

# ELRA Validation Methodology and Standard Promotion for Linguistic Resources

Hanne Fersøe

Center for Sprogteknologi, Københavns Universitet

hanne@cst.dk

Monica Monachini

Istituto de Linguistica Computazionale (ILC) –  
Consiglio Nazionale delle Ricerche

Monica.monachini@ilc.cnr.it

---

# Content

1. The Validation Methodology for WLR
  2. Lessons learned from applying it
  3. Standards – in lexicon production and in validation
  4. Future work
-

# Definitions

## *Validation*

- the quality assessment of a Language Resource against one or more checklists of relevant criteria

## *Absolute criteria*

- Derived from the specification of a specific resource
- Derived from the recommended or de facto standard for the creation of such a resource

## *Relative criteria*

- Derived from the requirements of an application, a class of applications etc., e.g. a BLARK specification
-

## Validation scenarios

The producer scenario, in which

- the producer develops the lexicon specifications, produces the lexicon, and specifies the validation criteria

The user scenario, in which

- the resource already exists and where the potential user defines the validation criteria

The agency scenario, in which an agency

- distributes existing lexical resources and promote their reuse
  - wants resources to have been validated according to standard criteria, which make the validated resources comparable
  - maintains a catalogue of available resources preferably with standard declarations of contents and quality
-

## The agency scenario at ELRA

ELRA has developed methodologies and checklists for resource validation in the agency scenario

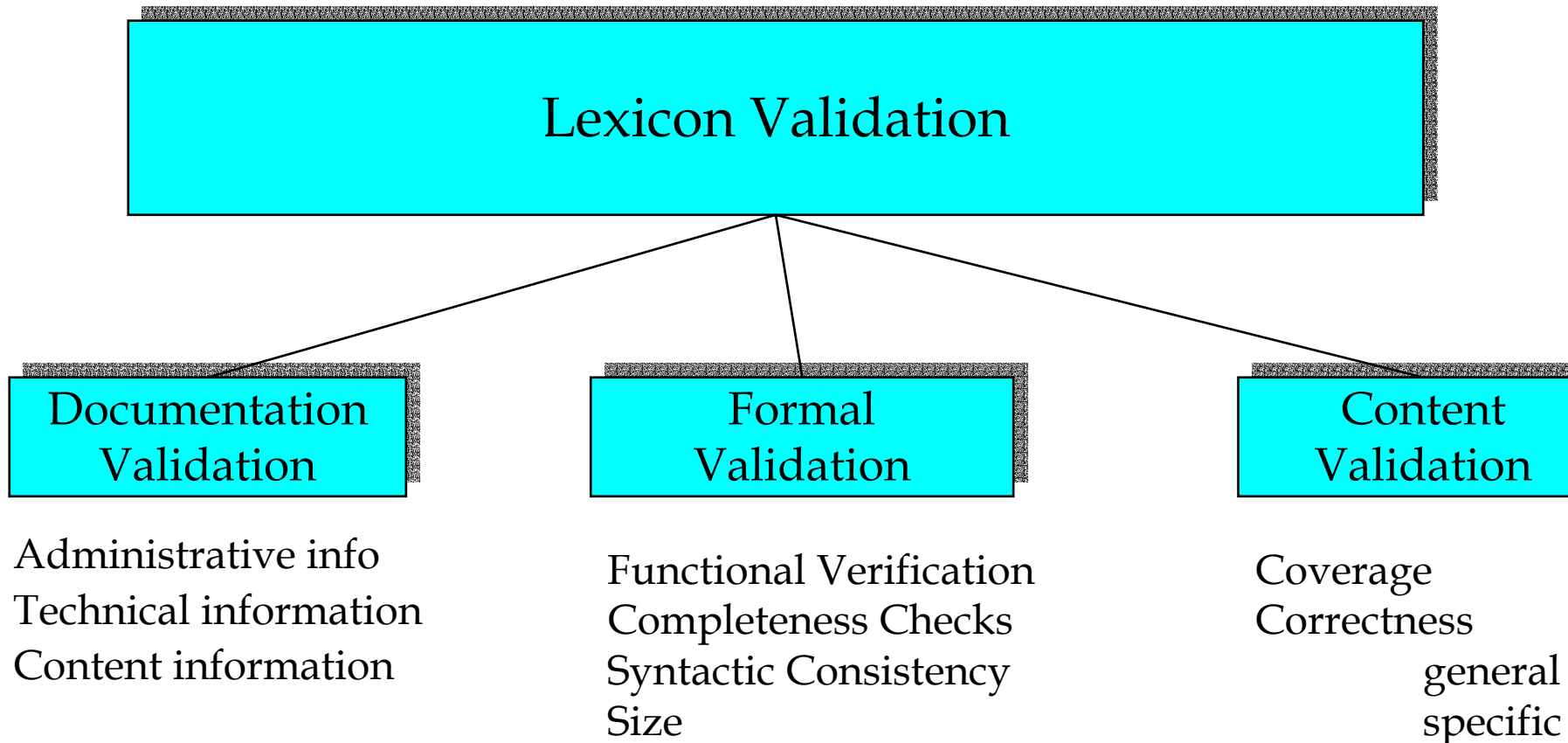
- Written Language Resources (WLR)
  - Corpora
  - Lexica
- Spoken Language Resources (SLR)
  - Speech databases
  - Phonetic lexicons

See [www.elra.info](http://www.elra.info)

---

# The ELRA Validation Manual for Lexica

## Three levels of validation



# Documentation Validation – standard checklists

## basic information

## administrative information

Is there any documentation?  
Is it in English?  
Is there documentation in other languages?  
Is there a 'read me' file in .txt format?  
Is the editor for reading the documentation specified?  
Does the documentation contain administrative information?  
Does the documentation contain technical information?  
Does the documentation contain content information?

Contact person details?  
Number and type of physical media  
The content of each piece of physical media?  
Copyright issues?  
IPR issues?

---

# Documentation Validation – standard checklists

## technical information      content information

Directory structure?

File list?

Accompanying files?

Procedures for unpacking,  
installing, viewing, accessing?

Format and character set?

Data structure of an entry?

Fields of an entry?

Order of fields?

Obligatory fields?

Number of entries?

Mono-, bi- or multilingual?

Language(s) covered?

Types of entry and sorts of  
information with entry?

Legal attributes and their values?,  
their dependencies?

Coverage: Domain/text type?  
degree of? principles for? per POS?

Open/closed classes?

Intended application/system?

Syntactic theory/formalism?

Principles for POS-assignment?

Special word types? foreign words?

## Formal Validation (conformance) – checklists

manual checks                      (semi-)automatic checks

Media?

Handling of media?

Completeness?

Directory structure?

Format and character set?

Readability?

Undocumented files?

Completeness of data model?

Only legal attributes and values?

Are all legal attributes and values used?

Are all obligatory fields filled?

Number of entries OK?

Number of entries for each category OK?

Types of entries OK?

---

## Content Validation

Language dependent, language specific, resource specific

Sample based due to the size of the data material

The validator creates checklists and samples

Checklists for coverage are different for

- General language
- Specific sub-language
- Corpus based coverage

Checklists for linguistic content/correctness should be based on standards, such as the recommendations developed through the EAGLES, PAROLE-SIMPLE, ISLE standardisation initiatives.

---

# Lessons learned

## Questions to consider

- Are the checks appropriate w.r.t. quality level?
- Are the checking results informative?
- Is the amount of details manageable and meaningful?
- Are the checks generic enough for all types of lexica?

## Answers

- Documentation: yes to all questions
  - Formal properties:
    - for syntax and size checks: probably
    - other checks: need more work
  - Content: too early to say, more experience with producing validations is needed because the diversity is huge, but the general direction seems to work
-

## Future work

Produce validation reports for as many lexica and types of lexica as possible

Refine methodology and checks based on the lessons learned from these

Develop a methodology for a quick quality check (QQC) that can be produced in max one day based on a subset of the checks

Continue the promotion of linguistic standards and validation standards in order to encourage that they be used by future resource producers

---