

SPEECHDAT

## Deliverable D3.1.3

Validation of Spoken Language Databases

Commercial in Confidence

Contract Code	LRE-6331410072
Date	February 21, 1995
Version	Initial Draft Version 1.0



## DELIVERABLE IDENTIFICATION

Identification Number	LRE-6331410072
Type	Report
Title	Validation of Spoken Language Resources
Status	Draft
Deliverable	D3.1.3
Work Package	Work Package 3
Task	Task 3.1
Period Covered	9 Dec 94 — 9 Mar 95 (Project Months 1 - 3)
Date	February 21, 1995
Version	Initial Draft Version 1.0
Number of Pages	13
Author	Reinhold Häb-Umbach
Task Responsible	Jean-Marc Dolmazon ICP Direction Recherche RP 45 rue de Londres F-75379 Paris Cedex 08
Project Contact Point	Harald Hoege Siemens AG Abt. ZFE SE SN 53 Otto Hahn Ring 6 D-81730 München Tel.: +49-89-636-3374 Fax.: +49-89-636-48000 E-mail: hh@habicht.zfe.siemens.de
CEC Project Officer	Jose Soler
Status	Confidential
Actual Distribution	Consortium, CEC
Supplementary Notes	
Key Words	SPEECHDAT, Speech data collection, validation
Abstract	Spoken Language Resources (SLR) can only be distributed successfully if the products can be guaranteed to comply with minimum quality requirements. Thus, new SLR can only be entered in the catalogue of the European Linguistic Resource Agency (ELRA) after a set of minimum quality checks have been passed. These checks will be carried out by a central validation center. This document provides a set of quality checks to be carried out by such a validation procedure.
Status of Abstract	Confidential

Received on	
Recipient's Catalogue No.	



**Contents**

<b>1</b>	<b>Introduction</b>	<b>7</b>
<b>2</b>	<b>General Remarks</b>	<b>7</b>
2.1	Validation versus Transliteration . . . . .	7
2.2	The Importance of Common Standards . . . . .	7
2.3	How to Define Common Standards . . . . .	8
<b>3</b>	<b>Validation Procedures</b>	<b>8</b>
3.1	General Remarks . . . . .	8
3.2	Documentation . . . . .	9
3.3	Speech Data Files . . . . .	10
3.4	Annotation . . . . .	10
3.5	Database Structure . . . . .	12
3.6	CD-ROM Printing and Mastering . . . . .	12
<b>4</b>	<b>Summary</b>	<b>13</b>



## 1 Introduction

Spoken Language Resources (SLR) can only be distributed successfully if the products can be guaranteed to comply with minimum quality requirements. Thus, new SLR can only be entered in the catalogue of the European Linguistic Resource Agency (ELRA) after a set of minimum quality checks have been passed. It has been agreed by the parties contributing to ELRA that these quality checks will be carried out by a central validation center. Language-dependent validation, i.e. validation procedures that require in-dept knowledge of the recorded language can be subcontracted to institutions that are able to provide that service.

Depending on its scope validation can easily be a very time consuming and cost-intensive task. The desire for comprehensive quality checks has to be assessed in light of the accompanying costs. This document therefore proposes a set of quality checks which guarantee a reasonable quality standard of the validated databases and which can be carried out at reasonable costs. Optional additional validation procedures will, however, also be treated. From the minimum quality requirements a suite of tests and procedures can be derived to design and monitor the production process of SLR. The validation center will develop and maintain these guidelines. It will also develop and maintain software packages to support the collection and quality assurance of SLR. Validation can be carried out at the ‘semantic’ and at the ‘syntactic’ level. Validation at the semantic level means that the contents of the spoken language resource is checked against its specification. This requires in general a lot of manual work, e.g. listening to data files to check whether the transliteration is correct. On the other hand validation at the syntactic level checks the formal adherence to agreed standards and can often be automatized. In this document we will describe these two types of validation in more detail.

Validation checks whether a database adheres to an agreed set of working standards and thus depends on the contents of those standards. Those standards deal with data protocols, normalization of data, collecting protocols, recording tools, storage formats and media etc. In this document those working standards, however, will not be presented. They will be described in detail in another deliverable (Subtask 3.1.1)

## 2 General Remarks

### 2.1 Validation versus Transliteration

In this document validation is defined as ascertaining that a spoken language resource (SLR) meets a specified set of quality requirements in terms of accuracy, completeness and consistency. Sometimes the term ‘validation’ is confused with what we call ‘transliteration’. Transliteration, however, is defined as adding an accurate verbatim description of the speech to the sampled data file.

### 2.2 The Importance of Common Standards

To a considerable extent these validation procedures are independent of the exact details of the procedures followed to create the corpus. Of course, one should strive towards common procedures and common standards for building corpora, but a corpus built in

an idiosyncratic fashion can still be very worthwhile, provided it meets the formal quality requirements.

Speech file headers are an excellent example to demonstrate the importance of common standards: it is absolutely true that every well-documented header that includes the minimum required information can be translated into each other header format by means of a simple programme. Yet, experience shows that corpora with idiosyncratic headers are much less widely distributed and used than corpora with standard headers. Each conversion step appears to constitute a threshold, and a small number of consecutive thresholds are very soon unsurmountable.

For the ELRA corpora with idiosyncratic formats of data files, file headers, transliteration, etc. will add dramatically to the cost of validation, because the validation center may have to write software to convert each new format to a standard one before validation can take place. However, since each conversion is very likely to cause some loss of information, the ELRA should distribute the corpus in the original format. Consequently, the level of validation of an idiosyncratic corpus is likely to be (substantially) lower than what can be the case for a corpus adhering to the standard (unless, perhaps, ELRA spends an undue amount of effort in the validation).

### 2.3 How to Define Common Standards

There are as yet no commonly agreed standards for spoken language corpora. It is probably not productive to try and impose an extensive set of very strict standards for corpora to be developed in the next five years. It is better to define a growing set of “best practice” guidelines, that will develop into de fact standards.

ELRA can make substantial contributions by formulating a set of minimum quality requirements, even if for practical purposes (to make useful corpora more widely available) it might be willing to accept existing corpora that do not meet all requirements. New corpora commissioned by ELRA or supported by EU funding should meet all requirements.

## 3 Validation Procedures

### 3.1 General Remarks

The purpose of validation is to check how well a database conforms with a commonly agreed set of standards. Validation has to be applied to the following entities:

- Documentation
- Speech data files
- Annotation, including pronunciation lexica
- Database accompanying an SLR
- CD-ROM printing and mastering

In most cases the validation procedures proposed in the following can be carried out automatically, i.e. without much human interaction. Those procedures that require a lot of manual work, e.g. listening to speech data files, have to be reduced to a minimum in order to limit the overall cost of the validation procedure.

### **3.2 Documentation**

Quality checks start at the level of the written documentation which comes with a SLR. The documentation should comprise at least unambiguous descriptions of

- name, address, contact person in the institution which originally produced the data
- reasons and goals for original corpus collection
- the number of CD's comprising the corpus (or the type and number of tapes carrying the data, in which case the layout of the CD-ROMs must be specified in addition to a specification of the tape structure and format)
- the directory structure of the CD's
- the number of speakers on each CD
- criteria used to select speakers
- specification of speaker data and characteristics available and the way in which these are stored and made accessible
- the number of items on each CD, usually specified in terms of number of items per speaker
- file naming conventions used for directories and files
- specifications of the format and file header structure of speech files
- specifications of the format and file header structure of annotation files, if any
- recording conditions (microphone, acoustic environment, etc.)
- signal characteristics (number of bits per sample; bandwidth; coding, if different from PCM; compression)
- prompting material and prompting method
- criteria used to specify the prompting material, i.e., linguistic specification of the prompting material
- presence of phonemic lexicon pertaining to the spoken text
- procedure used to obtain phonemic forms from orthographic input
- transcription conventions
- transcription procedure
- procedure used to monitor the transcription process and quality assurance

- IPR issues related to speech files and/or prompting material

The validating institute will check the existence of the minimum documentation. The check of the correctness of the documentation, is however, for many items beyond the scope of the validating center. To give an example, it will not be possible to verify that a stated speaker selection procedure has actually been applied. The validation procedures outlined in the following sections will touch some of the documented properties/features. Those will then be accordingly verified.

### 3.3 Speech Data Files

The files declared to contain sampled data in the documentation must be checked on a number of topics.

The following checks can be done automatically, i.e. without listening to the data:

- check file headers (if present) for legality of contents
- check whether the data are likely to be sampled data (against the background of documentation data on bits/sample; amplitude distributions can be made and assessed)
- Establishing the proportion of clipped samples
- Checking the compression (if applied)
- Check number of samples per file against rough estimates derived from the documentation

There are at least two different ways in which the set of files can be incomplete:

1. speakers did not produce all items requested.  
The absence of part of the items should be specified in a table of "Missing items" included in the documentation. Missing files comprised in that table are irrecoverably lost.
2. files were lost during the processing of the data.  
Missing files not contained in the table of "Missing items" should be reported to the corpus developer, who must then decide whether they were inadvertently left out of the table of "Missing items" or of the set of files provided to the validation center. In the latter case, the missing files must be added.

One of the results of this set of checks is a table of items for which annotation files must be present (unless all annotation is included in the file headers).

### 3.4 Annotation

Annotation should consist of

1. verbatim transliteration of the speech

2. transcription of additional sounds and noises, including their position relative to the words
3. assessment of the speech items in terms of signal-to-noise ratio, presence of additional noises, adherence to prompting text, etc. Tokens with acceptable signal-to-noise ratio and devoid of disfluencies will obtain an “O.K.” rating
4. lexicon with graphemic and phonemic forms of all words in the corpus
5. lexicon containing all lexicalised non-standard forms (e.g. in English forms like “gonna”, “I’d”, etc.) used in the transliteration
6. list with all symbols used to transcribe additional sounds/noise (syntax used to distinguish intrusive sounds from speech, e.g. square brackets), and listing of all terms and abbreviations that may be contained within the brackets, plus their meaning/definition

Some of these items (files) may be empty. For instance, if all tokens containing audible extra sounds have been discarded the list of symbols to transcribe additional sounds will be empty.

Best practice should specify the minimum set of annotation files. It is proposed that each corpus should include a graphemic/phonemic lexicon file in addition to the transliterations.

Annotation files must be checked in a number of ways.

The following checks can be carried out automatically, i.e. without listening to the data

- Establish whether they are ASCII files
- Spell checking (automatically)
  - Check the existence and format of a lexicon file
  - Check the existence and format of a file with accepted non-standard spellings/words
  - Distinguish between annotation files that may or may not have deviant spellings (e.g. in POLYPHONE files with an O.K. label can only contain correctly spelled word forms)
- Checking annotations (brackets etc. (automatically))
- Check annotation for expected number of characters (e.g., if a token is supposed to be a sentence containing at least four words, it should contain at least four non-blank characters)
- Check the phonemic word list (if present, but it should be)
  - Check whether SAMPA is used consistently (or whether all tables necessary to translate a non-standard computer readable phonetic alphabet to SAMPA are present and readable)
  - Check for occurrence of characters not present in the phoneme label definition file (or in the SAMPA alphabet)

- Check the phonemic word list for completeness: is every word occurring in the transliteration present in the phonemic word list?

Note that the correctness of phonemic forms in terms of adequate pronunciation will not be checked.

Other checks require that the speech samples are listened to:

- Correctness of verbatim transliteration: did the speaker actually speak the transliterated words?
- Correctness of the annotation of non-speech acoustic events

These checks are very time consuming and can thus only be carried out on a random sample test basis, such that only a small percentage (5% – 10%) can actually be checked.

Another possibility to automate these checks is to employ an automatic speech recognizer and manually check only those data files where the recognition result is different from the transliteration. To be efficient this requires a sufficiently good performance of the speech recognizer. This is probably only attainable if the unchecked transliteration is already correct to a great extent.

There are techniques to automatically check the correctness of phonetic transcriptions. They require the availability of a good grapheme-to-phoneme conversion system. This system can propose a pronunciation of a word given in orthographic form, which then has to be compared with the entry in the pronunciation dictionary. If the two pronunciations are different, this may indicate an inadequate pronunciation in either of the lexica. A human then has to decide which pronunciation to choose (e.g. by comparing with the pronunciations of orthographically similar words).

For the validation of the databases to be collected in workpackage 1.4 this technique will, however, not be applied.

### 3.5 Database Structure

If the corpus contains information about its contents in the form of a database, the integrity and completeness of that database must be checked. This can be done by loading the tables in a database management system and check for counts, etc. Also, by loading the tables it will be established whether all necessary files are available.

It is recommended that each corpus includes a database which covers at least all speaker-related information. A database containing information about the words/utterances spoken should also be included.

### 3.6 CD-ROM Printing and Mastering

Before printing master copy the following checks should be carried out:

- Check the data on the PC (standard for pressing CDs)
- Is the physical volume name present (11 chars max)?

- Are the directory names correct (8 chars max, no extension, no hyphen)?
- Is there, in each sub-directory, the right number of files per speaker per corpus?
- Is there the same number of signal files and of associated description files ?
- Is the correct documentation present on the right disk ?
- Check correctness and completeness of the information on the CD's label and on the jacket

Check the printed CD-ROMs

- Are all files readable and can they be copied?
- Check print quality of label, jacket and documentation.

The steps related to the media are completely automatic. Checking the printing of label, jacket and documentation is a manual process.

## 4 Summary

This document presents a list of guidelines for validation procedures to be carried out in order to ascertain a certain quality standard of spoken language resources to be distributed by the ELRA. The methods proposed are chosen such that they are a good balance between achievable quality standards and associated costs of the validation procedure.