VALID VALIDATIONS: BARE BASICS AND PROVEN PROCEDURES

Henk van den Heuvel, Eric Sanders

CLST/SPEX, Radboud University Nijmegen, Netherlands

CLST, Radboud University, Erasmusplein 1, 6525 HT Nijmegen, Netherlands

E-mail: {H.vandenHeuvelEric}@let.ru.nl

Abstract

Language resources (LRs) are essential for research and application development. In this article we outline relevant principles for LR validation. We argue that the best way to validate LR is to implement it all along the way of LR production and have it carried out by an external and experienced institute, so that this institute can help define the validation criteria in terms of LR specifications and tolerance margins. We address which tasks should be carried out by the validation institute, and which not. Further, a standard validation protocol is shown, illustrating how validation can prove its value all along the production phase in terms of prevalidation, full validation and pre-release validation.

1. INTRODUCTION

This paper deals with the validation of LRs, more specifically of spoken language resources (SLRs). SLRs are annotated collections of speech data. The difference between a mere collection of speech and an actual SLR is "the fact that the latter is augmented with linguistic annotation (i.e. a symbolic representation of the speech)", as is attested in the EAGLES handbook (Gibbon, Moore& Winski, 1997, p. 146). On the other hand, collections of annotations without accompanying speech data cannot strictly be called SLRs, even when these annotations clearly refer to spoken versions of the database entries, as is the case for e.g. phonemic transcriptions.

By validation of a Language Resource (LR) we refer to a quality assessment of the resource by way of a systematic comparison with its specifications, augmented by a set of tolerance margins for these specifications (e.g. 50% of the speakers should be male, with a permitted deviation of 5%). The specifications (the full set or a subset) and the corresponding tolerance margins are the validation criteria for an LR. The criteria may also come from a set of minimal requirements set by a validation centre which are not explicitly part of the specifications. Output of a validation is a report that lists all checks performed together with an account of the results of the checks.

The relevance of validation of large SLRs emerged when the SpeechDat project (Höge, et al., 1997) was started around 1995. The SLRs within this project were produced in a European framework according to design and recording specifications similar to the American-English Macrophone corpus (Bernstein, Taussig & Godfrey, 1994) and the Dutch Polyphone corpus (Den Os, et al., 1995). The SpeechDat SLRs were, however, produced by a large consortium, the idea being that each consortium member would produce from one to three SLRs and obtain the SLRs produced by the other partners at the end of the project. Because of its experience in the production of Polyphone, and because SPEX was not involved in the production of SpeechDat SLRs, SPEX was included in the consortium as the validation centre with the task to monitor the quality of data and to ensure that all databases would be of comparable quality. The other objective of SpeechDat was that the SLRs become available to third parties after the end of the project. This was another reason for the involvement of an independent validation centre to monitor and ascertain data quality.

Since SpeechDat, SPEX has been involved as validation centre in many projects, particularly in data collections supported by the EU, such as SpeechDat Car, SpeeCon, and OrienTel. The experience on SLR validation gained over the years has been reported at conferences, tutorials and summer schools. This paper presents a comprehensive and up–to-date overview of our experience in the field. Although the paper focuses mainly on the validation of SLRs of the SpeechDat type, experience in validations of other SLRs and pronunciation lexicons will be touched upon where considered appropriate.

In this paper we will address basic principles of validation (section 2) and proven procedures (section 3) and we conclude with lessons learnt from our experiences (section 4).

2. VALIDATION PRINCIPLES

Basic aspects of SLR validation have been addressed in Van den Heuvel, Boves & Sanders 2000), Schiel & Draxler (2003), Van den Heuvel, Iskra, Sanders, De Vriend (2004). A brief overview of SLR validation is also presented by Maegaard, et al. (2005).

2.1 Purposes

Result of a SLR validation is a validation report. This report presents a systematic survey of the validation criteria and the degree in which they were met by the SLR. The report can be used for a variety of purposes:

1. Quality assurance: in this case the validation report attests that the SLR meets the minimum of required specifications and is therefore approved;
2. Quality improvement: the validation report shows to what extent the specifications are achieved. Even if the minimum required criteria are met, the validation report can still be used to improve the SLR to meet the full specifications.
3. Quality assessment: since the validation report describes the extent to which the SLR meets the specifications, it can be added as an appendix to the SLR itself, even if remaining errors have not been corrected.

2.2 Strategies

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 validation criteria are already formulated. (b) A SLR is created, and based on the specifications the validation criteria and validation procedure are defined afterwards. In this way the risk is increased that the validation of some parts of the specification may become infeasible, because in retrospect there is no meaningful way to check these specifications.
Furthermore, validation can be done in house (internal validation) or by another organisation (external validation). The two dimensions thus identified are shown in Table I.

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

Table I: Four types of validation strategies

(1) in this table is in fact essential for proper database production. Each LR producer is responsible for 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. These principles are employed by the Linguistic Data Consortium (LDC) (Cieri, Liberman, 2000; Strassel, et al., 2003). Alternatively (or additionally) an external organisation can be contracted to carry out the validation of an SLR. This is important if the production of database is (sub)contracted or if LR-production is carried out in a consortium where an independent validation institute has to monitor that all SLR are of sufficient quality. In fact, this strategy was adopted by many EU-funded 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. An overview of these projects is presented in Table II (see final page). In that context, all four validation activities shown in Table I are carried out.

This two-dimensional view of the SLR validation process is obviously valid for other types of LRs as well, cf. Fersøe (2004) for lexica.

2.3 The role of validation institute

Validation is just one element in the process of quality control of SLRs. Validation is an instrument to make a diagnosis about the quality of a SLR. It is important to distinguish between the validation and correction of a SLR. The two tasks should not be performed by one and the same institute; a conflict of interest may arise when the validation institute is, in the end, checking its own corrections. The appropriate procedure is that the producer corrects the deviations found and that the validation institute again checks the correctness of the adjustments.

The best position for a validation institute is when it is involved from the very beginning of the design of SLRs. Throughout the design phase, it can contribute its expertise to defining and fine-tuning specifications. It can also make clear from the start which of these specifications can be reliably validated by the institute. During the specification phase the validation institute is responsible for addressing the definition of the tolerance margins for deviations of the specifications.

When the specifications have been agreed upon, the contribution of the validation institute can be of great value by carrying out quality checks at strategic moments during the production process (see section 3 below).

It is important that the validation institute provides efficient feedback on data submissions, and keeps all communication channels open for consultation and feedback on the results found. In practice, this means that:

- The arrival of a data set at the validation office is reported to the producer instantaneously
- The data set is immediately checked for readability and completeness in terms of required files. This is of major importance if the SLR cannot be validated straight away. Readability and completeness issues can be resolved by the provider while the SLR is awaiting its turn in the validation queue.
- If possible in a reasonable time frame, the producer should be allowed to resubmit defective files on the fly during validation.
- The validation report is first reviewed by the producer before it is disclosed to anyone else.

This is correct diplomacy and necessary to avoid and remove any misunderstandings on the text of the report. For instance, a reported error may in fact be a lack of clarity in the documentation, and should be repaired there, not in the database itself. Furthermore, a validation institute can make errors, too! Based on the producer's comments a final report is edited which can be distributed to others.

The validation institute should thus be flexible, and open for communication. However, it must also be determined and assertive. The open communication channel is not meant to wipe out or reason away errors, but to obtain a proper view on their nature and cause.

2.4 Approval Authority

When the validation takes place internally, the approval authority is with the producer. Another situation arises when the producer is not the owner of the SLR (e.g. production is (sub)contracted), or when the SLR is produced within a consortium of partners producing similar SLRs with the aim of mutual exchange, as in SpeechDat. In these cases an external validation institute can play an important intermediary role. The institute can perform an objective test to ascertain whether a producing party has fulfilled the requirements set out by the patron/consortium. In these cases the tasks of the validation institute are typically twofold:

1. Checking a SLR against the predefined validation criteria;
2. Putting a quality stamp on a SLR as a result of the aforementioned check.

In these cases, the validation institute can obtain, as a third task, approval authority. However, this is not a desirable situation. The task of the external validation institute is to provide a comprehensive report in which the remaining deficiencies of an LR, if any, are clearly described. Based on this report the patron (resp. consortium) should decide upon the acceptability of a LR In SpeechDat like projects, the approval of a SLR is commonly arranged in another wat, viz. by a voting procedure. The arrangement and execution of the voting procedures is a task that can very well be delegated to the validation institute.

If a SLR is rejected, the owner will have to correct the deficiencies (re-annotate, or make new recordings) and have the corrected SLR validated once more.

3. VALIDATION TYPES AND PROCEDURES

Over the years SPEX has developed a standard validation protocol for SLR in SpeechDat-like projects, which is, apart from details, also applicable to other types of LR. The protocol is developed along three validation milestones: prevalidation, full validation, pre-release valdation.

3.1 Prevalidation

Prevalidation of a SLR is carried out before the stage of extensive data collection is entered. The main objectives of prevalidation is to detect design errors before serious data collection starts. Secondary objectives are:

- to enable the producer to go through the whole stage of documenting and packaging at the beginning so that missing information, ambiguity and errors in the documentation are avoided at the end
- to develop and fine-tune software for validation of the full database

At the prevalidation phase three components are assessed: prompt sheets, lexicon, mini database. The producer can deliver these components together as one package, or one-by-one, submitting a new component after the previous has been validated.

Prompt sheet validation
Before embarking on recording speakers, the producers design reading scripts. These scripts should be an ideal representation of the content of the corpus items and the number of repetitions for each item. Since in practice not all intended material is recorded 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, etc., the reading scripts contain the (theoretical) upper bounds of types and tokens of what is achievable in a database. You will not get more!

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 maximum achievable number of repetitions at word or sentence level calculated for the complete database. For phonetically rich words and sentences, if included, it can also be checked if a fixed minimum number of tokens per phoneme can be collected, provided that a lexicon containing all the words and their phonemic transcriptions is delivered as well.

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 that they 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 uncover parts which are underspecified and need further clarification.

Lexicon validation
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. However, the quality of the phonemic transcriptions has to be checked as well. Since this work needs to be done by phoneticians familiar with each language, the validation institute contracts this task to external experts. There are two conditions for the selection of these experts: they have to be native speakers of the language and must have a phonetic training. They

obtain the relevant parts of the documentation describing the principles of the phonemic transcriptions employed by the producer. The experts obtain a sample (normally 1000 entries) of the entire lexicon which they have to check manually. They are instructed to give the provided pronunciation the benefit of the doubt and only to mark 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.

Mini database validation
10 initial recordings are made in different environments and annotated. The data is formatted and packaged as if it were a final completed SLR, including documentation, and submitted to the validation institute. The purpose of this part of the prevalidation is to check if all items as specified in the prompt sheets are recorded and, if relevant, in the correct order. Further, the format, and the annotations are inspected, all with the aim of preventing errors during large-scale production. Since the documentation is included as well, the producers are forced to start documenting at an early stage. This may be felt as annoying at that time, but the advantages are clearly felt in the final production phase; the burden of documenting in that phase is greatly reduced to some final text editing and modifications of numeric tables.

3.2 Full validation

When all recordings are collected and annotated, the database is packaged and shipped to the validation institute for what is called full validation. The purpose of the full validation is a quality assessment of the end product. At full validation, all checks are carried out.

The validation institute may have a queue of SLRs to be validated. This queue is typically handled on a First-In First-Out (FIFO) basis. Nonetheless, a more efficient procedure is possible. Upon receiving the SLR, the validation institute can perform a so-called Quick Check: this is a quick formal test running the validation scripts to find out if all required files are included in the SLR and if they have the correct formal structure. If so, the SLR can remain in the queue as it is. If not, the producer is requested to submit updated versions of defective or missing files. Quick Checks avoid discovering, for instance, missing files a few weeks later when the SLR is at the end of the queue. Since action can be taken in the meantime, further delays for both the producer and the validation centre can be avoided. Quick Checks allow the producer and the validation institute to work efficiently in parallel.

Since the validation of the (orthographic) transcriptions is restricted to a sample of all recordings, not all speech data is needed during full validation. For large SLR such as those collected in SpeeCon, copying of all speech files onto a hard disk would use up the main part of the validation effort. For this reason, in SpeeCon and similar projects, the validation institute selected a list of 2000 items during the Quick Check, for which the producer instantly had to provide the speech files. Thus, the producer submitted only a subset of the speech files, so that these were available at the validation institute by the time the SLR reached the top of the queue. Note that all orthographic transcriptions were already delivered for the quick check and that updates of the transcriptions were not accepted at this stage. This avoids that new transcriptions were just made for the subset of files selected for validation.

In case all speech data is needed for validation (e.g. for acoustic quality measurements), submission of the database on DVDs or on a hard disk is a sensible alternative.

If substantial shortcomings are found during validation, rectification and a subsequent re-validation of an SLR may become necessary. This is decided by the owner or the consortium in charge of the SLR production. Since mostly not all parts are defective, re-validation is normally of a partial nature. Re-validations are, as rule, carried out at additional costs for the producing party, so as not to encourage sloppy behaviour. Re-validations may iterate until approval of the LR is achieved.

3.2 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 (manual) annotation and/or 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 reparable errors which were found during complete validation are fixed and that no new errors have been introduced.

After full validation the documentation file is augmented with an additional section: "Modifications after validation". It is checked if all changes agreed upon are included in this section and if they have been implemented in the submitted pre-release version. The validation software is run, so that all formal checks on the data are carried out once more.

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, since these corrections can introduce new (possibly greater) errors.

Also the pre-release phase may have one or more iterations until the LR is approved for distribution.

4. CONCLUDING REMARKS

In this article we have clarified the concept of SLR validation. A standard validation protocol has been shown illustrating how validation can prove its value all

along the production phase in terms of prevalidation, full validation and pre-release validation.

From our experience as validation centre in many (mainly European) projects we have learnt a number of valuable lessons:

- External validation is an important quality safeguard
- If the validation institute is involved during the specification phase of a SLR it can advise in the specification of the design and setting the validation criteria.
- The validation institute can provide important input at strategic points along the data collection and annotation, not only after the completion of the SLR. A good prevalidation procedure can avoid mistakes that would not be reparable at the end.
- The validation institute needs to keep open communication channels to the SLR provider
- Clear validation protocols help structuring the work and effective quality control
- A relevant part of the work of the validation institute is to find a proper balance between developing automatic checks by scripts and hand labour.
- The validation institute, as a rule, does not claim the approval authority for a SLR.
- The validation institute, as rule, does not perform any of the required corrections itself to avoid the situation in which it is checking its own work.

5. REFERENCES

Bernstein, J., Taussig, K., Godfrey, J. (1994). Macrophone: An American English telephone speech corpus for the Polyphone project. Proc. ICASSP-94, Adelaide, pp. 81-83.

Cieri, Chr., Liberman, M. (2000). Issues in Corpus Creation and Distribution: the Evolution of the Linguistic Data Consortium. Procedings LREC2000, Athens, pp. 49-56.

De Vriend, F., Maltese, G. (2004) Exploring XML-based Technologies and Procedures for Quality Evaluation from a Real-life Case Perspective. Proceedings ICSLP-Interspeech 2004, Jeju, Korea

Den Os, E.A. den, Boogaart, T.I., Boves, L., Klabbers E. (1995). The Dutch Polyphone corpus. Proceedings Eurospeech 1995, Madrid, Spain, pp. 825-828.

Fersøe, H. (2004). Validation Manual for lexica. http://www.elra.info

Gibbon, D., Moore, R., Winski, R. (eds) (1997) The EAGLES Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter.

Höge, H., Tropf, H.S., Winski, R., Van den Heuvel, H., Haeb-Umbach, R. & Choukri, K. (1997) European speech databases for telephone applications. Proc. ICASSP 97, Munich, pp. 1771-1774.

Höge, 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, pp. 2699-2702.

Iskra, D., Grosskopf, B., Marasek, K., Van den Heuvel, H., Diehl, F., Kiessling, A. (2002) SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation. Proceedings LREC2002, pp. 329-333.

Iskra, D., Siemund, R., Jamal Borno, J., Moreno, A., Emam, O., Choukri , K., Gedge, O., Tropf, H., Nogueiras, A., Zitouni, I., Tsopanoglou, A., Fakotakis, N. (2004) OrienTel - Telephony Databases Across Northern Africa and the Middle East. Proceedings LREC 2004. Lisbon, pp.591-594.

Maegaard, B., Choukri, K., Calzolari, N., Odijk, J. (2005) ELRA – European Language Resources Association – Background, recent developments and future perspectives. Language Resources and Evaluation (39), pp. 9-23.

Moreno, A., Lindberg, B., Draxler, Chr., Richard, G., Choukri, K., Euler, S., Allen, J. (2000a) SpeechDat Car. A large speech database for automotive environments. Proceedings LREC 2000, Athens, pp. 895-900.

Moreno, A., Comeyne, R., Haslam, K., Van den Heuvel., H., Horbach, S., Micca, G. (2000b). SALA: SpeechDat across Latin America. Results of the First Phase. Proceedings LREC 2000, Athens, Greece, Vol. II, pp. 877-882

Moreno, A., Choukri, K., Hall, Ph., Van den Heuvel, H., Sanders, E., Tropf, H. (2004) Collection of SLR in the Asian-Pacific area. Proceedings LREC 2004, Lisbon, Portugal, pp. 101-104.

Schiel, F., Draxler, Chr. (2003) The production and validation of speech corpora. Bavarian Archive for Speech Signals. Bastard Verlag München.

Strassel, S., Miller, D., Walker, K., Cieri Chr. (2003). Shared Resources for Robust Speech-to-Text Technology. Proceedings EUROSPEECH 2003, Geneva, pp. 1609-1612.

Van den Heuvel, H. Boves, L., Sanders (2000) Validation of content and quality of existing SLR: Overview and Methodology. ELRA Technical report D1.1.

Van den Heuvel, H., Boudy, J., Bakcsi, Z., Cernocky, J., Galunov, V., Kochanina, J., Majewski, W., Pollak, P., Rusko, M., Sadowski, J., Staroniew, P., Tropf, H.S. (2001) SpeechDat-E: Five Eastern European Speech Databases for Voice-Operated Teleservices Completed. Proceedings EUROSPEECH 2001, Aalborg, Denmark, Vol. 3, pp. 2059-2062.

Van den Heuvel, H., Hall, Ph., Moreno, A., Rincon, A., Senia, F. (2004a). SALA II across the finish line : a large collection of mobile telephone speech databases from North & Latin America completed. Proceedings LREC 2004, Lisbon, Portugal, pp. 97-100

Van den Heuvel, H., Iskra D., Sanders, E., De Vriend F. (2004b). SLR Validation : Current Trends & Developments. Proceedings LREC 2004, Lisbon, Portugal, pp. 571-574

Van den Heuvel, H., Choukri, K., Gollan, Chr., Moreno, A., Mostefa, D. (2006) TC-STAR: New language resources for ASR and SST purposes. Proceedings LREC 2006, Genova.

Project	Type of SLR	Number of SLR	Period	Ref.
SpeechDat(M)	Fixed telephone network, for voice-driven teleservices, European languages	8	1994-1996	Höge et al. (1997)
SpeechDat(II)	Fixed and cellular telephone network, for voice-driven teleservices, European languages	28	1995-1998	Höge, et al. (1999)
Speechdat-Car	Car recordings incl. GSM channel, European languages	9	1998-2001	Moreno, et al (2000a)
SpeechDat-East	Fixed telephone network, for for voice-driven teleservices, Central and East European languages	5	1998-2000	Van den Heuvel, et al. (2001)
SALA	Fixed telephone network, for for voice-driven teleservices, Latin America	5	1998-2000	Moreno, et al. (2000b)
SALA II	Cellular telephone network, for for voice-driven teleservices, America (full continent)	16	2002-2005	Van den Heuvel, et al. (2004a)
Speecon	Broadband recordings for commanding consumer devices (major world languages)	28	1999-2002	Iskra et al. (2002)
Network-DC	Broadcast News (Arabic)	1	2000-2001	http://www.elda.org/article45.html
OrienTel	Fixed & Mobile telephone network, for for voice-driven teleservices (Oriental region)	23	2001-2003	Iskra et al. (2004)
TC-STAR	Parliamentary speeches & TTS	3	2004-2007	Van den Heuvel, et al. (2006)
LILA	Mobile telephone network, for for voice-driven teleservices (Asian & Pacific region)	6+	2005-	Moreno et al. (2004)

Table II. Overview of SLR data collection projects with an external validation component. Information about all projects can be obtained via http://www.speechdat.org. For TC-STAR see: http://www.tc-star.org.