Give me a bug: a framework for a bug report service

Henk van den Heuvel (1), Khalid Choukri (2), Harald Höge (3)

(1) SPEX, Nijmegen, The Netherlands, (2) ELRA/ELDA, Paris, France, (3) Siemens AG, Munich, Germany

H.v.d.Heuvel@let.kun.nl; Choukri@elda.fr; Harald.Hoege@mchp.siemens.de

Abstract

Recently, ELRA initiated the development of a bug report mechanism for the speech databases in its catalogue. This paper reports on the framework of this new service and its practical implementation. Topics dealt with are bug administration, communication with the reporters, formal error listings, corrections of databases, and the release of corrective patches and updated versions of databases. The bug report service is now up and running at http://www.spex.nl/validationcentre/bugreport.html.

KEYWORDS: Speech Databases, Quality Control, Validation, Language Resources

1. Introduction

A glance at the catalogues of database distribution agencies such as ELRA (the European Language Resources Association) [1] and LDC (the Linguistic Data Consortium) [2] shows that language resources (LRs) in general and spoken language resources (SLRs) in particular have grown rapidly in number and in size over the last ten years. Such developments pose a growing demand on LR maintenance, quality control and improvement.

Nowadays, a quality check (also termed 'validation' [3,4]) is integrated in the production of many European SLRs. Validation entails that, during production and immediately after completion, the SLRs created in a project are checked against a set of criteria based on the original specifications and accompanying tolerance margins; a SLR can only be released if it passes the validation. Typical examples of such validated SLRs are the databases in the SpeechDat family [5].

However, this type of validation can only be one slice in the cake of a comprehensive LR quality control procedure. Firstly, many existing SLRs were produced in a project that did not have a validation component. Secondly, bugs may also be found when a validated LR is actually used, e.g. if a SLR is used for training an automatic speech recognizer. An adequate way of reporting the bugs gives way to a wealth of possible LR improvements that otherwise remain unaccomplished.

Bug report services through the internet are offered by reputable software houses like Adobe, Microsoft, Java, and Netscape. For automatic speech recognition products, bug report facilities via internet are offered by e.g. HTK at the University of Cambridge, and Sphinx Speech Recognition Engines at CMU. Bug report services for SLRs in the web already exist at some places: LDC Online allows users to report errors found in LDC SLRs [6], and so do BAS [7] and IDIAP [8].

Recently, ELRA initiated the development of a bug report mechanism for the SLRs in its catalogue. This paper reports on the framework of this new service. This framework is devised for SLRs, but can be tailored to other types of LR where appropriate.

The framework was developed by SPEX under supervision of ELRA's Validation Committee (http://www.icp.inpg.fr/ELRA/services/valcom.php3). At present the bug report service is up and running at http://www.spex.nl/validationcentre/bugreport.html.

2. A framework for a bug report service

Bug report services must be embedded in an effective framework of bug administration, communication with the reporter, error listing, correction and (possibly) re-validation of the database, and issuing new SLR releases. A bug report service should be of greater value than merely allowing a frustrated user of a database to ventilate his or her grievances. If nothing appropriate is done with the bugs, then frustration will only increase!

A proper framework for a bug report service provides a sequence of satisfactory actions (both to ELRA and the customer) to various types of errors.

2.1. Type of errors

The first distinction to be made is that between small and severe errors that are reported. Severe errors refer to substantial deficiencies in elementary properties of the database:

- the quality of the speech files
- the quality of (orthographic) transcriptions
- the lexicon (with phonemic transcriptions)

Reparation of these errors is typically time-consuming because it involves a relatively large amount of human effort.

(Relatively) Small errors typically refer to errors in:

- file names and directories
- annotation/label files
- metadata (e.g. speaker table)

In general, these errors can be repaired without substantial human effort.

The boundary between both types is not very marked. We note that "small" errors may be considered as severe if they show up in huge quantities. Conversely, a "severe" error may be regarded as small if there is only very few of them.

2.2. Appropriate actions to bug reports

In principle, only errors in text files of the SLR are repaired; speech files are not touched. The following procedure for the processing of bug reports is used:

1. Bug reports are sent to SPEX via the public validation page of SPEX (http://www.spex.nl/validationcentre/). SPEX acknowledges the receipt of the report.
2. After the reported bugs are verified by SPEX, then they are added to the formal error list (FEL) maintained by SPEX. The updated list is sent to ELDA[1].
3. The FEL is linked to each SLR in the catalogue (the list may be empty), provided the owner of the SLR allows ELRA to do so (action ELDA).
4. Based on an update of the FEL, the owner of the SLR is asked by ELDA to correct the faulty part. ELDA sends the corrected part to SPEX.
5. If the owner, for any reason, does not rectify the incorrect files, ELDA or other institutions selected by ELDA produce the corrected part.
6. ELDA sends the corrected part to SPEX. SPEX produces a patch from the corrected part. This patch converts the old version of the SLR into the corrected version. The version of the patch and the version of the SLR have to be consistent. SPEX checks that the patch properly integrates the corrected part of the SLR into the latest version of the SLR. SPEX sends the patch to ELDA.
7. The patches can be ordered through ELDA. The corresponding information (cost, version) is then included into the catalogue.

Various details of the procedure are elucidated below.

2.2.1. Formal error list (FEL)

Before a reported error is included in the FEL, it should be verified. The verification of a reported bug is performed by SPEX, if the error is not language-specific. If the error is language-specific (e.g. errors in the orthographical transcriptions), then SPEX consults a qualified institution to check the errors. Such an external check is typically done if a series of such language-specific errors are collected for the SLR (not when just one error is reported). ELRA will pay a reasonable remuneration to the external validator if so required.

SPEX maintains FELs for all SLR in ELRA's catalogue. For each SLR a separate FEL exists. The access to the FEL is free of charge and allows bug reporting users to check the status of the bugs of an SLR.

As long as there are only a relatively small number of errors reported and verified for an SLR, the users should consult the FEL and use this information to their benefit.

2.2.2. SLR correction

When severe (or many small) errors are reported, then rectification of the erroneous file becomes necessary. The rectification of erroneous files is coordinated by ELDA. Minor changes can be performed by ELDA itself. If major changes are needed, ELDA contacts its (language-specific) production centres to fix the files. The owner of the SLR is asked first. Alternatively, if customers (e.g. the reporting one) already made the necessary corrections, then these could be purchased by ELRA (and validated by SPEX). The reporting customer could also be subcontracted to carry out the work. Once corrected, the files are sent to SPEX. SPEX compares the updated files with the formal bug report, and makes a corresponding patch file.

2.2.3. Patch files

The patch is a tar file containing all the files that need to be replaced in order to correct the SLR. A patch file has the following properties.

- The patch adds/substitutes text files; it leaves the signal files unchanged;
- If several patches have to be made for a specific version of an SLR; then they are made in an additive, not in a cumulative way;
- The patch is owned by ELRA and maintained by SPEX;
- The patch files may be used by the receivers for internal use onlyand not be distributed further;
- A patch is associated only with a specific version of the SLR, not with any other version. It should not be supplied with any other version than the one for which it was made.

2.3. Validation

If severe errors are found in more than one elementary property (see section 2.1) of a SLR, then a full validation of the database can be considered. If a (partial or full) validation is deemed necessary by ELRA's Validation Committee, SPEX includes the database in its general validation queue.

SPEX does not carry out any rectifications of SLRs, since a conflict of interests emerges when the corrections need to be validated. In essence, this implies that correction and validation should be iterated until a satisfactory result is achieved.

If validation shows that the errors observed render the database below minimum quality standards, then this information is added to the FEL of the database. In that case ELRA decides what to do with the SLR until the errors are corrected.

2.4. Time schedule

If the time between bug reporting and appropriate action is short, then this will probably encourage SLR users to use the service and make them feel positive about it. Error verification time will be short, presumably about two weeks; however a validation may take longer depending on the length of the general validation queue at SPEX. The progress can then be monitored via the

[1] ELDA (The European Language resources Distribution Agency) is the executive office of ELRA (see http://www.elda.fr).

publicly accessible validation status table that SPEX maintains (http://www.spex.nl).

2.5. Ownership issues

In principle the reporter of the bug is the owner of this information. Therefore, s/he should be aware that s/he transfers all (non-exclusive) exploitation rights on this information to ELRA.

The original SLR and the patches remain strictly separated. The SLR is owned by the owner; the patch is owned by ELRA.

When the patch is run by a user, the original version can be restored by copying the original CDs back.

3. Implementation

3.1. Bug reports

The bug report sheet is a slot-based html-page (see http://www.spex.nl/validationcentre/bugreport.html and the appendix). The tool has slots for the following information:

- SLR name
- Code in ELRA's catalogue
- Coordinates (name, affiliation, e-mail address) of the reporter
- Errors to report
- Desired prize (see section 3.4)

The bug report sheet explicitly states that the bug reporter transfers all rights on the reported information to ELRA (on a non-exclusive basis).

The bug report page also contains a brief explanation of the procedure for bug report handling as presented in section 2.2, together with a few examples of bug reports.

After completion the bug report sheet is (automatically) sent to

1. the validation centre (SPEX)
2. ELDA staff

SPEX created the html page and maintains it at its own validation portal. A link to the page is established from ELRA's web pages (http://www.icp.inpg.fr/ELRA/services/validat.php3).

3.2. Formal error lists

After verification of a reported error, SPEX updates the formal error list for an SLR and sends notification to ELDA. Formal error lists for all SLR in ELRA's catalogue are maintained by SPEX. They have a fixed (but protected) place on an FTP site, from where ELDA can access them.

3.3. Archiving

Each formal error list and each patch should be administered as belonging to a specific version of an SLR. This is reflected in the file name of a formal error list and of a patch file. They are not valid for any other versions. Especially, if the owner/producer of the SLR releases a new version, this becomes relevant. SPEX can be given instructions to update the formal error list for the new version, and ELDA can make a new patch file based on SPEX's findings.

3.4. Rewards for bug reporters

The reporter of the errors should be stimulated to be as precise as possible in the bug reports; s/he should report file names, errors, and suggested corrections. Helpful and attractive essays on how to write good bug reports are those by Black [9] and Tatham [10].

To stimulate the submission of bug reports, two prizes (PDA' s in the range of–600-800 Euros) will be given once a year. One goes to the best contributor, i.e. the person who reports the most, serious, true bugs in a clear manner. The other goes to one of the other contributors by means of a random draw. They are presented to the winners at one of the major conferences, e.g. LREC.

SPEX proposes the best bug reporter to the ELRA's Validation Committee. The Validation Committee makes the final decision.

4. Acknowledgement

The html bug report sheet for ELRA was implemented by Eric Sanders of the Speech Processing Expertise Centre, SPEX.

5. References

[1] ELRA: http://www.icp.grenet.fr/ELRA/home.html

[2] LDC: http://www.ldc.upenn.edu/

[3] Van den Heuvel H. (2000) The Art of Validation. ELRA Newsletter, vol. 5(4).

[4] Van den Heuvel, H., Boves, L., Choukri, K., Goddijn, S., Sanders, E. (2000) SLR Validation: Present State of Affairs and Prospects. Proceedings LREC' 2000, Athens, Greece, Vol. I, pp. 435-440.

[5] SpeechDat projects: http://www.speechdat.org

[6] Bug report service LDC: http://www.ldc.upenn.edu/ldc/service/index.html

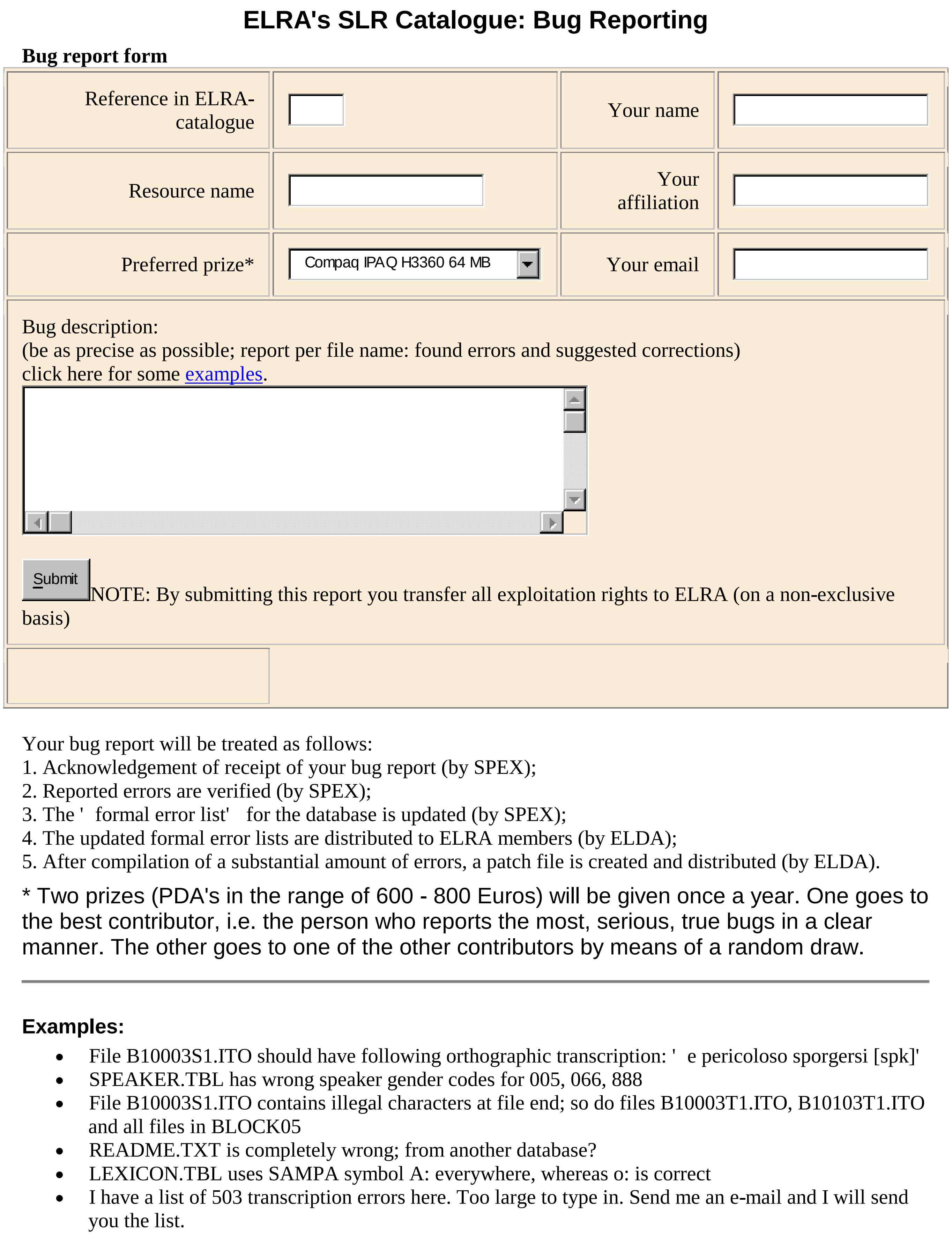
[7]Bug report service BAS: http://www.phonetik.uni-muenchen.de/Bas/BasUpdateeng.html

[8]Bug report service IDIAP: http://www.idiap.ch/cgi-bin/w3-msql/system/dbbug/bugview.html

[9] Black, R. (2000) The fine art of writing a good bug report. http://www.rexblackconsulting.com/publications/Fine%20Art%20of%20Writing%20a%20Good%20Bug%20Report%20(Paper).pdf

[10] Tatham, S. (1999) How to report bugs efficiently. http://www.chiark.greenend.org.uk/~sgtatham/bugs.html

6. Appendix: Screen Shot of Bug Report Sheet (http://www.spex.nl/validationcentre/bugreport.html)



ELRA's SLR Catalogue: Bug Reporting

Bug report form

Reference in ELRA-catalogue		Your name	
Resource name		Your affiliation	
Preferred prize*	Compaq IPAQ H3360 64 MB	Your email	

Bug description:
(be as precise as possible; report per file name: found errors and suggested corrections)
click here for some examples.

NOTE: By submitting this report you transfer all exploitation rights to ELRA (on a non-exclusive basis)

Your bug report will be treated as follows:
1. Acknowledgement of receipt of your bug report (by SPEX);
2. Reported errors are verified (by SPEX);
3. The ' formal error list' for the database is updated (by SPEX);
4. The updated formal error lists are distributed to ELRA members (by ELDA);
5. After compilation of a substantial amount of errors, a patch file is created and distributed (by ELDA).

* Two prizes (PDA's in the range of 600 - 800 Euros) will be given once a year. One goes to the best contributor, i.e. the person who reports the most, serious, true bugs in a clear manner. The other goes to one of the other contributors by means of a random draw.

Examples:

- File B10003S1.ITO should have following orthographic transcription: ' e pericoloso sporgersi [spk]'
- SPEAKER.TBL has wrong speaker gender codes for 005, 066, 888
- File B10003S1.ITO contains illegal characters at file end; so do files B10003T1.ITO, B10103T1.ITO and all files in BLOCK05
- README.TXT is completely wrong; from another database?
- LEXICON.TBL uses SAMPA symbol A: everywhere, whereas o: is correct
- I have a list of 503 transcription errors here. Too large to type in. Send me an e-mail and I will send you the list.