

The “Standard Deviation” of LR Quality

Henk van den Heuvel
SPEX/CLST, Radboud University Nijmegen
H.vandenHeuvel@let.ru.nl

Two approaches have been developed to assess the quality of Language Technology (LT) and Language Resources (LRs) during the last decade: Evaluation and validation. The term *evaluation* is used for the quality assessment of LT (systems and tools). The term *validation* is used for quality assessment of LRs. In that context, validation is traditionally defined as the check of a LR against its specifications (often derived from the documentation). However, as I see it, LR validation will increasingly take the shape of LT evaluation in the next decades. The crucial elements in this shift are the increased facilities for standardisation and the improved performance of LT itself. I'll briefly explain this.

The traditional notion of LR validation was developed in the SpeechDat framework. Protocols and procedures were devised to produce LRs that were of equal quality. Uniformity had to be achieved by standardisation. What we have learnt from that experience is that standardisation is a means to warrant:

- (re)usability
- data merging / interoperability
- multi-linguality
- automatic validation

Thus, standardisation is a key to LR quality. Consequently, there is direct relation between validation and standardisation. **Standardisation will become a dominant means for validation** in the future.

The components of a LR are (apart from the data proper, such as audio-files and texts):

1. its metadata
2. its documentation
3. its contents (transcripts, annotations)

The keys to standardisation are the definition of appropriate metadata sets and automatic content generation. Existing LRs usually contain most of their *metadata* (such as speaker information, recording characteristics) in some formal structure that is machine readable and that can easily be converted and/or aggregated into a standard metadata set formalism. Standards for metadata sets are developed by such initiatives as OLAC, IMDI, ISOcat, and most recently CLARIN. Efforts should be made to standardise the pointers to fragments of data as well preferably through the use of persistent identifiers.

The *documentation* is crucial for the proper use of a LR. Therefore, standardising the documentation is one of the primary challenges for the future. As I see it, this standardisation of documentation is directly related to the metadata challenge. After all, documentation is informal metadata, and metadata is formalised documentation. Thus creating standardised metadata sets is the way to standardise documentation as well.

If we have the *content* in some (standard) formalised framework, then this does not say anything about the quality of the content itself. The creation of content (annotations) is essentially a matter of human effort at present. And so is its validation. What we should aim at

is to deploy language and speech technology to generate this content automatically, e.g. ASR for transcripts. The use of tools will also make it easier to adhere to annotation standards. Therefore, in the future, content validation will more and more boil down to the evaluation of the tools that created the annotations. In this way, LR validation will increasingly evolve into LT evaluation.

As a result of this shift the “standard deviation” will become the fundamental measure of LR quality.