Validation and Distribution of Speech Corpora

Henk van den Heuvel

SPEX: Speech Processing Expertise Centre

CLST: Centre for Language and Speech Technology

Radboud University Nijmegen, Netherlands

SPEX: Mission statement

The mission statement of SPEX is:

1. to provide and enrich spoken language resources and concomitant tools which meet high quality standards
2. to assess spoken language resources
3. and to create and maintain expertise in these fields

SPEX aims to operate:

- for both academic and commercial organisations
- as an independent academically embedded institution

SPEX: Organisation

Employees (in chronological order):

Lou Boves	(0.0 fte)
Henk van den Heuvel	(0.6 fte)
Eric Sanders	(0.5 fte)
Andrea Diersen	(0.7 fte)
Dorota Iskra	(1.0 fte)
Folkert de Vriend	(1.0 fte)
Micha Baum	(1.0 fte)

SPEX: Activities

- **SPEX's main activities at present are the creation, annotation and validation of spoken language resources.**

- SPEX has been selected as the ELRA's primary Validation Centre for speech corpora. Further, SPEX acts as validation centre for several European projects in the SpeechDat framework.

- SPEX is also involved in the creation and/or annotation of SLR.
- SPEX fulfilled several tasks in the construction of the Dutch Spoken Corpus (CGN).

- Publication of results in proceedings, journals

Overview of the presentation

- Validation
  - What is SLR validation
  - Overview of validation checks
  - History of SLR validation
  - Aims of validation
  - Dimensions of validation
  - Validation flow and types
  - What can be checked automatically
  - Validation software
  - On the edge of SLR validation: phonetic lexica
  - SPEX and SLR validation
  - Validation at ELRA & LDC
- Distribution
  - Models of distribution
  - ELRA & LDC

What is SLR Validation? (1)

- Basic question: What is a "good" SLR?
  - "good" is what serves its purposes
  - Evaluation and Validation
- Validation of SLRs:

1. Checking a SLR against a fixed set of requirements;
2. Putting a quality stamp on a SLR as a result of the aforementioned check. If the database passes the check, then we say that it has been "validated"

What is SLR Validation? (2)

- Validation criteria
  - Specifications
  - Tolerance margins
- Specs & Checks
  - have a matrimony in validation
- Validation and SLR repair are different things:
  - Diagnosis and cure
- Dangerous to combine !

Overview of checks

- Documentation
- Database format
- Design
- Speech files
- Label files
- Lexicon
- Speakers and recording environments
- Transcriptions
  - Example: template report SALA II

History of SLR validation (1)

- Production of similar SLRs in (European) Consortia
  - SpeechDat family
- Principle of "Put in one, pull out many"
- "E-quality" (Equality in quality) of SLRs becomes of paramount importance
- Demand for independent validation institute

History of SLR validation (2)

- ELRA has a similar demand for quality control for the SLRs in the catalogue: customers value a quality stamp
- The same is true for the LDC

Aims of validation

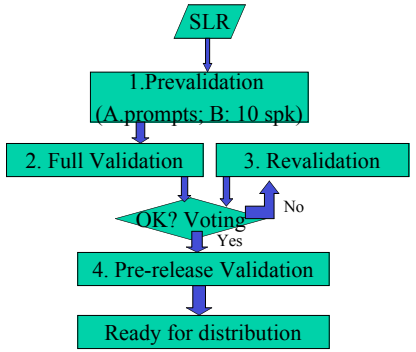
- Quality assurance
- Quality improvement

Dimensions of validation

- Two dimensions:
  - Dim. 1: checks vs specs
  - Dim. 2: subjective vs objective

Validator	**Validation scheduling**	
	During production	After production
Internal	(1)	(2)
External	(3)	(4)

Validation Flow & Types (1)



Validation Flow & Types (2)

- Objectives of prevalidation
  - Detect major shortcomings before recordings start
  - Develop software for:
    - Database formatting (producers)
    - Database validation (SPEX)
- Types of prevalidation
  - Check of all prompt sheets and lexicon (before any recording): is db potentially OK?
  - Mini database of 10 speakers

Validation Flow & Types (3)

- Full validation
  - On *complete* database
  - Preceeded by a Quick Check on formats
  - All checks, incl. transcriptions/completeness checks
- Voting procedure
  - Provider obtains validation report with request to comment to the report and to the list of irreparable shortcomings if any (design/transcription errors)
  - (Updated) report together with main shortcomings & reply provider is sent to consortium with request to vote
  - In case of rejection rectification of the corpus and revalidation is necessary

Validation Flow & Types (4)

- Purpose of Pre-release-validation
  - Final check on master CD before distribution
- Procedure
  - Check if all files are there
  - Check if most recent versions of files are there
  - One more run of validation software to preclude any hidden format defects
- At remaining errors: rectification and revalidation necessary

Validation Flow & Types (5)

- Evaluation:
  - close involvement in the specification phase desired / recommended
  - How to avoid a full revalidation
    - resubmission of files "on the fly"
    - include minor corrections in the documentation file
  - Gap between validation and CD mastering should be kept minimal
  - Validation costs (paradox)

What can be checked automatically ?

Automatic	***By hand/ear***
	Documentation
Database format	
Design	
Speech files	Speech files
Label files	
Lexicon	Lexicon
Speakers and recording environments	
Transcriptions	Transcriptions
	Interpretation of output software
	Editing the validation report

Validation software

- Is it advantageous to distribute the validation software to database providers?
  - **Yes**
    - they can check in advance
  - **No**
    - No double check
    - Validation centre becomes helpdesk
      - Platforms,prog.languages,errors…
    - Delays in database delivery

On the edge of SLR validation: Phonetic Lexicons

- What is an SLR:
  - Speech database
  - Phonetic lexicon
- LC-STAR as example:
  - 12 lexicons with common/application words and names for ASR & TTS: Lemma, phon.transcriptions, POS tags
  - SPEX: XML-format, documentation, phon.transcriptions (see LREC 2004 paper)
  - CST: POS-tags
  - Bilingual lexicons?
  - Corpora?

SPEX & SLR validation (1)

- Checks, specs & SPEX
  - Internal validation of data productions
  - External validation these data: by client or by another institute (CGN)
  - External validator in SpeechDat projects & successors and for ELRA

SPEX & SLR validation (2)

Project	SLR	Period
SpeechDat(M)	8 FDB	1994-1996
SpeechDat(II)	20 FDB 5 MDB 3 SDB	1995-1998
Speechdat-Car	9 CDB	1998-2001
SpeechDat-East	5 FDB	1998-2000
SALA	4-5 FDB	1998-2000
SALA II	12 MDB	2001-2004
LILA	?? ?DB	2004-
SpeechDat-AT	1 FBD,1MDB	2000
SpeechDat-AU	1 FDB	2000
Speecon	18 HDB	1999-2003
NET-DC	1 BCNDB	2002
OrienTel	23 FDB	2001-2004
LC-STAR	12 LEX	2002-2005

SPEX & SLR validation (3)

Principles:

- SPEX validates SLR (not WLR)
- SPEX aims at involvement in the specification phase of a project in order to avoid backward engineering and other infeasibilities afterwards
- SPEX never creates a database that it has to validate itself
- SPEX only checks databases, but does not modify them, to avoid that we check our own work

Validation at ELRA

- Quality assessment of LR in catalogue
- VCOM with two validation centres
  - SPEX for SLR
  - CST (Copenhagen) for WLR
- Tasks
  - Validation manual
  - Bug report handling
  - Quick Quality Checks (QQCs)

ELRA's bug report service

- Accessible via http://www.elra.info
- Bug reports
- Formal error lists
  - Made by validation centre after verification
  - Accessible via web after approval provider
- Correction (by provider)
- Patches

ELRA's QQC procedure

- A QQC is a quick validation restricted to formal properties of a database and the documentation
- Done on LR in ELRA's catalogue or entering it
- Takes about 6 working hours
- Results in two reports:
  - For provider or end-user (about LR proper):
    - Based on check-list minimal requirements
    - Accessible via web after approval provider
  - For ELDA:
    - about information on description forms
- Updates of LR and/or description forms

Validation at LDC/BAS

- Self-produced corpora
- Internal validation
- External corpora are upgraded and reformatted to LDC's own quality standards / BAS has no external corpora
- There are no validation reports for LR available
- Bugs can be reported via website

So much for validation …

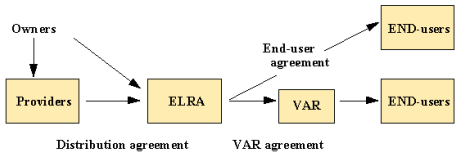
Distribution

- Do it yourself
- Do it via broker (ELRA or LDC)

- Advantages broker: a central place for
  - LR identification
  - Contracts/licenses
  - Marketing/pricing
  - Packaging/shipping
  - Quality maintenance

Distribution at ELRA

- Steps:
  1. Description of LR (by description forms)
  2. Licensing
     - By tailoring generic contract models
     - Usage/pricing/royalties
  3. QQC (if not validated before)



Membership of ELRA/LDC

	ELRA	LDC
Fee	EUR 750 - 5,000	$2000 - 20,000
LR price	Reduced for members	Free for membership year
Member binding	Fidelity program	On-line service

ELRA Sales