THE ART OF VALIDATION

Henk van den Heuvel

SPEX, Nijmegen, The Netherlands

e-mail: H.v.d.Heuvel@spex.nl

• Introduction

An increasing number of Spoken Language Resources (SLRs) in ELRA's catalogue contains a remark such as: "The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications." Some may read such a sentence in "dustbin-mode", so without paying attention to it, but others may be interested in the background and contents of such a validation procedure. This article serves to satisfy the curiosity of the latter group of readers, at least to some extent.

Validation of SLRs may refer to a variety of actions:
1. checking a SLR against a fixed set of requirements;
2. putting a quality stamp on a SLR as a result of the aforementioned check. If the database passes the check, then we say that it has been "validated";
3. the evaluation of a SLR in a field test, thus testing the usability of the LR in an actual application.
4. …

SLR validation, as carried out by SPEX (acronym for Speech Processing Expertise Centre) , typically refers to the first type of action: the quality evaluation of a database against a checklist of relevant criteria. These criteria are typically the specifications of the databases, together with some tolerance margins in case deviations are found.

The validation of language resources in general, and SLRs in particular, is a rather new type of activity in the area of language and speech technology. As more and more SLRs are entering the market, the need for validation of these resources increases, and therefore the best ways to accomplish validation need to be established. Validation of SLRs is of particular interest to the European Language Resources Association and its distribution agency ELDA (http://www.elda.fr/). ELRA offers a wide range of SLRs in its catalogue. Before distribution can proceed, the products must be subjected to quality control and validation. ELRA has established manuals for validation and has been actively persuading producers of Language Resources to adopt these as a means of adding value to the marketability of their products. ELRA, therefore, has started instituting a system that, in the long term, will yield a specification and quality control document to be issued with every product that ELRA sells or licenses. In order to evaluate the quality of the SLRs in the ELRA catalogue, a procedure to describe and validate these SLRs has to be developed. ELRA entrusted this task, after an open call, to SPEX. SPEX constitutes the first SLR validation unit of ELRA's Validation Network.

In this contribution I will give an overview of various aspects of SLR validation and present some future directions in this field, especially with respect to SPEX's validation mission for ELRA.

- ## **What is there?**

The first SLRs that were formally validated were the databases of the collaborative EC funded SpeechDat(M) project. An important internal motivation for this SLR validation was the idea that all partners should exchange equivalent databases within a project. For this reason, validation also was used in the sense of the second interpretation given above: validation as a binary quality stamp: *pass* or *reject.* Only databases which passed the validation were released by the consortium. SpeechDat has created an impressive off-spring. Table 1 presents an overview of the projects in, what is nowadays called, the SpeechDat "family".

Project	SLR	Period	Ref.
SpeechDat(M)	8 FDB	1994-1996	Höge & Tropf (1996)
SpeechDat(II)	20 FDB 5 MDB 3 SDB	1995-1998	Höge, et al. (1999)
SpeechDat-Car	9 CDB	1998-2001	Van den Heuvel, et al (1999)
SpeechDat-East	5 FDB	1998-2000	Pollak, et al. (2000)
SALA	4-5 FDB	1998-2000	Moreno, et al. (2000)

Table 1. Overview of SpeechDat projects. CDB = Car databases; FDB = Fixed (telephone) Network databases; MDB = Mobile network (telephone) databases; SDB = Speaker Verification databases.

The SpeechDat formula was, in addition, also used for a number of other data collections, as shown in Table 2. Also here, a formal SLR validation was carried out by SPEX.

Language	SLR	Producing Company	Ref.
Russian	1 FDB	Auditech (for Siemens), Petersburg, Russia	Pollak, et al. (2000)
Austrian German	1 FDB 1 MDB	FTW, Vienna, Austria	Baum et al. (2000)
Australian English	1 FDB	Callbase, Isleworth, UK	www.callbase.com

Table 2. Overview of projects collecting data according to SpeechDat protocols.

Also the SLRs collected in the Speecon project (Siemund et al., 2000) will be collected more or less according to the SpeechDat standards. All SLRs mentioned above will be offered to ELRA for distribution.

- ## **How do we do it?**

As I see it, SLR validation operates along two dimensions with two points on the axis of each dimension. The first dimension concerns the integration of validation into the specification phase. Along this axis 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, throughout the

definition of the specifications, the feasibility of their evaluation and the criteria to be employed for such an evaluation are taken into account. (b) A SLR is created, and the validation criteria and procedure are defined afterwards. In this way, validation may boil down to reverse-engineering and the risk is faced that the validation of some parts of the specification may become infeasible. As for the second dimension, validation can be done (a) in-house by the SLR producer (internal validation) or (b) by another organisation (external validation). The two dimensions thus identified are shown in Table 3.

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

Table 3: Four types of validation strategies

Compartment (1) in this table points to an essential element 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. In this way, each producer is his own validator. An internal final check (2) should be an obvious, be it ideally superfluous, part of this procedure. Alternatively, or in addition, an external organisation can be contracted to carry out the validation of a SLR. In that case the best approach is that the external validator is closely involved in the definition of the specifications (in order to assess the feasibility of corresponding validation checks), and performs quality checks for all phases of the production process (3), followed by a final check after database completion (4). (3) and (4) are more objective quality evaluations, and should be considered important for that reason.

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.
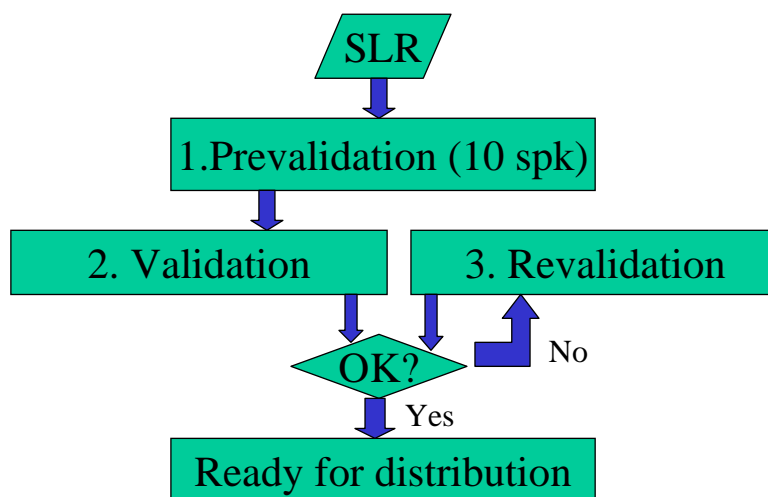
Validation Procedure

Figure 1. SLR validation procedure in SpeechDat-related projects.

As shown in Figure 1, validation in "SpeechDat style" proceeds in three steps:
1. Prevalidation of a small database of about 10 speakers shortly after the design specifications have been established and the recording platforms installed. The objective of this stage is to detect serious (design) errors before the actual recordings start. This stage also allows partners to build their database compilation software in an early stage of the project. This corresponds to strategy (3) in Table 3.
2. Validation of the complete database. The database is checked against the SpeechDat specifications and a validation report is edited. This stage corresponds to strategy (4) in Table 3.
3. Revalidation of a database. In case the validation report shows that corrections of a database are necessary or desirable, then (part of) the database can again be offered for validation, and a new report is written. In horrendous cases this phase may show some iterations.

In SpeechDat projects the eventual decision about the approval of a database is not made by SPEX, but by the consortium concerned. In fact, the consortium performs validation in the second interpretation mentioned in the introduction: putting a quality stamp on a product.

Back to Table 3. For obtaining the highest SLR quality the numbers in the compartments in the table reflect the order of importance of validation strategies: 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.

ELRA resources are distributed "as-is with all defects" as stated in the licenses. The databases are created (and sold), but a thorough validation has yet to be carried out for the majority of the SLRs in the catalogue. Of course, one may have some faith that internal quality checks in the spirit of (1) and (2) took place for individual databases, but an objective external validation is a valuable, if not necessary, additional means of quality assessment.

- ## Validation and improvement

A principal issue concerns the difference between validation and improvement of a SLR. At first sight, both seem closely intertwined. Who could better rectify the errors in a database than the person (or institute) that was smart enough to detect the errors? Nonetheless, a principal stance should be taken here. In SPEX's view, validation and improvement should be clearly distinguished. There are differences with respect to:
1. Nature of the actions: Validation is a quality assessment procedure and therefore a diagnostic operation.
2. Chronology: Validation yields the diagnosis; the improvement is the cure. Therefore, SLR validation should obviously precede SLR improvement.
3. Responsible institutes: In principle, the validator and the corrector should be different institutes, in order to avoid the undesirable situation that the validating institute assesses its own work. The correction of a SLR is accordingly in principle a responsibility of the SLR owner.

- ## What is checked?

SLR validation criteria come in the following categories:

1. Documentation. It is checked if all relevant aspects of a SLR (see 2-8 below) are properly described in terms of the three C's: clarity, completeness and correctness.
2. Database format. It is checked if all relevant files (documentation, speech files, label files, lexicon) are present in the appropriate directory structure and with the correct format.
3. Design. The appropriateness of the recorded items for the purpose of the envisaged application(s) and the completeness of the recordings should be checked.
4. Speech files. The acoustical quality of the speech files is measured in terms of (e.g.) (average) duration, clipping rate, SNR, mean sample value. Also auditory inspection of signal quality belongs to this category.
5. Label files. The label files should obey the correct format. Ideally, they can be automatically parsed without yielding erroneous information.
6. Phonemic lexicon. The lexicon should contain appropriate phonemic (or allophonic) transcriptions of all words in the orthographic transcriptions of a SLR.
7. Speaker & environment distributions. The recorded speakers should present a fair sample of the population of interest in terms of (typically) sex, age and dialectal background. Also the recording environments should be representative for the targeted applications.
8. Orthographic transcriptions. A (native) speaker of the language should check a sufficiently large sample of the orthographic transcriptions by comparing these to the speech in the signal files and the transcription protocol.

An example of an extensive list of validation criteria in terms of specifications and tolerance intervals is given in Van den Heuvel (1996).

• **Rank order of validation check points**

The acoustic quality of the speech files is of utmost importance. Although the desired quality may to a great deal depend on the wishes of the customer or on the targeted applications, it is obvious that recordings containing rubbish disqualify for being included in a speech database. Further, the clarity, completeness and the correctness of the documentation is a first order requirement for any SLR that deserves this name. Also, only a proper transcription of the speech qualifies the database as more than a mere collection of speech signals. In summary, at SPEX we consider documentation, transcription, and good speech signals as the core ingredients of a SLR, which should have the highest validation weight.

On the second level in the validation rank order follow: completeness criteria for the design of the SLR and for the recordings actually contained in the database, and completeness criteria for distributions of speakers and environments, etc.

The third level of priority concerns SLR aspects that can be easily corrected afterwards, such as the phoneme lexicon, 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 fixed. 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.
Quality labels can be attached to each aspect of the database. Our quality labels have three possible values: 1. not acceptable; 2. not OK, but acceptable; 3. OK.

Table 4 gives a summary of the priority weights and quality values that can be attached to the SLR characteristics. SPEX regards this scheme as the key framework to validate SLRs in the ELRA catalogue.

Database part	Rank order	Quality value		
		1	2	3
Documentation	1			
Transcription	1			
Speech signal	1			
SLR completeness	2			
Speaker distributions	2			
Recording conditions	2			
Annotation files	3			
Lexicon	3			
Formats & file names	3			

Table 4: Quality assessment methodology for existing SLRs in ELRA's catalogue. See the text for clarifications for rank orders and quality labels.

• **Who is responsible for what?**

The validation and improvement of a SLR involves two players: (1) The validation institute which assesses the quality of a database and reports its deficiencies; (2) the database owner taking care of the improvements that become necessary after such a report. In the specific case of SPEX performing the validation for ELRA, ELRA is the third player. As a matter of fact, SPEX as validation institute acts as the intermediary between ELRA and the database owner. The Board of ELRA is represented by the Speech College members of the Board. The ELRA Board strives for a validation of the SLR in its catalogue; the database owner may be asked to supply an improved database if deficiencies of the database show up, and SPEX carries out the validations and takes care of the communication between ELRA and the database owner. Further, the ELRA Board decides or affirms the priority list with which SLR have to be validated (i.e. priority in time); it determines the corrections that have to follow after a validation and the sanctions to incur if a SLR owner refuses rectification of the database.
The procedure can be captured by the action list given in Table 5. In vertical direction this table reflects a rough time axis. For SPEX, the role of intermediary between A and C holds for the full validation process.

A. ELRA	B. SPEX	C. SLR owner
	Makes priority list (see section 8 below)	
Decision of SLR validation		
	Intermediary between A and C	
	Performs validation and makes report	
		Reaction to validation report/results
Decision on necessary corrections		
		Corrects and updates the SLR

Table 5: General procedure and responsibilities for the validation and improvement of SLRs in the ELRA catalogue

• **Bug reports**

Errors in a database do not only emerge during the validation procedure. Errors are also typically detected by clients once they use the database. An efficient means of bug reporting and appropriate procedures for updating a SLR and distributing a new release should, therefore, be an integral part of permanent quality maintenance.

Below I present the procedure for ELRA that I see as the most promising for the time being, and which SPEX intends to start with. This procedure can easily be combined with the validation/correction procedure presented just before.

1. A link to a *bug report sheet* is created at ELRA's WWW home page
2. The bug report sheet is a frame based sheet, with slots for the information like: Database name; Code in ELRA's catalogue; Coordinates (name, affiliation, e-mail address) of the reporter; Errors to report.
3. Lists of all reported bugs for each SLR in the catalogue are made available through ELRA's home page and can be accessed by ELRA members.
4. Depending on the seriousness and the number of the bugs reported, SPEX recommends a SLR for validation and/or correction. The decision is made by the ELRA Board, and the steps indicated in Table 5 are followed.

• **Who comes first?**

The order in the priority list of SLRs to be validated is driven by several factors. First of all the number of copies sold through ELRA gives a good indication of the market value of a database and hence of the need to have this database in an optimal condition. On the other hand, if this database has already been validated before (as is the case with the databases in the SpeechDat projects), then a (new) validation should have lower priority (but this is something that practice should prove).

Furthermore, the bug reports are also indicative of the condition of a database. If many and serious bugs are reported for a SLR, then rapid action should be taken. In that case, we recommend to give a database a thorough validation first in order to have the major shortcomings detected at once. This is in agreement with the general strategy pointed out above to precede SLR improvement by a validation. To insert a validation between bug reports and SLR improvements serves two purposes:

1. Verification of the reported bugs
2. Guarantee that the most serious other bugs are found in one action

Therefore, in summary, the following determinants for prioritising SLR validation are considered:

- The numbers of copies sold / expected to be sold through ELRA
- The number and seriousness of errors reported via bug reports
- Availability of reports of previous validations

• **Future plans**

SPEX has established a first priority list of SLRs in ELRA's SLR catalogue that need validation. The idea is to validate various SLRs this year, following the quality chart presented in Table 4. Plans are being developed in order to make a validation protocol for Broadcast News databases, as part of the new MLIS project NETWORK-DC.

• **References**

Baum, M., et al. (2000) SpeechDat AT: Telephone speech databases for Austrian German. Proceedings of the LREC'2000 Satellite workshop on XLDB - Very large Telephone Speech Databases, Athens, Greece, pp. 51-56.

Höge, H., Tropf, H.S. (1996) Final Report. SpeechDat(M) Technical Report D0.6 & 0.7. http://www.icp.grenet.fr/SpeechDat/home.html

Höge, H., et al. (1999): Speechdat multilingual speech databases for teleservices: across the finish line. Proceedings EUROSPEECH'99, Budapest, Hungary, 5-9 Sep. 1999, Vol. 6, pp. 2699-2702

Moreno, et al.: SALA: SpeechDat across Latin America.. Results of the first phase. Proceedings LREC2000, Athens, Greece, pp. 877-882.

Pollak, P., Czernocky, J., Boudy, J. et al. (2000) SpeechDat(E)- Eastern European telephone speech databases. Proceedings of the LREC'2000 Satellite workshop on XLDB - Very large Telephone Speech Databases, Athens, Greece, pp. 20-25.

Siemund, R., et al. (2000) SPEECON - Speech Data for Consumer Devices. Proceedings LREC2000, Athens, Greece, pp. 883-886.

Van den Heuvel, H. (1996): *Validation criteria*. SpeechDat Technical Report SD1.3.3. http://www.speechdat.org/SpeechDat.html

Van den Heuvel, H., et al. (1999): The SpeechDat-Car multilingual speech databases for in-car applications: Some first validation results. Proceedings EUROSPEECH'99, Budapest, Hungary, pp. 2279-2282.

Dr Henk van den Heuvel
SPEX / A2RT
Dept. of Language and Speech
University of Nijmegen
P.O.Box 9103
NL-6500HD Nijmegen
H.v.d.Heuvel@spex.nl
http:/lands.let.kun.nl