SABC2TXT – High-quality Text Corpora Creation for Low-Resource South African Languages Using Speech Recognition Tools

Project Proposal

Kristen Jodie Basson
Department of Computer Science
University of Cape Town
Cape Town, South Africa
bsskri003@myuct.ac.za

Fardoza Tohab
Department of Computer Science
University of Cape Town
Cape Town, South Africa
thbfar002@myuct.ac.za

Mosamat Sabiha Shaikh
Department of Computer Science
University of Cape Town
Cape Town, South Africa
shkmos004@myuct.ac.za

1. Project description

In South Africa, several African languages - nine out of eleven official languages - lack electronic linguistic resources such as documents or books [7][11]. This lack of resources negatively impacts computational and statistical systems, like language models, that rely on these articles. High-quality text corpora could facilitate text-based research experiments, and one way to produce high-quality text documentation in these languages is to transcribe already available audio in these languages. This transcription could be done by speech recognition tools, such as Automatic Speech Recognition (ASR), which is a mechanical method of decoding oral speech through a microphone, analysing the data with a pattern, model, or algorithm, and generating an output, often text [30][31].

The project aims to evaluate if it is possible to automatically transcribe structured and unstructured audio to create a high-quality textual corpus using standard speech tools and models. The CMUSphinx speech recognition toolkit which has been shown to provide the best results for the transcription of isiXhosa will be used for the audio transcriptions, and the data used to train the acoustic model in this toolkit will be obtained from the publicly available South African Centre for Digital Language Resources (SADiLaR) website.

The project is split into three sections. The first section involves transcribing structured audio, such as SABC broadcast news, publicly available on YouTube. These recordings are seen as a structured environment, as the speech styles present will be formal and read from a script. This audio will be used to evaluate a transcription system that will, if successful, transcribe broadcast news recordings. These transcriptions can be used for further text-based research and to increase the number of electronic documents available in low-resource languages.

The second section involves the evaluation of the quality of using mobile devices for the transcription of unstructured audio, such as a casual conversation between two people in a noisy environment. Mobile device transcription is important since it can aid in the creation of electronic resources for low-resource South African languages. The percentage of people in sub-Saharan Africa who use mobile phones has increased significantly over the past ten years, reaching 60% of the total population [4]. This means that even if they do not have access to more expensive equipment, speakers of low-resource languages can capture and transcribe audio using their mobile devices.

Lastly, a gold standard corpus will be created for the evaluation of the accuracy of the transcribed audio. Gold standard corpora are manually annotated collections of text used as dependable sources of information regarding languages [19]. They are necessary for the training and meaningful evaluation of algorithms. The success of the project would mean that text corpora can be accurately developed using standard speech tools and models, which is essential for natural language processing as it is a primary source of data.

2. Related Work

2.1 Structured Audio Transcription

There have been many studies on the transcription of broadcast news [16]. This research had initially focused on North American English but has since been expanded into other languages such as Japanese [5], Turkish [6], Italian [8], French [10], and many more. These transcription systems require a pronunciation dictionary, a language model, and the training of the acoustic model. This is accomplished by using audio resources to train the model as well as using standard, already-developed, speech recognition tools. There are multiple open-source toolkits available to use. Naidoo and Tsoeu [14] tested English and isiXhosa with three open-source toolkits: Kaldi, HTK, and CMUSphinx. Although the overall results were the best for the Kaldi toolkit, the CMUSphinx performed the best for isiXhosa.

Research on transcription almost always mentions forms of audio segmentation that are used to further divide acoustic signals into homogeneous segments in terms of speaker, gender, and bandwidth [10][13]. The content is then labelled and structured and non-speech segments are identified and removed. Acoustic segmentation is done using the Bayesian information criterion (BIC) framework [8]. Labels for segments are generally classified as F0-baseline broadcast speech, F2-spontaneous broadcast speech, F3-speech over telephone channels, F4-speech with background music, F5-speech from non-native speakers and other segmentations are labelled FX [9][8]. The results of Brugnara *et al.* [8] showed that F0 produced the lowest WER (word error rate).

This is due to the conditions being less noisy and having higher audio quality. There are multiple advantages to using segmentation [12]. Firstly, valuable information can be extracted such as speaker turns and speaker identities. Secondly, speech discontinuity at speaker turns can be avoided. Lastly, identifying and removing non-speech segments as well as dividing the data into smaller segments simplifies decoding and significantly reduces computation time. Furthermore, it is emphasized that model accuracy is greatly dependent on the size of the audio and textual corpora. The larger the corpora to work from the greater the accuracy.

2.2 Mobile audio transcription

Numerous studies have been conducted on how mobile devices can aid in the creation of electronic resources and documents, which can then be used to support studies in search engines, machine learning, and other language models. The following studies were conducted using the CMUSphinx and the PocketSphinx speech recognition toolkits in mobile device applications to transcribe audio. This is the toolkit that will be used to conduct transcriptions of unstructured audio to test the quality of mobile audio transcription.

Lakdawala *et al.* [17] presented an offline voice-to-text transcription system for healthcare organisations. It can be used by counsellors and non-governmental groups to capture talks during surveys, convert them to text, and then save the messages. This system includes an open-source application. The CMUSphinx toolkit is utilized for speech recognition. The system can recognize multiple languages. The language model, phonetic dictionary, and acoustic model are all utilized by the CMUSphinx toolset. The user captures their voice using the mobile application, and the CMUSphinx toolkit analyses and transcribes it. The transcription file will be saved as a text file in the device's memory; the user can use the application to upload and download data to and from the database server. Although the CMUSphinx speech recognition toolset is dependable and accurate, mistakes can nevertheless happen. Background noise, speaker accents, and speech speed, among other things, can affect how accurately a transcript is made [17].

Liu and Zhou [3] introduced a Chinese small-vocabulary offline voice recognition system based on the PocketSphinx toolkit. The language model is built through the online tool LMTool, and the acoustic models are renewed by enhancing the Sphinx models that already exist. Then an offline voice recognition system that might function on an Android smartphone was created. The outcomes of the experiment demonstrated that the system used to recognize voice commands for mobile phones performs well in terms of recognition [3].

These studies highlight how critical it is to recognize the difficulties associated with the transcription of conversations in natural settings and the requirement to thoroughly examine them.

2.2 Gold standard corpus

Sabou *et al.* [21] propose a set of best practice guidelines for crowdsourcing. The steps can be broken down into project definition, data preparation, project execution, and data evaluation and aggregation. Salam *et al.* [22] constructed a representative monolingual corpus in its first phase which consisted of collecting raw data, encoding adjustments, filtering, word segmentation and tokenizing, and annotation. These steps will be used with modifications to suit this project. Hughes *et al.* [23] recruited university students for crowdsourcing. It was deemed efficient as university students have good social networks and technology expertise.

Amazon's Mechanical Turk (MTurk) has been used for crowdsourcing for many low-resource languages [24][25][26]. Due to Amazon not employing and paying workers in South Africa for crowdsourcing tasks, it is not an option. However, practices that deemed the results accurate can be employed. The GATE Crowd plugin introduced by Sabou *et al.* [21] was built using best practice guidelines. It is an open-source plugin that offers crowdsourcing features and can be used in a South African context. ELAN is a media annotation tool that has been used to develop a multi-lingual corpus in South Africa [27]. Data collected can be segmented using one of the many annotation tiers available.

It is integral that the gold standard corpora are of high quality and accuracy to be able to assess the results of the structured and unstructured experiments. Quality assessment techniques such as the Inter-transcriber agreement applied by Munyaradzi and Suleman [28], as well as Zipf's law, token-to-type ratio, and Heap's law as applied by Mustafa and Suleman [29] will be considered. Packham and Suleman [30] found that monetary payments play a large role in motivating corpora crowdsourcing participation and hence why payment will be a part of the project.

3. Problem Statement

3.1 Research problem and aim

The lack of electronic linguistic resources in several African languages in South Africa is a significant problem that negatively impacts computational and statistical systems that rely on these resources. These languages lack documents or books, which makes it difficult to produce high-quality text documentation. Creating high-quality text corpora for these languages is essential for facilitating text-based research experiments.

Because the process of transcribing audio to create text documentation is resource-intensive and often not feasible for low-resource languages, we aim to evaluate the possibility of automatically transcribing structured and unstructured audio to create a high-quality textual corpus using standard speech tools and models such as CMUSphinx, which is essential for natural language processing as it is a primary source of data.

3.2 Research Questions

The main research questions are:

1. How accurate is the use of CMUSphinx for transcribing structured audio such as SABC broadcast news in low-resource South African languages?
2. How accurate is speech-to-text transcription of unstructured audio using the PocketSphinx speech recognition toolkit on mobile devices?
3. Can a gold standard corpus be developed using crowdsourcing for low-resource languages?

4. Procedures and Methods

The procedures and methods will be broken down into three categories: structured audio transcription, unstructured audio transcription, and the creation of the gold standard corpus.

4.1 Structured Audio Transcription

To test the quality of transcription for structured audio using a desktop computer, the CMUSphinx Sphinx-4 toolkit will be used. Sphinx-4 is a pure Java speech recognition library. It provides a quick and easy API to convert speech recordings into text with the help of CMUSphinx acoustic models. It can be used on servers and in desktop applications. Besides speech recognition, Sphinx4 helps to identify speakers, to adapt models, to align existing transcription to audio for timestamping, and more [2]. The Eclipse integrated development environment (IDE) that supports modern build tools such as Apache Maven or Gradle will be used. The process involves the list of steps below.

4.1.1 Data gathering

A dataset of audio recordings that will be used to do the transcriptions on the CMUSphinx-based system will be gathered from publicly available SABC broadcast news on YouTube. These recordings will be structured audio since it is formal scripted speech in a controlled studio environment.

4.1.2 Training CMUSphinx

CMUSphinx will be trained with the resources provided by the SADiLaR website such as the language models, pronunciation dictionaries, and training data for the acoustic model. The training data and model parameters may need to be adjusted over time by using audio with its transcription as a reference to optimize parameters to increase accuracy.

4.1.3 Transcription and testing

Transcription of the audio data obtained will be done on the CMUSphinx-based system and the results will be compared to a standard dataset available online. Other tools supported by the CMUSphinx speech recognition toolkit will be made use of such as tools used for segmenting the audio into speakers as well as speech and non-speech segments [15]. This could greatly increase the accuracy of results and decrease computation time [12]. There are tools supported by CMUSphinx used for speaker diarization, an example being the LIUM toolkit.

The word error rate (WER) metric will be used to test and evaluate the accuracy of CMUSphinx. WER is a common metric used to measure the accuracy of speech recognition systems, and it is calculated as the percentage of words that are incorrectly transcribed by CMUSphinx [14]. The results from testing will be analysed to identify areas where the system can be improved. This may involve tweaking the parameters of the CMUSphinx system or collecting additional training data. The testing and analysis of the system will be done in three iterations.

The gold standard corpus, which will comprise the audio and transcription data utilized for evaluation, will also be used to measure the accuracy of the transcriptions.

4.2 Unstructured Audio Transcription

To test the quality of transcription for unstructured audio on a mobile device, the PocketSphinx speech recognition toolkit will be used. PocketSphinx is built on top of CMUSphinx II. In essence, it is an improvement of CMUSphinx-II to make it compatible with portable mobile devices, which have far lower processing capabilities than desktop PCs, which is where CMUSphinx-II is utilized. It was created in Java and functions mostly on Unix systems [1].

PocketSphinx will be trained and used on Android. The type of audio that will be used for the speech-to-text process is unstructured audio, which includes casual conversations, unstructured interviews, and audio that contains a noisy background or interference from the environment.

The PocketSphinx toolkit will be used and trained on Android. The process involves the list of steps below.

4.2.1 Data gathering

A dataset of audio recordings that will be used to do the transcriptions on the PocketSphinx-based Android application will be obtained from various sources including YouTube, and if possible, recorded from a mobile device.

4.2.2 Training the PocketSphinx toolkit

This step involves choosing a suitable speech corpus from SADiLaR, which includes language models, acoustic models, and dictionaries in a low-resource South African language to train PocketSphinx. The training data will then be used to train the PocketSphinx system. The CMUSphinx training tools will be used to do this. The trained model's correctness can be verified on a different test set to confirm it. To increase accuracy, the training data and model parameters may need to be adjusted over time by running recognition on a pre-recorded reference database to optimize parameters.

4.2.3 Android app development and integration

A simple Android application that does speech recognition using PocketSphinx will be created. The Android Studio IDE will be used to construct the application and include the PocketSphinx libraries. An Android device will be used to test the application to ensure it is working correctly.

4.2.4 Transcription and testing

The audio data obtained will be fed to the decoder in small chunks, and the transcription results are processed as they become available. Furthermore, the separation of audio into speech and non-speech portions is also useful since doing so makes decoding easier and cuts down on computation time [12].

The system will be tested and evaluated using a standard dataset available online to measure its accuracy. The word error rate (WER) metric will be used to gauge how well the audio has been transcribed. The results from testing will be analysed to identify areas where the system can be improved. This may involve tweaking the parameters of the PocketSphinx system or collecting additional training data. The testing and analysis of the system will be done in three iterations.

Because the environment in which the audio data is collected is less controlled than in the case of studio-based or laboratory recordings, it can pose a significant barrier to the quality of the transcription. However, recent developments in CMUSphinx contain a noise cancellation feature and language models need to be retrained with noise cancellation [2].

Similar to CMUSphinx, PocketSphinx offers tools and speaker diarization algorithms that may be used to recognize distinct speakers in an audio recording and give each speaker a special identity. This will be helpful when transcribing audio that contains multiple speakers and increase the accuracy of results [15]. However, it's crucial to note that due to its lightweight design and constrained computing capabilities, PocketSphinx's speaker diarization accuracy could be lower than that of CMUSphinx.

The gold standard corpus, which will comprise the audio and transcription data utilized for evaluation, will also be used to measure the accuracy of the transcriptions.

4.3 Gold Standard Corpora Creation

The corpus development process will be broken down into data gathering and preparation, participant recruitment, project execution, and finally data evaluation.

4.3.1 Participant Recruitment

Participants are needed for gathering data and annotating the data gathered. Consequently, participants will be recruited through the Department of Student Affairs (DSA). This will allow students from the University of Cape Town (UCT) to be part of the research. UCT has a diverse population with students from many backgrounds speaking a variety of languages. Additionally, students are in proximity and as a result are prime and suitable targets for participation.

4.3.2 Data Gathering and Preparation

Two types of data will be gathered corresponding to the structured and unstructured experiments that will be carried out. Structured/controlled audio will be sourced from YouTube, specifically news reports in the respective languages. Unstructured/casual audio will be gathered via participant involvement. Participants will be chosen in groups of two or three and will be given casual conversational topics from a list. The conversations will commence and be recorded to form the data.

The goal is to gather approximately 15 hours each of controlled and uncontrolled audio. The gathered data will then be tagged with relevant information such as whether it is controlled or uncontrolled and the number of speakers. The data will be prepared and presented in a manner that can be used in validating the structured and unstructured experiments.

4.3.3 Project Execution

A standard crowdsourcing tool will be used to collect the transcriptions of the audio datasets. The participants selected for this portion will be trained initially on how to use the platform and then will be given a set of tasks to annotate a selection of audio. Participant progress will be monitored. The data obtained from this will form the gold standard corpora.

4.3.4 Project Evaluation

The reliability of the corpora has to be assessed to ensure that it is of high quality. This will be assured by measuring inter-transcriber similarity. During the project execution stage, each bit of audio will be transcribed by 3 randomly assigned participants. The similarity of the transcriptions will be measured as percentages. The higher the percentage accuracy the more reliable it will be. Statistical tests are needed to assess the validity of a corpus and its adequacy in evaluating experiments. Zipf's law and token-to-type ratio are two tests that can be performed on the corpus.

5. Ethics

Ethics clearance will be needed to recruit and work with participants from UCT for the development of the gold standard corpora. The general procedure entails applying for approval from the Ethics Committee. Once it is approved, an application for permission must be submitted to the DSA.

The project entails minimal risks, and no identifying or sensitive data will be required for the experiment itself. Consequently, privacy will be preserved. The participants will be given all the necessary information to make an informed decision to consent to participate. Voluntariness will be emphasized. The payment amount will not be to a degree that could be considered coercive.

We will report the findings with accuracy alongside limitations and alternative interpretations to ensure a high degree of validity. Unbiased language will be maintained and media with people will be handled with necessary ethical and legal procedures.

6. Anticipated outcomes

6.1 Expected impact of the project

The success of the project would mean that text corpora can be accurately developed using standard speech tools and models. The development of electronic documents is integral in natural language processing as it is a primary source of data.

6.2 Key success factors

- The key success factor for the development of the gold standard corpora would be a high average inter-transcriber similarity. This would indicate that the corpora collected have high transcription accuracy [28].
- The key success factor for determining the accuracy of the structured audio transcription would be a low word error rate (WER).
- The key success factor for determining the accuracy of the unstructured audio transcription would be a low word error rate (WER).

7. Project Plan

7.1 Risks

Please see **Appendix A** for a risk matrix.

7.2 Resources

7.2.1 Hardware

- An Android device for unstructured audio transcription and application testing. Must be capable of running applications with PockeSphinx.
- Desktop PC for structured audio transcription capable of running CMUSphinx, Sphinx-4.
- Recording device

7.2.2 Software

- CMUSphinx toolkit with Sphinx-4 and PocketSphinx (Sphinxbase and PocketSphinx libraries).
- Eclipse IDE which supports Apache Maven and Gradle packages.
- Android Studio
- Additional supported tools for CMUSphinx used for segmentation, speaker diarization, and training.
- Standard/custom crowdsourcing tool

7.2.3 Other

- SADiLaR resources such as language models, pronunciation dictionaries, and training data.
- Participants
- Funding to pay participants

7.3 Timeline

Appendix B, which can be found below, illustrates the overall project timeline. It includes a list of all tasks, deliverables, and milestones for the project.

7.4 Deliverables

Some of the key deliverables for the remainder of the project include:

- Project proposal presentation
- Project proposal
- Ethics application
- Project progress demonstration
- Final project paper draft
- Final project paper
- Final project code
- Final project demonstration
- Project poster
- Website

7.5 Milestones

The following milestones relate to the three sections of the project.

- PocketSphinx and CMUSphinx trained.
- The third iteration of the structured audio transcription experiment is complete.
- The third iteration of the unstructured audio transcription experiment is complete.
- Android app development complete.
- Ethics application complete.
- Structured and unstructured data gathering is complete.
- All data is transcribed manually.

Please see **Appendix B** for the Gantt chart that contains an illustration of the timeline including tasks, milestones, and deliverables.

7.6 Work Allocation

All the research related to the creation of the gold standard corpora will be done by Sabiha. The research and experiments regarding the transcription of structured, broadcast audio will be performed by Kristen. The research and experiments for the unstructured, conversational audio transcription on a mobile device will be performed by Fardoza.

References

[1] B. Lakdawala, F. Khan, A. Khan, Yash Tomar, Rahul Gupta, and Ashfaq Shaikh, "Voice to Text transcription using CMUSphinx A mobile application for healthcare organization," in *Second International Conference on Inventive Communication and Computational Technologies*, April 2018.

[2] N. Shmyrev. 2023. Frequently Asked Questions (FAQ). CMUSphinx Open-Source Speech Recognition. Retrieved April 14, 2023, from https://cmusphinx.github.io/wiki/faq/

[3] X. Liu, and H. Zhou, "A Chinese Small Vocabulary Offline Speech Recognition System Based on Pocketsphinx in Android Platform," in *Applied Mechanics and Materials*, pp. 267–273, August 2014.

[4] J. C. Aker, and I. M. Mbiti, "Mobile Phones and Economic Development in Africa", in *Journal of Economic Perspectives*, pp. 207–232, August 2010.

[5] A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, "Real-time transcription system for simultaneous subtitling of Japanese broadcast news programs," in *IEEE transactions on broadcasting*, vol. 46, no. 3, pp. 189-196, 2000.

[6] E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraçlar, "Turkish broadcast news transcription and retrieval," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 874-883, 2009.

[7] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," in *Speech communication*, vol. 56, pp. 85-100, 2014.

[8] F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani, "A baseline for the transcription of Italian broadcast news," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1667-1670, 2000.

[9] G. D. Cook, D. J. Kershaw, J. Christie, C. W. Seymour and S. R. Waterhouse, "Transcription of broadcast television and radio news: The 1996 ABBOT system," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[10] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMUSphinx III-based system for French broadcast news," in *Interspeech,* 2005.

[11] R. Eiselen, and M. J. Puttkammer, "Developing Text Resources for Ten South African Languages," in *Language Resource and Evaluation Conference,* pp. 3698-3702, 2014.

[12] J.-L Gauvain, L. Lamel, and G. Adda, "Transcribing broadcast news for audio and video indexing," in *Communications of the ACM*, vol. 43, no. 2, pp. 64-70, 2000.

[13] L. Lamel, J.-L. Gauvain, G. Adda, M. Adda-Decker, L. Canseco, L. Chen, O. Galibert, A. Messaoudi, and H. Schwenk, "Speech transcription in multiple languages," in *IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 757-760, 2004.

[14] Naidoo, A. and Tsoeu, M. "Evaluating Open-source Toolkits for Automatic Speech Recognition of South African Languages," in *IEEE SAUPEC/RobMech/PRASA Conference*, 2019.

[15] T. Patel, D. Krishna, N. Fathima, N. Shah, C. Mahima, D. Kumar, and A. Iyengar, "An Automatic Speech Transcription System for Manipuri Language," in *Interspeech*, pp. 2388-2389, 2018.

[16] H. Kamper, F. De Wet, T. Hain, and T. Niesler, "Resource development and experiments in automatic SA broadcast news transcription," in *Workshop on Spoken Technologies for Under-Resourced Languages,* pp. 102-106, 2012.

[17] B. Lakdawala, F. Khan, A. Khan, Y/ Tomar, R. Gupta, and A. Shaikh, "Voice to Text transcription using CMUSphinx A mobile application for healthcare organizations" in *Second International Conference on Inventive Communication and Computational Technologies*, April 2018)

[18] M. G. Rathando, "Representation of official languages on South African Broadcasting Corporation (SABC) television: A study of selected Tshivenda Programmes,", 2022.

[19] Corpora - English Language: a short guide to online resources. Retrieved March 23, 2023 from https://libguides.bodleian.ox.ac.uk/english-language/Corpora

[20] L. Wissler, M. Almashraee, D. Monett, A. Paschke, "The Gold Standard in Corpus Annotation," in *Proceedings of the IEEE GSC,* Passau, Germany, 26–27 June 2014.

[21] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus annotation through crowdsourcing: Towards best practice guidelines," in *Proc 9th Int. Conf. Lang. Resour. Eval.*, pp. 859–866, 2014.

[22] K. M. A. Salam, M. Rahman, and M. M. S. Khan, "Developing the bangladeshi national corpus-a balanced and representative bangla corpus," in *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1–6, IEEE, 2019.

[23] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010), International Speech Communication Association,* pp. 1914-1917, 2010.

[24] M. Post, C. Callison-Burch, and M. Osborne, "Constructing parallel corpora for six Indian languages via crowdsourcing," in *Proceedings of the 7th Workshop on Statistical Machine Translation. Association for Computational Linguistics*, pp. 401–409, 2012.

[25] M. Marge, S. Banerjee, & A. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 5270– 5273, 2010.

[26] Gelas H., Abate S.T., Besacier L. and Pellegrino F, "Evaluation of crowdsourcing transcriptions for African languages," in *Conference on Human Language Technologies for Development,* Alexandria, Egypt, 2011.

[27] E. van der Westhuizen, and T. Niesler, "A first South African corpus of multilingual code-switched soap opera speech," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

[28] N. Munyaradzi, and H. Suleman, "Quality assessment in crowdsourced indigenous language transcription," in *International Conference on Theory and Practice of Digital Libraries*, Valletta, Malta, Proceedings 3, pp 13-22, 22-26 September 2013.

[29] M. Mustafa, and H. Suleman, "Building a Multilingual and Mixed Arabic-English Corpus," in *Proceedings Arabic Language Technology International Conference*, Alexandria, Egypt, 9-10 October 2011.

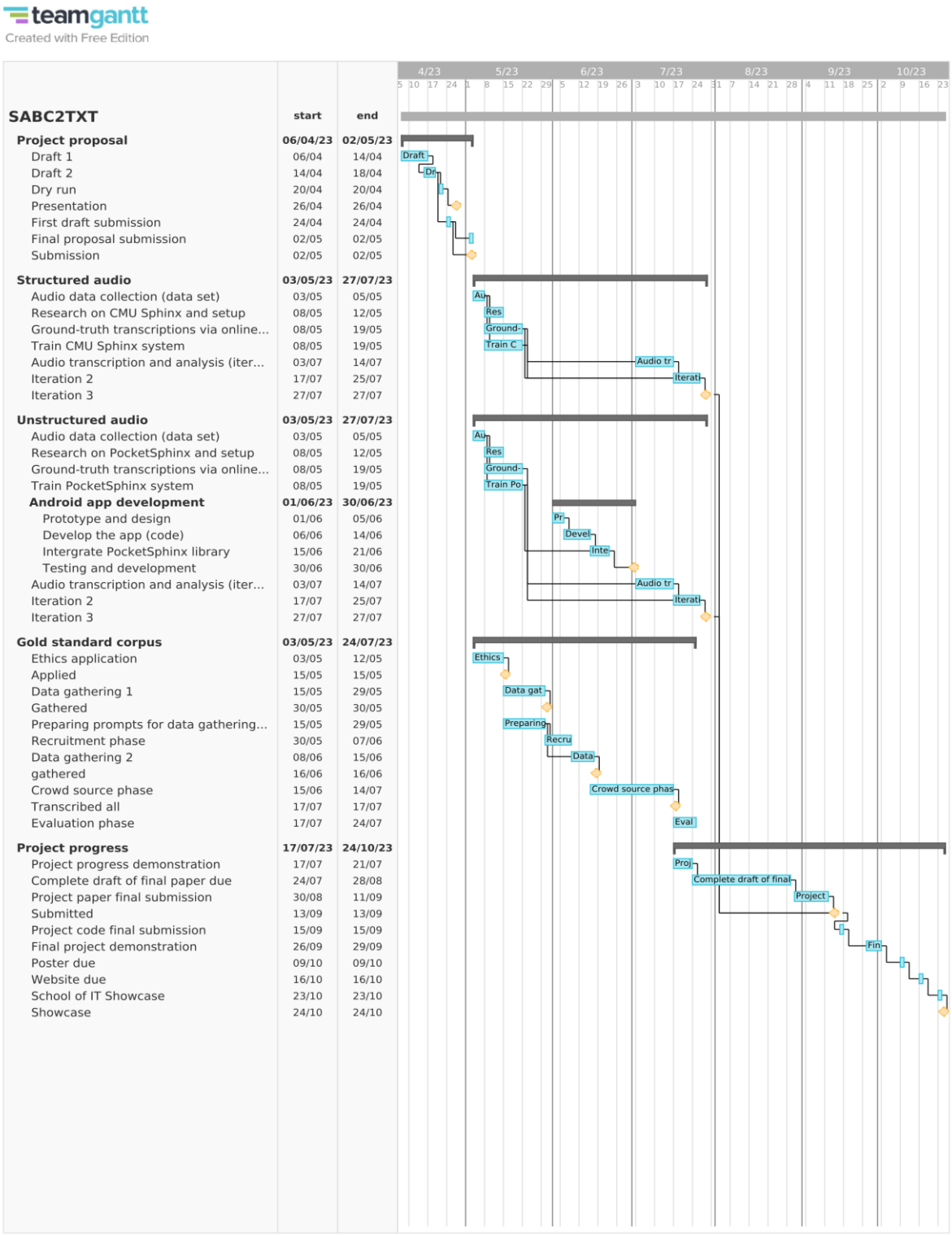
[30] S. Packham, and H. Suleman, "Crowdsourcing a Text Corpus is not a Game," in *Digital Libraries: Providing Quality Information: 17th International Conference on Asia-Pacific Digital Libraries*, ICADL 2015, Seoul, Korea, 9-12 December 2015.

[31] J. Lai, Karat, C. M. Karat , and N. Yankelovich, "Conversational speech interfaces and technologies," In *Human-Computer Interaction,* CRC Press, pp. 71-82, 2009.

[32] J. Levis. and R. Suvorov, "Automatic speech recognition," in *The encyclopedia of applied linguistics,* 2012.

Appendix A						
Risk	**Consequence**	**Probability**	**Impact**	**Mitigation**	**Monitoring**	**Management**
Ethics clearance is not approved/ acquired on time	Will not be able to get participants to create the gold standard corpus therefore there will be no test data to test the system	Low	High	Apply for ethics clearance well in advance to commencement of project	Email notifications to and from committee	Start project where possible and if it is not approved apply again or apply elsewhere
An insufficient number of research participants is acquired	Gold standard corpora will not be of high validity	Low	High	Start recruitment as early as possible	Keep track of participants acquired	Search for potential participants outside of UCT
An insufficient amount of transcribed audio data for gold standard corpora is acquired	Gold standard corpora will not be of high validity	Medium	High	Pay participants as they complete their tasks for motivation	Keep track of the corpora	Recruit new participants/Offer participants who completed tasks on time incomplete tasks with further pay offer
Chosen toolkits prove insufficient for the requirements of the project	Failure to meet project goals and answer research questions	Low	High	Research toolkits sufficiently to ensure the one used is the best one to ensure the project's success. Listen to supervisor advice on which is the best toolkit to use.	Start working early enough to allow for time to research issues that may arise.	Where toolkits are insufficient look for libraries and other tools that can be integrated.
Lack of adequate skills to complete the project	Parts of the project will be incomplete which could jeopardize the entire project	Medium	High	Research and study relevant areas	Check if deadlines and milestones are being met adequately with supervisor	Discuss issues with other members and supervisor and obtain assistance as appropriate
Failure to meet project requirements on time	Incomplete project	Low	High	Attempt to meet milestones and deadlines in advance	Follow Gannt chart	Work overtime and hand some tasks (appropriate amounts) to other members if schedule allows
Loadshedding	Disruption in the progression of the project	High	Low	Monitor and work gradually so sudden interruptions don't mess with schedule massively	Allow notifications for loadshedding application such as Eskom sepush.	Work around schedule and working in areas not affected when yours is for example in campus when there's loadshedding in a members residential area
Team members(s) fall ill	Parts of the project will not be completed which could jeopardize the entire project	Low	Medium	Members will comply with standard health regulations	Members will keep track of their health.	Notify other members and supervisor if ailment is of a degree of seriousness
Team members do not complete their part of the project	Parts of the project may be reliant on other team members which could result in an incomplete project	Medium	High	Members who rely on the gold standard corpus will find other audio and their transcriptions to use to evaluate their part.	Members will research other data to use in the evaluation of their project	Make a list of other sources of data that can be used to evaluate the project

Appendix B


teamgantt
Created with Free Edition
4/23
5/23
6/23
7/23
8/23
9/23
10/23
SABC2TXT
start
end
Project proposal 06/04/23 02/05/23
Draft 1 06/04 14/04
Draft 2 14/04 18/04
Dry run 20/04 20/04
Presentation 26/04 26/04
First draft submission 24/04 24/04
Final proposal submission 02/05 02/05
Submission 02/05 02/05
Structured audio 03/05/23 27/07/23
Audio data collection (data set) 03/05 05/05
Research on CMU Sphinx and setup 08/05 12/05
Ground-truth transcriptions via online... 08/05 19/05
Train CMU Sphinx system 08/05 19/05
Audio transcription and analysis (iter... 03/07 14/07
Iteration 2 17/07 25/07
Iteration 3 27/07 27/07
Unstructured audio 03/05/23 27/07/23
Audio data collection (data set) 03/05 05/05
Research on PocketSphinx and setup 08/05 12/05
Ground-truth transcriptions via online... 08/05 19/05
Train PocketSphinx system 08/05 19/05
Android app development 01/06/23 30/06/23
Prototype and design 01/06 05/06
Develop the app (code) 06/06 14/06
Intergrate PocketSphinx library 15/06 21/06
Testing and development 30/06 30/06
Audio transcription and analysis (iter... 03/07 14/07
Iteration 2 17/07 25/07
Iteration 3 27/07 27/07
Gold standard corpus 03/05/23 24/07/23
Ethics application 03/05 12/05
Applied 15/05 15/05
Data gathering 1 15/05 29/05
Gathered 30/05 30/05
Preparing prompts for data gathering... 15/05 29/05
Recruitment phase 30/05 07/06
Data gathering 2 08/06 15/06
gathered 16/06 16/06
Crowd source phase 15/06 14/07
Transcribed all 17/07 17/07
Evaluation phase 17/07 24/07
Project progress 17/07/23 24/10/23
Project progress demonstration 17/07 21/07
Complete draft of final paper due 24/07 28/08
Project paper final submission 30/08 11/09
Submitted 13/09 13/09
Project code final submission 15/09 15/09
Final project demonstration 26/09 29/09
Poster due 09/10 09/10
Website due 16/10 16/10
School of IT Showcase 23/10 23/10
Showcase 24/10 24/10