THE SYSTEMATIC COLLECTION OF SPEECH CORPORA FOR ALL ELEVEN OFFICIAL SOUTH AFRICAN LANGUAGES

Marissa van Rooyen, Cecile van Zyl , Nico Oosthuizen

Centre for Text Technology (CTexT), North-West University (Potchefstroom Campus), Potchefstroom 2531, South Africa

ABSTRACT

In this paper we outline the methods and best practices when collecting speech data for under-resourced languages. The focus of this discussion is on showing ways of improving the quality of the collection and turnaround time. This paper shows how to deal with matters concerning assistants and technical problems, as well as suggesting ways in which data management may be optimised with the use of certain techniques. This article aims at providing the reader with a total overview of improvements made during the course of a real data collection project with tangible problems and results.

1. INTRODUCTION

In a country like South Africa with eleven official languages, there is a large amount of work to be done when collecting and building language resources for indigenous languages. In this paper, one of these efforts will be discussed in a bid to help other under-resourced languages to acquire speech data in a fast and effective manner. The project aims at collecting a speech database from 2200 first-language speakers – no small feat by any means. Currently the project is at the beginning of Phases 3 of 4 and we have already made great improvements in managing and collecting the data and resources – the time it took to complete Phase 1 was halved in Phase 2 and the next phases also show promise of timely completion.

In this paper we will show what the project entailed (Section 2), technical aspects such as software tools that needed consideration at the conceptualisation of the project (Section 3), our processes and methods for completing the data collection will be discussed (Section 4), as well as data management in particular (Section 5). The last section of this paper (Section 6) will give some recommendations as to the further improvement of our system.

2. SCOPE OF PROJECT

The scope of the current project entails the collection of speech data for all eleven South African official languages to be utilised in ASR-applications. The languages have been divided into the following four phases and the process described in this article was followed for every language with the languages in each phase being processed simultaneously:

- Phase 1: Afrikaans, Setswana, isiZulu.
- Phase 2: English, isiXhosa, Sesotho sa Leboa.
- Phase 3: Sesotho, SiSwati, Xitsonga.
- Phase 4: Tshivenda and isiNdebele.

The collecting process requires the collection of data from 200 first-language speakers of each language, thus 2200 in total, for the development of a speech-driven, telephone-based information system for all official languages of South Africa. Each phase also entailed the annotation and the quality control of the final data.

Criteria for the collection of the data were taken and adapted from De Wet *et al*. [1], who suggested certain divisions are important when collecting speech data. Firstly, it was important to make a 50:50 split between cell phones and fixed lines, because, as De Wet *et al*. state, "if telephone-based dialogue systems are to be developed, telephone data from a variety of speakers and handsets must be gathered". Following the discussion in De Wet *et al*, a 50:50 split was made between male and female speakers and finally, two age categories were devised, namely an 18-35:36-65 split.

The prompt sheets distributed to speakers contained additional questions to verify the above data division (see more details in Section 3.3). These questions acted as a means of verifying the speaker demographics as well as being useful speech data. Thus, the first question would confirm the speaker's first-language, the second would ask the speaker's age, the third asked the gender of the speaker and the fourth asked the speaker to state whether he/she was being recorded over a fixed line or a cell phone.

The scope of this leg of the project therefore entailed:

- the recruitment of speakers,
- the collection of speech data,
- the splitting and orthographic transcription of the speech data, and
- the continuous quality control and management of enormous amounts of data.

3. SOFTWARE TOOLS

3.1 One-A-LOG

The recorded data needs to be of the highest quality possible and we have found that it can be hard to achieve this when collecting a large amount of speech data. OmniLOG Voice Solutions (PTY) LTD develops software to assist in this regard.
This project utilised their One-A-LOG Single line Voice Recording System [2] in the recording process.

One-A-LOG can be set up to record to MP3 or WAV-format. We selected WAV-format for this project as the MP3-format is a file format that would cause a loss of audio information. Furthermore, One-A-LOG can be connected to digital or analogue telephone lines. We used analogue lines for this project. All of the assistants were instructed to use the mute button on the telephone handset

whenever the person being recorded answers a question. We had only one large office to our disposal, which had to house five recording stations. The use of the mute button eliminated noise from this very busy centre.

3.2 Praat

Praat [3] is a platform independent, open source software package that is used to transcribe the recorded phone calls orthographically. Praat has a built-in function to annotate the WAV-file. The annotated waveform display makes it easy to distinguish silence from speech, thus making the cutting process quicker and more accurate. The main purpose is to eliminate the operator's voice. This is done by demarcating each utterance and numbering it. A Praat script is used to break the WAV-file into smaller utterances (chunks) so it can be transcribed individually. Praat generates a corresponding TEXTGRID-file for each utterance, which contains information on the WAV-file, as well as the transcribed text.

3.3 PromptSheetGen

For this project PromptSheetGen was developed by CTexT to generate prompt sheets for every language according to the specifications mentioned in Section 2. Based on these demographics a reference number is also assigned to each prompt sheet and thus to each speaker. It was necessary to develop software that could generate 200 unique prompt sheets for each language so as to get randomised coverage in the answers from the speakers. These prompt sheets were incorporated into the database (Section 5.2) and could easily be accessed via the database.

The software receives the required fields in text format. UTF-8 was used to accommodate the diacritic symbols present in some of the languages. The text for the prompt sheets comprises of dates, times and the names of state departments in South Africa for each language. These were chosen to represent the domain for which the data will be utilised. There are also sentences that the speaker must read. These were specifically chosen for diaphone coverage of the language. It is important that all possible combinations of phones are covered in the sentences so that they are useful in ASR applications.

PromptSheetGen can be set to generate sections in the prompt sheet in three different modes:

- *Fixed section:* display all the text contained in a file under a specified heading (headings and instructions to the reader were incorporated in this way).
- *List section:* display a specified number of random lines from an input file under a specified heading. Therefore, different combinations of lines are picked for each prompt sheet (the sentences were added using this mode).
- *Number generator:* display a random number of a specified length under a specified heading. One can choose a prefix if one decides to use numbers in the form of an area's dialling code (we included a telephone number with prefix 0).

The use of this sheet is explained explicitly in Section 4.3.

4. METHODS AND PROCESSES

4.1 Recording Assistants

Many other projects have used automated prompts for making recordings. Another popular option is the call-centre approach in which people were asked to phone in to a centre set up for the recording ([4] and [5]). In this project we found that people in rural areas where some of the under-resourced languages are found, were (a) not able or willing to phone in to the centre, or (b) felt very uncomfortable when talking to a machine. The human touch was needed to secure good quality speakers and recordings.

This choice also meant that we did not overshoot the number of calls needed. In the FRESCO project [4], an over sampling of 25 % was made. This means that they recorded 25 % more data than was ultimately needed. Since this would influence our budget tremendously, we felt it crucial that assistants working on this project knew exactly how many recordings were needed at all times (Section 5.1.). This practice meant that we showed no over sampling. Furthermore, assistants could immediately judge whether a recording was up to the standard because they were hearing it as the recording was made; resulting in a 98 % success rate (the number of finished recordings that in the end passed quality control). To accommodate the 2 % of calls that were not usable, we rerecorded specific calls as it became apparent during quality control (Section 5.2)

It proved to be a challenge to find people willing and able to make the calls. It is especially difficult if you take into consideration the fact that these assistants have to speak and read a variety of 11 official languages (preferably each assistant should be capable of more than one), plus they need to have some degree of computer literacy. When working with university students, it is also true that some will have more time than others, due to academic responsibilities. All of these factors had to be taken into consideration with the selection of assistants.

When starting with this project, we contracted only two assistants per language – one would be doing recordings and the other annotation. This, however, proved impractical. Because the two assistants were constantly waiting on one another, it took over 8 months to complete Phase 1. For Phase 2 we assigned up to seven assistants per language, and also looked for people who could speak all three of the languages in that phase. This resulted in completion of the 600 calls and transcripts in 90 days.

Another change that reduced the time was the payment method. In Phase 1 the assistants were paid on an hourly basis. This hampered productivity to a great extent – we found that assistants were not eager to work faster, because the monetary reward did not increase. The solution to this problem was a change in payment structure. In a meeting with all of the assistants, it was agreed that piecework would be a better option since faster workers would be rewarded accordingly and slower workers would not receive the same compensation for less work. Assistants were, however, only paid for work that adhered to the quality guidelines.

A self-motivational approach was thus introduced. This immediately solved the problem of going over budget, as there was a set figure for 200 calls per language. Assistants also realised that they could increase their payment with increased effort, so

productivity increased drastically. Another positive outcome was the quality of the recordings – as we paid only for good quality work, assistants took greater care in ensuring the usability of their work. They also felt a sense of pride and work satisfaction at the end of the day.

4.2. Recruitment of speakers

Recruitment of speakers had to be managed carefully. It was important that the speaker's mother tongue aligned with the phase we were recording. The speakers themselves got no compensation for their time, so the assistants had to convince them to contribute to the development of their language. South Africa has focussed on the empowerment of smaller language groups for some time now, and assuring a speaker that they were helping in this matter more often than not convinced them to participate. The recording assistants were also asked to recruit speakers within their own families, friends or whatever means they could come up with. Each assistant had access to a telephone and called anyone they knew. Some assistants also took the initiative to call big department stores, government offices and schools and ask for their cooperation.

Management also launched a competition in which the group with the most successful recordings would win a party to the value of R 15 000. In conjunction with this effort, people could act as recruiters only – for every 10 successful recordings we made, they received a R 300 gift voucher at one of five big stores in South Africa. Especially the last incentive received great response. People were eager to send lists of potential speakers in their offices or families. The management of this process was mostly done in the database (Section 5.2). With this approach, we had a response rate of 60 %. This meant that for every 10 potential speakers we phoned, six were recorded – much better than the 12 % in the FRESCO project [4].

4.3. Solicitation Protocol

Using One-A-LOG and the generated prompt sheet, the recording assistant would first recruit a speaker as described above, and then proceed with the recording. The first step is to gather certain information from the speaker, like their telephone number, age and gender. These are all inserted in the database to form a complete record of every one we recorded. Next the assistant will make an appointment with the speaker to phone them at a time suitable for both to do the recording. The assistant will then send the unique prompt sheet to the speaker via fax or e-mail a day or so before the date set for recording. This gives the speaker some time to familiarise him or herself with the content.

On the day of the recording, the assistant will phone the speaker and start the process. The entire conversation is recorded via One-A-LOG and takes the shape of a two-way conversation based on the prompt sheet. The recording takes on average 7 minutes to complete. The assistant will read all instructions (printed in bold on the sheet) and also help the speaker should he or she not understand a question.

It is important that the recording assistant listen to the quality of the recording while it is being made to minimise background noise or interruptions. The assistant must also listen to the geographical information as the speaker gives them so as to insure the correct language, age and so on.

Once the speaker has recorded the entire prompt sheet, the assistant will ask them to repeat unclear sections or simply thank them for their time. The recording is then marked with the reference number as assigned by PromptSheetGen and saved for transcription. This reference number is connected to the recording in every step of the process. Even when the recording is transcribed, the folder has this number as name.

4.4 Transcription of recordings

As this paper has only the collection of data in mind, it is sufficed to say that the recordings were further orthographically transcribed on sentence level in Praat (Section 3.3). This meant that transcribing assistants wrote down what the speaker said during the recording and breaks the recording down into separate sentences or responses from the speaker. Noises in the WAV-file had to be noted and marked as such and the recording assistant's voice was cut out. Noise here means anything not part of the intended recording, such as background noise, other voices or non-speech sounds from the speaker.

This annotation had to be done uniformly according to a protocol or other predefined rules depending on the project that the data will be used for. In this project we had rules not only for marking noise, but also capitalisation and the use of punctuation marks. Both were kept to a minimum as this is influenced by the writing style of the transcriber.

5. DATA MANAGEMENT

5.1 Database design

The amount of data collected made it difficult to keep track of the number of successful recordings. Initially we used a printed list for each language but it soon became impractical. We decided to develop an MS Access [6] database for this purpose. It contained, in addition to the fields required for the transcription and recording, fields to keep track of personal details of the speaker in case queries should arise. The main fields included were the age and gender of the speaker, as well as the type of phone (cell phone or fixed line) used for the recording. A filter dialog was constructed to allow one to easily search for a specific category.

After each recording or transcription had been completed, the assistant would open the database, browse to the specific reference number and mark the required field as completed.

5.2 Common storage location

A common, central storage location is used to improve the management of the resources, as computers are not reserved for specific individuals. Therefore, an assistant is not bound to a single computer and can use any available computer, ensuring that the maximal amount of computer resources is available at any time.

The common storage location is backed up daily and weekly by the server. All of the data is protected in this way from loss or damage, but we felt it necessary to burn the data to DVD's as soon as a phase was complete. All recordings, transcripts and the database for that language was then burnt to disks.

5.3 Transfer of data

The common storage location described above (Section 5.2) simplified the task of data tracking. It meant that data could now be moved from the hard drives of the computers fitted with One-A-LOG to the centrally accessible drive before continuing to the next step in processing. This also meant that version control had to be practiced at all times so as to eliminate double work or confusion over the stage in which what data is. To eliminate this problem was as easy as firstly moving data in batches – at the beginning of each day the previous day's work would be moved to a folder on the central drive. The name of the folder included information about the language of the recordings and the date that it was moved.

To further improve the process, a spreadsheet was created with all of the reference numbers (a unique number, generated by PromptSheetGen, and assigned to a speaker) and the status of each recording, for example, if the recording was moved to the common storage drive, the date of that move was noted, if the recording was transcribed, the date and assistant's name who did the transcript was noted.

These methods worked very well, as one could, at a glance, see the progress of each individual recording as well as keep track of the overall flow of data.

5.4 Quality Control

The quality of the recordings in this project was crucial as the quality of the final speech recogniser is directly related to the quality of the recordings. Quality control was done throughout the process of collecting and annotating data, as well as in a final stage of processing after transcribing.

- During the recording of data, all of the assistants were constantly on the lookout for data exceeding the maximum allowed amount of background noise, or incomplete recordings.
- They were also aware of the fact that first-language speakers were needed and immediately eliminated the speakers that were obviously not first-language speakers. Since each recording had the name of an assistant coupled to it, mistakes could easily be traced back to the person who would have to redo the recording.
- The next stage in processing is annotation. Here assistants can also point out any errors with the recordings. Again the transcriber's name is connected to the data by means of the spreadsheet (Section 5.3).
- As a final stage, quality control was formally done on all 200 of the recordings and transcripts. The quality control was performed by four assistants on every aspect of the data, according to the protocol set at the beginning of the project as point of reference. Quality control basically involved two steps – listening to a section of the recording and checking its orthographic representation.
- Spelling errors were corrected and noises had to be marked. It was important that the assistants doing this detailed work had to be familiar with both the transcription and recording of data. A thorough knowledge of the language was also required.

After this stage, the data was ready to be utilised further to build voice recognition systems.

6. RECOMMENDATIONS

1. The database had to rely on each assistant's honesty as no user rights or logins were set up. It can become difficult when you have up to seven people working on a single language to keep track of "who did what?" Using a relational database could be a better option:
    - Most of these systems have access control by default.
    - It will be faster and more secure than using one database on a common storage location.

2. Other points to consider include appointing additional recruiters whose only job it is to bring names and details of speakers who are willing to be recorded into the centre. This would mean that the possible speakers have already been sifted and may further increase productivity.

7. CONCLUSION AND OUTLOOK

In this paper we outlined the methods and practices we used for collecting speech data for under-resourced languages with the purpose of showing ways with which we improved the quality of the recordings and turnaround time. After successfully recording six of the eleven official South African languages, the project managers are confident in starting the next phase of the project wherein the remaining five languages will be recorded. We have learnt a great deal about fast resource acquisition in an environment that does not always promote under-resourced languages. The knowledge and experience gained in the first phase of the project had the effect of halving the time that recording and annotation took during the second phase.

As we start with the third and fourth phases, working hours may diminish even more, as the implementation of the new methods and processes described in this article would have taken full effect. Delivering high quality data will, however, remain our top priority, since quality cannot be compromised for increased speed of delivery.

8. ACKNOWLEDGEMENTS

The authors would like to thank:

- The entire staff of CTexT for their support and valuable insights.
- All of the almost twenty assistants who worked on this project at some stage – thank you for your loyalty.
- The Meraka Institute (CSIR) for support and funding.
- All of the 1200 speakers and all of the recruiters who found them.

9. REFERENCES

[1] F. de Wet, P. Louw & T. Niesler. 2006. *The design, collection and annotation of speech databases in South Africa.* Proceedings of the Pattern Recognition Association of South Africa (PRASA 2006). Available at www.dsp.sun.ac.za/~trn/conference_papers.html.

[2] http://www.omnilogsa.co.za/Voice LoggingProducts/ onealog.htm

[3] http://www.praat.org

[4] D. Langmann, R. Haeb-Umbach, L. Boves & E. den Os. *FRESCO: The French Telephone Speech Data Collection*. Proceedings of the 4th International Conference on Spoken Language Processing (ISCLP 1996). Available at www.asel.udel.edu/icslp/cdrom/vol3/831/a831. pdf.

[5] H. Mögele, M. Kaiser & F. Schiel. *SmartWEB UMTS Speech Data Collection.* Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). Available at www.phonetik.uni-muenchen.de/Publications/LREC2006_ Moegele.pdf

[6] http://office.microsoft.com