SLR VALIDATION: PRESENT STATE OF AFFAIRS AND PROSPECTS

Henk van den Heuvel (1), Lou Boves (1),

Khalid Choukri (2), Simo Goddijn (1), Eric Sanders (1)

(1) SPEX, Nijmegen, Netherlands
(2) ELRA/ELDA, Paris, France

e-mail: H.v.d.Heuvel@let.kun.nl

Abstract

This paper deals with the quality evaluation (validation) and improvement of Spoken Language Resources (SLR). We discuss a number of aspects of SLR validation. We review the work done so far in this field. The most important validation check points and our view on their rank order are listed. We propose a strategy for validation and improvement of SLR that is presently considered at the European Language Resources Association, ELRA. And finally, we show some of our future plans in these directions.

1. Introduction

Validation, as we will use the term here, refers to the quality evaluation of a database against a checklist of relevant criteria. The validation of language resources in general, and spoken language resources (SLRs) in particular, is a rather new type of activity in the area of language & 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 to the marketability of their products. The users of LR, ELRA's "customers," need to know about the product they are purchasing: they need to know its specification and quality. ELRA must build up a reputation for the product it sells, though there may be a role for products of limited quality or coverage that have properties that are of interest to the research community.

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. A body that creates an LR and enters into an agreement or contract with ELRA for its distribution is required to provide some basic information about that LR product.

However, ELRA cannot rely solely on that specification, for it is the reputation of the Association that will be at stake. 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 the Speech Processing EXpertise centre. SPEX constitutes the first SLR validation unit of ELRA's Validation Network.

In this paper we will present an overview of the state of the art in SLR validation and show some future directions in this field, especially with respect to SPEX's validation mission for ELRA.

2. SLR Validation and Improvement

SLR validation operates along two dimensions. 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.

Second, validation can be done in-house by the SLR producer (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. 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 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 already 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 (Draxler, et al., 1998; Hoege, et al. (1999); Van den Heuvel, et al. (1999); Moreno, et al. (2000); Czernocky, et al. (2000)), 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 reduced validation approach the numbers in Table 1 above reflect the order of importance of quality assurance: 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, though ELRA resources are distributed "as-is with all defects" as stated in the licenses. The databases are created (and even sold), but the validation has yet to be carried out. Of course, one may have some faith that internal quality checks in the spirit of (1) and (2) took place for individual databases. The validation report by SPEX can then serve as a valuable starting point for SLR improvement, if necessary.

Validation and improvement should be clearly distinguished. They differ with respect to:
- Nature of the actions: Validation is a quality assessment procedure and therefore a diagnostic operation
- Chronology: Validation yields the diagnosis; the improvement is the cure. Therefore, SLR validation, as a general rule, precedes SLR improvement
- Responsible institutes: In principle, the validator and the corrector should be different institutes, in order to avoid the undesirable situation that the validating institute should assess its own work. The correction of an SLR is accordingly in principle a responsibility of the SLR owner.

3. Other SLR Validation activities

The market for large SLRs has been strongly growing over the past years. At several places and in various projects large speech databases are being produced. The reason for this growth is that large collections of speech material can nowadays be collected due to fast CPUs and huge storage capacities of modern media. These large databases are needed to build reliably working automatic speech recognisers, even for 'simple' applications like digit recognition. As a consequence, quality assessment of such databases is becoming a very important topic in the area of SLR production. However, as it appears from our inquiries, many organisations who are active in disseminating information on SLRs, and guidelines to produce them, are considerably less active in (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 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 consultations of some of their representatives.

Also a query for SLR validation activities at CSLU (Center for Spoken Language Understanding), see http://cslu.cse.ogi.edu/, was not successful. Validation of SLR was not part of the Survey of the State of the Art in Human Language Technology (Cole, et al., 1996), 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. As a rule, LDC does not validate SLR produced by others (external speech databases). In the exceptional case, when a corpus from an outside source is published, a limited quality control protocol is followed in which it is checked:
- if all components of the corpus mentioned in the documentation are present;
- if all components are formatted as stated in corpus documentation;
- if all supporting components (like tables of speaker attributes) are consistently formatted.

In the normal case, however, LDC produces SLRs 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. 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 (Van den Heuvel, 1996; Van den Heuvel, 1999a, 1999b).

4. Validation check points

SLR validation criteria come in the following categories:
1. Documentation. It is checked if all relevant aspects of an 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, addressing the appropriateness and the completeness of the recorded items for the purpose of the envisaged application(s).
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. It is checked if the label files obey the correct format, and if they can be automatically parsed without yielding erroneous information or even system crashes.
6. Phonemic lexicon. The lexicon should contain appropriate phonemic (or allophonic) transcriptions of all words in the orthographic transcriptions of an 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.

5. 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 in fact 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 recordings (Gibbon, et al. 1997: 146). Next, hardly any automatic speech recogniser can be sensibly trained or tested if a phonemic lexicon is missing in the database. In summary, we consider documentation, transcription, lexicon, and good speech signals as the core ingredients of an SLR, which should have the highest validation priority.

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 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 2 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			
Lexicon	1			
Speech signal	1			
SLR completeness	2			
Speaker distributions	2			
Recording conditions	2			
Annotation files	3			
Formats & file names	3			

Table 2: Quality assessment methodology for existing SLRs in ELRA's catalogue. As for the quality labels: 1: Not acceptable; 2: Acceptable (with minor corrections); 3. OK (no corrections needed).

6. Validation procedure

As stated in the introduction, the validation and improvement of an 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, the ELRA Board is the third player. As a matter of fact, SPEX as validation institute acts as the intermediary between the ELRA Board and the database owner. 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 an SLR owner refuses rectification of the database.

The procedure can be captured by the action list given in Table 3.

Step	A: ELRA Board	B: Validator	C: SLR Owner
1		- Propose to A a priority list of SLR to validate	
2	- Confirmation / modification of proposed list		

3		- Validation of an SLR - Notify A and C of this activity - Send validation report to: C	
4			- Reaction to validation report
5		- Finalisation of validation report - Send report to A and C	
6	- Decision on things to correct and sanctions in case of refusal, if needed - Communication of the decisions to B		
7		- Inform C of A's decisions	
8			- Correction of deficiencies
9		Revalidation of (part of) SLR and report to A and C. Re-validation report to C	
10			- Reaction to revalidation report
11		- Finalisation of re-validation report - Send report to A and C	
12	- Approval/rejection of rectified SLR		
13		- Create new SLR patch to be distributed by ELDA	- or: Create new SLR version to be distributed by ELDA

Table 3: General action plan for the validation and improvement of a SLR in the ELRA catalogue

7. 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 an appropriate procedure for updating a SLR and disseminating a new release should, therefore, become an integral part of permanent quality maintenance.

Below we present the procedure that we see as the most promising for the time being, and which we prefer to start

with. This procedure can easily be combined with the correction procedure presented in the previous section.

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 following information: Database name; Code in ELRA's catalogue; Coordinates (name, affiliation, e-mail address) of the reporter; Errors to report.
3. SPEX takes care that a list of all reported bugs for each SLR in the catalogue is available via ELRA's home page and can be viewed by ELRA members.
4. Depending on the seriousness and the number of the bugs reported, SPEX recommends SLR for validation and/or correction. The decision is made by ELRA's Board, and steps 3-13, as indicated in Table 3, are followed.

8. SLR priority listing

The order in the priority list of SLR 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.

Furthermore, the bug reports are also indicative of the condition of a database. If many and serious bugs are reported for an 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 through ELRA
- Availability of reports of previous validations
- The number and seriousness of errors reported via bug reports

9. Future plans

As required by the ELRA-SPEX agreement, SPEX has established a first priority list of SLRs in ELRA's SLR catalogue that need validation. Various SLRs will be validated this year, following the quality chart presented in Table 2. Plans are being developed in order to make a validation protocol for Broadcast News databases, as part of the new MLIS project NETWORK-DC that aims at developing close collaboration actions between ELRA and LDC.

10. References

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.

Pollak, P., Czernocky, J., Boudy, J. et al. (2000) SpeechDat(E)-Eastern European telephone speech databases. Proceedings of the XL-DB workshop, 29 May 2000, Athens.

Christoph Draxler, Henk van den Heuvel, Herbert S.Tropf (1998) SpeechDat Experiences in creating Large Multilingual Speech Databases for Teleservices. Proceedings LREC98, 28-30 May 1998, Granada, Spain, Vol. I, pp 361-366.

Gibbon, D., Moore, R., Winski, R., (Eds) 1997. *Handbook of standards and resources for spoken language systems.* Mouton, de Gruyter. Berlin, New York.

Hoege, H., Draxler, C., Heuvel, H. van den, Johansen, F.T., Sanders, E., Tropf, H.S. (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. These proceedings.

Van den Heuvel, H. (1996): *Validation criteria.* SpeechDat Technical Report SD1.3.3.

Van den Heuvel, H. (1999): *Validation criteria.* SpeechDat Car Technical Report D1.3.1, 1999.

Van den Heuvel, H. (1999): *Validation criteria.* SpeechDat East Technical Report ED1.4.2, 1999.

Van den Heuvel, H., Boudy, Comeyne, R., Euler, S., Moreno, A., Richard, G. (1999): The SpeechDat-Car multilingual speech databases for in-car applications: Some first validation results.Proceedings EUROSPEECH'99, Budapest, Hungary, 5-9 Sep. 1999, Vol.5, pp. 2279-2282.