6% - 6.4%
50%	46.9% - 53.1%	46.9% - 53.1%
95%	93.6% - 96.4%	93.6% - 96.4%

Table 4: 95% Confidence intervals for a 5% sample containing 1000 short items and 1000 long items.

4 PRIORITY AND CLASSIFICATION SCHEME

Looking at all parts of an SLR that can be validated as listed in section 3 above, it can be seen that validation of some aspects should have higher priority than others.

Essential for an SLR that can be classified as such is the quality of the documentation coming with it. Furthermore, only a proper transcription of the speech qualifies the database as more than a mere collection of speech recordings. Next, any automatic speech recogniser cannot sensibly be trained nor tested if a phonemic lexicon is missing in the database. On the same level in the priority listing is the acoustic quality of the speech files. Although the desired quality may to a great deal depend on the wishes of the customer, it is obvious that recordings with rubbish disqualify for being included in a speech database. In summary, we consider documentation, transcription, lexicon, and good speech signals as the core ingredients of an SLR. These four ingredients then have the highest validation priority.

On the second level in the priority listing follow: completeness criteria and distributions of speakers and environments, etc.

The third level of priority concerns SLR aspects that can be easily corrected afterwards, such as the formatting of the annotation files and the directory tree structure and file nomenclature of the database. Of course, errors on this level may be very frustrating when one uses the database, but the important thing for database validation is that they can be relatively easily repaired. In fact, also the documentation files could be considered as part of this third priority level, since they can be easily modified as well. The reason why we in contrast consider documentation as a priority 1 matter is that a good documentation is a prerequisite for a sensible database validation.

In a next step, we can attach quality labels to each aspect of the database. Our quality labels have three possible values: 1. not acceptable; 2. not OK, but acceptable; 3. OK.

Table 5 gives a final overview of the priority weights attached to the SLR characteristics discussed in section 3. SPEX regards this scheme as the key framework to validate the existing databases in the ELRA catalogue.

Database part	**Priority level**	**Discussed in section**	**Quality value**		
			1. Not acceptable	**2. Acceptable**	**3. OK**
Documentation	1	3.1			
Transcription	1	3.9			
Lexicon	1	3.6			
Speech signal	1	3.4			
SLR completeness	2	3.3			
Annotation files	2	3.5			
Speaker distributions	2	3.7			

Recording conditions	2	3.8			
Formats & file names	3	3.2			

Table 5: Quality assessment methodology for existing SLR in ELRA’s catalogue

In practice this methodology entails that we validate all SLR parts dealt with in section 3. For each database part we look at the relevant validation criteria on the basis of the order of importance reflected in the sequence of the bullets in the section mentioned. After the check of this part is completed a quality value 1, 2, or 3 is given. Thus the complete table is filled out. This table will be supplied together with an detailed validation report to the database owner/producer (see [8]). The table can also be included in the SLR information given in the ELRA catalogue. If a category with priority level 1 has a value 1 (not acceptable) immediate action should be taken to have the deficiencies repaired.

5 REFERENCES

[1] L. Boves: *Validation Manual for SLR (Spoken Language Resources)* . ELRA Deliverable D6.1.1., 1998.

[2] ISO 9000-3 Guidelines translated into plain English. http:// www.connect.ab.ca/~praxiom/9003.htm

[3] R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, V. Zue (eds*): Survey of the State of the Art in Human Language Technology*. 1996, http://cslu.cse.ogi.edu/HLTsurvey.

[4] H. van den Heuvel*: Validation criteria*. SpeechDat Technical Report SD1.3.3., 1996.

[5] H. van den Heuvel : *Validation criteria*. SpeechDat Car Technical Report D1.3.1, 1999.

[6] H. van den Heuvel : *Validation criteria*. SpeechDat East Technical Report ED1.4.2, 1999.

[7] F. Senia:*Specification of speech database interchange format*. SpeechDat Technical Report SD1.3.1, 1996.

[8] Henk van den Heuvel, Louis Boves, Eric Sanders: *Procedures to validate and improve existing SLR.*. ELRA/9901/VAL-1 Deliverable 2.1, 2000.

APPENDIX A : UPDATED SLR DESCRIPTION FORM

Description form - PROVIDER

SPEECH

PRODUCER/PROVIDER

Organisation:			
Department:			
Representative:			**Position:**
Contact person:			**Position:**
Address:			
Postal code:		**City:**	**Country:**
Telephone:		**Fax:**	**E-mail:**

COPYRIGHT HOLDER

Organisation:			
Department:			
Representative:			**Position:**
Contact person:			**Position:**
Address:			
Postal code:		**City:**	**Country:**
Telephone:		**Fax:**	**E-mail:**

GENERAL INFORMATION[2]

Full name of data collection:

Short name of data collection:

Resource:

□ Acoustic □ Aero-dynamic □ Physiologic □ Other:

Speech style:

□ Spontaneous □ Read ; from screen, sheet : …………

□ Prepared □ Prompted ; voice prompt, screen prompts : ………………..

Specification (e.g. interview, casual conversation, phonetically balanced sentences, isolated words, etc.)**:**

Speech setting:

□ Dialogue □ Monologue □ Multilogue

[2] Throughout this form multiple blocks may be ticked

Recording environment (e.g. public place, moving vehicle, quiet room)**:**

Recording medium (e.g. digital audio tape, audio cassette, etc)**:**

Microphone type:

Telephone type:

Network type:

□ Fixed □ Mobile GSM □ Mobile other:

Language(s):	**Domain(s):**	**Duration (hours):**	**Number of word types:**

Speaker Specific Information

Sex and number of speakers:

□ Male Number:

□ Female Number:

Total number:

Age class:

□ Children (up to 12) □ Teenagers (12-18) □ Adults (over 18) □ Unknown

Comments:

Origin of speakers:

□ Native □ Non native

Comments:

Information included about:

□ Place of living □ Place of birth □ Place of (secondary) education □ Dialect/accent

Comments:

Other speaker information included:

□ Speaking/hearing impairments □ Length □ Weight □ Smoking habits

□ Trained/untrained speakers □ Education level □ Other:

Linguistic Information and Segmentation

Linguistic annotation:

□ Orthographic □ Phonemic □ Phonetic □ Syntactic □ Semantic

□ Prosodic □ Morphological □ Other:

Level of segmentation:

Lexicon

Lexicon included (yes/no) :

Size (number of lexicon entries):

Format:

□ ASCII □ SGML □ TEI □ Other:

Pronunciation lexicon:

□ Available □ Not available

Transcriptions:

□ Canonical only □ Canonical + alternative pronunciations

□ Automatically generated □ Checked manually □ Generated fully manually

Phoneme set:

□ IPA □ SAMPA □ CPA □ Other:

Technical Information

Signal coding:

□ A-law □ μ-law □ linear □ Other:

File format:

□ AIFF □ wave □ headerless

□ SAM □ NIST/sphere □ Other:

Sampling rate (kHz):

□ 8 kHz □ 16 kHz □ Other:

Number of quantisation bits:

□ 8 bit □ 16 bit □ 32 bit □ Other:

Compression tool:

□ None □ zip □ shorten □ Other:

Number of recording channels:

□ 1 (mono) □ 2 (stereo) □ Other:

Annotation standard:

□ SAM □ SGML □ XML □ Other:

□ NIST/LDC

Distribution media:

□ CD-ROM □ DVD □ Other:

Number: Full size in bytes:

Quality Information

Sound quality information included:

□ SNR □ Cross talk □ Background noise

□ Clipping rate □ Other:

Tools used to measure sound quality:

Validation procedure followed (if any):

ADDITIONAL INFORMATION

Documentation available:

□ Design specification □ Other:

On-line documentation (WWW, ftp):

Related tools:

Application purposes:

AVAILABILITY

Date of availability:

Price for research use:

Price for commercial use:

Special conditions/offers:

SAMPLE(S)

Accompanying this description form, a sample of the resource:

□ on floppy disk □ by e-mail □ on other medium (cartridge, ftp, etc.):

Description of the content of the sample(s):

□ Hardcopy enclosed hereby

FURTHER COMMENTS

ELRA/ELDA
55-57 rue Brillat Savarin
F-75013 PARIS, FRANCE
Tel +33 1 43 13 33 33 - Fax +33 1 43 13 33 30 - E-mail mapelli@elda.fr
http://www.icp. grenet.fr/ELRA/home.html

6 APPENDIX B: VALIDATION SOFTWARE

SPEX uses a battery of software tools to validate a SLR. This tools can be offered to ELRA customers in order to carry out part of the validation prior to delivering the SLR to ELRA's validation centre. At present, the tools are oriented at the internal configurations of SPEX. With some efforts they can be made available to ELRA's customers. ELRA should consider the investment that would go into such efforts compared to the benefit of a valuable extra service to its customers.

The following software could be made available by SPEX:

1. Checks on the completeness of the pronunciation lexicon in terms of

- Inclusion of all words in the orthographic transcriptions (section 3.6)
- Frequency of each phone (sections 3.1.6, 3.6):
  - Per item
  - In total

2. Frequency of each word in the database (section 3.1.7):
   - Per item
   - In total
3. Creation of contents list from label files (section 3.2.3)
4. Set of acoustical characteristics of each speech file: minimum sample value, maximum sample value, clipping rate, average sample value, signal duration, SNR (section 3.4).