SLR Validation: Current Trends and Developments

Henk van den Heuvel, Dorota Iskra, Eric Sanders, Folkert de Vriend

SPEX (Speech Processing Expertise Centre), Department of Language and Speech, Nijmegen, the Netherlands

e-mail:{henk,dorota,eric,folkert}@spex.nl

Abstract

This paper deals with the quality evaluation (validation) of Spoken Language Resources (SLR). The current situation in terms of relevant validation criteria and procedures is briefly presented. Next, a number of validation issues related to new data formats (XML-based annotations, UTF-16 encoding) are discussed. Further, new validation cycles that were introduced in a series of new projects like SpeeCon and OrienTel are addressed: prompt sheet validation, lexicon validation and pre-release validation. Finally, SPEX's current and future activities as ELRA's validation centre for SLR are outlined.

1. Introduction

Validation, as we will use the term here, refers to the quality evaluation of a database against a checklist of relevant criteria (Van den Heuvel et al., 2003; Fersøe, 2003; Schiel and Draxler, 2003). For the validation of language resources (LR) in general, and spoken language resources (SLR) in particular the relevant criteria are dependent on the application domain targeted with the SLR at hand and the setting in which the criteria are developed. Basically, two settings should be distinguished. The first setting is the situation in which SLR are completed in a framework where, validation by an external (i.e. non-producing partner) is an integral part of the SLR production process and the validation centre is involved from the beginning of the specifications of the databases. Therefore, the validation criteria are closely linked to the specifications. Examples of such validation scenarios are SpeechDat (II) (Van den Heuvel, 1996), SpeechDat-Car (Van den Heuvel, 1999), SALA (Van den Heuvel, 1997), SpeeCon (Van den Heuvel et al., 2001), OrienTel (Iskra et al., 2002)., and more recently LC-STAR (Shammass and Van den Heuvel, 2003).

The other setting is that in which validation is not an integral part of the SLR production and should be done post-hoc. The European Language Resources Association (ELRA) faces this situation for part of the LR in its catalogue. ELRA regards quality assessment as an important element for the LR that it distributes. For this reason, ELRA is developing a set of minimum requirements which the various types of resources in its catalogue should fulfill. Obviously, these minimum requirements do not simply coincide with the specifications of the database proper (Van den Heuvel et al., 2003).

In this paper relevant issues as experienced by SPEX in both validation settings are presented. We start with an overview of the current situation and the new challenges encountered and then deal with new developments in more detail.

2. Current Situation and New Challenges

For a SLR the validation criteria typically comprise the following elements:

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. The appropriateness and the completeness of the recorded items are addressed 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 a SLR.
7. Speaker & environment distributions. The recorded speakers should present a fair sample of the population of interest in terms of (typically) gender, age and dialectal background. Also the recording environments should be representative for the targeted applications.
8. Orthographic transcriptions. A native speaker of the language checks a sufficiently large sample of the orthographic transcriptions by comparing these to the speech in the signal files and the transcription protocol.

Formats and formal criteria can be tested automatically. The content of a database such as the correctness of the orthographic or phonemic transcriptions, but also the contents of the documentation require manual labour.

The associated criteria can be found in detail in the validation deliverables given in the reference section for individual projects as mentioned in section 1 above.

The annotation of SLR in the SpeechDat-family is rather flat and is captured by label files following the SAM-standards (SAM, 1992). However, for SLR with more complex annotation layers the SAM concept is not well suited. More appropriate annotation formalisms for such SLR are ATLAS (Laprun et al., 2002), and IMDI (Broeder at al., 2001). Examples of hierarchically structured annotation layers recently validated by SPEX

are broadcast news databases developed for TC-STAR_P (http://www.tc-star.org) and the phonetic lexicons produced in the LC-STAR project (Hartikainen et al, 2003). Annotation of these databases is in XML-based encodings. The new challenges that such new formats pose for validation are discussed in section 3.

Traditionally, the validation scenario in a SpeechDat approach consisted of two phases: 1) prevalidation, 2) full validation. During prevalidation, the recordings of the first 10 speakers are evaluated in order to find systematic errors at an early stage of the speech collection. For these 10 speakers identical checks are carried out as will be the case later for the complete database. These checks are executed on the speech files, label files, and documentation files and refer to the aspects mentioned above. For a full validation, all the checks which were executed in the prevalidation phase are carried out again, this time, however, on a complete database. Furthermore, orthographic transcriptions are evaluated by native speakers and the database is checked against a number of distribution criteria, such as gender or environment distributions, which is only possible when all the database recordings are available.

This scenario was followed in projects such as SpeechDat (II), SpeechDat-East, SpeechDat-Car and SALA (I & II). (http://www.speechdat.com). The experience of both producers and the validation centre was that the two validation stages were not sufficient to detect certain errors both in the early design phase and in the very final phase when, after full validation, some final corrections were made by the producing parties without these corrections being re-checked. Therefore, in more recent projects like SpeeCon and OrienTel, new validation stages were introduced in order to minimize such risks. These new stages are presented in section 4 of the paper.

As mentioned in section 1, there is also the setting in which validation has to be done post-hoc. In section 5 we provide a concise update of the latest validation activities at ELRA, where SPEX acts as the validation centre for SLR.

3. New Data Formats

As mentioned above, annotations in other than SAM oriented formats require new validation approaches. Such annotations are found in, e.g., broadcast news databases (BCN) from the TC-STAR_P project and phonemic lexicons, from the LC-STAR project. These databases differ from the SpeechDat format in a number of ways.

	SpeechDat-family	BCN (TC-STAR_P)	LC-STAR
Speech	Many short utterances	Very long items with complete broadcast	No speech
Database Structure	Many files in relatively complex directory structure	Few files in simple directory structure	Very few files in very simple directory structure
	Meta files in SAM or tab separated	Meta files in XML	Files in XML

Character coding	ANSI/other	ANSI/other	Unicode UTF-16

Table 1: Differences between LR types validated by SPEX

Table 1 gives the most important differences between SpeechDat, TC-STAR_P and LC-STAR databases. In the following we will discuss how these differences influence the validation procedure.

Speech

Where SpeechDat-like databases have many items containing short phrases like numbers, names or dates, typically lasting between two and ten seconds, the BCN databases from TC-STAR_P have huge speech files up to half an hour or longer. LC-STAR contains only lexicons and no speech files at all.

Because of the length of the speech files in BCN databases, it is impossible to make a straightforward random selection of speech files for the validation of orthographic transcriptions. Therefore, a semi-random selection (accounting for all sorts of distributions, like gender and accent of speaker) of the transcriptions is made, and checked against the corresponding speech segments. In order to do this, the time stamps of the selected parts are searched automatically in the XML label files and used to cut the speech segments out of the large speech files. Speech segments of the same speaker are grouped together in order to allow the validator to assess if subsequent segments come from the same speaker as indicated by the producer. For this more sophisticated procedure new software tools were developed.

The quality of the speech is checked by computing a number of statistics of the signal, like clipping rate and signal-to-noise ratio (see section 2). These statistics, however, are only meaningful in relatively short speech clips up to a few minutes at most. To compute meaningful signal statistics on very large files, these files have to be divided in smaller segments first, so that portions with bad signal quality can be detected and are not averaged out.

Database Structure

SLR in the SpeechDat-family have a relatively complex directory structure accompanied by simply structured information files. These label files are encoded in SAM, a scheme that was standardised by the EAGLES group (Gibbon et al., 1997). Other metafiles in the SpeechDat-approach usually contain just tab-separated fields. The TC-STAR_P BCN databases and LC-STAR lexicons have a simple directory structure but more complex structured information files. The multi-layered annotations in the BCN and the lexicons in LC-STAR are in XML format.

For validation this means that the relevant information has to be extracted by parsing XML-format files. That implies that the validation software for automatic checks either has to be adapted or, alternatively, already existing off-the-shelf tools can be used. These tools are typically freely or commercially available parsers, like XML-Spy, which can, for instance, check the XML against a Document Type Definition (DTD). This means that instead of writing new software only a set of proper DTD rules have to be

stated. The definition of these rules forms part of the specifications of the database and are not directly developed for validation. They have to be used, however, by the validation centre to carry out the check against the DTD.

Nonetheless, additional smart parsing procedures were needed to check for instance sufficient coverage of certain POS-tags in the LC-STAR XML-based lexicons (De Vriend et al. 2004).

Character coding

For European languages plain ANSI character encoding was sufficient, but with databases in all kinds of other languages appearing, a lot of other character encodings are needed. In Orientel and SpeeCon languages like Mandarin, Arabic, Hebrew and Korean are recorded, to name a few. For transcription validation of more 'exotic' character codings tools are required that are able to handle codings other than those in the ISO-8859 series.

Unicode is becoming a new standard and is also used in LC-STAR. In this case the software has to be able to cope with UTF-16, in which characters are coded in two bytes. This poses special challenges for comparing strings, inserting characters in strings, and generating validation output.

4. New Procedures

Since its specification in the early nineties, the validation procedure as described in section 2 has undergone a number of changes. These are due to, on the one hand the experience of the validation centre, but on the other hand the needs of the producers. The current procedure which has been applied in the more recent projects such as SpeeCon and OrienTel comprises the following validation stages:

1) *prompt sheet validation*
2) *lexicon validation by an external expert*
3) pre-validation of the first 10 recorded speakers
4) validation of a complete database
5) *pre-release validation*

The stages 1, 2 and 5 are new and were not applied in the first SpeechDat projects. In the following section these new stages are presented in more detail together with a motivation for their introduction.

1) Prompt sheet validation

Before embarking on recording speakers, the producers design reading scripts. These scripts should be an ideal reflection of the specifications with regard to the content of the corpus items and the number of repetitions for each item. Since things are bound to go wrong during the recordings due to problems with the recording platform, of speakers omitting certain items altogether, not reading them correctly, stuttering or speaking in an environment with high background noise, the reading scripts have to meet the upper bounds of what is achievable in a database. The validation of the prompt sheets comprises checks with regard to the presence of the corpus items, adherence of their design to the specifications as well as the number of repetitions at word or sentence level calculated for the complete database.

If at this stage the prompt sheets do not fulfil the validation criteria (the absolute minimum which is required in the end), measures can still be easily taken to repair the errors since no recordings have been made yet. Database producers indicate to highly appreciate this part of validation which allows them to spot and repair errors in an early design stage.

The prompt sheet validation is also a test for the specifications as it pinpoints parts which are underspecified and need further clarification.

2) Lexicon validation by an external expert

A formal check of the lexicon with regard to the format and the use of legal phoneme symbols is part of all the validation stages and can be carried out by the validation centre itself. From experience in the SpeechDat projects, however, a need to check the quality of the phonemic transcriptions has arisen. Since this work needs to be done by phoneticians of each language, the validation centre delegates this task to external experts. There are two conditions for the selection of these experts: they have to be native speakers of their language and must have a phonetic training. These experts check manually a relevant sample of the lexicon. They are instructed to give the provided pronunciation the benefit of the doubt and only to correct transcriptions that reflect an overtly wrong pronunciation. This is in order to prevent marking as errors differences which are due to different phonetic theories or different ideas about what the 'most common' or 'best' pronunciation is.

5) Pre-release validation

The validation of a complete database results in a report containing a list of errors which were found in the database. Some of them are irreparable and related to flaws in the design of the database or the recordings themselves. However, a large number are usually minor and refer to the documentation, label files or other text files which are produced during post-processing. These errors can easily be repaired and the producers are willing to do that. The danger, however, is the introduction of new errors or format inconsistencies during the rectification. Therefore, a pre-release validation has been introduced so that the envisaged master disks can be checked again by the validation centre. The purpose of this validation is to make sure that the minor errors which were found during complete validation are repaired and that no new errors have been introduced. If the pre-release validation is finished with a positive result, the database is ready for distribution and the producers are not allowed to make any more changes, however minor.

It may seem that with these new validation stages the procedure becomes more complex. Our experience, however, is that it also becomes more structured and more efficient with as a result a higher quality of the final product. It should be stressed that the extra stages 1 & 2

are of importance for data collections of which the contents are predictable in advance, whereas the pre-release validation is of relevance for all SLR that need an update after validation.

5. SLR validation and ELRA

SPEX is ELRA's official validation centre for SLR. This work is typical for a setting in which quality assessment and LR repair is performed on a post-hoc basis. SPEX maintains a bug report service for SLR and conducts Quick Quality Checks (QQC) for SLR that are in ELRA's catalogue or are about to enter it.

For the bug report service we refer to http://www.elra.info/ (Services around LRs > Validation > Bug report service). Attractive prizes are offered at a regular basis for the best bugs reported.

A QQC is a shortened version of a full validation still addressing all 8 relevant aspects listed in the introduction section, but only at a formal level for which mainly automatic format checks can be defined and applied. Exception is the documentation which is always manually checked. A QQC can be carried out in say 6 hours whereas a normal full validation takes at least 25 hours.

Depending on the type of application domain of the SLR a set of minimal requirements is formulated (Van den Heuvel et al., 2003). Different sets have now been defined for SLR for Automatic Speech Recognition and phonetic lexicons. Sets of minimal requirements are currently under development for speech synthesis SLR. The QQC will consist of two parts in the future. The first report will present validation results on the SLR proper and will contain comments to the provider; the second report will present validation results on the description forms that ELRA provides with the SLR, and will be directed to ELRA.

ELRA has a validation centre for written language resources (WLR) as well, being CST in Copenhagen. Also CST is developing templates for QQCs and a bug report service for WLR (Fersøe, Monachini, 2004).

6. Conclusion and prospects

Validation of SLR is not static, neither in content nor in procedure. Validation criteria are dynamically adapted to the application domains of the SLR at hand and to the settings in which validation is required.

Apart from that, new data formats require new checks or new implementations of existing checks. This was illustrated on the basis of recent validation work in the TC-STAR_P and LC-STAR project.

The procedures for validation have not reached an end-point either. The introduction of new validation stages at the very beginning and at the very end of database production allows us to more closely assist SLR producers in making a better product.

For existing databases for ELRA's catalogue, new quick check templates are under development to allow for rapid and efficient validation of a SLR at the formal level.

7. References

Broeder, D., Offenga, F., Willems, D., Wittenburg, P. (2001) The IMDI meta-data set, its tools and accessible linguistic databases. Proceedings IRCS Workshop on linguistic databases, 11-13 December 2001. Philadelphia USA. http://www.ldc.upenn.edu/annotation/database/papers/Broeder_etal/32.3.broeder.pdf.

De Vriend, F., Castell, N., Giménez, J., Maltese, G. (2004). LC-STAR: XML-coded Phonetic Lexica and Bilingual Corpora for Speech-to-Speech Translation Proceedings LREC'2004 Workshop on XML-based richly annotated corpora, 29th May 2004.

Fersøe, H. (2003). Validation Manual for lexical. ELRA/0209/VAL-1 Deliverable D1.1A.

Fersøe, H., Monachini, M. (2004). ELRA Validation Methodology and Standard Promotion for Linguistic Resources. In: Proceedings LREC 2004, Lisbon.

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

Hartikainen, E., Maltese, G., Moreno, A., Shammass, S., Ziegenhain, U. (2003). Large lexica for Speech-tp-Speech Translation: from specification to creation. Proceedings Eurospeech 2003, Geneva, Switzerland, September 2003.

Iskra, D., Van den Heuvel, H., Gedge, O., Shammass, S. (2002). Specification of Validation Criteria. OrienTel Technical Report D6.2. http://www.orientel.org.

Laprun, C., Fiscus, J.G., Garofolo, J., Pajot, S. (2002) A practical introduction to ATLAS. Proceedings LREC 2002, Las Palmas.

SAM (1992). User guide to ETR tools. SAM: Multi-lingual speech Input/Output Assessment, Methodology and Standardisation. Ref: SAM-UCL-G007.

Shammass, S., van den Heuvel (2004). Specification of validation criteria for lexicons for recognition and synthesis. LC-Star Technical Report D6.1. http://www.lc-star.com.

Schiel, F., Draxler (2003). Production and validation of speech corpora. Bastard Verlag München. http://www.phonetik.uni-muenchen.de/Bas/.

Van den Heuvel, H. , Choukri,K., Höge,H., Maegaard, B., Odijk, J., Mapelli, V.(2003). Quality Control of Language Resources at ELRA. Proceedings Eurospeech 2003, Geneva, Switzerland, pp. 1541-1544.

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

Van den Heuvel, H. (1999): *Validation criteria*. SpeechDat Car Technical Report CD1.3.1, 1999. http://www.speechdat.com

Van den Heuvel, H., Shammass, S., Moyal, A. (2001): Definition of validation criteria. SpeeCon Technical Report D4.1. http://www.speecon.com/