Quality Control of Language Resources at ELRA

Henk van den Heuvel[a], *Khalid Choukri*[b], *Harald Höge*[c],

Bente Maegaard[d], *Jan Odijk*[e], *Valerie Mapelli*[b]

[a]SPEX, Nijmegen, the Netherlands; [b]ELRA/ELDA, Paris, France; [c]SIEMENS AG, Munich, Germany; [d]CST, Copenhagen, Denmark; [e]ScanSoft Belgium & Utrecht University, the Netherlands

H.v.d.Heuvel@let.kun.nl

Abstract

To promote quality control of its language resources the European Language Resources Association (ELRA) installed a Validation Committee. This paper presents an overview of current activities of the Committee: validation of language resources, standardisation, bug reporting, patches of updates of language resources, and dissemination of results.

1. Introduction

Quality control of language resources (LR) pertains to an ongoing process of collecting errors, validation of data, creating LR updates, and converting experiences into better specifications and procedures.

For a clearing-house of LR such as ELRA (European Language Resources Association) quality control of its inventory of LR is of utmost importance. The satisfaction of its customers is a direct function of the quality of the LR offered. Thus, for ELRA, the key of commercial survival is a well-defined infrastructure for the quality control of its LR. Obviously, this care for quality does not only involve the LR in ELRA's catalogue but also new LR that are offered for distribution to ELRA.

According to its statutes (article 1), one of the primary activities of ELRA is to give advice, coordinate, and carry out Language Resources validation at a European level. To perform this task, a validation committee, VCom, was set up by the Board, on 23rd October 2000. This Committee takes care of all issues related to quality control that are perceived as important by ELRA's customers and the Board.

At present the VCom addresses the following topics as essential elements of quality maintenance:

- Validation of LR: relevant checks and procedures
- Collection of specifications for various kinds of LR and making an effort in disseminating best practices and guidelines for LR production.
- Bug report services for various types of LR via the internet
- Patches for corrected versions of LR
- Dissemination of work on LR quality via the web.

This contribution presents an overview of the various elements of quality control currently available and partly under development at ELRA. The focus will be on Spoken Language Resources (SLR), since most experience was collected on this type of resources. The activities at ELRA on quality control of Written Language Resources (WLR) will be briefly addressed in section 6 of this paper.

2. ELRA's VCom

The aim of the VCom is to maximize the "ease of use" and "suitability" of the LR needed for Human Language Technologies. For promoting "ease of use", the VCom pushes forward the quality of LR, i.e. helping ELRA to make available validated language resources with optimal documentation and minimal errors. For promoting "suitability", the VCom ensures that ELRA supports standards and best practices for LR leading to best performance of state of the art LE-systems.

Presently the VCom has 9 members, who belong either to the subcommittee for spoken (VCom_SLR) or to the subcommittee for written language resources VCom_WLR). The head of the VCom, Harald Höge, coordinates the work in both subcommittees. The CEO of ELRA takes care that the interests of ELRA are endorsed in all these committees.

Due to the generic work of VCom, most issues are usually handled by all of the VCom members. Currently the operational units of the VCom are ELDA (Evaluations and Language resources Distribution Agency, Paris), and the validation centres SPEX (Speech Processing Expertise Centre, Nijmegen) which coordinates the network for validation of SLR, and CST (Center for Sprogteknologi, Copenhagen) which coordinates the network for validation of WLR.

The tasks of the VCom are:

- Define and supervise tasks performed by the operational units
- Define the validation criteria to be implemented by the validation networks
- Ensure that bug reports are exploited to improve the quality of LRs being distributed (e.g.

production of patches for corrected versions of LRs).
- Dissemination of work on LR quality via the web.
- Report to the board of ELRA

The tasks of the validation centres are:
- Produce validation manuals
- Promote standards and best practices
- Describe the quality of existing LR
- Improve the quality of existing LR
- Maintain the LR validation portals

Specific tasks for ELDA are:
- Communicate with users and producers of LR
- Improve the quality of existing LR
- Maintain the ELRA web pages concerning validation, according to the progress achieved within the VCom.

3. Validation

Validation refers to the evaluation of a LR against a set of quality criteria. These criteria are the specifications of the database and direct derivatives in terms of tolerance margins for deviations of the specifications. Validation checks typically include the following elements of a SLR:
- Documentation: correctness & clarity
- Formats: directory structure & formats and names of files
- Design: completeness of recordings
- Speech files: quality in terms of clipping, SNR, etc.
- Lexicon: completeness & correctness of formats and transcriptions
- Speakers: realistic distributions over gender, age, accents
- Recording environments
- Orthographical transcriptions: format & correctness

Extensive validations were carried out for the SLR collected in the SpeechDat framework, such as SpeechDat II, SpeechDat Car, SpeechDat-East, SALA. Similar extensive validations are also integrated in the projects SPEECON and OrienTel. A characteristic of these projects is that the validation criteria and procedures were set up during the project and that the validations were performed within the lifetime of the project. The SLR created in these projects thus entered the ELRA catalogue in a validated state. Other types of resources have been validated recently, extending the validation criteria from basically read speech databases (the SpeechDat-family) to a new genre which is based on broadcast news speech corpora, as developed in the framework of Network-DC project.

Many other SLR in the catalogue were not subjected to such an extensive (external) validation scenario. Since extensive validations are time-consuming and costly, the VCom instructed SPEX to develop a method for a quick validation of a database. As a result, SPEX introduced the Quick Quality Check (QQC) based on two principles:
1. The QQC mainly checks the database contents against its documentation. The main purpose of a QQC is to check if the documentation of the SLR gives a correct account of the contents of the SLR, in other words if the SLR meets the internal standards set up in the documentation.
2. Generally, the QQC of an SLR should take about half a day's work (for one person at SPEX)

The topics checked in a QQC are basically the same as those in the list of validation elements presented above. The crucial difference with a full validation is that a QQC only comprises a number of formal checks to see if the database contains what the documentation promises. There are no checks on the contents, that is on the correctness of, say, orthographic and phonemic transcriptions. A detailed account of checks is provided in [2].

At present, 64 (out of 228) SLR in ELRA's catalogue are validated, and 12 other SLR have undergone a QQC. After each QQC, the resulting report is sent to the provider of the database for comments. If the provider allows, the QQC is made available via ELRA's web catalogue. A QQC takes about 6 hours in average, which is somewhat longer than the envisaged half day.

4. Standardisation

Standardisation plays an important role in the area of validation. First, *quality standards* directly affect validation, since they describe standard criteria and procedures for validation. A uniform validation procedure according to some pre-defined quality standard makes the quality of different language resources comparable. Second, other standards and guidelines (e.g. for the contents, format and structure of LR) help achieve better and comparable quality of different resources, they make the validation procedure easier and more efficient (since tools can be reused, the experts carrying out the validation are familiar with the contents, structure and format of the resources, etc.) and, of course, they optimise reusability of the language resources proper.

To our knowledge, no official validation standards exist. However, a number of *de facto* standards have emerged in the last decade. The prime example is the validation procedure adopted in the SpeechDat project, in which a wide range of acoustic databases for speech recognition over the telephone has been developed. The

validation specifications and procedures adopted in this project have been applied to all 28 databases developed in the project. More importantly, the general approach to and the methodology for validation have been applied in a wide range of other projects (see section 3 above). The SpeechDat validation methodology has also been applied in various national and industrial projects for the development of telephony speech databases for specific languages. Evidence for the latter point can be obtained from several resources available in the ELRA catalogue (e.g., S0119, S0138, S0142). Furthermore, the methodology for validation developed and applied in the SpeechDat project has formed the basis for developing closely related validation methodologies for speech databases intended for other types of recognizers. Examples of these include the validation methodologies developed in the SpeechDat-Car and SPEECON projects. Though the validation procedures have been adapted to the specific requirements of the database types developed in these projects, the core methodology is the same in these projects as the one used in the SpeechDat project. Of course, the methodology is also continuously being refined. For example, the OrienTel and SPEECON projects have introduced a procedure of pre-validation of prompt sheets, i.e., a validation of all intended material before any recording has actually taken place. Pre-validation eliminates or at least reduces potential errors and other quality issues in a very early stage of the database creation process.

The validation methodology developed in the SpeechDat family of projects has also been taken up in national projects. For example, the Dutch-Flemish intergovernmental Spoken Dutch Corpus project (CGN) has included, from the start of the project, both internal and external validation procedures heavily influenced by the validation methodology developed in the SpeechDat family of projects.

The SpeechDat family of projects has also played an important role in emerging *de facto* standards regarding the contents, structure and format of the databases. For example, the relation between speech and annotation files is based on the same principles in all of these projects. All these projects have used SAMPA as a standard for phonetic notation, thus promoting SAMPA as a standard and extending SAMPA's coverage to new languages, etc. Such standards optimise the usability of resources but also contribute to validation since they make validation easier, more efficient, and therefore cheaper.

5. Bug reporting & Patches

The checks carried out by the validation centre are one source of finding errors in an SLR. However, there is another important source of knowledge about bugs in databases: the people who actually use the LR. In order to get access to this knowledge, ELRA activated a bug report service. Users of SLR who find errors in a database can report these via http://www.spex.nl/validationcentre/bugreport.html/. At regular intervals the most valuable bug report over the period is selected and an attractive prize (ranging from a PDA to a digital camera) is offered to the winner.

The bug report service is described in more detail in [3] and [4]. Find below a brief account of the actions that follow to a bug report to the validation centre of SLR, SPEX.

1. Bug reports are sent to SPEX; SPEX acknowledges the receipt of the report.
2. The bug report is verified by SPEX and, if accepted, added to the formal error list (FEL) maintained by SPEX (for each SLR a separate FEL exists). The updated list is sent to the provider for feedback (by ELDA).
3. ELDA links the formal error list to each SLR in the catalogue if the provider of the SLR allows to do so. The access to the FEL is free of charge and allows bug reporting users to see the status of the bugs of an SLR.
4. Based on an update of the FEL the provider of that SLR is asked by ELDA to correct that part of the SLR which was reported to be faulty. If the provider refuses to correct the files, ELDA or other institutions selected by ELDA produce the corrected part.
5. SPEX produces a patch from the corrected part. This patch produces a new version of the SLR from the old version. ELDA puts the patch into the catalogue.
6. ELDA produces a new version of the SLR with this patch, if the provider of the SLR agrees. This new version of the SLR is put in the catalogue.
7. The patches may be ordered through ELDA.

So far, 6 bug reports have been verified, and 3 prizes have been awarded. The VCom considers the number of bug reports received rather low and is developing new strategies to encourage database users to report the bugs they find. The verified bugs have resulted in 5 FELs. A FEL is included in the web pages with previous consent by the provider.

For one of the current FELs a first patch file will be made in order to test (and fine-tune, if needed) the procedure outlined above.

6. Work on WLR

As most speech applications will require access also to written language resources, e.g. to lexicons or to large corpora, we have decided to include a short description of the validation of WLR.

The work on WLR validation in ELRA started late 2002, but is already in reasonably good shape. First, a validation manual had to be developed. ELRA had already sponsored the elaboration of validation manuals in 1998, both for lexicons and for corpora (cf. [5], [6]). The EU PAROLE project had made a project specific validation manual for lexicons on this basis. The PAROLE project developed lexicons for 9 languages, and 2 of them, Italian and Danish, were validated as part of the PAROLE project. This experience, combined with various experiences from national projects, as well as the experience made at SPEX, have been exploited in the preparation of the ELRA WLR Validation manual for lexicons which now exists in its preliminary version. The manual for corpora is underway.

The first versions of validation manuals for WLR will be put to a test, by using them for the validation of 3-5 LR. After having been validated this way, the manuals will be promoted on the web sites, and we hope that many LR producers will take them and use them in their production work flow. This will be the most important step forward for quality control in WLR production and dissemination.

7. Spreading the news

The VCom is presently also in charge of structuring ELRA's validation web pages in such a manner that all information about SLR maintenance and quality control is accessible in a well readable way. In order to achieve this, the following scheme for the organisation of the web pages was created.

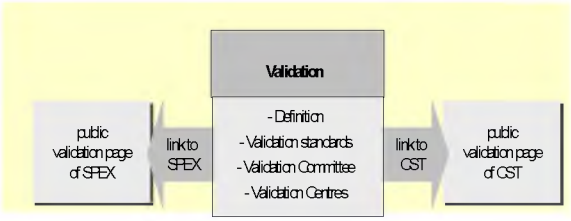

Figure 1: Public part of ELRA's web pages on validation.

At ELRA's website (http://www.elra.info) one can enter the validation part by clicking "Services around SLR" > "Validation". Information can be found regarding: definition of validation, validation standards, the validation committee, and the validation centres. From these subpages the portals of the individual validation centres can be entered. These portals contain status information about the tasks of an individual validation centre. Also the bug report forms are situated at the portals.

As far as possible and granted by the providers, validation reports, QQC reports and FELs are added to the SLR description in the ELRA catalogue. For example, a look at the German SpeechDat II SLR via the catalogue (ID: S0063) shows the options to open the documentation file with the specifications, the validation report and the formal error list (under "imperfection file").

Other channels for spreading the news are the validation networks created by the validation centres. Contacting the most important resource providers, promoting the validation standards to them (and getting feedback), can prove to be the most efficient mechanism for integrating validation in new resource creation projects.

8. Future work

In future validation will be an integrated part in the process to produce new LR. Nevertheless there always will be non-validated LR ready for distribution. For those LR a quick quality check is a useful instrument. The VCom will make an effort in providing various QQCs for selected SLR in its catalogue in the next years. At present ELRA has set-up services like bug report forms, validations and QQCs mainly for SLR; similar activities are now being developed for Written LR, and in the mid-term future for terminology LR. For example, a bug reporting service for WLR will be implemented in the second half of 2003.

Due to the progress in Human Language Technology new kinds of LR will be specified (e.g. multi-modal resources), for which new validation standards have to be developed.

9. References

[1] Höge, H.,"Spoken Language Resources for Voice Driven Man Machine Interfaces", *Proc. LREC98 Granada,* pp.209-216, 1998

[2] Van den Heuvel, H., "Methodology for a Quick Quality Check of Existing SLR", ELRA VCom Deliverable D1.2, 2002.

[3] Van den Heuvel, H., "A Bug Report Service for ELRA", ELRA VCom Deliverable D2.3, 2001.

[4] Van den Heuvel, H., Höge, H. & Choukri, K. "Give me a bug: a framework for a bug report service", *Proc. LREC'2002, Las Palmas*, pp. 569-572.

[5] McEnery, T., Burnard, L.,Wilson, A., Baker "Validation of Linguistic Corpora", ELRA, 1998.

[6] Underwood, N.L., Navarretta, C. "A Draft Manual for the Validation of Lexica", ELRA, 1998.

Institutions:
ELRA: http://www.elra.info
ELDA : http://www.elda.fr
CST: http://www.cst.dk/validation/index.html
SPEX: http://www.spex.nl/validationcentre/