

The “Standard Deviation” of LR Quality

Henk van den Heuvel



Presentation overview

1. Validation and Evaluation
2. Validation and Standardisation
3. Roads towards
 - A. Standardisation of LR metadata
 - B. Standardisation of LR documentation
 - C. Standardisation of LR content
4. Conclusion & Perspective



1. Validation and Evaluation

Two approaches for assessing the quality of LR and LT

1. Evaluation

- For LT
- Fixed (evaluation) LRs
- Benchmarking and competition

2. Validation

- For LRs
- Reference: specs, documentation
- Tolerance margins

This presentation is about LR validation





2. Validation and Standardisation

Lesson from LR validation in SpeechDat framework:

- Standardisation is key to:
 - Re-usability
 - Datamerging through interoperability
 - Multi-linguality
 - Automatic validation
- There is a link between validation and standardisation

The link between validation and standardisation will only become stronger in the future.



2. Validation and Standardisation

How?

1. By using standard meta data sets
2. By automatic content generation

Apply these to the main constituents of LRs:

1. Meta data
2. Documentation
3. Content (speech, transcriptions, annotations)



3A. Standardisation of LR metadata

- Convert LR’s metadata into standard metadata sets
- So that they become known in the semantic web and can be harvested from repositories
- Initiatives such as OLAC, IMDI, ENABLER, ISOcat, CLARIN
- Make the metadata sets fine-grained for various LR-types
- Referencing through persistent identifiers



3B. Standardisation of LR documentation

- Documentation = metadata
- Ideally:
 1. Text of the documentation can be generated from metadata
 2. Metadata can be derived from documentation text



3C. Standardisation of LR content

- Create annotations automatically by LT
 - E.g. transcripts by ASR
- Fosters adherence to annotation standards
- Validate annotations automatically
 - And minimize manual validation
- If LR content is generated/validated automatically, then LR validation will increasingly boil down to LT evaluation

To think about:

- Let's go for standardised references to fragments of data & annotations as well





4. Conclusion & Perspective

The “standard deviation” will become the fundamental measure of LR quality

- For metadata
- For documentation
- For content