Validation of third party Spoken and Written Language Resources – Methods for performing Quick Quality Checks

Hanne Fersøe[1], Henk van den Heuvel[2], Sussi Olsen[1]

[1]Center for Sprogteknologi (CST) – Københavns Universitet
Njalsgade 80, Copenhagen, Denmark
hanne, sussi@cst.dk
[2] SPEX/CLST – Radboud University Nijmegen
Erasmusplein 1, Nijmegen, Netherlands
H.vandenHeuvel@let.ru.nl

Abstract

This paper presents the experience and insights gained from developing and applying methodologies for quick quality checks (QQC) of third party language resources based on the existing methodologies for full validation, which were documented in validation manuals under contract for ELRA during 2003-2004. The types of resources are Spoken Language Resources (SLR) and Written Language Resources (WLR). The experience gained from applying the QQC methodologies to a number of the resources in ELRA's catalogue is described and on the basis of this, recommendations to the producers of language resources are given. The authors point to the strengths and weaknesses of the current practices, and the similarities and differences between the QQC method and its usefulness for SLR and WLR, respectively, are discussed. Finally a short account of future work is given.

1. Full Validation versus Quick Quality Checks

1.1. Background

The ever increasing importance of easily available language resources for industrial and research purposes is a well established fact, and so is the key importance of the quality of such resources. A validation report resulting from a commonly accepted and standardized validation procedure adds value to a resource as a safeguard of quality, and supports sharing, interchange, availability and reusability of resources.

About a decade ago, ELRA, having as its paramount objective to promote and distribute high quality language resources, found itself in the situation that validation reports were only provided for a part of the SLRs in its resource catalogue, namely those produced in the SpeechDat context (Höge et al, 1999). Validation procedures were not in place for the other resources they distributed. For that reason, ELRA's board decided to help drive and support the creation of quality measures for language resources by setting up a validation committee to handle generic validation issues and to select operational units, validation centres, to be in charge of validation of spoken and written resources, respectively. SPEX, SPeech EXpertise centre in Nijmegen is responsible for spoken resources, while CST, Center for Sprogteknologi in Copenhagen is responsible for written resources. A documented methodology for full validation of third party SLRs was developed first, (van den Heuvel, 2003), and applied to a number of resources (van den Heuvel et al, 2003). Later a methodology for third party WLRs, specifically lexical resources, was developed based on this approach (Fersøe, 2004; Fersøe & Monachini 2004).

ELRA's resource catalogue is available online offered by their distribution agency, ELDA, http://www.elra.info. The catalogue is organized according to type of resource, e.g. spoken, written, multimodal, terminological, and a resource is an entry with an identifier, a name, a description, a price and, possibly, a validation report or a QQC report. The description in the catalogue derives from the Description Form (DF) filled in by the owner or producer of the resource. The description form cannot be accessed online, but it can be obtained on request and is included in the package that ELDA delivers to a buyer.

1.2. The Cost of Validation

Each procedure was created in such a way that a full validation, would ideally take only about 30-40 hours. The larger the resource is, in terms of e.g. words or levels and complexity of annotation, the smaller the selection of samples for content checking is, and vice versa. This amount of hours for a full validation is indeed very low, and it should be seen as the cost a distributor allows for an external validation of a third party resource, which was not and would not be validated by the original producer. The goal of the distributor is to obtain a quality description, and they will therefore accept a certain cost, but will try to minimize it.

The idea behind a quick quality check (QQC) is to minimize the cost even more by describing only the most basic quality aspects of a resource. The goal is that a trained validator by applying tools to automate most of the checking must be able to complete a QQC report in 6-7 hours. For some potential buyers such basic quality measures will be sufficient, for others they may serve as a starting point.

Resource producers that include internal validation in their production plan followed by external validation by an independent validation centre are not very likely to adopt this kind of approach. They will usually allocate and be willing to pay for more manpower because their goal is to make sure that the resource meets the specifications. For other producers, the QQC paradigm offered by and at the cost of ELRA presents a valuable alternative quality assessment to a full external validation of the resource.

2. The QQC Method for SLR

2.1. Content of the Method

As points of departure for the QQC the following principles are adopted:

A. The QQC mainly checks the database contents against a number of minimal requirements. These requirements are of a formal nature which enables a quick, i.e. automatic, check. Content checks are not included because this would involve substantial language-dependent effort.

B. Generally, a QQC should take about 6-7 hours work at maximum.

For each SLR two QQC reports are produced: One for the provider and users on the quality of the database proper (QQC_DB); one for ELDA on the quality of the information on the description forms (QQC_DF). For the templates of the QQC_DF the division as made by ELDA into Speech and Lexicon is maintained.

2.1.1. QQC_DB

The QQC report contains a quality assessment of the resource with respect to a number of minimum formal requirements to specific parts of the resource for example documentation format, transcriptions, lexicon. A star notation is used for this.

Meaning of the quality stars:

* : The minimal criteria for this part of the resource are not met.

**: The minimal criteria for this part of the resource are not completely met.

***: The minimal criteria for this part of the resource are all met.

Other values:

Not Included: This part is not relevant for this resource and not included in the QQC.

Missing: This part is missing in the resource, but relevant.

The QQC_DB checks the documentation regarding completeness and correctness of the SLR description, along similar lines as explained in 3.1.1. for WLR. Further, the QQC concentrates on a series of formal checks regarding:

- directory structure, file names and data integrity
- design in terms of types and tokens of materials contained in the database
- acoustic quality of the speech signals
- formal quality of transcriptions and other annotations (incl. meta-data)

The QQC_DB report is intended for ELRA's database users if the database is already in the ELRA catalogue and for the database providers if the database is new and not in the catalogue yet. Prior to publication, ELDA forwards QQC reports to providers for comments. The final QQC report is made available via ELRA's web pages (catalogue).

2.1.2. QQC_DF

Each database at ELRA is accompanied by one or two description forms: a general description form and/or a specific description form. These description forms contain the basic information about a database according to ELRA. The description forms are filled out by ELDA in cooperation with the LR provider. The form is used to inform potential customers about the database. The information provided on the description form should be correct. The general description form contains information about e.g. the provider (coordinates), price and availability, information on documentation and validation of the resource, and the distribution media. The specific description form contains more detailed information, e.g. for SLR, about the number of speakers and their distribution in terms of gender, age, accent, about included annotation layers and data encoding, and so on.

The QQC_DF report contains a quality assessment of the correctness of the information on the description forms. A star notation is used for this as well.

Meaning of the quality stars:

* : The information provided is insufficient/incorrect.

**: The information provided is close to sufficient/correct.

***: The information provided is complete and correct.

Other values:

Not Included: This information is not relevant for this resource and not included in the QQC.

Missing: This part is missing in the resource, but relevant.

2.2. Applying the QQC Method: Experience gained

SPEX experiences with the QQC method for SLR can be summarized as follows:

- Data collections with many and/or voluminous speech files pose administrative difficulties in the sense that copying the material to hard disk may take a large proportion of the allocated time.
- There is no sensible way to define minimal QQC validation criteria that apply to all kinds of SLR. Currently, SPEX has developed different QQC templates. There are templates for different application domains: ASR, phonetic lexicons, TTS. Templates for multimodal LR are planned.
- The star assessment system needs a good explanation to producers. The QQC departs from the idea that a three star product (highest quality) is provided. Less stars are only provided for serious deviations of the minimal requirements. Small deviations are reported but not penalized in the star assessment.
- An action point is to complete the description forms for the resources in the catalogue. ELDA is currently working on a new procedure to fill in missing information on existing resources.

3. The QQC Method for WLR

3.1. Content of the Method

The QQC method for WLR makes use of the same star notation as the SLR method. A score of one, two or three stars is given for documentation suitability and completeness, formal properties, and reliability of content, respectively. So a few content checks are included here as opposed to the SLR method.

One QQC report only is produced for each written resource and not two, as described in section 2.1. The existing resources in the catalogue targeted by this method

in most cases do not have description forms, partly because there is a stronger tradition for metadata in the SLR area. Spoken resources constituted the main focus for ELRA's resource distribution for a long time, both because many resources of this kind were available for distribution, and because they were more in demand than written resources. The routines and procedures developed for SLRs could not simply be copied, they had to be redeveloped or at least adapted first. This process is complete now, so new WLRs offered for distribution do have description forms, and in a foreseeable future QQC reports for new WLRs will include an assessment of the DF.

3.1.1. Documentation

The documentation is checked manually for suitability, i.e. whether it is clear and to the point and whether it is written in either the language of the resource or in English, the only two possibilities accepted. It is checked for completeness of the information regarding

- copyright and contact persons
- format and character set of the resource files, naming conventions and how to handle them
- languages of the lexical resource, mono-, bi- or multilingual
- type and structure of the entries, lists of legal attributes with mutual dependencies

Ideally, the documentation should specify coverage of the resource, of the domain type, and of the specific information types in the resource. Information on intended applications should also be part of the documentation.

3.1.2. Formal properties

The formal properties concern the usability of the lexical resource. Here the conformance with the specifications is checked, mostly automatically but partly manually, too. Even properties left undocumented can be checked, like e.g. size of the resource, structure of entries etc. These are checked and reported, leading to an added value of the resource.

3.1.3. Content

Finally, a few manual checks on the reliability of the resource content are performed. This is where a QQC differs most from a full validation. About 30 entries are sampled randomly, keeping in mind that different word classes and the different information types must be represented. The sample is checked for correctness of the information types present in the resource in question, be it PoS tag, morphological, syntactic, semantic information or translational equivalents.

3.2. Applying the QQC Method: Experience gained

A summary of CST's experience with the QQC method for WLR is given below.

Documentation may vary a lot in size from one page to several hundred pages. Very short documentation with little information complicates the validation process. Reading very long and detailed documentation takes up a rather large proportion of the time allocated. The extraction of the relevant parts of such documentation is not always straightforward.

Resources are of quite different size and structure, and for large resources or resources with annotation layers in separate files the handling and manipulation of the data is very time consuming.

Lack of conformance with the specifications is a general problem. In nearly all cases the inconsistencies concern the structure of the entries, the attributes and the values allowed. In the worst cases we have checked multilingual resources with two sets of specifications, a general, very detailed and comprehensive one and a language specific one, where the data turned out to be annotated with a combination of the attributes and values from both specifications mixed with other values not documented at all. For other resources we have seen inadequate documentations full of errors where data, if documented at all, do not correspond to the documentation. These examples are of course extreme but very few of the resources checked so far can claim to be fully conformant with their specifications.

Very few content errors are found in the QQCs due to the small number of entries checked but sometimes general and systematic errors are in fact detected. Lexical resources can be of very different nature, ranging from full form wordlists with PoS and morphological information through multilingual resources with semantic information to bilingual collocational resources, and it is indeed important for the credibility of the QQC to check the reliability of the content information stated in the documentation, i.e. to check that the lexical resource is what it claims to be and to give future users an impression of the quality of the resource. The discrepancy between the desire to check the content and the limited time available is quite a dilemma. The credibility of a QQC of a smaller resource is higher than for a larger resource since the percentage of the content checked is higher. Here the methodology still needs further development and a point of revision could be the discussion of whether the star notation should be used for content checks. It is hardly fair to give three stars to a resource of half a million words or more based on a sample of 30 words while this would be far more reasonable for a smaller resource.

4. Comparison of Methods

4.1. Strengths

The strength of the QQC for SLR is that it allows for a good impression of the quality of a SLR, at least at the formal level. A QQC constitutes a good test bed to assess the directory and file structure of a SLR, and it allows for testing of technical completeness and consistency of annotations at various levels. A QQC also gives a good idea of completeness and correctness of the documentation. Further, the procedure provides a general impression of the quality of the signal files by applying a series of acoustic measures on the data.

The strength of the QQC for WLR is that it gives a first quality impression of the basic properties of a resource. It gives some insight into the documentation and the formal properties together with a hint of what problems or shortcomings may exist.

4.2. Weaknesses

An inherent weakness in the SLR QQC is that content correctness is not checked. Within the objectives and time limitations of a QQC the correctness of annotations, e.g. transcriptions, cannot be checked. Especially when hand-crafted annotations are the main part of the LR (such as in phonetic lexicons), the limitations of the QQC approach are felt stronger. However, the alternative of appropriate content checks would lead to substantial amounts of labour by relatively expensive experts, which would exceed the very objective of a QQC.

For WLRs the sparse content checks represents a main weakness both because the quality of linguistic annotations, i.e. the content itself, is frequently the core concern of for the buyer and because the method does not reveal but a small part of the content errors unless these are systematic, in which case they may be detected. Furthermore the content checks are less representative for large resources than for small since the samples, due to the time limit, have to be of the same size.

Another weakness concerns the differing sizes and the differing complexity of the resources, which result in QQCs of varying quality. For some resources the QQC assessments are sound because it was possible to check thoroughly every aspect involved within the limit of the allocated time. But for very large or complex resources the manual checks can only be performed on a rather superficial level. Lists of discrepancies produced automatically are useful for the producer of a resource, but are of less value to a future user.

4.3. Similarities

The QQC approaches are to a large extent parallel and similar, because the underlying assumption is that, regardless of the classification into types such as spoken or written, language resources as such have many features in common, and both the validation and the QQC methodologies should reflect this.

They both build on the same basic assumption that a QQC report provides a valuable quality assessment with a high level of credibility because it is provided by an organization independent of the producer. They use the same simple star notation system to grade the quality and the same criteria for applying the stars. They also proceed through the same steps of checking documentation and formal aspects. Further, the procedures hardly require any language-dependent knowledge from experts, this reducing validation time and costs considerably. Finally, the same amount of time is allocated to QQC a resource, whether spoken or written.

4.4. Differences

The differences are to a high degree, although not completely, motivated by the longer tradition for resources evaluation in the speech community.

There are two QQC templates for each spoken resource, one for the resource itself and one for the description form, and there are variants of the resource template depending on the intended application areas. For written resources only one template exists. There are no variants of this template along the lines of SLR, because it is seldom declared what the intended application area is. The experience up till now shows that variants for mono-, bi-, and multilingual resources are likely to be more useful for written resources than application oriented variants. The linguistic properties of the annotations of bi- and multilingual written resources differ a lot from the monolingual ones, and splitting up the template in two variants would make it possible to skip issues irrelevant to one kind of resource perhaps making it possible to go more into some other issues. But this improvement of the methodology will concern content only, emphasizing the importance of this aspect. For written resources, furthermore, the QQC methodology applies to lexical resources only, while corpora still have to be included or rather have their own variant or their own template altogether.

Ideally most of the QQC work should be done automatically using tools, but currently this is much more the case for spoken than for written resources where more manual checks are made. However tools for WLR are under development.

Along the dimension from SLR to WLR, with phonetic lexicons residing somewhere in between, the proportional contribution (and thus value) of manually encoded annotations increases. Since content checks are not part of the QQC methodology at all for SLR and only absolutely sporadically for WLR, the limitations of the approach as true 'validation' of the data manifest themselves stronger for WLR than for SLR. Nonetheless, also for WLR, the QQC approach offers sufficient means to test the consistency, completeness and formal correctness of the linguistic annotations to acknowledge it as a valuable contribution to data quality assessment.

4.5. Recommendations for Producers

4.5.1. SLRs

In order to maximize the information provision to (potential) customers, SPEX recommends producers:

- To put some effort in completing the LR documentation where required, since complete and correct information substantially enhance the usability of a LR against relatively little costs.
- to complete the description forms for the resources they offer through ELRA
- to provide feedback to QQC reports that are offered. This is to the benefit of the quality of the QQC-report. In addition a good QQC report is a recommendation for the database as a product.

4.5.2. WLRs

All the observations documented in the QQC reports should be taken into account by the producers. To future producers CST has these general recommendations:

It is extremely important that a resource has a good and suitable documentation, not too detailed but clear and to the point.

The coverage of the vocabulary is indeed of interest to a potential user and it should be documented. Very few of the resources which have been QQCed, document the principles for coverage, neither the coverage of domain type, nor the coverage of different word classes or other categories. This is a weakness in quality.

For a potential user lack of conformance between the data and the specifications is a major flaw. Producers are

recommended to establish internal quality procedures during production to prevent this kind of problems. And it is of great importance for both users and producers that such inconsistencies once indicated in a QQC report are repaired, resulting in a more correct resource and subsequently a better QQC.

5. Future Directions

5.1. Consolidation

In the course of the next year the work already done will be consolidated through QQC-work on more spoken and written resources on the one hand, and through subsequent fine tuning of methodologies and templates on the other hand. This will happen in areas where QQC experience reveals that fine tuning is needed, and the methodologies will be extended with more templates where necessary. This paper shows that there are still questions left unanswered, particularly about template variation according to resource type, about the degree of automation of the QQC task, and about the role of content checks at least for lexical resources.

In ELDA's catalogue of LRs a validation report column with links to the reports has been introduced for all SLRs. The values listed in the column are N for no validation, Y for a full validation, and QQC for a quick quality check. A validation report column will also be created for WLRs giving access to existing reports. It is also expected that more resources will have description forms and it will be investigated if and how the WLR template should include or treat these.

5.2. New areas

The major new areas that will be the object of attention in the future are on the one hand the development of a methodology for validation of multimodal resources and on the other hand the creation of a methodology and a QQC template for written corpora.

For written corpora the major challenge is the total size of the data and the number of files. File handling alone may well take more than 5-6 hours, and inspection of the documentation just to get an overview may also be relatively complex. Other new aspects are for example the sources, the selection of texts, the metadata, the principles for transcription and organization of spoken text corpora, the principles of alignment for multilingual corpora, multiple levels of annotation, e.g. text, chapter, paragraph, sentence, word level.

For multimodal resources the objective is to have a closer look at resources produced in the context of the CHIL project[1]. For various modalities, quick checks will be formulated to assess the formal correctness of annotations in individual modalities and mutual consistency between modalities.

Both of these areas will build on the previous work done for ELRA as described above and on other work done lately where both SPEX and CST have acquired experience with these areas.

6. References

Fersøe, H., S. Olsen (2005): Methodology for a Quick Quality Check for WLR-Lexica. Report submitted to ELRA under the ELRA/0209/VAL-1 contract.

Fersøe, H., M. Monachini (2004): ELRA Validation Methodology and Standard Promotion for Linguistic Resources. In *Proceedings LREC 2004, International Conference on Language resources and Evaluation*, Lisboa 2004, page 941-944.

Fersøe, H. (2004). Validation Manual for Lexica. Report submitted to ELRA under the ELRA/0209/VAL-1 contract.

Höge, H., Draxler, Chr., Van den Heuvel, H., Johansen, F.T., Sanders, E., Tropf, H. (1999): Speechdat multilingual speech databases for teleservices: across the finish line. In *Proceedings EUROSPEECH'99*, Budapest, pp. 2699-2702.

van den Heuvel, H. (2003): Methodology for a Quick Quality Check of SLR and Phonetic Lexicons. Report submitted to ELRA under the ELRA/0201/VAL-1 contract.

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

van den Heuvel, H., D.J. Iskra, E. Sanders & F. de Vriend (2004): SLR Validation: Current Trends & Developments. In *Proceedings LREC 2004*, Lisbon, Portugal, pp. 571-574.

[1] http://chil.server.de/servlet/is/101/