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

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

6 May 2008; SLTU, Hanoi

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

PromptSheet

Generated by PromptSheetGen

Ref :: Sesotho_1.1.1_1

Section [1]

- Thusa o eme ho fihlela mohlophisi a o kopa ho araba dipotso tse ka tlase tse leshome le metso e mene (14).
- O se ke wa bala dipotso hahole, di arabe feela.

Section [2]

- 1. Puo ya hao ya lapeng ke efe?
- 2. Dilemo tsa hao di kae?
- 3. Na o motho e motona kapa e motshehadi?
- 4. Na o bua ka selefounu kapa mohala wa fatshe?
- 5. Thusa o bale nomoro e lehlakoreng le letshehadi hodimo hukung ya leqephe lena.
- 6. Le reng lebitso la hao la tswalo?
- 7. Na o na le molekane?
- 8. O tswaletswe kae?
- 9. O dula toropong / motseng ofe jwale?

Section [3]

Thusa o bale nako e latelang: ::

- 1. Metsotso e leshome le metso e mehlano ka morao ho hora ya bosupa.

Section [4]

Thusa o pelete lentswe le latelang: ::

- 1. salletseng

Section [5]

Thusa o bale letsatsi le latelang: ::

- 1. 23 Hlakola

Section [6]

Thusa o bale nomoro e latelang ya founu: ::

- 005 057 858 5

Section [7]

Thusa o bale ho latelang: ::

- 1. Lekala la Matlotlo a Naha Rephaboliki ya Afrika Borwa

Section [8]

- Thusa o bale dipolelo tse latelang jwalo ka ha di hlophisitswe mona ka tlase. Ema nako e ka bang metsotswana e mmedi pakeng tsa polelo ka nngwe.

Section [9]

Dipolelo :: ::

- 1. Sefate se ommeng se thibile tsela e patisaneng.
- 2. Nngo nngwe le be le nahane ho kgaohana.
- 3. Ke ka hona re buang ka leetsetli.
- 4. Ho kgethwa ha Mopresidente.
- 5. Makgowang sonoto e na le morumo o tlamang ka hohlohohlo.
- 6. Moeetho o re bodiba bo utluwa ka moeapa.
- 7. Ha ke utlwa seboko seo ka tsebe hore ho senyehile.
- 8. Kajeno o fumane monyetla wa ho buisetsa molamu sefateng.
- 9. Hlokomelang kgafutso e pakeng tsa mokga e mmedi.
- 10. Pjhapjhathehe, qadika, hlahlatheba, shashara le nyayanyaya.
- 11. Lebopi la leetsehi lona le totobatsa moetsuwi letsahalong.
- 12. Mohlomong ho teng baka tse ding tseo re di sileng.
- 13. Bashanyana ba andha boroku lebotong.
- 14. Ka ya moo ba dutseng hobane e ne e le haufi le keraka.
- 15. Phapang ke efe mahareng a ho bosoeda le ho tshoha?
- 16. Tshepo o tsejwa ke batho ba bangata.

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

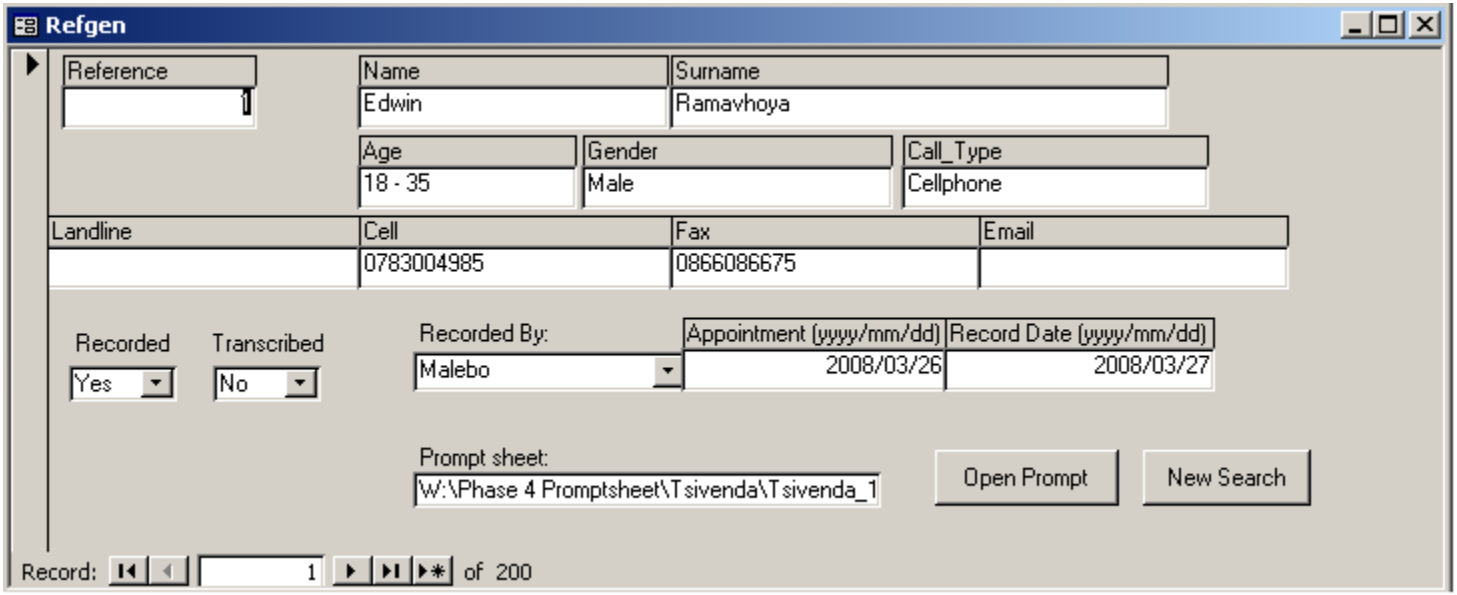
- Advantages of assistants:
  - Personal touch necessary in rural areas
  - No overshooting
  - QC throughout stages
- Skilled in language and use of computer
- University students

Management of assistants

- Proved difficult
  - Skill
  - Time
  - Motivation
- Improvement from Phase1
  - 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 1st to 2nd phase – might be further reduced
- Quality remains no. 1 priority

Acknowledgements

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

THANK YOU!