The systematic collection of speech corpora for all eleven official South African languages

Marissa van Rooyen
Centre for Text Technology (CTexT™)
Research Unit: Languages in South African Context (School of Languages)
North West University, Potchefstroom Campus (PUK)
South Africa
E-mail: 13017527@student.nwu.ac.za
Introduction

- South African Situation
  - 11 official languages that need to be treated equally
  - Completed resources
    - Spell checkers and CALL-products mostly
    - No workable solution for ASR
- Technology can ease lives and bridge gap
- Goal of this paper: to show a way in which data collection can be organised and managed effectively and quickly
Outline

• Scope of project
• Basic method (5 steps)
• Management of assistants
• Data management
• Suggestions
• Conclusion
Scope of this project

• Developing a speech-driven, telephone-based information system
• All languages in 4 phases
• Phases include recording, transcription and quality control for each language
• 200 speakers per language (male:female, cell:landline, 18-35:36-65)
Basic method

• 5 Basic steps to collecting data
2. Recruit speakers and make appointment
3. Send prompt sheet
4. Record speaker
5. Transcribe recording
6. Final quality control
1. Recruit speakers

• Where to find them:
  – Family and friends
  – Competition
  – Language Boards etc.

• Get demographics and verify first language

• Appointment
2. Send prompt sheet

- Automatically generated by PromptSheetGen
- Includes questions, dates, times, numbers and sentences
- Each one is unique – reference number
Marissa van Rooyen
6 May 2008; SLTU Hanoi

Section [1]
- Thusa o mme ho fihleta mohophi e o kopula ho alaba di potlo ho ka tlabe e e ka tshemo ka motsa o mme (9-4).
- O se ka wa ba di potlo, halole, di alaba festa.

Section [2]
- 1. Pha ya hae ya leaping ho dc?
- 2. Ole lela la ba di kalet?
- 3. Na o mtho e motse ka kopula e mdelela ho?
- 4. Na o ba la ba telefana ka kopula mihale ya fihleta ya tlaletla?
- 5. Thusa o bale mme na lehla e le fihleta ho ho lehla ho lekapa ya fihleta ya tlaletla.
- 6. Le mbala ho lela la ba lekaletla?
- 7. Na o na la molaledi ho?
- 8. O tselelelele la kaletla ho?
- 9. O duale tselelele ho molaledi ho lekaletla?

Section [3]
Thusa o bale nako e tselelela: ::
- 1. Motsoe o tshemo le metsa o molaledi ka mme ho lela ya tselele.

Section [4]
Thusa o golele le tselelela le kaletla: ::
- 1. safaliwe

Section [5]
Thusa o bale le tselelela le kaletla: ::
- 1. 23 Mphakga
Basic method

• 5 Basic steps to collecting data
• Recruit speakers and make appointment
• Send prompt sheet
• Record speaker
• Transcribe recording
• Final quality control
3. Record speaker

- One-A-LOG
- Mute button
- First unofficial QC
  - Listen to answers
  - Listen to quality of voice and reading ability
  - Listen for noise or interruptions that would reduce usability
4. Transcription

• Praat
• According to predefined protocol
  – Other voices in surroundings
  – Non-speech sounds
  – Filled pauses
• Cut out all noisy parts if possible
• Second unofficial QC
5. Final QC

• Listen to recording and read transcript
• Adhere to transcription and recording protocols
• Must be assistant’s first language
Assistants to do the job

• Advantages of assistants:
  – Personal touch necessary in rural areas
  – No overshooting
  – QC throughout stages
• Skilled in language and use of computer
• University students
Management of assistants

- Proved difficult
  - Skill
  - Time
  - Motivation
- Improvement from Phase 1
  - More assistants (from 2 to 7)
  - More languages (trilingual)
  - More pay (hourly vs piecework)
- Only paid for quality work
- From 8 months to 3
Data management

- Database with criteria as fields
Data management (2)

• Common storage location
  1. Keep recordings in batches (date it was recorded)
  2. Move to assistant’s folder on common drive for transcription
  3. Move to folder for QC
  4. Bring everything back to one folder per language for delivery

• Connected to server – automatic backup
• Assistant can work from any station
• Spreadsheet for easy tracking
Suggestions for further improvement

- Relational database
  - User rights on common storage location
  - Faster
- Dedicated recruiters
Conclusion

• Overview of method and practices we used with success
• Time halved from 1\textsuperscript{st} to 2\textsuperscript{nd} phase – might be further reduced
• Quality remains no. 1 priority
Acknowledgements

• CTexT-staff
• The Meraka Institute (CSIR)
• Every assistant, speaker and recruiter

THANK YOU!