Using question-specific vocabularies to support speech data collection with SALAAM

By Kayokwa Nick Chibuye

Supervised by Dr. Brian DeRenzi

A dissertation submitted in partial fulfilment of

the requirements for the degree of

Master of Science in Computer Science

Department of Computer Science

University of Cape Town

February 11, 2019

Contents

Declaration

The work presented in this dissertation is based on research carried out at the Centre in Information and Communication Technology for Development (ICT4D), Department of Computer Science, University of Cape Town. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is solely my own work unless referenced to the contrary in the text. I hereby declare that this written work I have submitted is original work which I alone authored and is written in my own words. With the signature below, I declare that I have been informed regarding normal academic citation rules and I conform to citation conventions customary to the sciences. This written work may be tested electronically for plagiarism.

Signed by candidate

Signature

11-02-2019

Date

Abstract

There has been an increasing use of small-vocabulary spoken dialogue systems in low-resource settings for information dissemination and data collection. This provides an opportunity to reduce the information gap in low-resource settings in which low-literacy is a huge hindrance to the adoption of Information Communication Technologies (ICTs). Since the languages spoken in these areas are computationally low-resourced, they rely on techniques such as cross-language phoneme mapping to facilitate fast development of small-vocabulary speech recognisers. Despite the success of this technique, there has been a lack of guidance on how to deploy such systems across a range of languages.

This study presents a systematic exploration of the suitability and limitations of using cross-language phoneme mapping for the development of small-vocabulary speech recognisers for computationally low-resource languages, particularly Bantu languages. Five target languages and four source languages were used in the study. Speech-based Accent Learning And Articulation Mapping (SALAAM), a cross-language phoneme mapping algorithm was used to aid the study based on its implementation in an open-source tool Lex4All. The following research questions guided our investigations: i) What impact does source language choice have on recognition accuracy, ii) What impact does gender composition of the training data set have on recognition accuracy and iii) What impact do varied alternative pronunciations per word type have on recognition accuracy.

Data for the target languages was collected from 104 university student volunteers consisting of 58 female and 46 male students. The results showed that target and source language phonetic similarity as well as gender composition of the training datasets affects recognition accuracy of speech applications developed using cross-language phoneme mapping techniques. They also showed that increasing the number of alternative pronunciations per word in the vocabulary generally increases recognition accuracy although with a slower system response time. This study provides evidence that a careful selection of the source language, gender composition of the training data and the number of alternative pronunciations per word can improve the recognition accuracy of speech recognisers developed using cross-language phoneme mapping.

Prior Publications

Some early versions of the content in this dissertation appeared in the following publications

1. **Nick K Chibuye**, Todd Rosenstock, and Brian DeRenzi: Cross-language phoneme mapping for low-resource languages: An exploration of benefits and trade-offs. Proc. Interspeech 2018, pages 2623–2627, 2018

2. (Extended Abstract) **Nick K Chibuye**, Todd Rosenstock, and Brian DeRenzi: Using question-specific vocabularies with Salaam to support speech data collection, HCI Across Borders Symposium, ACM conference on Human Factors in Computing Systems (CHI), Denver, Colorado, US, CHI 2017.

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Brian DeRenzi for all the insights and experiences shared through my research and the writing of this dissertation. Secondly, I would especially like to thank Joan Byamugisha for her relentless efforts to help become a better researcher, writer and scholar. I will forever be grateful for the friendship, time, advice and counsel. Last but not least, a huge thanks to all my colleagues from ICT4D lab at the University of Cape Town and my family for the support and for being a source of inspiration whenever I was lacking of it.

Above all, I thank God for His constant provision, the gift of life and His wisdom.

Dedication

Dedicated to the memory of my mother, Queen Elizabeth Kasewe Shamputa.

Glossary

ALISP	Automatic Language Independent Speech Processing
ASR	Automatic Speech Recognition
SALAAM	Speech-based Accent Learning And Articulation Mapping
CHW	Community Health Worker
CPU	Central Processing Unit
DTMF	Dual-Tone Multi-Frequency
GUI	Graphical User Interface
HMM	Hidden Markov Models
IBR	Institutional Review Board
ICT	information Communication Technology
ICT4D	Information Communication and Technology for Development
IVR	Interactive Voice Recognition
LVCSR	Large Vocabulary Continuous Speech Recognition Systems
MSP	Microsoft Speech Platform
SDK	Software Development Kit
SDS	Spoken Dialogue Systems
SMS	Short Message Service
UCT	University of Cape Town
VUI	Voice User Interface
XML	Extensible Markup Language

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Across the world, 750 million people are considered low-literate. 27% of the world's low-literate adults come from sub-Saharan Africa [1]. Low-literacy has been identified as a contributing factor to the slow adoption of technology among low-literate populations [2]. Traditionally, there have been three main approaches to dealing with this: mediated input [3], graphical user interfaces [4], and speech-based systems [5, 6, 7]. We focus on the latter as a way of reaching people directly without the need of a smartphone for a graphical interface or the addition of an enumerator as a mediating user.

Voice data collection has become popular amongst research projects in low-income countries over the past decade [6, 8, 9]. The rapid penetration of mobile phones in developing regions [10, 11, 12] has contributed to it's popularity, especially through spoken dialogue systems. A spoken dialogue system is a computer system that interacts with a user using voice in order to achieve a task [13]. Declining prices of mobile phones and increases in network coverage in many developing countries [10, 11, 12] has seen the adoption of mobile phones as tools for collecting high frequency and often low cost survey data. Mobile surveys also often offer flexibility and short turnaround times, making quick responses to unexpected data needs possible [11]. One can collect voice data from respondents over a phone call on a one-on-one basis, through a call centre or through a spoken dialogue system. In a call centre setup, an interviewer can use different languages during a session and ask complex questions if need be [6]. This setup also allows an interviewer to clarify matters that the respondent might find confusing and also accommodates varying literacy levels amongst respondents, making voice data collection robust and flexible [6]. These reasons and the ability to accommodate respondents

[1] http://uis.unesco.org/sites/default/files/documents/fs45-literacy-rates-continue-rise-generation-to-next-en-2017_0.pdf

owning low-end phones has made voice data collection the data collection method of choice for most of sub-Saharan Africa [14, 15]. For example, the World Bank's Listening to Africa project uses this mode of data collection as a complement of paper-based household surveys in Madagascar, Malawi, Mali, Senegal, Tanzania, and Togo[2].

Some of the biggest benefits of using voice as a means of human-computer interaction are that it is the natural form of human communication [16] and it is not limited by someone's ability to read [17, 18]. This makes voice human-computer interaction very important especially in the developing world where varying literacy levels exist and text-based interfaces may not be usable by large segments of the population [6, 9, 15]. Spoken dialogue systems have been used widely for information dissemination [18], data collection [19, 20], real-time monitoring [21] and community building [22]. Specifically, spoken dialogue systems have been used in a number of domains such as agriculture [7], health [5], entertainment [8], education [23] and journalism [17].

Automatic speech recognition has played a key role in the design and development of speech-driven interfaces that spoken dialogue systems use. Unfortunately, the majority of languages spoken in developing regions typically lack adequate computational resources needed to train speech recognition engines [24]. The process to train a speech recognition engine is expensive and demands expert knowledge in speech technology and linguistic expertise in the local language of interest, all of which are lacking in developing regions [24, 25, 26]. This makes it difficult to develop applications suitable for these regions.

Recent advances, however, suggest that one can use an existing speech recogniser trained in a high-resource language, such as French, to achieve small-vocabulary automatic speech recognition tasks of words in a low-resource language such as ciBemba, a language indigenous to Zambia [27]. This can be achieved by leveraging the similarity of sounds (phonemes) between the two languages through a technique called cross-language phoneme mapping. A phoneme is often regarded to be the smallest unit of sound that can, by itself, distinguish one utterance from another [28]. By low-resource language, we refer to all languages that lack computational language resources, have a small economically disadvantaged user base and are not supported in commercial speech recognition products and services [29]. Conversely, high-resource languages refer to languages that have computational resources, they are supported in commercial speech recognition products and services and generally have a large user base.

Using cross-language phoneme mapping, one can generate a pronunciation lexicon repre-

[2]https://blogs.worldbank.org/africacan/measuring-the-pulse-of-africa-one-phone-call-at-a-time

senting the pronunciation of target language words or phrases based on the phonetic alphabet of the high-resource language, and achieve speech recognition over the low-resource language vocabulary [26]. A pronunciation lexicon is a collection of words or short phrases, their written form and mappings to their respective pronunciations specified using an appropriate pronunciation alphabet [30, 31]. Lexicons can be manually generated but they demand the use of an expert linguist who is fluent in both the source and target languages, yet they do not often yield high quality recognition accuracy [26]. Therefore, processes of automatically creating cross-language phoneme mappings between languages were developed [32], omitting the need for linguist experts. In the context of this study, we shall refer to target language words or short phrases as word types.

1.2 Problem statement

Cross-language phoneme mapping has been successfully used in a number of Information Communication Technologies for Development (ICT4D) [33] projects in the health sector [5], agriculture sector [7, 15] and for research purposes [26, 34]. Despite the development and use of this technique in these and other projects, there remains little guidance of how the technique behaves in different conditions, i.e., different source-target language pairings, and training techniques. This is especially true for African languages in general and Bantu languages in particular.

1.2.1 Research questions

To address this problem, we devised the research questions below to guide our work:

1. R1: What impact does source language choice have on recognition accuracy?

   Using a cross-language phoneme approach, we generated lexicon files using Speech-based Accent Learning And Articulation Mapping (SALAAM) and developed four different speech recognisers based on four source languages, Mandarin, English (US), German and French whose choice was driven by the availability of phonetic alphabets. We then evaluated the recognition accuracy of each individual speech recogniser with respect to the source and target language pairings. We observed a source language dependent recognition accuracy and established that the source language choice has a significant impact on recognition accuracy for speech recognisers developed using cross-language mapping.

2. R2: What impact does gender composition of the training data set have on recognition accuracy?

   We developed three gender-based training sets, male-only, female only and multi-gender and developed speech recognisers for each source-target language pair. We established that speech recognisers developed using cross-language mapping are also affected by the gender composition of the training data. We also established that one could employ gender separation to achieve better recognition accuracy, although a multi-gender dataset produces a more robust speech recognition against gender bias in speech recognition.

3. R3: What impact do varied alternative pronunciations per word type have on recognition accuracy?

   (a) How does recognition accuracy compare across different pronunciation sizes?

   For each source-target language pair, we generated lexicon files containing a variable number of alternative pronunciations per word type. These lexicons consisted of 5, 10, 20, 40, 60, 80 and 100 alternative pronunciations per word type. We evaluated the accuracy of each of these lexicon files and established that recognition accuracy generally improves with an increase in the number of alternative pronunciations.

1.3 Research approach

In this study, the SALAAM technique was used to help us answer the aforementioned research questions. SALAAM is a technique that can automatically generate the aforementioned mapped pronunciations from a handful of training data and achieve high quality recognition accuracy [32]. The pronunciation lexicon produced by SALAAM can then be used with a speech recogniser to support speech recognition of low-resource language word types they contain. This makes it easier for developers with no speech technology expertise to quickly develop small-vocabulary speech driven applications for low-resource languages. The SALAAM technique has been used to support speech driven applications for different developing regions and low-resource languages with high speech recognition accuracy [5, 9, 15, 34].

1.3.1 Target languages

To aid our study, we used five target languages: ChiShona, Kiswahili, Afrikaans, English (South Africa) and seSotho chosen as a focus of this study. Three of these languages, English

(South Africa), seSotho and Afrikaans are part of the 11 official languages of South Africa [35]. Afrikaans and English have Germanic roots [35, 36]. English provided us with an upper bound of how we expected the system to perform since it served both as a source and target language. We hypothesised that Afrikaans, which primarily has Germanic roots, would similarly perform better than the other languages whose roots are not Germanic because it shared similar roots with English and German source languages. seSotho, ChiShona and Kiswahili are indigenous to Africa and are representative of the Bantu language family [37, 38, 39, 40, 41]. This helped us investigate the performance differences between languages that have similar roots, both target and source languages, from those that do not in the context of cross-language phoneme mapping and recognition accuracy.

This dissertation presents the results of our efforts to establish the limitations and suitability of using cross-language phoneme techniques for speech data collection with a focus on Bantu languages. Our findings will contribute to the understanding of the applicability of cross-language phoneme mapping to low-resource languages, particularly Bantu languages, to support the development of speech-driven applications. It will also provide a guide that other researchers or practitioners can follow to develop high quality small-vocabulary speech-driven applications for low-resource languages using cross-language phoneme mapping.

1.4 Outline

In chapter two, we discuss the background and related literature that characterises our work. We begin by looking at traditional data collection practices and the role of mobile phones in effective and efficient data collection. We proceed to discuss voice data collection and the role of speech recognition and Voice User Interfaces (VUIs) in voice data collection. We then proceed to discuss speech technology in developing regions, cross-language phoneme mapping and related methods.

Chapter three discusses the system we used to carry out our research. In the chapter, we begin by discussing the system overview, looking at how each individual component of the system works. We proceed to discuss how one can train and generate a lexicon using the system. We further discuss the structure of the lexicon file and the process of recognition accuracy evaluation using these lexicon files.

In chapter four, we discuss the methodology employed in our research. We begin by discussing the rationale behind the target language, source language and vocabulary choices. Participant recruitment and data collection procedures are discussed next. Lastly, we end the

chapter by describing the methods employed during each experiment to address each of our research questions.

We describe the experiments we conducted and discuss the results obtained in chapter five. We achieve this by looking at one experiment at a time, each experiment addressing one research question. We end the chapter by discussing the results obtained from each experiment and their implications.

We conclude this dissertation in chapter six, bringing together our findings from the three experiments we conducted. We reflect on these findings, summarise the contributions of our study, discuss the limitations of our work and opportunities for future work.

1.5 Summary

We began this chapter by introducing the area of our research focus and the motivation for our study. We then went on to discuss the problem statement that characterises our work and the research questions we set out to answer. We ended the chapter with an outline of the rest of the work presented in this dissertation.

Chapter 2

Background and Related work

2.1 Overview

In the previous chapter we described the motivation of our work and introduced our area of focus. This chapter presents relevant literature regarding the background upon which we build our work as well as the body of work that relates to our area of focus.

2.2 Data collection practices in developing countries

Data collection in developing countries is predominantly done through paper-based face-to-face household surveys [11, 42]. These surveys are complex in nature and are often characteristic of high costs, low frequency and long turnaround times of no less than one year [11, 42]. As a consequence, the data collected are often incomplete, less reliable or outdated and fail to meet the urgent data needs [11]. Consequently, the last decade has seen a growth in interest for using new technologies for gathering high quality, high frequency survey data on the living conditions and perceptions of citizens in developing countries [11, 12, 21, 43]. The drop in prices of mobile phones and the increase in network coverage in a lot of developing countries has seen the adoption of mobile phones as tools for economic development, governance and tools collecting high frequency and often low cost survey data [10, 12, 42]. Mobile surveys also often offer flexibility and short turnaround times making quick responses to unexpected data needs possible [11].

2.3 The role of mobile phones in effective and efficient data collection

In recent years, the efficiency and effectiveness of data collection in resource-poor environments has been improved through the use of mobile phones [21]. Unlike the paper-based data collection model, they allow for immediate digitisation of collected data at the point of a survey. As a result, fast and automated data aggregation is achievable [11]. Additionally, mobile phones afford the ability to enforce adherence to complex or context-dependent logic within questionnaires. This allows for clarification of ambiguity in the questions if any, consequently contributing to the collection of accurate and complete survey data [11].

Data collection on a mobile phone can be done via several interfaces, voice [44], Short Message Service (SMS) [45], electronic forms [46] or a mobile application [47]. Voice data collection can further be divided in two categories, operator based, where an operator calls a participant and data is collected over the call and Interactive Voice Recognition (IVR) [48], where a participant enters data using using a touchtone or voice user interface. Some of the factors influencing the choice of a data collection interface include, but are not restricted to, the means of interaction, effectiveness and cost [6].

It is evident that mobile phone surveys have improved the way research is conducted and certainly achieved milestones paper-based face-to-face surveys failed. Some of the mobile applications that researchers and practitioners can choose from include Open Data Kit (ODK) [47], surveyCTO [49], MagPi [50], CommCare [51], KoBoCollect [52] and medicMobile [53]. However, research further shows that mobile phone surveys are best suited for short interviews [11]. They are more effective for monitoring rapidly changing conditions and obtaining real-time feedback from households, notwithstanding some considerably large projects that have successfully collected data using mobile phones such as the Listening to Africa project [54]

Mobile phones also enable collection of more complex data than paper-based data collection such as video, audio and geolocation and offer the potential to improve data quality by specifying automatic quality checks before data is submitted [11, 55]. Examples of some of these quality checks include: ensuring all questions have been answered before submission, providing predetermined ranges for specific values and making sure the answers provided are consistent with previous responses. The feasibility of such systems also generally hinges on the type and quality of network connection available [55]. The use of mobile phones for data

collection is commonplace in household surveys [56], clinical trials [57], surveillance [42] and spatial / geographical data collection [58] among other research areas. Some of the benefits that come with the use of mobile phones for data collection are that they are economically and environmentally friendly, they are flexible, support faster reporting with accuracy and offer the potential for enriched data collection [59]. Other benefits of mobile data collection for researchers as listed below:

1. Ability to gather data in volatile and high risk environments - for example, mobile phone surveys were used for contact tracing and provided real-time data to assess the impact of the ebola outbreak in affected countries [60].

2. Support for monitoring and impact evaluation efforts - due to the short turnaround nature of mobile data collection, they can be used to meet the data demands for up-to-date information on the living conditions of a country's citizens [43, 60] .

3. High flexibility - mobile data collection is flexible in that implementers can react to a new or unexpected data need. For example mobile phones were used by the 'Listening to Malawi' to establish the severity of Malawi's January 2015 floods [11].

4. Automatic data digitisation - data can be automatically digitised as soon as it is collected, supporting consistent data formats amongst enumerators and providing real-time access to data [56].

2.4 Voice Data Collection

Over the past decade, the collection of data via voice has become popular amongst research projects in low and middle income countries [14, 54]. Voice data collection is especially popular because of the flexibility live interviews offer. An interviewer can conduct interviews in different languages, ask complex questions and accommodate participants of varying literacy levels. A supervisor or interviewer can also re-call, in lieu of revisiting, respondents for quality control purposes. In most cases, setting up for voice data collection, involving an enumerator, is relatively easy [14]. For example, if one intends to collect data from participants directly, setup can be as easy as obtaining a mobile phone and negotiating time that best suits the respondent's schedule. This type of setup does not require custom software, it has less cognitive load on the respondent and there is often no need to train neither the interviewer nor the respondents on how to use a phone. It is also likely to have fewer operational risks such as accidental deletion

of a data form or forgetting an SMS cue card by a respondent, making voice data collection relatively robust [6].

The acceptance and popularity of the use of voice-based data collection is evident in much of the research carried out across Africa. For example, a study in Kenya revealed that farmers still preferred voice over SMS communication because of its ease of use and their lack of practice of using SMS [61]. Research further affirms that voice-based systems may seem to be the only practical option for self-completion surveys where a substantial proportion of respondents are illiterate [55]. This is because spoken language serves as the primary means of human communication [16]. If autonomous voice-based data collection through spoken dialogue systems is to be achieved in low-resource settings, the systems would need to support voice input as a means of human-computer interaction.

2.4.1 Voice User Interfaces and Dual-Tone Multi-Frequency

A Voice User Interface (VUI) is a way through which a person is able to interact and control a computer system or device using voice input [48]. Typically, VUIs consist of prompts, grammars and control/dialogue logic [48]. The prompts are the utterances the system makes, either synthesised or pre-recorded audio, to elicit input from the user [48]. The grammar defines the all possible utterances the users can make in response to the prompts. The control logic defines the action that the system can take based on a users responses to system prompts [48].

Voice interfaces have the potential to cater for the information needs of speakers of low resource languages, whether or not they have a formal writing script [15]. For example, in [9], a telephony-based spoken language interface was developed to provide a means through which low literate users could access healthcare information in Urdu, one of the low-resource languages of Pakistan. This was achieved by developing an application that one could interact with using Interactive Voice Response (IVR). The IVR menu could be navigated using either speech recognition or dial tone. In [15] a Hindi Speech recognition module was developed for a mobile video search application for Indian farmers called VideoKheti. Both of the speech recognition applications mentioned in [9] and [15] were developed using the SALAAM method and are perfect examples of successful development projects that leveraged the benefits of speech technology, the Hindi agricultural mobile application achieving over 90% recognition accuracy [15].

In [62] a study was conducted to compare isolated-word speech and Dual-Tone Multi-Frequency (DTMF) input Voice User Interfaces (VUIs) for farmers in rural Gujarat, India.

There were 45 participants most of whom only had an education up to eighth grade. A telephony application called Avaaj Otalo was developed and it enabled farmers would call-in to listen to archived agricultural radio programs of interest. The application also allowed farmers to record their own questions, if any, for review and response by experts. The main aim of the study was to compare performance, ascertain the user's preferred input model and correlate the results to users' education levels as well as their age. At the end of the study, DTMF outperformed speech in terms of the task completion rate and learnability. Users also had less difficulty providing input using DTMF. A similar study was conducted by Delogu et. al in which DTMF was compared to speech recognition systems for telephony technologies. The results showed no difference in performance, however, a user preference for DTMF over an isolated word interface was found [63].

Looking at literature further, it is interesting to note that studies yielded different results when comparing between DTMF and speech driven interfaces in an effort to establish which one is the most preferred input modal. For example, the results obtained in [9] contradicted those found in [62] because a speech driven interface was found to be the most preferred input modal. Shewani et. al. stated a number of factors that may have had been responsible for the contradiction and they are listed below:

1. The design of the speech-input by Shewani et. al. was more conversational e.g. "What would you like to hear more information about, diarrhoea, malaria, or hepatitis?" as compared to "To ask a question, say 'question'; to listen to announcements, say 'announcements'; to listen to the radio program, say 'radio' by Patel et. al.

2. The combination of restrictive keyword-based grammars with open-ended "say anything" recording segments. This makes it particularly difficult for uses since it is not obvious when (or even why) it is not possible to speak in open sentences in one part of the interaction yet it is required to do so in another.

3. Lack of user training prior to deployment and testing. Shewani et. al. argues that even a limited amount of training can make a significant difference to the usability of an interface.

Lee and Lai also compared a natural language system to a dial interface and concluded that user input-modal was dependent on the task to be completed. One preferred the use of DTMF when completing linear tasks such as listening to voice-mails in the order in which they were received and speech for non-linear tasks such as listening to voice-mails from a specific acquaintance in random order [64].

The above findings suggest that a user's preference of using either speech or DTMF as input methods is influenced by a number of factors highlighted in the studies described above. Some of which are the quality of the design of the interfaces, availability of user training or the lack of it, literacy and education levels of the users as well as the nature of the tasks to be completed. Of the two input methods, speech input would be more useful in low-literate settings such as most of rural developing regions, to support speech based data collection. Therefore this calls for an understanding of the current state of speech technology in developing regions.

2.5 Speech technology in developing regions

Previous research has shown that Spoken Dialogue Systems (SDSs) and Automatic Speech Recognition (ASR) technologies can be used as tools to bridge the gap between the low-literate populations of developing regions and information technology [2, 24]. With the widespread adoption of mobile phone use in underdeveloped regions we discussed in section 2.2, the use of speech technology is feasible in an attempt to reach a large low-literate population of the developing world where text-based interfaces may not be very useful [15]. However, these regions face several challenges that hinder the adaptation of speech technology. We discuss some of these challenges in the next section.

2.6 Challenges faced with speech technology in developing regions

The effort to establish the effectiveness of speech technology, particularly SDSs, in developing regions began with projects such as the Tamil market [65] and Carnegie Melon University's Healthline [5]. Several other case studies and experiments followed [9, 62, 66]. These studies demonstrated that the use of speech technology in developing regions especially among low-literate users was effective. However, they also brought to light a number of challenges pertaining to the development of high quality speech recognisers for languages spoken in underdeveloped regions.

One of the major challenges was the need for tens of speakers and up to hundreds of training audio samples per speaker in order to develop competent speech recognition technology as suggested by experimental results obtained at Meraka Institute[67, 68]. These findings emphasised the lack of language resources as one of the major hindering factors in the development of high quality speech recognition technology for developing regions. Low literacy was also

identified as challenge as it affects audio data collection, speaker recruiting and user testing-therefore requiring novel ways to get the desired outcome [2]. Other issues included the lack of speech technology experts and linguists to aid in the development of the high quality speech technology for these regions [24].

These findings motivated the development of several general solutions to address some of the major challenges faced when developing high quality speech recognition technologies for under-developed regions and low-resource languages [25, 32]. The next subsection discusses some of the early efforts to generally quicken the development of speech recognisers for low-resource languages.

2.7 Approaches towards quick speech recogniser development

This subsection discusses some of the general solutions that focused on methods that could potentially get rid of the need for the involvement of experts and reduce training data and training time needed for the development of speech recognisers for low-resource languages.

2.7.1 Reduction of training data and training time

We begin our discussion by looking at a series of work done by Schultz and Waibel whose aim was to reduce the amount of training data needed to develop acoustic models for new languages by leveraging a large amount of data from several source languages [69, 70, 71, 72]. This work began with the collection of a multilingual high quality speech corpus suitable for the development of multilingual large vocabulary continuous speech recognition systems (LVCSR) called GlobalPhone [69]. It consisted of 9 out of the 12 most spoken languages in the world. This included Arabic, Russian, Spanish, Japanese, Korean, Portuguese, Chinese (Mandarin), Croatian and Turkish, leaving English, French and German out [69, 71]. Using GlobalPhone, Schultz and Waibel developed a multilingual speech recognition system which served as a language identification system as well as a source engine from which other new language acoustic models could be modelled [71]. Their work established that cross-language bootstrapping was an efficient technique to use for fast bootstrapping new LVCSR even in cases of phonetic inventory mismatches. They also established that, with respect to cross-language transfer, multilingual acoustic models performed better than monolingual acoustic models [72]. In spite of the findings, Schultz and Waibel's data-driven models were unable to outperform the models that were generated using a heuristic approach [25]. Consequently, there were still no

satisfactory solutions at this time that eliminated human involvement in the development of recognisers for low-resource languages.

Following work by Schultz and Waibel, a phonetic analysis of Afrikaans, English, isiZulu and isiXhosa was carried out with a view of furthering progress in multilingual acoustic modelling for automatic speech recognition in general [35]. Their findings revealed a significant phonetic overlap between isiZulu and isiXhosa to use for the development of phone models for multilingual speech recognition. This further supported the plausibility of exploiting phonetic similarity of languages to support the fast and efficient development of speech recognition technology especially for the developing regions [35].

Further, work done by Constantine and Chollet [73] was one of the earliest that used a data-driven approach and cross-language pronunciation transcription for speech processing. Using Automatic Language Independent Speech Processing (ALISP) as described in [74], they were able to achieve automatic phoneme transcriptions using a multilingual database and a simple genetic algorithm. Their work showed a correlation between phonemic similarity and cross-language usability of sub-unit words for different languages [73].

Bansal et al developed a joint Viterbi decoding algorithm, based on a method described in [75], to automatically determine the pronunciation of target language training audio [76]. They achieved this by using a modified version of the Carnegie Mellon University's Sphinx-2 semi-continuous density Hidden Markov Model (HMM) [77] based speech recognition engine [78]. The engine was modified to support correspondence-constrained decoding on multiple audio samples of a single word type to determine the best pronunciation of a word type based on multiple audio samples [76]. Two other commonly used approaches to determine the best pronunciation of a word type from multiple audio samples were used to investigate the performance of this joint algorithm. The first approach was one that generated pronunciations by voting from amongst the recognition outputs from individual audio samples [79]. The second approach generated N-best hypotheses from the provided audio recordings and jointly rescored the cumulative set with all the recordings [80, 81]. The results revealed that the joint algorithm significantly outperformed the other two commonly used approaches [76]. However, the need to modify the decoding algorithm of a speech engine had two implications: the modification demanded expert knowledge and the technique excluded the prospects of using commercial off-the-shelf speech recognition engines as baselines in which training with the source language(s) had already been done [25]. Therefore, this approach did not eliminate expert involvement either and was therefore limited in its applications in low-resource settings.

The techniques discussed above aimed at the reduction of training time and training data to reduce the development of speech recognisers for new languages. Though useful, they still depended on expert involvement, one of the major hindrances in the development of speech recognisers for under-resourced languages. Therefore, they were not practical techniques for the quick development of small-vocabulary speech recognisers that are much needed for languages spoken in low-resource settings. However, these techniques provided a starting point and further motivated efforts to develop techniques that could reduce training time, amount of training data as well as expert involvement in general and small-vocabulary speech recogniser development. One such technique is discussed in the next section.

2.7.2 Reduction of training data, training time and expert dependency

Poor Man's Recogniser

The "Poor Man's Speech Recogniser" was a tool developed, by employing cross-language phoneme mapping using existing acoustic models, to support speech recognition for telephony health information access for low-literate Community Health Workers (CHWs) in Pakistan [5, 9, 82]. By using cross-language phoneme mapping, one can avoid training new acoustic models which often the most complex and costliest part of speech recogniser training [32]. In order to achieve the cross-language phoneme mapping, a lexicon file containing hand-coded mapped pronunciations of word types based on the English (US) phonetic alphabet was used to support the recognition of Urdu word types [5]. Despite developing and successfully deploying and testing a spoken dialogue system using this approach, there was human involvement during the generation of the aforementioned mapped pronunciations which demanded linguistic expertise [5]. This problem lead to the development of an improved version of the "Poor Man's Speech Recogniser" called Speech-based Accent Learning And Articulation Mapping (SALAAM) [32]. In the new version, the new speech recogniser was used to semi-automatically decode audio samples of each target word type to obtain more accurate pronunciation transcriptions, eliminating the need of a linguistic expert altogether [25, 32]. The SALAAM technique is discussed further in the next section.

The SALAAM technique

The primary idea behind the SALAAM technique is to find the best pronunciation sequence for a given word in a target language from one or more audio samples by using a source language speech recogniser to perform phone decoding (decoding by phoneme) [26]. Since most commercial speech recognisers do not directly support phone decoding, the SALAAM technique

uses a specially-designed grammar to mimic phone-decoding [26, 29]. This is achieved by creating a recognition grammar representing a phoneme *super wildcard* to guide pronunciation discovery.

The grammar enables the speech recogniser to break down a word in the target language into a series of one to ten 'sounds'. Each of these 'sounds' are then matched a sequence of one to three source language phonemes [26, 32]. The SALAAM heuristic accepts, as input, a set of audio samples of the same word or short phrase and a requested number of k pronunciations. Using an iterative process, the heuristic builds a set of phoneme strings, returning the top-k performing pronunciations based on the phonetic inventory of the underlying speech recognizer [26, 32]. This results in a ranked list of pronunciation(s) for each word type being represented as a set of phoneme sequences. For example, using SALAAM with the English (US) source language to generate the top three pronunciations for *Mkate*, the Kiswahili word for bread, would result in the following phoneme sequences: *M K AA T I*, *M K AA CH I*, or *M K AH CH E*, which are then written to a lexicon file that is used later during the speech recognition process.

Therefore, using the SALAAM technique, the need for linguistic expert involvement in the development of mapped pronunciations is completely eradicated. The training time is also reduced from days to a couple of minutes or hours, depending on the vocabulary size. Additionally, the amount of training data is reduced from hours of audio data to minutes, requiring a minimum of one audio file per word type [32].

SALAAM proved to be a solution that comprehensively addressed most of the major challenges that come with speech recognition support for low-resource languages. It is for this reason that we decided to use the SALAAM technique for our study. It is a practical technique to use for the development of speech-recognisers in low-resource settings, an area in which our work is characterised. Establishing the benefits and trade-offs of cross-language phoneme mapping particularly with regards to the SALAAM technique would allow the development of high quality small-vocabulary speech recognisers, allowing more people to be reached and understood. The practicality of the technique has been demonstrated in a number of low-resource speech-based ICT4D projects [15, 26, 34] which we briefly describe in the following section.

SALAAM-based research and projects

There exists research that has used and evaluated several aspects of the SALAAM technique [15, 26, 29, 32, 34]. Using vocabulary sizes ranging from 3 to 10, Sherwani's test of the method

using several languages yielded more than 90% recognition accuracy [83]. Similarly, a speech-driven agricultural video search application developed for farmers in rural India using the SALAAM technique also yielded recognition accuracy greater 90% [15]. The SALAAM method reported in [32] was further improved by introducing discriminative training [29]. The idea behind discriminative training was to choose a subset of phoneme sequences to minimize any conflict between word types of acoustic similarity which was observed in [32]. A less correct phoneme sequence representing a word type's pronunciation was therefore acceptable as long as it prevented the word type from being identified as another with similar acoustic properties. This in turn reduced the errors in recognition accuracy and substantially increased the performance of the SALAAM technique [29].

Some other projects that built new recognizers on top of recognisers trained on different languages include a comparative study of on speech-driven interfaces conducted in India, as described in [62], and a study of the development of basic spoken dialogue systems conducted by Meraka Institute [84]. Both projects recorded over 90% recognition accuracy for most of their experiments. These projects further affirm that the SALAAM method is a viable approach for the development of small-vocabulary applications in spite of it understandably falling short when compared to recognisers trained directly on resources of a specific target language.

There exist several other tools and utilities that one could consider for the development of speech recognisers and speech recognition research such as HTK Toolkit [85, 86], CMU Sphinx Toolkit [87], Kaldi Tookit [88] and WebMaus [89]. Though these tools are powerful and open source, we did not consider them for our study mainly because they demand high technical knowledge and a considerable amount of training and evaluation data in order to use them [90], which is not a realistic expectation for most under-resourced settings and languages.

2.8 Summary

This chapter covered the background of our work as well as the related work. We began the chapter by discussing the current data collection methods in most developing countries, particularly rural Africa. This was followed by the role of mobile phones in effective and efficient data collection, where we discussed the suitability of using mobile phones for data collection and the benefits they offer to rural developing regions. We then looked at voice data collection in general and also discussed related research that has been conducted regarding the use of voice user interfaces as tools to aid data collection and how they compare to DTMF as a means of human-computer interaction in spoken dialogue systems. We proceeded to discuss

speech technology in developing regions, the challenges faced and the efforts that have been made to mitigate the challenges. This specifically focused on efforts in trying to quickly develop high quality speech recognisers, spanning acoustic model development that leverages cross-language phoneme transfer and cross-language phoneme mapping. We ended the chapter by discussing the SALAAM technique, how it works and its relevance to our work. The next chapter discusses the tools and utilities used to conduct our study.

Chapter 3

Tools and Utilities

3.1 Overview

This chapter describes the tools we used to aid our study. For the purposes of this dissertation, Lex4All [34] and the underlying algorithm were largely treated as a black box. The intuition of the algorithm and the architecture of the implementing tool are presented here as background for the reader.

We begin this chapter with an introduction to the Lex4All tool and this is followed by a description it's architecture. We proceed to discuss how one can 'train' a speech recogniser and evaluate it's recognition accuracy using the tool. We also briefly describe the use of a custom Android app called TimestampLogger and an open-source audio editing software called Audacity in our study before ending the chapter.

3.2 Introduction

Lex4All [34] is an easy-to-use Microsoft Windows based open-source tool that implements the SALAAM technique. The tool depends on Microsoft Speech Platform (MSP)[1] for its functionality and allows both skilled and unskilled users to create pronunciation lexicons for words in any language, using a small number of audio files and a well-trained pre-existing speech recognition engine in a high-resource language. The pronunciation lexicons can then be fed back into a speech recogniser to specify the application vocabulary and use it to support speech recognition [34]. The tool was selected because it was the only open-source implementation of the SALAAM technique that was found, it is well documented and easy to use. The tool and its

[1] https://docs.microsoft.com/en-us/previous-versions/office/developer/speech-technologies/hh361572(v=office.14)

source code are freely available and can be found on Github [91]. Presented here are the core components of the tool, as implemented by the original authors of the tool [34].

3.3 System Architecture

Figure 3.1 shows the three main components of the system architecture, the front-end, consisting on the Graphical User Interface (GUI) which facilitates user input and program output, the back-end responsible for the logic behind phoneme discovery, lexicon generation and recognition accuracy evaluation and the Microsoft Speech Platform which provides the speech recognition engines.

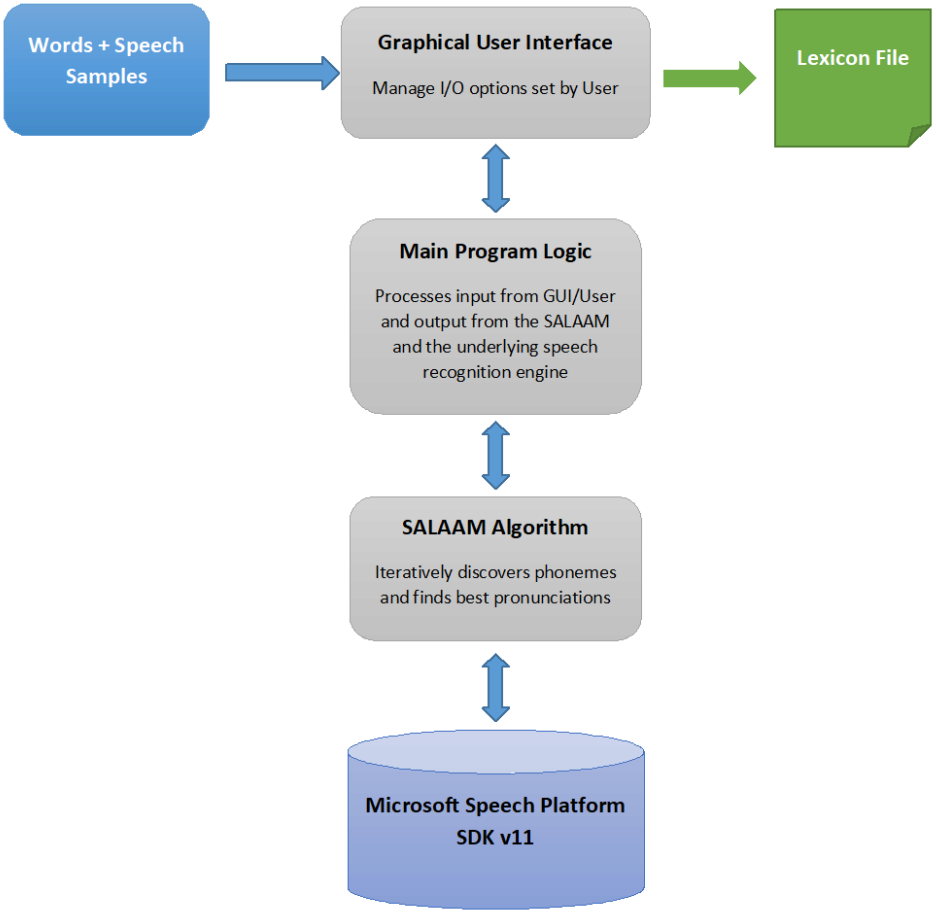

Figure 3.1: **System architectural overview** - The GUI supports easy configuration and a means to provide the application with input (audio files and words in text format). The input is then passed to SALAAM to discover best pronunciations which are later given as output in an Extensible Markup language (XML) format [1].

3.3.1 Training and lexicon generation

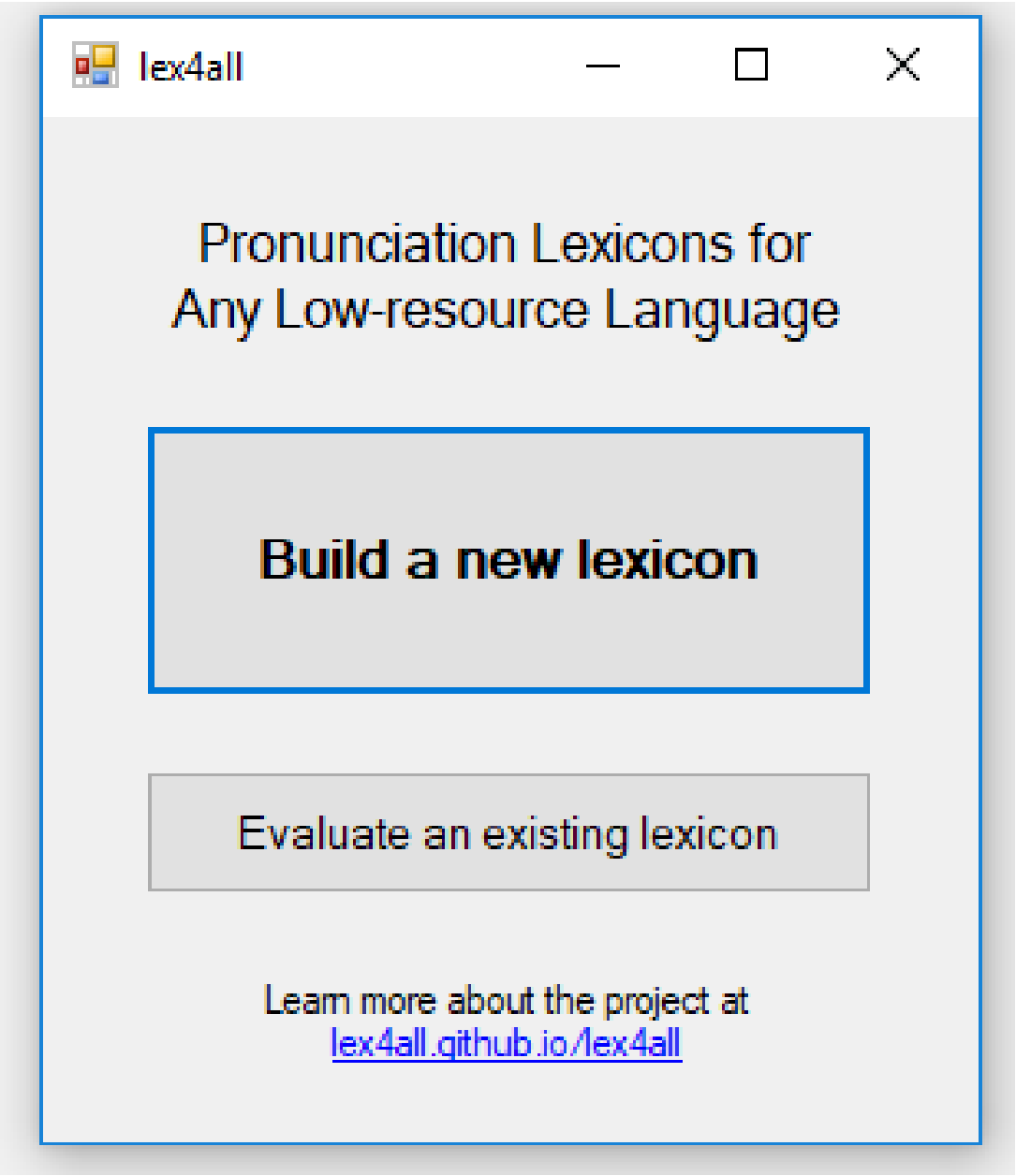

Figure 3.2: **Home screen** - Here the user is presented with two options, to build a new lexicon or evaluate an existing one using the systems evaluation module.

When the application starts, the user is first greeted with a screen from which they can choose either to generate a new lexicon or evaluate a new one as shown in Figure 3.2 above. When a user chooses to build a new lexicon, the Lexicon Builder is opened.

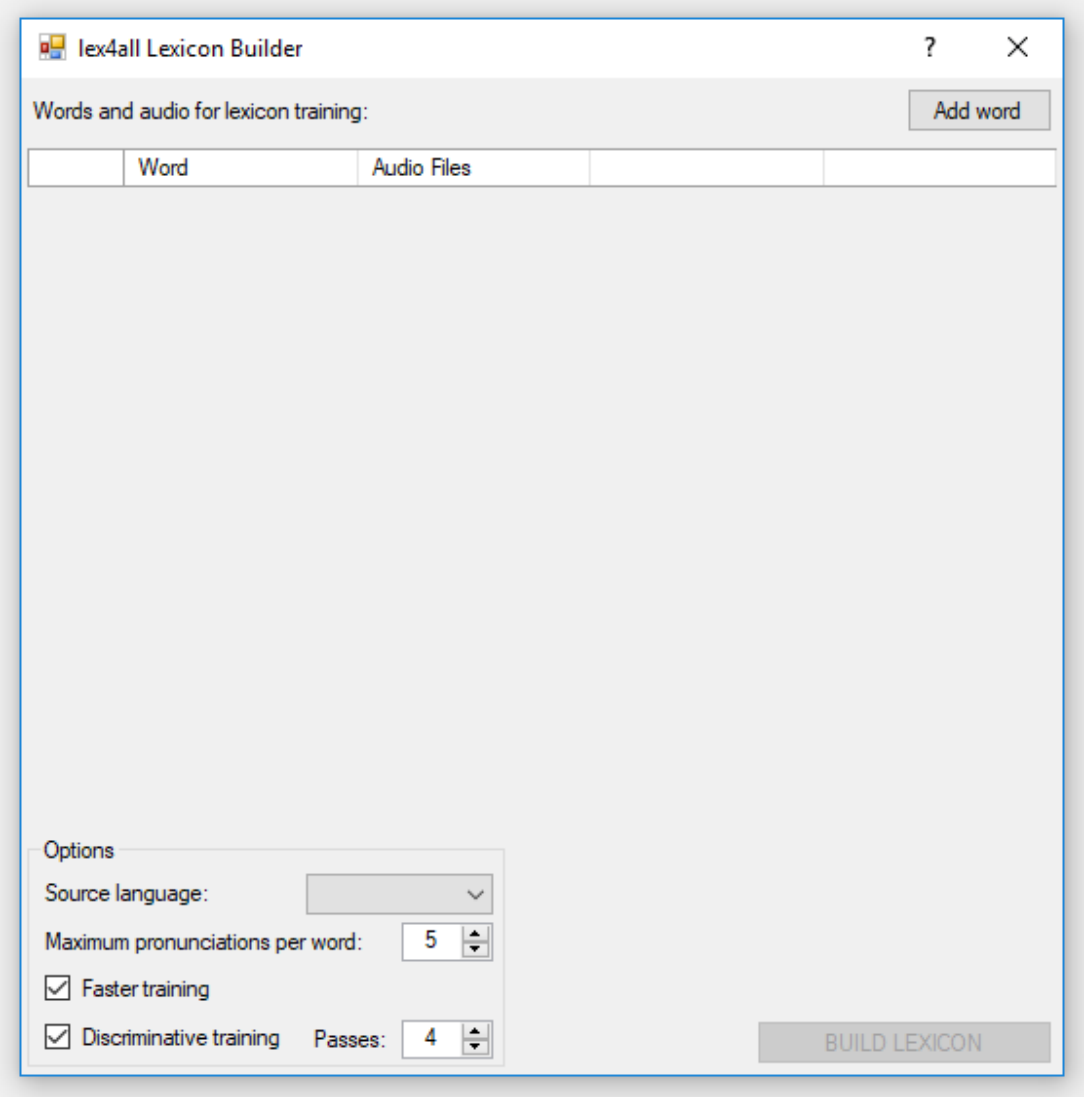

Figure 3.3: **Lexicon Builder** - Here the user can specify the source language to use using a drop-down menu and the number of alternative pronunciations, one being the minimum and one hundred the maximum. They can also choose to use discriminative training and whether or not the they would like to use fast training

With reference to Figure 3.3, the user is able to specify various parameters before generating a lexicon. In the bottom-left corner, one can specify the number of pronunciations to generate per word, the source language to use, type of training and whether or not to use discriminative training. Discriminative training in the context of the SALAAM technique refers to a modification of the original implementation of SALAAM in which instead of only matching phonemes that best suit a word type regardless of other word types in the vocabulary, the SALAAM algorithm would select a subset of matching word type phonemes to reduce potential conflicts with other acoustically similar word types in the vocabulary [29]. Fast training refers to the training procedure based on the modified version of training algorithm of the SALAAM technique as implemented in Lex4All [34]. If fast training is not selected, the training procedure would default to using the original implementation of the SALAAM technique which is much slower [25, 29, 32, 34], as such, we used ‘fast training’ in our study. In order to add the vo-

cabulary, the user would have to click the 'add word' button shown in the top-right corner. The user would then have to provides the textual representation of the word type and the associated audio files of the word type's pronunciation.

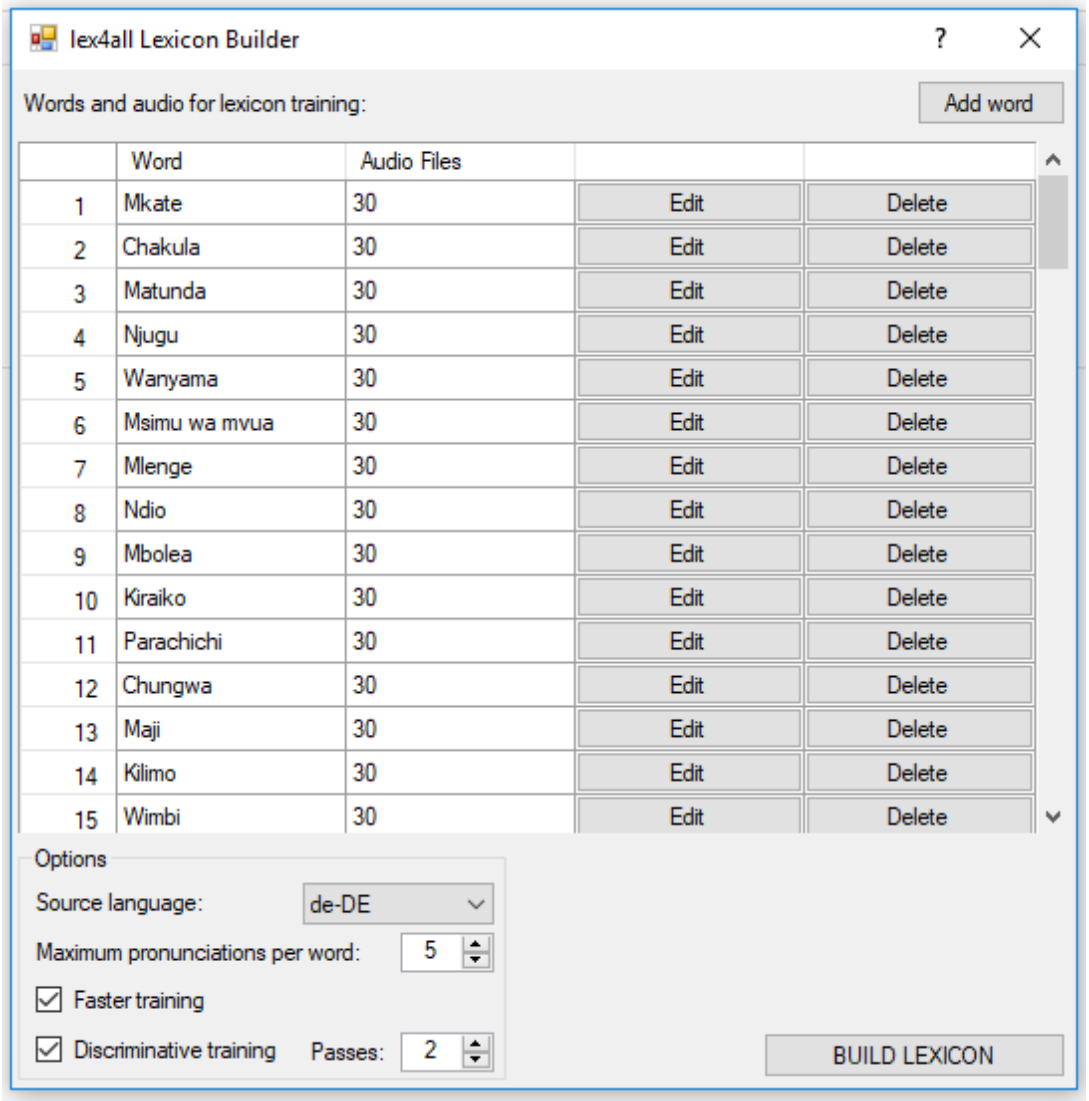

Figure 3.4: **Lexicon generation** - The *Word* column represents the textual representation of target language word types with the number of audio files containing their respective pronunciation shown in the *Audio files* column

Figure 3.4 shows a typical screen one would see right before generating a lexicon based on the specified vocabulary. The user would have to click the 'Build Lexicon' button on the bottom-right of the window and wait until the process is done. The program exits with a summary of training time.

During the lexicon generation, the audio is passed to the underlying speech recognition engine of a high-resource language such as English. The SALAAM algorithm then finds the best pronunciation(s) for each word in the low-resource language vocabulary based on the phonetic alphabet of the high-resource language being used. The final output of the program is a lexicon (.pls XML file) file containing automatically generated pronunciation for each word type in the vocabulary. Figure 3.5 shows what a typical lexicon file looks like.

```xml
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<lexicon version="1.0" xml:lang="zh-CN" alphabet="x-microsoft-sapi" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon">
  <lexeme>
    <grapheme>Bohobe</grapheme>
    <phoneme>mo hou wen si</phoneme>
    <phoneme>bu huo gei</phoneme>
    <phoneme>mo he wen si</phoneme>
    <phoneme>mu he wen si</phoneme>
    <phoneme>mu hou wen si</phoneme>
</lexeme>
  <lexeme>
    <grapheme>Dijo</grapheme>
    <phoneme>bi xiu</phoneme>
    <phoneme>di xiu</phoneme>
    <phoneme>pi xiu</phoneme>
    <phoneme>pi shou</phoneme>
    <phoneme>ji shou</phoneme>
</lexeme>
  <lexeme>
    <grapheme>Ditholwana</grapheme>
    <phoneme>ji chu luan</phoneme>
    <phoneme>ji cu luan</phoneme>
    <phoneme>ju chu luan</phoneme>
    <phoneme>zui chu luan</phoneme>
    <phoneme>zhei chu luan</phoneme>
</lexeme>
  <lexeme>
    <grapheme>Makotomane</grapheme>
    <phoneme>mai zhou gun mai lin</phoneme>
    <phoneme>mai zhong guo mai lin</phoneme>
    <phoneme>mao guo zhong ma nei yi</phoneme>
    <phoneme>huang huo dong ma li</phoneme>
    <phoneme>huang huo dong ma xi</phoneme>
</lexeme>
```

Figure 3.5: **seSotho lexicon file** - The first two lines in the lexicon file represent describe the metadata of the document, specifying the file type, version and XML schema to use. We also see the name of the source language used, *zh-CN* in this case and the type of phonetic alphabet used, *x-microsoft-sapi*. The rest of the document contains lexeme entries that contain the target language word types wrapped in grapheme tags and their respective alternative phonetic pronunciation wrapped in phoneme tags. Each word type in this lexicon contains five alternative pronunciations.

3.3.2 Lexicon usage and accuracy evaluation

After successful training and lexicon generation, the resulting lexicon file can be fed back into a speech recognizer trained in the same source language as that used to generate the lexicon to support speech recognition of word types it contains. This is how the evaluation module is designed to work. One specifies the lexicon file whose recognition accuracy they wish to evaluate, feed it into a speech recognizer and provide a source of audio for speech recognition, either directly using a microphone of choice or through Wav audio files.

3.3.3 TimestampLogger

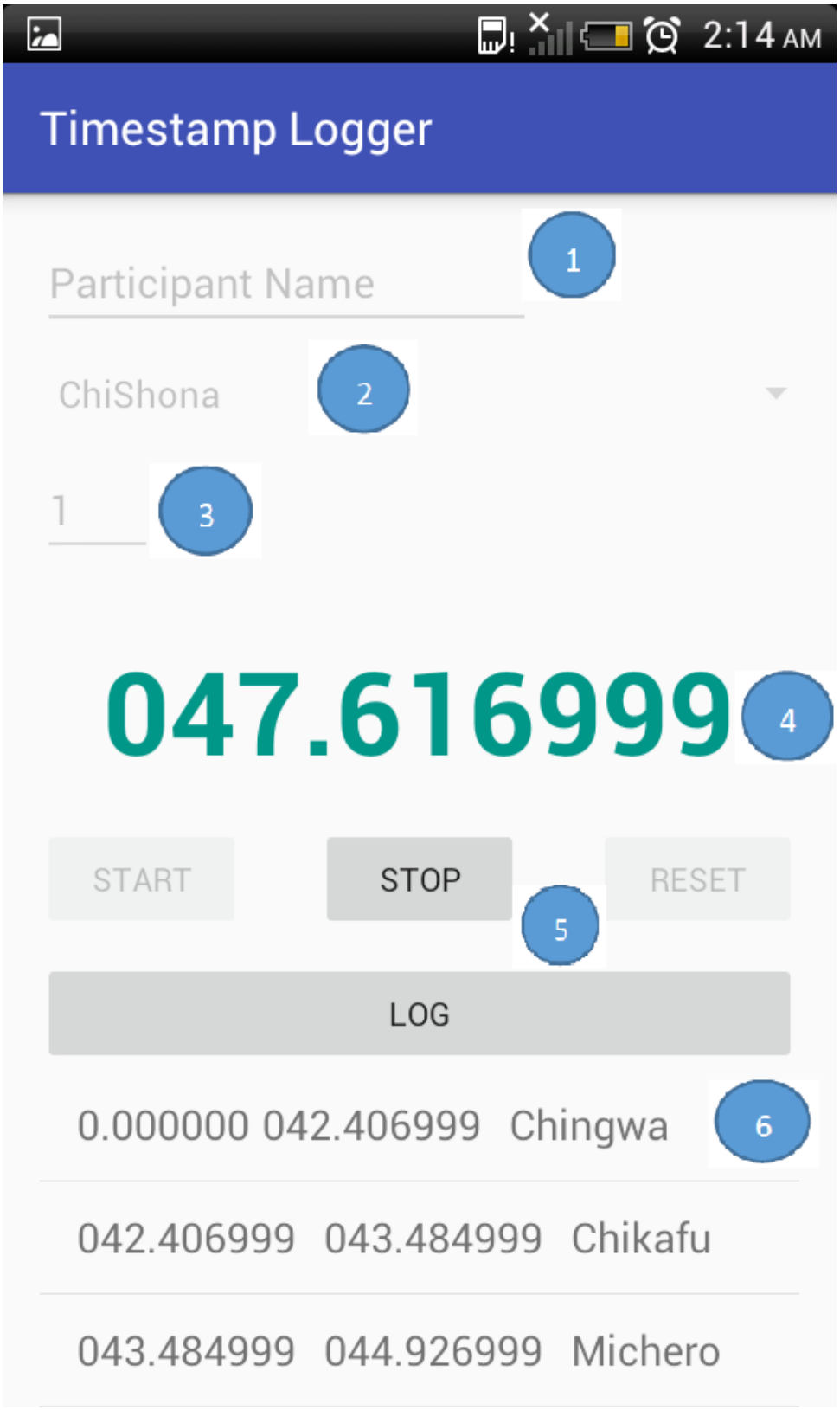

Figure 3.6: **TimestampLogger App User Interface** (The researcher would enter the participant's name in field 1 then choose a target language from the menu 2 and specify the session number in field 3. Portion 4 shows a timer in milliseconds, followed by 5, a set of control buttons and lastly part 6 shows the entries in the resulting transcript.)

TimestampLogger, whose User Interface (UI) is shown in Figure 3.6, is a utility android application that was developed to aid the truncation of the continuous audio steams of word type

pronunciations into individually labelled audio files based on each word type. The application constituted of five main parts: fields to enter participant and session details (1,2 and 3 in Figure 3.6), a timer, a menu listing target languages to select from (part 4), button controls (part 5) and a snippet of transcript entries as shown in part 6 of Figure 3.6.

During each recording session, the researcher would begin by entering session and participant data, specifying the participants name, the target language of interest and the session number. The researcher would then press the start button. The timer on the app would start running and the researcher would tap the 'Log' button to log a timestamp entry indicating the start and end timestamps when a word type was read out. At the end of each session, the researcher would reset the timer and save the resulting transcript as a text file. Since all the sessions were recorded in a single audio stream, the transcript would be used for audio stream word type segmentation, which means that the audio stream would be split into multiple smaller word-type-specific audio files. Figure 3.7 shows how the transcript would be used in Audacity, a free, open-source, cross-platform audio editing software [92]. As can be seen in the figure, each wave form representing a word type had its name correctly labelled.

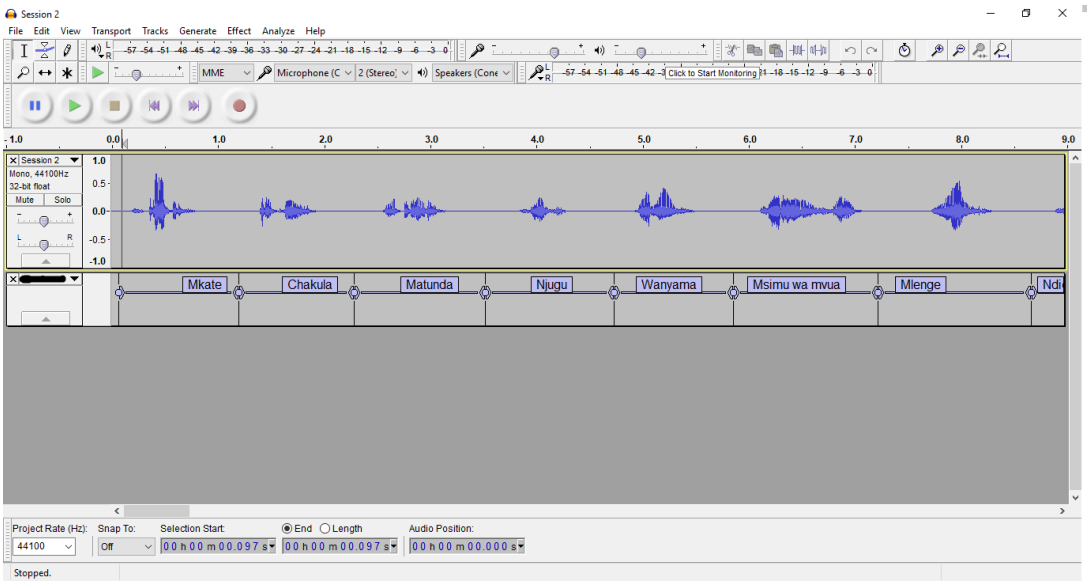

Figure 3.7: **TimestampLogger app transcript being used in Audacity for Kiswahili audio segmentation** (The first track contains audio whose word types are aligned against their respective labels with the aid of a vocabulary transcript)

3.4 Summary

This chapter described the overall system design of Lex4All, an open-source tool whose implementation of the SALAAM algorithm was used to aid our study. We began by discussing the system overview, focusing on its main components. This was then followed by a description of the training process one follows to generate lexicon files based on a source-target language pair

of choice. The structure of a resulting lexicon file and how the lexicon file is used to support speech recognition and evaluation of recognition accuracy was then discussed. Before ending the chapter, we also described a custom utility app called TimestampLogger that we used for audio stream segmentation and how the vocabulary transcript obtained from it was used with Audacity. The next chapter describes the rationale behind the decisions we made regarding participant recruitment, source language choice, target language choice, choice of target language vocabulary and the methodology we followed in conducting our experiments.

Chapter 4

Methodology

This chapter discusses the methodologies employed in our study as well as the tools and utilities used. It focuses on three main aspects: the rationale behind the choice of tools and utilities; selection of target language vocabulary; methods used during study participant recruitment and data collection procedures.

4.1 Target and Source Languages

Five target languages were used in our study, namely: English (South Africa), Afrikaans, seSotho, Kiswahili, and ChiShona. The target languages were chosen out of convenience, seeing as they were well represented amongst University of Cape Town students. This gave us confidence that we had a high chance of collecting high quality data from native speakers of these languages. English (South Africa) served as control language seeing that it shares the same phonetic inventory as English (US) which was one of our four source languages. As stated in section 1.3.1 of chapter one, English provided us with an upper bound of how we expected the system to perform since it served both as a source and target language.

4.1.1 Afrikaans

Afrikaans is a Germanic language that has its origins from 17th century Dutch birthed from the contact of the Dutch with other languages in South African after landing at the Cape of Good Hope in 1652 [35, 36]. The development of Afrikaans is characterised by borrowing mainly from Dutch, German, Bantu, Malay and Khoisan languages. It also borrows from Portuguese and other European languages such as English and French [35, 36]. Afrikaans is spoken by an estimated 17 million people around the world, 7 million as a first language and 10 million as a second language [93, 94]. Afrikaans is predominantly spoken in South Africa and it's

immediate neighbours which include Namibia, Botswana and Swaziland.

4.1.2 Kiswahili

Kiswahili is a Bantu language spoken by over 80 million people and counts as one of the most widely spoken languages in Africa [41, 95]. Kiswahili retains 44% of the protoBantu word roots, similar to Ki-Kongo, abantu language spoken in Democratic Republic of Congo, Congo-Brazzaville and Angola [95]. Kiswahili is widely spoken in East and Central Africa, specifically in Kenya, Tanzania, Uganda, Democratic Republic of Congo, Zambia, Malawi, Rwanda, Burundi and Comoros. It is also widely used for academic purposes in several places such as Europe, United States of America and the far east [41, 95].

4.1.3 ChiShona

ChiShona is a Central Bantu language that is predominantly spoken in Zimbabwe by a population of about 9 million people [40, 96]. Guthrie classifies ChiShona as S10 [39] and it belongs to the Southern Bantu cluster. ChiShona constitutes of five dialects, Karanga, Zezuru, Manyika, Korekore, and Ndau. The Karanga and Zezuru dilaects being the principal dialects [40].

4.1.4 seSotho

seSotho is a Southern African Bantu language that is predominantly spoken in South Africa (Northern seSotho) and Lesotho (Southern seSotho). Southern seSotho is the main language of Lesotho [37] whilst Northern seSotho, which is what was used in this study, is spoken in the northern province of South Africa and it is one of the eleven official languages of the Republic of South Africa [97]. It is spoken by more than 3.5 million speakers across the republic of South Africa [98]. According to Guthrie's classification, seSotho belongs to the group S30 [38] and it is mutually intelligible with all languages within the same group, Tswana and Southern seSotho.

4.1.5 English (South Africa)

English is one of South Africa's 11 official languages [35]. According to the 2016 community survey, it was spoken by 4.6 million people, 8.3% of the country's population at the time [99]. The language's history may be traced back to 1795 when South Africa saw its first British occupants [35].

4.2 Vocabulary development

We developed a vocabulary of 100 word types in English, based primarily on words commonly used with agriculture and nutrition, in addition to standard words such as numbers, months and days of the week. A vocabulary size of 100 word types per target language provided a basis for ease of statistical significance assessment as done in previous studies [29]. The vocabulary included a mixture of words and short phrases (word types). These word types were chosen to reflect the type of words likely to be used by our intended use case, a nutrition and agriculture related survey. The vocabulary was then translated from English to the other target languages with the aid of Google Translate and other sources. We substituted words or phases that either did not have a direct translation or word from English to a particular target language with word types in the same domain. In some cases, some words that were unfamiliar to the native speakers of certain languages were replaced with more familiar words, for example the words *Millet (giers)* and *Sorghum* were replaced by *cucumber (Komkommer)* and *lettuce (Blaarslaai)*. Please refer to appendix B for a list of vocabularies with respect to each target language. All vocabulary lists were checked for accuracy by native speakers before being used to prompt pronunciation audio recordings during the data collection phase.

4.3 Participant Recruitment

Participation in the study was voluntary, and the study was granted ethics approval, FSREC 077 – 2016, through the Institutional Review Board (IRB) at the university, see appendix A. An email was sent out to all University of Cape Town (UCT) [1] students through the Department of Student Affairs, inviting them to voluntarily participate in our study. If interested, the students would make an appointment with us via a web application whose link was included in the invitation email. The participants entered their contact details, selected a language they spoke with native proficiency and picked a suitable appointment date and time. Therefore, participants were considered on a first-come-first-serve basis.

Other participants were obtained using snowball sampling. Snowball sampling or chain referral sampling is a sampling methodology in which a study sample is obtained through referrals made amongst people who share or know of others who possess some characteristics that are of research interest [100]. This approach was used in an attempt to maximise the number of sign-ups in the shortest period of time. Some students generally ignore or dismiss research study participant recruitment emails or treat them as Spam. Therefore, snow ball

[1] http://www.uct.ac.za/

sampling served as an alternative way to recruit those students who either ignored the email, treated it as Spam, missed it or would only go at a friends recommendation. We recruited 104 native speakers for the five target languages consisting of 58 female and 46 male participants in total. These participants were mostly from the undergraduate population at the University of Cape Town with an exception of a few postgraduate students.

4.4 Data Collection

The recordings were done in a quiet room using a mobile phone recording at 44.1kHz in WAV format. A mobile phone was used for data collection because they have been widely adopted as a technology in low and middle income countries and the audio quality is similar to what one would expect when users are interacting with a spoken dialogue system in this context [32]. As such, our study did not employ the use of close-talking microphones which are often used for recordings that are done for training speech recognisers. The use of a mobile phone proved to be more realistic as a tool for collecting speech training data while performing similar work in the field where one may not have a sonically controlled laboratory. In spite of not replicating the various audio transformations performed by telecommunication channels, our set-up is a valid simulation of a spoken dialogue system that is running locally. Specifically, a Lenovo Moto E XT1700 [101] running Android 6.0 was used as our recording device. This was because it was readily available as it was owned by the researcher. The audio was recorded using a free android application called Voice Recorder [102]. The application was chosen because it was configurable, free and it was easy to use. The application allowed us to select the sample rate at which audio would be recorded as well as the file format in which to save the data. This provided us with control on the type of audio data we sought to collect. This allowed us to maintain the same audio quality throughout the data collection process.

Upon arrival, the objective of the study was briefly shared with the participant. They would then be given a consent form to sign, indicating that they voluntarily agreed to participate in the study and understood what was expected of them. A vocabulary list with respect to the target language the participant spoke natively was given to them . This was to allow them to get familiar with the vocabulary and make clarifications, if needed. They were then asked to read the vocabulary out loud once. Afterwards, the researcher would start the recording and the participant would then read the 100 word types, one after the other, with small breaks between each utterance. The breaks in-between each word were used to mimic natural human speaking speed and they also allowed the researcher to carefully log the time after each utterance using a custom android tool, TimestampLogger, as described in section 3.3.3 of chapter three. This

procedure was repeated five times to provide a variety of pronunciations of the same words by the participant as done in previous studies [32]. We also avoided consecutive word repetition during our data collection sessions because this may affect one's pronunciation of that word [32].

Figure 4.1: **Researcher and participant during a data collection session** (While the participant reads out one word type after another, the researcher, running TimestampLogger, taps a button after each utterance to create a transcript for use later during audio segmentation)

After recordings were done, the participant was asked to sign a form acknowledging receipt of monetary compensation for their time. Each participant was given 40 ZAR (2.8 USD). They were thanked for their participation and encouraged to tell their friends that may have not signed up for the study to participate in it if they met the requirements.

4.5 Data Cleaning

The data cleaning process involved aligning the generated audio transcript obtained from the TimestampLogger app with the audio recorded from an individual recording session with a participant using Audacity [92]. The alignment entailed listening to the audio and making sure that the labels for each word type matched the audio wave they captured during the timespan they defined. Additionally, all noise before and after a word type was also removed to prevent that from affecting the training and evaluation process. No other filtering or noise reduction techniques were used during this process to retain some of the background noise one

would normally have if they interacted with a spoken dialog system while in a quiet room. Essentially, the researcher performed word segmentation with the TimestampLogger app and hand-trimming.

4.6 Experiments

Using the data collected, there were three experiments we conducted, based on the three main research questions we raised:

1. **What impact does source language choice have on recognition accuracy**: The generation of pronunciation lexicons that map each term from a target language to one or more sequences of phonemes in the source language depends on the phonemes the high resource language speech recogniser can model [25, 26, 32]. Therefore, for this research question, we hypothesised that if the target and source languages are of similar linguistic properties then the overlap between the source language's phoneme inventory and that of the target language shall be maximised. This would in turn reduce the difficulty of phoneme mapping by finding better pronunciations and yielding better recognition accuracy.

2. **What impact does gender composition of the training data set have on recognition accuracy**: For this research question, we hypothesised that, for applications developed using cross-language phoneme mapping, gender also has a confounding effect on recognition accuracy, as it has in previous studies [103]. This is due to the different acoustic properties between the genders and the effect on interpretation by the underlying speech recogniser [103]. We evaluated recognition accuracy across three experimental setups: *same-gender pairs* (training and testing datasets comprised of a single gender), *multi-gender* pairs (mixed-gender training and testing datasets) and *cross-gender* pairs (training with a single gender and testing with the other gender). Each dataset consistently constituted of data from four randomly selected participants with respect to the gender. Our experimental setup for this question is unique in that unlike previous studies [103], we focus on speech applications developed using cross-language phoneme mapping and not a traditionally developed acoustic model for speech recognition as was the focus in [103]. Additionally, previous studies [26, 32] did not separate their training datasets as we did and their sample sizes were smaller than ours.

3. **What impact do varied alternative pronunciations per word type have on recognition accuracy**: For this research question, we hypothesised that increasing the number

of alternative pronunciations would improve recognition accuracy, as demonstrated in previous work [32], up to an inflection point, after which recognition accuracy would decrease. We expected that this drop would occur due to the inevitable overlap of alternative pronunciations for words with similar phonetic structure, as hypothesised in prior work [26]. Unlike previous studies [26, 32] which used a maximum of five alternative pronunciations, we investigated this with up to 100 alternative pronunciations per word type.

We used the SALAAM method as implemented in the open source tool Lex4All [34]. Concerning hypothesis (1), we used four source language recognisers: English (US), French, Mandarin and German, chosen because of availability of phonetic alphabets. We accessed these recognisers through Microsoft Speech Platform Software Development Kit (SDK) 11 [38], a technology developed by Microsoft for server-side recognition of telephone-quality audio. This system was used because of its robustness and it also allowed us to mimic the experimental environment of previous studies [26, 32, 34]. No additional modifications to the underlying models of this system were made –our goal was to test a system that was feasible for groups to implement quickly.

4.6.1 Training and lexicon generation

For each target language, we created three training datasets: *male-only*, *female-only* and *mixed-gender*. The single gender datasets were created by randomly selecting a sample of four participants per gender. The *mixed-gender* dataset was made up of two male speakers and two female speakers from the *male-only* and *female-only* datasets, all of which were also randomly selected. All test datasets were created by randomly selecting two female and two male speakers whose data was not used to form any of the training datasets. These datasets were then used during the evaluations stage of individual experiments. The total number of speakers per dataset was capped to four to ensure uniform testing conditions across all target languages. The randomised selection of speakers for each of these datasets was achieved using the 'sample' function from the R software environment [104].

For each source-target language pair, using Lex4All's 'fast training' [34] feature, we generated lexicon pronunciation files for the *female-only (single-gender)*, *male-only (single-gender)* and *mixed-gender (multi-gender)* training sets using the SALAAM algorithm as implemented in Lex4All. All generated lexicon files had a maximum of 100 alternative pronunciations per word type originally. We further segmented the generated lexicon files into other lexicon files

with different number of alternative pronunciations per word type as per our experiment setup: five, ten, twenty, forty, sixty and eighty alternative pronunciations. To generate a lexicon file with five alternative pronunciations we picked the top five alternative pronunciations per word type from the original lexicon file obtained during training. Likewise, to obtain a lexicon file with ten alternative pronunciations, we picked the top ten pronunciations from the original lexicon file and so on. Not all word types retained 100 alternative pronunciations after training, therefore we took all alternative pronunciations of each of these word types if they had less than the desired number of alternative pronunciations during segmentation. All the resultant lexicon files were used for evaluation.

During evaluation, the underlying speech recogniser would match the audio data provided against any of the alternative pronunciations per word type without making any distinction or preference among them as per design [26]. We used the R software environment [104] for statistical analysis and Seaborn [105] for data visualisation using the Python programming language [106].

It must be noted that although discriminative training was shown to further improve recognition accuracy of speech applications developed using SALAAM [29], we did not employ it in our study. This was due to the long training time it demands, as also observed in previous studies [26] and the loss of some alternative pronunciations and vocabulary word types from the resulting lexicon file for some target languages which was observed during our preliminary experiments, resulting in inconsistent vocabulary sizes across the different target languages.

4.7 Summary

This chapter discussed the methodology used in our study. The chapter began by discussing the target and source language choices as well as the rationale behind them. We then proceeded to briefly discuss the background of each target language, the countries in which each of these target language are predominately spoken and the average number of speakers. We went on to discuss the development of the vocabulary used, starting with the context within which the vocabulary would be commonly used and the rationale behind the size of the vocabulary and source of translations. This was followed by our discussion of our participant recruitment and data collection approaches. We ended the chapter by discussing the methods employed in investigating the research questions that guided our study, specifying the specific approach undertaken with respect to each research question.

In the next chapter, we extensively describe the experiments we undertook, applying the

aforementioned methods above, as well as the results we obtained from them. We will then proceed to discuss the results and their implications.

Chapter 5

Experiments and Results

5.1 Overview

This chapter discusses the individual experiments we performed in our study, the setup of each experiment, the measure, methods employed and the results obtained. Section 5.2 describes the first experiment in which we investigated the effect of source language choice on recognition accuracy. The second experiment, we investigated the effect of training technique on recognition accuracy with respect to gender, is described in section 5.3. Section 5.4 describes the third and last experiment in which we investigated the effect of the number of alternative pronunciations per word on recognition accuracy. The results are discussed and the chapter ends with a summary.

5.2 Experiment 1: Source Language Effect on Accuracy

Experiment one investigated the effect of source language choice on recognition accuracy for speech-driven applications that use cross-language phoneme mapping to support speech recognition. The generation of pronunciation lexicons that map each term from a target language to one or more sequences of phonemes in the source language is dependent on the phonemes the high resource language recognizer can model [25, 26, 32]. To this effect, we hypothesized that if the target and source languages are of similar phonetic properties then the overlap between the source language's phoneme inventory and that of the target language shall be maximized, hence reducing the difficulty of phoneme mapping, finding better pronunciations and yielding better recognition accuracy. The following sections describe the experiment setup, procedure and the results obtained.

5.2.1 Experiment Setup and Procedure

Table 5.1 shows a summary of the parameters and values used in this experiment. The first column shows the parameters and the second column shows the values each of these parameters took. We used all five target languages: ChiShona, Afrikaans, Kiswahili, English (South Africa) and seSotho and four source languages: English (US), French, German and Mandarin. Each of the target languages had a vocabulary size of 100 words.

Table 5.1: **Experiment 1: Summary of parameters and values used**

Target Language (s)	Afrikaans, ChiShona, English (South Africa), seSotho and Kiswahili
Source Language(s)	English (US), Mandarin, French and German
Vocabulary Size	100 word types
Variable	Source Language
Measure	Recognition accuracy

This experiment followed the training and lexicon generation procedure described in section 4.6.1. We evaluated recognition accuracy with respect to the source language used. This evaluation was run against each source-target language specific lexicon file across all alternative pronunciations per word type, five, ten, twenty, forty, sixty, eighty and one hundred and training techniques, target language training datasets consisting of either *female-only*, *male-only* or *mixed-gender* training data. The results obtained were recorded and are discussed in the next section.

5.2.2 Results

The overall results showed that using English (US) as the source language produced the best results across all target languages with a mean recognition accuracy of 71%, this was followed by French with 66%, German with 65% and Mandarin with 64%. Looking at source-target language pairs, the best source language choice for English (South Africa) target language was English (US) followed by German, French and then Mandarin. For ChiShona, English (US) was the best source language choice followed by Mandarin, French and German. Kiswahili also performed best with English (US) as the source language followed by French, German and Mandarin. Lastly, seSotho performed best with Mandarin as the source language choice followed by French, German and then English (US).

These results were further analysed using several statistical methods. A Shapiro-Wilk test determined our data was not normally distributed. The Kruskal-wallis test, a non-parametric test, was used for statistical analyses to determine the statistical significance of the differences observed. The first evaluation looked at the results irrespective of the target language, we

achieved this by aggregating the results based on the source language only. The tests revealed a significant overall effect of source language on recognition accuracy ($x^2(3) = 110.29, p < 0.01$), confirming our hypothesis.

Performing Post-hoc pairwise comparisons using the Wilcox sum rank test with Bonferroni correction showed a significant effect of source language on recognition accuracy among all source languages except between French and German. See Table 5.2 for the source language post-hoc pairwise comparison test results. The table shows the source language pairs with their respective medians in brackets in the first column then the p-value obtained from the tests in the next column. If the p-value is not less than 0.05, it is represented by *n.s* which stands for non-significant.

Table 5.2: **Experiment 1: Overall post-hoc pairwise tests across source languages**

Source language pairs	**P-value**
French (66) - English (71)	< 0.0001
German (65) - English (71)	< 0.0001
German (65) - French (66)	n.s
Mandarin (64) - English (71)	< 0.0001
Mandarin (64) - French (66)	< 0.0001
Mandarin (64) - German (65)	< 0.01

5.2.3 Experiment 1 - Target language specific findings

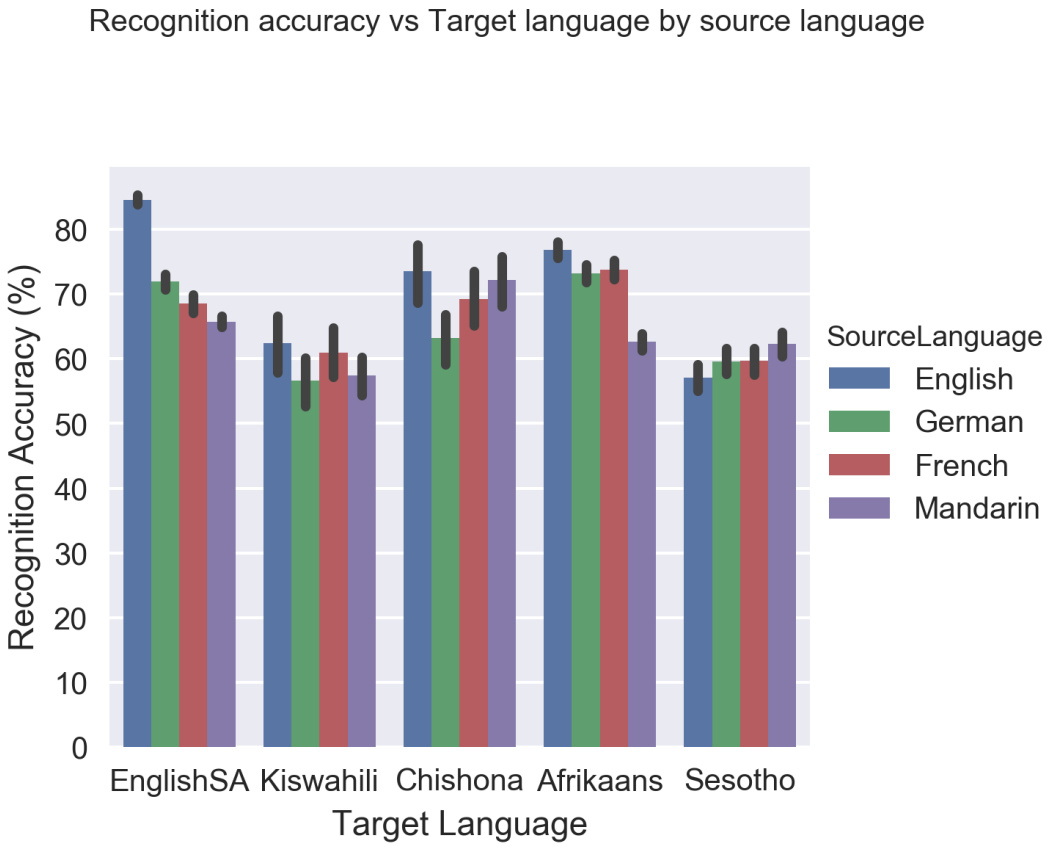

Figure 5.1: **Recognition accuracy vs Target language by source language** (The x-axis shows the target language and the y-axis shows the percentage of the number of correctly recognised words. The coloured bars represent individual source languages.)

Figure 5.1 shows an overview of the recognition accuracy recorded with each source-target language pair. We repeated the statistical analyses described in section 5.2.2 but this time focusing on individual target languages. The results from these analyses also showed significant overall effects of source language on recognition accuracy for all target languages. Our findings are summarised and shown in table 5.3.

Table 5.3: **Experiment 1: Overall Statistical findings per target language**

Target language	**Chi-Squared**	**Degrees of freedom**	**P-value**
Afrikaans	131.37	3	< 0.0001
Kiswahili	24.084	3	< 0.0001
English (South Africa)	203.95	3	< 0.0001
ChiShona	49.136	3	< 0.0001
seSotho	15.727	3	< 0.01

The post-hoc pairwise analyses performed across all source languages with respect to individual target languages produced the results shown in Table 5.4, represented in the same fashion as the results in Table 5.2 with the medians for each source language in round brackets.

Table 5.4: **Experiment 1: Post-hoc pairwise tests across source languages per target language**

Kiswahili	
Source language pairs	**P-value**
French (61) - English (62)	n.s
German (57) - English (62)	< 0.001
German (57) - French (61)	< 0.05
Mandarin (57) - English (62)	< 0.01
Mandarin (57) - French (61)	n.s
Mandarin (57) - German (57)	n.s
ChiShona	
Source language pairs	**P-value**
French (69) - English (73)	< 0.001
German (63) - English (73)	< 0.0001
German (63) - French (69)	< 0.001
Mandarin (72) - English (73)	n.s
Mandarin (72) - French (69)	n.s
Mandarin (72) - German (63)	< 0.0001
seSotho	
Source language pairs	**P-value**
French (60) - English (57)	n.s
German (60) - English (57)	n.s
German (60) - French (60)	n.s
Mandarin (62) - English (57)	< 0.001
Mandarin (62) - French (60)	n.s
Mandarin (62) - German (60)	n.s
Afrikaans	
Source language pairs	**P-value**
French (74) - English (77)	< 0.05
German (73) - English (77)	< 0.01
German (73) - French (74)	n.s
Mandarin (63) - English (77)	< 0.0001
Mandarin (63) - French (74)	< 0.0001
Mandarin (63) - German (73)	< 0.0001
EnglishSA	
Source language pairs	**P-value**
French (67) - English (85)	< 0.0001
German (72) - English (85)	< 0.0001
German (72) - French (67)	< 0.01
Mandarin (66) - English (85)	< 0.0001
Mandarin (66) - French (67)	< 0.05
Mandarin (66) - German (72)	< 0.0001

For Kiswahili, *German - English*, *German - French* and *Mandarin - English* source language pairs recorded statistically significant differences in recognition accuracy while ChiShona recorded statistically significant differences in recognition accuracy across all source language pairs except for *Mandarin - French* and *Mandarin - English*. seSotho only recorded a statistically significant difference in recognition accuracy for the *Mandarin -English* source language pair and Afrikaans recorded all but the *German-French* source language pair as having

statistically significant differences in recognition accuracy. English (South Africa) recorded statistically significant differences in recognition accuracy between all its source language pairs.

5.3 Experiment 2: Effect of Gender on Accuracy

Here, we describe an experiment in which we investigated the effect of training technique on recognition accuracy with respect to gender. We hypothesized that gender would have a confounding effect on recognition accuracy for applications developed using cross-language phoneme because each gender has different acoustic properties which in turn affect the way a gender's audio signals are interpreted by the underlying speech recognizer [103].

5.3.1 Experiment Setup and Procedure

Table 5.5: **Experiment 2: Summary of parameters and values used**

Target Language (s)	Afrikaans, ChiShona, English (South Africa), seSotho and Kiswahili
Source Language(s)	English (US), Mandarin, French and German
Vocabulary Size	100 word types
Variable	Training technique (Multi Gender, Same Gender, and Cross Gender)
Measure	Recognition accuracy

As shown in the table 5.5, we used all five target languages and source languages, with each target language having a 100 word vocabulary size. The variable for the investigation was training technique which was defined by the gender composition of the training dataset. As described in section 4.6.1, we prepared three training datasets for each target language; *female-only*, *male-only* and *mixed-gender* training datasets. Each training dataset contained audio data from four randomly selected speakers; four female speakers, four male speakers and two female and two male speakers respectively. Each target languages evaluation dataset constituted of data from four randomly selected speakers, two female and two male speakers. We trained and evaluated recognition accuracy across all lexicon files with five, ten, twenty, forty, sixty, eighty and one hundred alternative pronunciations per word type as we did in experiment one.

During evaluation, we used three labels to tag our results, *same-gender*, *cross-gender* and *multi-gender*. *same-gender* meant that the gender of the speakers whose data was used for training matched that of the speaker whose data was currently being used for evaluation. On the other hand, *cross-gender* meant that the gender of the speakers whose data was used for training was different from that of the speaker whose data was currently being used for evaluation.

multi-gender meant that the *mixed-gender* training dataset was used to generate the current lexicon file used during the evaluation process.

5.3.2 Results

Presented here are the results obtained from this experiment. We begin by presenting overall results followed by language specific findings.

The overall results showed that the *multi-gender* training technique produced the best results with a median of 68%. This was followed *single-gender* with 67% and lastly *cross-gender* with 65% recognition accuracy medians.

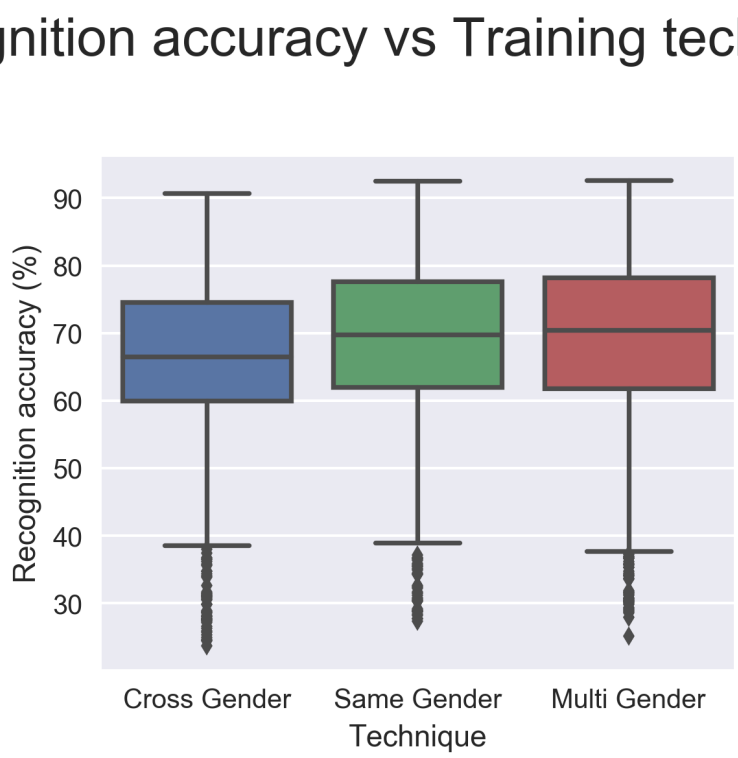

Figure 5.2: **Overall - Recognition accuracy vs Training technique**

Figure 5.2 shows the overall results obtained from this experiment. The x-axis represents the training technique and the y-axis represents the recognition accuracy in percentages. The figure shows how the three training techniques compare in terms of recognition accuracy. We can see that the *multi-gender* technique performed the best followed by *same-gender* and *cross-gender* respectively.

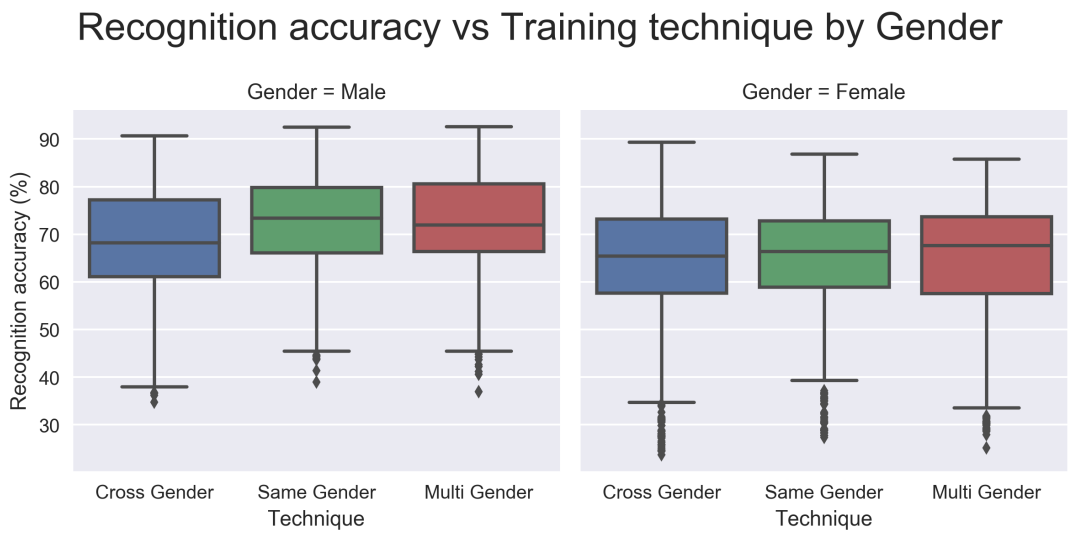

Figure 5.3: **Overall - Recognition Accuracy vs Training Technique by Gender**

Figure 5.3 above shows the same results, however, with emphasis on how each of the genders performed with respect to the training technique used. We see from Figure 5.3 that the performance of the male gender are consistent with the overall results presented in Figure 5.2; the *multi-gender* technique performed the best followed by *same-gender* and *cross Gender* respectively. For the female gender, the order of training technique performance is such that the *multi-gender* technique performed best followed by the *cross-gender* and lastly the *same-gender*.

The results were further analysed using the statistical tests described in experiment one and they revealed an overall statistically significant impact of training technique on recognition accuracy, $(x^2(2) = 17.77, p < 0.001)$, confirming our hypothesis. Performing Post-hoc pairwise showed statistically significant differences in recognition accuracy between the *Cross-gender - multi-gender* and *cross-gender - same-gender* training technique pairs, $(p < 0.001)$ and $(p < 0.01)$ respectively.

5.3.3 Experiment 2 -Language specific findings

Further investigations revealed that only Kiswahili and English (South Africa) recorded significant differences in recognition accuracy with respect to training technique. A summary of source-target language findings for this experiment are shared in table 5.6 below.

Table 5.6: **Experiment 2: Target language specific statistical findings**

Target language	**Chi-Squared**	**Degrees of freedom**	**P-value**
Afrikaans	2.3836	2	n.s
Kiswahili	14.605	2	< 0.001
English (South Africa)	36.296	2	< 0.0001
ChiShona	4.7317	2	n.s
seSotho	0.47859	2	n.s

5.4 Experiment 3: Effect of Alternative Pronunciations on Accuracy

In experiment three, we aimed to establish the effect of the number of alternative pronunciations per word type on recognition accuracy. We used all five target languages, each with a 100 word vocabulary size to aid this investigation.

Figure 5.4 below shows a snippet of a lexicon file generated by training Kiswahili audio on a recognizer trained for American English. The two Kiswahili words (graphemes) shown in the figure *Mkate* (bread) and *Chakula* (Food) each have five different pronunciations. Each of these pronunciations are a unique combination of phonemes that represent the pronunciation of a word based on the phonetic alphabet of the underlying speech recognition engine.

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<lexicon version="1.0" xml:lang="en-US" alphabet="x-microsoft-ups" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon">
  <lexeme>
    <grapheme>Mkate</grapheme>
    <phoneme>M K AA T I</phoneme>
    <phoneme>M K AA T I I</phoneme>
    <phoneme>M T AE K I</phoneme>
    <phoneme>M K AA K I</phoneme>
    <phoneme>M H AE T I</phoneme>
  </lexeme>
  <lexeme>
    <grapheme>Chakula</grapheme>
    <phoneme>CH AE P L I AX</phoneme>
    <phoneme>CH AA K L I AX</phoneme>
    <phoneme>CH EH P L I AX</phoneme>
    <phoneme>CH AH K L I AX</phoneme>
    <phoneme>CH AA K W I AX</phoneme>
  </lexeme>
```

Figure 5.4: **Snippet of Kiswahili Lexicon file**

In this experiment, we hypothesised that increasing the number of alternative pronunciations will improve recognition accuracy, as demonstrated in previous work [32]. However, we further hypothesise that this improvement in recognition accuracy would only be observed up to an inflection point, after which recognition accuracy will decrease. We expect that this drop will occur due to the inevitable overlap of alternative pronunciations for words with similar phonetic structure, as hypothesised in prior work [26]. This is due to the phonetic mismatch (confusion) that the speech recognition engine undergoes when trying to distinguish words of similar phonetic structures using the SALAAM technique [29, 32].

5.4.1 Experiment Setup and Procedure

Table 5.7 shows a summary of the parameters and values used in this experiment. The first column shows the parameters and the second column shows the values each of these parameters took. We used all five target languages: ChiShona, Afrikaans, Kiswahili, English (South Africa) and seSotho and four source languages: English (US), French, German and Mandarin. Each of the target languages had a vocabulary size of 100 word types.

Table 5.7: **Experiment 3: Summary of parameters and values used**

Target Language (s)	Afrikaans, ChiShona, English (South Africa), seSotho and Kiswahili
Source Language(s)	English (US), Mandarin, French and German
Vocabulary Size	100 word types
Variable	Alternative pronunciations per word type (5,10,20,40,60,80,100)
Measure	Recognition accuracy

Like the two experiments above, this experiment also followed the training and lexicon generation procedure described in section 4.6.1. We evaluated the recognition accuracy of each generated source-language-specific lexicon file with respect to the number of alternative pronunciations per word type, five, ten, twenty, forty, sixty, eighty and one hundred alternative. The results obtained were recorded and are discussed in the next section.

5.4.2 Results

We present results obtained from investigating the impact of number of alternative pronunciations on recognition accuracy.The results revealed that the highest mean recognition accuracy recorded across the different alternative pronunciation sizes was 68% which was true for alternative pronunciation sizes of forty, sixty, eighty and one hundred. The mean recognition accuracy for the other alternative pronunciation sizes was as follows, 66% for 20, 65% for 10 and 63% for 5 alternative pronunciations per word type.

Figure 5.5 below shows a box plot of how recognition accuracy varies across all target and source language pairs with variable alternative pronunciations per word type. The x-axis represents the number of pronunciations per word, 5p meaning 5 alternative pronunciations, 10p meaning 10 alternative pronunciations and so on. The y-axis represents recognition accuracy in percentage.

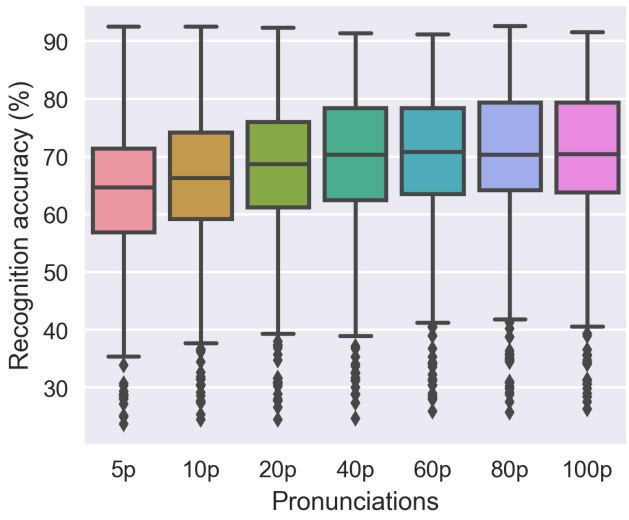

Figure 5.5: **Recognition Accuracy vs Number of Pronunciations**

As shown in Figure 5.5, recognition accuracy generally improved as the number of alternative pronunciations per word type increased. From a pronunciation size of five we see a steady improvement in recognition accuracy up to size forty after which we do not see much improvement.

Statistical analyses on the results further revealed that there existed an overall statistically significant difference in recognition accuracy across all alternative pronunciation sizes, $(x^2(6) = 51.31, p < 0.0001)$. Table 5.8 shows the alternative pronunciation size pairs whose recognition accuracy was found to have a statistically significant difference. All other alternative pronunciation size pairs did not record statistically significant differences with regards recognition accuracy.

Table 5.8: **Experiment 3: Post-hoc pairwise tests across alternative pronunciations**

Alternative pronunciation pairs	**P-value**
5p (63) - 20p (66)	< 0.01
5p (63) - 40p (68)	< 0.0001
5p (63) - 60p (68)	< 0.0001
5p (63) - 80p (68)	< 0.0001
5p (63) - 100p (68)	< 0.0001
10p (65) - 60p (68)	< 0.05
10p (65) - 80p (68)	< 0.01
10p (65) - 100p (68)	< 0.01

All the results presented for this experiment thus far were obtained across all source-target language pairs and training techniques, overall target language specific results are presented under section 5.4.3.

We also recorded the time, with respect to the number of alternative pronunciations, it took to run each evaluation session. We observed that, the evaluation time increases as the number of alternative pronunciations increases as well. Presented in Table 5.9 is a summary of our observations of mean evaluation time for 2000 English (South Africa) word type evaluations with a varying number of alternative pronunciations.

Table 5.9: **Experiment 3: Evaluation time vs number of pronunciations**

Number of pronunciations	Mean evaluation time (Minutes)
5	2.08
10	2.08
20	4.08
40	8.75
60	13.25
80	17.41
100	21.08

The experiment whose results are shown in Table 5.9 was conducted using a Lenovo G50-80 laptop running a 64-bit Microsoft Windows 10 operating system, then machine was equipped with 6GB RAM and it was running on an Intel Core i5-500U Central Processing Unit (CPU).

5.4.3 Experiment 3 - Target language specific findings

Further exploring the overall effect of the number of alternative pronunciations on recognition accuracy per target language revealed that only three target languages, Afrikaans, English (South Africa) and seSotho, as shown in Table 5.10 below, recorded overall statistically significant differences in the recognition accuracy across the different alternative pronunciation sizes.

Table 5.10: **Experiment 3: Target language specific statistical findings**

Target language	Chi-Squared	Degrees of freedom	P-value
Afrikaans	27.138	6	< 0.001
Kiswahili	6.6002	6	n.s
English (South Africa)	16.328	6	< 0.05
ChiShona	8.6358	6	n.s
seSotho	41.53	6	< 0.0001

5.5 Discussion

Experiment one results revealed an overall significant difference of recognition accuracy across the four source languages used. These results were also reflected in the subsequent analyses that focused on individual target languages. The results supported our hypothesis that the source

language choice has an impact on recognition accuracy when cross-language phoneme mapping is used as a technique to support speech recognition. However, our findings are contrary to those of previous work [26]. Vakil et al investigated the impact of source language choice on recognition accuracy using a 25-word size vocabulary of Yoruba, the target language, and English (US) and French as the source languages. Their results showed that source language had no statistically significant impact on recognition accuracy [26]. However, there are several possible ways one could explain the differences in our findings and theirs:

Firstly, the training and testing datasets used in previous research could have been insufficient to facilitate the observation of results similar to our findings. In their experiment, Vakil and Palmer used a 25-word subset of Yoruba audio data collected from by Qiao et al [26, 32]. The vocabulary size is especially important because the SALAAM technique is reported to perform very well with small vocabulary sizes [29], making it difficult to assess statistical significance of differences among test cases. As stated in the vocabulary development, section 4.2 of the methodology chapter, this knowledge motivated us to use a vocabulary size of 100 word types. Secondly, the sample size used was also small, consisting of only two speakers, one male and one female. A subset of the data from the two participants was used for training and the rest of it for evaluation which suggests that the statistically insignificant difference in recognition accuracy between the two source languages could have been due to the use of data from the same speakers for training and evaluation. Additionally, as suggested in [26], Yoruba shares a substantial phonetic overlap with both English (US) and French, English (US) and French also share a substantial phonetic overlap, therefore one could expect that the delta between Yoruba and English and that between Yoruba and French are similar, consequently contributing to the statistically insignificant differences in recognition accuracy observed between the two source languages.

The implications of these findings are such that researchers and practitioners should take the source-target language phonetic similarity into consideration if they are to employ cross-language phoneme mapping during the development of small-vocabulary speech recognisers or other speech related technology such as cross-language acoustic modelling to maximise the quality of the resulting recognition accuracy. If the SALAAM technique and the Microsoft Speech Platform SDK version 11 are used, we would recommend the use of English (US) as the source language in an event that the researchers or practitioners do not have the time and resources to establish the source language that would produce the best recognition accuracy. Otherwise, as shown in our study, it is worth establishing the source-target language pair that produces the best recognition accuracy. This is accentuated by the unexpected re-

sults we obtained for seSotho. Unlike all other four target languages, seSotho recorded the best recognition accuracy when Mandarin was used as the source language and not English (US). Though unexpected, we hypothesised that, since Mandarin and seSotho are both tonal languages, the Mandarin based speech recogniser was able to model seSotho words better than the other speech recognisers based on non-tonal source languages, English [107], French [108] and German [109]. Additionally, we further hypothesised that the overall better performance of English (US) as a source language could be because the English (US) acoustic models in MSP are the best-trained out of our choice of source languages. However, since MSP is a commercial platform, we were unable to ascertain this hypothesis.

The results obtained in experiment two suggest that gender has confounding effects on recognition accuracy recorded by speech applications developed using the SALAAM technique. Our findings show that using the *multi-gender* training technique produces the best speech recognition accuracy. The reason for this could be that the technique would likely produce a more robust application against gender-bias because data from both the male and female participants were used during training. This factors in the difference in signal interpretation of different genders by the underlying speech recogniser as reported in [103]. Despite the *multi-gender* technique outperforming the *same-gender* technique, the difference in performance is not statistically significant. This seems to suggest that one can still use either training technique with no substantial loss in performance.

The underlying matter, however, is that gender has to be carefully considered during the development of speech recognition technology that uses cross-language phoneme mapping to accommodate different acoustic properties of the genders. Our work supports the findings from those recorded by Abdulla et al, 2001. In their study, they aimed at improving speech recognition accuracy by using gender separation [103]. The criterion they used to achieve gender separation was average pitch frequency of the speakers, achieving 100% gender discrimination accuracy. They hypothesized that if a speech engine was trained using data from the same gender as it's users, it would produce better recognition results than if the genders were different or mixed. They created three Hidden Markov Models; a pooled model (trained using both male and female audio training data), male model (trained using only male audio data) and female model (trained using only female audio data). Their findings revealed that the use of gender separation was an effective technique to improve speech recognition [103].

Lastly, the results we obtained in experiment three showed that an increase in the number of alternative pronunciations per word type generally improved recognition accuracy reaching plateau of mean recognition accuracy at 68%. These results tire in with our initial hypothesis as

described in Section 4.6, Experiments. However, we did not observe a decrease in recognition accuracy after reaching plateau. This could be explained by a number of reasons. Firstly, there is a possibility that the vocabulary words we chose did not have enough phonetic overlaps to show the inflection point we hypothesised. If this is the case, it would imply that the SALAAM algorithm was able to uniquely represent most word types as a sequence of phonemes based on the underlying source language's phonetic alphabet, thereby having very few recognition conflicts. Secondly, increasing the number of alternative pronunciations per word type increased the search space thereby increasing the probability of the algorithm finding a unique combination of phonemes that matched a speaker's pronunciation of a word type. We believe the increase in the search space size is what caused the observed increase in the mean evaluation time shown in Table 5.9. Since Microsoft Speech Platform could not allow us to set more than 100 alternative pronunciations per word type, we were unable to determine if adding even more pronunciations could eventually result in a decrease in recognition accuracy, further increase in recognition accuracy or if there would be no change in the mean recognition accuracy recorded. It is, however, important for researchers and practitioners to consider the trade-off between improving recognition accuracy by increasing the number of alternative pronunciations per word type and the response time of the overall system. Based on our findings, having 40 alternative pronunciations per word type would be ideal to develop a high quality small-vocabulary speech recognition application with SALAAM.

5.6 Summary

This chapter described three experiments we conducted to understand the effect of source language, training technique and number of alternative pronunciations on recognition accuracy. We then went ahead and discussed the implications of the results obtained for each experiment.

In experiment one, we investigated the effect of source language on recognition accuracy. We used four source languages: English (US), Mandarin, German and French and five target languages: English (South Africa), ChiShona, Kiswahili, seSotho and Afrikaans. The experiment revealed that there exists a statistically significant difference on recognition accuracy with respect to source language choice.

Experiment two investigated the effect of training technique on recognition accuracy. Using four source languages and five target languages, we created three training datasets; *female-only*, *male-only* and *mixed-gender* to aid our investigation. Our findings revealed that gender had a

confounding effect on recognition accuracy for applications developed using cross-language phoneme mapping, confirming our hypothesis. The results showed that the *multi-gender* training technique yields the best performance. Statistical analyses revealed statistically significant differences in recognition accuracy between the *Cross-gender - multi-gender* and *cross-gender - Same-gender* training technique pairs.

Lastly, experiment three was aimed at establishing the effect of the number of alternative pronunciations on recognition accuracy. We used four source languages and five target languages to aid our investigations. Our results showed a steady improvement in recognition accuracy with an increase in the number of alternative pronunciations per word type. Statistical analyses revealed an overall statistically significant difference across the pronunciation sizes. We also observed that the response time of the system is inversely proportional to the number of alternative pronunciations per word thereby introducing a recognition accuracy to response time trade-off.

Chapter 6

Conclusion

Cross-language phoneme mapping has been successfully used in a number of information, communication, and technology for development (ICT4D) projects in the health sector [5], agriculture sector [7, 15] and for research purposes [26, 34]. Despite the development and use of this technique in these and other projects, there remains little guidance of how the technique behaves in different conditions, i.e. different source-target language pairings, and training techniques. This is especially true for African languages in general and Bantu languages in particular.

This dissertation described a series of investigations in which we tried to understand the limitations and suitability of using cross-language phoneme mapping for data collection in low-resource languages using spoken dialogue systems, particularly those built using the SALAAM technique. This work was motivated by the lack of adequate guidance of how cross-language phoneme mapping, as a technique to develop small-vocabulary speech recognisers, fares in different conditions such as different source-target language pairings and training techniques. This is especially true for African languages in general and Bantu languages in particular which were the focus of this study. In order to establish the aforementioned suitability and limitations of the technique, the following three research questions were raised:

1. What impact does source language choice have on recognition accuracy?

2. What impact does gender composition of the training data set have on recognition accuracy?

3. What impact do varied alternative pronunciations per word type have on recognition accuracy?

Before discussing the approach taken to answer the research questions above, the background work and the body of literature that provided the foundation of this work was discussed.

UCT students were recruited for the study through a campus-wide email on a voluntary basis. seSotho, Kiswahili, ChiShona, Afrikaans and English (SA) were used as the target languages while English (US), Mandarin, German and French were used as the source languages. the vocabulary used constituted of 100 word types and it was developed in English based on the agriculture and nutrition use case this study was based on. It was then translated to the other target languages with the help of native speakers and the Google translate service.

During data collection, participants were invited to a quiet room and presented with the vocabulary they were expected to read out. They were then provided with a consent form to confirm they voluntarily participated in the study and were in agreement with the terms. They were then asked to read each word type once during a session, participating in five sessions in total. The participants were then given thanked and given a incentive of 2.8 USD. The data was then cleaned and prepared for use in the three experiments setup to address the research questions.

6.1 Synthesis of experimental results

This section synthesises experimental findings from this study with respect to their research questions as well as the methodology employed in each experiment.

1. What impact does source language choice have on recognition accuracy?

   Using a cross-language phoneme approach, we generated lexicon files using SALAAM and developed four different speech recognisers based on the four source languages, Mandarin, English (US), German and French. We then evaluated the recognition accuracy of each individual speech recogniser with respect to the source and target language pairings. We observed that mapping English (US) phoneme to English (South Africa) phonemes produced the highest recognition accuracy. This result underscores the hypothesis that if the target and source languages are phonetically similar, the resulting speech recogniser developed using cross language phone mapping is going to record a higher recognition accuracy of target language vocabulary than if they the two languages were not phonetically similar.

2. What impact does gender composition of the training data set have on recognition accuracy?

   To address this question, we developed three gender-based training sets, *male-only*, *female-only* and *multi-gender* and developed speech recognisers for each source-target language pair trained on these datasets. We then evaluated recognition accuracy of target language vocabulary word types from both male and female participants. Through the experiment results, we established that the recognition accuracy recorded by speech recognisers developed using cross-language mapping are also affected by the gender composition of the training data. The results also showed that one could employ gender separation to achieve better recognition accuracy, though a *multi-gender* dataset produced a more robust speech recognition against gender bias in speech recognition tasks.

3. What impact do varied alternative pronunciations per word type have on recognition accuracy?

   To address this research question,we generated lexicon files containing a variable number of alternative pronunciations per word type for each source-target language pair. These lexicons consisted of five, ten, twenty, forty, sixty, eighty and one hundred alternative pronunciations per word type. We evaluated the recognition accuracy of each of these speech recognisers developed with these lexicon files. The results that recognition accuracy generally improved with an increase in the number of alternative pronunciation per word type. However, the response time of the speech recognisers increased as did the number of alternative pronunciations.

6.2 Summary of contributions

This dissertation provides a set of experiments which demonstrate the suitability and limitations of using cross-language phoneme mapping for the development of small-vocabulary speech recognition applications for low-resource languages, particularly using the SALAAM algorithm with Bantu languages. Through our findings, we provided a guide that other researchers and practitioners could adopt if they decided to use the SALAAM technique in their work. Our secondary contributions include the modification of the Lex4All tool to support German and Mandarin source languages and a way to generate lexicons across all four source languages at once instead of one source-target language pair at a time. We also intend to release the lexicon files generated during our study as part of the Lex4All open source project.

6.3 Limitations of the study

In spite of the work presented here being grounded on original research, there are significant limitations of the study that need to be highlighted. One of the most significant limitations is the lack of deployment of any of the developed speech recognisers in a production environment to measure their performance in the different settings we focused on in this study. Another important limitation in the study is the lack of actual phonetic comparison between the source and target languages. This was due to the fact that the phonetic alphabets for most of the target languages was not readily available, further underscoring their under-resourced nature, hence no formal phonetic overlap between the source and target languages was established. Had these resources been available, the comparison would have provided more insight into the extent to which the phonetic overlap of source and target languages actually have on recognition accuracy. The dependency of the SALAAM algorithm on Microsoft Speech Platform also limited the study. A comparison of how the algorithm fairs across different speech recognition engines would have been important to establish whether the underlying speech engine has an impact on recognition accuracy. Since Microsoft Speech Platform is a commercial product, access to phonetic alphabets for all the supported languages proved to be a challenge. This therefore limited the number of source languages we could use in our study.

6.4 Future work

Presented in the following subsections are the potential areas one could find interesting to pursue further:

6.4.1 Deployment in a real-world rural Bantu speaking region

The study we conducted was in a controlled environment. It would be beneficial to evaluate the performance of the resulting speech recognisers in a real-world Bantu-speaking region.

6.4.2 Use SALAAM with open-source speech recognition engines

In future, one could consider adapting the SALAAM technique to open-source speech recognition systems such as Kaldi [88] and CMU sphinx [110] to see how its performance compares. The other advantage that would come with the porting of the technique to open source speech recognition systems would be the availability of phonetic alphabets for all the languages supported by these systems which proved to be a challenge to acquire in our study. Availability

of source-target language phonetic alphabets would also allow for a more detailed investigation into the extent to which phonetic similarity between source and target languages impact recognition accuracy. Additionally, porting the algorithm to other platforms would allow one to leverage the speech recognition technology improvements that have been realised since the release of the Microsoft Speech Platform SDK 11 on December 30, 2011 [111].

Bibliography

[1] Tim Bray, Jean Paoli, C Michael Sperberg-McQueen, Eve Maler, and François Yergeau. Extensible markup language (xml). *World Wide Web Journal*, 2(4):27–66, 1997.

[2] Eric Brewer, Michael Demmer, Melissa Ho, RJ Honicky, Joyojeet Pal, Madelaine Plauche, and Sonesh Surana. The challenges of technology research for developing regions. *IEEE Pervasive Computing*, 5(2):15–23, 2006.

[3] Nithya Sambasivan, Ed Cutrell, Kentaro Toyama, and Bonnie Nardi. Intermediated technology use in developing communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2583–2592. ACM, 2010.

[4] Indrani Medhi, Aman Sagar, and Kentaro Toyama. Text-free user interfaces for illiterate and semi-literate users. In *Information and Communication Technologies and Development, 2006. ICTD'06. International Conference on*, pages 72–82. IEEE, 2006.

[5] Jahanzeb Sherwani, Nosheen Ali, Sarwat Mirza, Anjum Fatma, Yousuf Memon, Mehtab Karim, Rahul Tongia, and Roni Rosenfeld. Healthline: Speech-based access to health information by low-literate users. In *Information and Communication Technologies and Development, 2007. ICTD 2007. International Conference on*, pages 1–9. IEEE, 2007.

[6] Somani Patnaik, Emma Brunskill, and William Thies. Evaluating the accuracy of data collection on mobile phones: A study of forms, sms, and voice. In *Information and Communication Technologies and Development (ICTD), 2009 International Conference on*, pages 74–84. IEEE, 2009.

[7] Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S Parikh. Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 733–742. ACM, 2010.

[8] Agha Ali Raza, Rajat Kulshreshtha, Spandana Gella, Sean Blagsvedt, Maya Chandrasekaran, Bhiksha Raj, and Roni Rosenfeld. Viral spread via entertainment and voice-

messaging among telephone users in india. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*, page 1. ACM, 2016.

[9] Jahanzeb Sherwani, Sooraj Palijo, Sarwat Mirza, Tanveer Ahmed, Nosheen Ali, and Roni Rosenfeld. Speech vs. touch-tone: Telephony interfaces for information access by low literate users. *ICTD*, 9:447–457, 2009.

[10] Simplice A Asongu and Jacinta C Nwachukwu. The mobile phone in the diffusion of knowledge for institutional quality in sub-saharan africa. *World Development*, 86:133–147, 2016.

[11] Andrew Dabalen, Alvin Etang, Johannes Hoogeveen, Elvis Mushi, Youdi Schipper, and Johannes von Engelhardt. Mobile phone panel surveys in developing countries. 2016.

[12] Laura Hosman and Elizabeth Fife. The use of mobile phones for development in africa: Top-down-meets-bottom-up partnering. *The Journal of Community Informatics*, 8(3), 2012.

[13] Satinder P Singh, Michael J Kearns, Diane J Litman, and Marilyn A Walker. Reinforcement learning for spoken dialogue systems. In *Advances in Neural Information Processing Systems*, pages 956–962, 2000.

[14] Kevin Croke, Andrew Dabalen, Gabriel Demombynes, Marcelo M Giugale, and JGM Hoogeveen. Collecting high frequency panel data in africa using mobile phone interviews. *World Bank Policy Research Working Paper*, (6097), 2012.

[15] Kalika Bali, Sunayana Sitaram, Sebastien Cuendet, and Indrani Medhi. A hindi speech recognizer for an agricultural video search application. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, page 5. ACM, 2013.

[16] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100, 2014.

[17] Preeti Mudliar, Jonathan Donner, and William Thies. Emergent practices around cgnet swara, voice forum for citizen journalism in rural india. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, pages 159–168. ACM, 2012.

[18] Madeline Plauché and Udhyakumar Nallasamy. Speech interfaces for equitable access to information technology. *Information Technologies & International Development*, 4(1):pp–69, 2007.

[19] Marilyn Walker, J Aberdeen, J Boland, E Bratt, J Garofolo, Lynette Hirschman, A Le, Sungbok Lee, Shrikanth Narayanan, Kishore Papineni, et al. Darpa communicator dialog travel planning systems: The june 2000 data collection. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[20] Marilyn A Walker, Alexander I Rudnicky, Rashmi Prasad, John Aberdeen, Elizabeth Owen Bratt, John S Garofolo, Helen Hastie, Audrey N Le, Bryan Pellom, Alex Potamianos, et al. Darpa communicator: Cross-system results for the 2001 evaluation. In *Seventh International Conference on Spoken Language Processing*, 2002.

[21] Gretta Fitzgerald and Mike FitzGibbon. A comparative analysis of traditional and digital data collection methods in social research in ldcs-case studies exploring implications for participation, empowerment, and (mis) understandings. *IFAC Proceedings Volumes*, 47(3):11437–11443, 2014.

[22] Aparna Moitra, Vishnupriya Das, Gram Vaani, Archna Kumar, and Aaditeshwar Seth. Design lessons from creating a mobile-based community media platform in rural india. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*, page 14. ACM, 2016.

[23] Martha Larson, Nitendra Rajput, Abhigyan Singh, and Saurabh Srivastava. I want to be sachin tendulkar!: a spoken english cricket game for rural students. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1353–1364. ACM, 2013.

[24] Frederick Webera Kalika Balib Roni Rosenfeldc and Kentaro Toyamab. Unexplored directions in spoken language technology for development.

[25] Fang Qiao, Roni Rosenfeld, and Jahanzeb Sherwani. Layperson-trained speech recognition for resource scarce languages. 2010.

[26] Anjana Vakil and Alexis Palmer. Crosslanguage mapping for small-vocabulary asr in under-resourced languages: Investigating the impact of source language choice. In *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[27] Alex Kasonde and Margaret Dunham. A pedagogical study of tone neutralization in cibemba phonetics and phonology. *African Research Review*, 3(1), 2009.

[28] Alvin M Liberman, Katherine Safford Harris, Howard S Hoffman, and Belver C Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5):358, 1957.

[29] Hao Yee Chan and Roni Rosenfeld. Discriminative pronunciation learning for speech recognition for resource scarce languages. In *Proceedings of the 2nd ACM Symposium on Computing for Development*, page 12. ACM, 2012.

[30] Pronunciation lexicon reference (microsoft.speech). https://msdn.microsoft.com/en-us/library/hh378339(v=office.14).aspx. (Accessed on 06/23/2018).

[31] Paul Bagshaw, Daniel C Burnett, Jerry Carter, and J Scahill. Pronunciation lexicon specification (pls) version 1.0. *Recuperado de https://www. w3. org/TR/pronunciation-lexicon/[Links]*, 2008.

[32] Fang Qiao, Jahanzeb Sherwani, and Roni Rosenfeld. Small-vocabulary speech recognition for resource-scarce languages. In *Proceedings of the First ACM Symposium on Computing for Development*, page 3. ACM, 2010.

[33] Gerard C Raiti. The lost sheep of ict4d literature. *Information Technologies & International Development*, 3(4):pp–1, 2006.

[34] Anjana Vakil, Max Paulus, Alexis Palmer, and Michaela Regneri. lex4all: A language-independent tool for building and evaluating pronunciation lexicons for small-vocabulary speech recognition. In *ACL (System Demonstrations)*, pages 109–114, 2014.

[35] Thomas Niesler, Philippa Louw, and Justus Roux. Phonetic analysis of afrikaans, english, xhosa and zulu using south african speech databases. *Southern African Linguistics and Applied Language Studies*, 23(4):459–474, 2005.

[36] Hans Du Plessis. Afrikaans en die khoe. *Silenced voices: studies on minority languages of South Africa. Kaapstad: CASAS*, pages 129–139, 2003.

[37] Corene de Wet. Factors influencing the choice of english as language of learning and teaching (lolt)—a south african perspective. *South African journal of education*, 22(2):119–124, 2002.

[38] Microsoft speech platform. https://msdn.microsoft.com/en-us/library/office/hh361572(v=office.14).aspx. (Accessed on 03/19/2018).

[39] Jouni Filip Maho. Nugl online: The online version of the new updated guthrie list, a referential classification of the bantu languages. *Online file: http://goto. glocalnet. net/mahopapers/nuglonline. pdf*, 2009.

[40] Calisto Mudzingwa. *Shona morphophonemics: repair strategies in Karanga and Zezuru*. PhD thesis, University of British Columbia, 2010.

[41] Guy De Pauw, Gilles-Maurice De Schryver, and Peter W Wagacha. Data-driven part-of-speech tagging of kiswahili. In *International Conference on Text, Speech and Dialogue*, pages 197–204. Springer, 2006.

[42] Johanna Brinkel, Alexander Krämer, Ralf Krumkamp, Jürgen May, and Julius Fobil. Mobile phone-based mhealth approaches for public health surveillance in sub-saharan africa: a systematic review. *International journal of environmental research and public health*, 11(11):11559–11582, 2014.

[43] Amy Wesolowski, Caroline O Buckee, Linus Bengtsson, Erik Wetter, Xin Lu, and Andrew J Tatem. Commentary: containing the ebola outbreak-the potential and challenge of mobile network data. *PLoS currents*, 6, 2014.

[44] Adam Lerer, Molly Ward, and Saman Amarasinghe. Evaluation of ivr data collection uis for untrained rural users. In *Proceedings of the first ACM symposium on computing for development*, page 2. ACM, 2010.

[45] Katie Shilton. Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection. *Communications of the ACM*, 52(11):48–53, 2009.

[46] Melissa R Ho, Emmanuel K Owusu, and Paul M Aoki. Claim mobile: engaging conflicting stakeholder requirements in healthcare in uganda. In *Information and Communication Technologies and Development (ICTD), 2009 International Conference on*, pages 35–45. IEEE, 2009.

[47] Carl Hartung, Adam Lerer, Yaw Anokwa, Clint Tseng, Waylon Brunette, and Gaetano Borriello. Open data kit: tools to build information services for developing regions. In *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development*, page 18. ACM, 2010.

[48] Michael H Cohen, Michael Harris Cohen, James P Giangola, and Jennifer Balogh. *Voice user interface design*. Addison-Wesley Professional, 2004.

[49] Jessica Gladstone and Robert Boruch. Information communication technologies and research in developing countries. 2017.

[50] Magpi - advanced mobile data collection, messaging, and visualization. https://home.magpi.com/. (Accessed on 11/14/2018).

[51] T Svoronos, P Mjungu, R Dhadialla, R Luk, C Zue, J Jackson, and N Lesh. Commcare: Automated quality improvement to strengthen community-based health. *Weston: D-Tree International*, 2010.

[52] Nino Paichadze, Amber Mehmood, Andres Vecino Ortiz, Abdulgafoor M Bachani, and Adnan A Hyder. Pw 1797 digitizing data collection for roadside observational studies: the process and experience, 2018.

[53] Medic Mobile. Medic mobile. *Retrieved'January*, 28:2015, 2015.

[54] Measuring the pulse of africa one phone call at a time — africa-can end poverty. https://blogs.worldbank.org/africacan/measuring-the-pulse-of-africa-one-phone-call-at-a-time. (Accessed on 02/09/2019).

[55] Henry Lucas, Simon Batchelor, and Evangelia Berdou. Real time monitoring and the new information technologies1. *IDS Bulletin*, 44(2):31–39, 2013.

[56] Mark Tomlinson, Wesley Solomon, Yages Singh, Tanya Doherty, Mickey Chopra, Petrida Ijumba, Alexander C Tsai, and Debra Jackson. The use of mobile phones as a data collection tool: a report from a household survey in south africa. *BMC medical informatics and decision making*, 9(1):51, 2009.

[57] Elodia Cole, Etta D Pisano, Gregory J Clary, Donglin Zeng, Marcia Koomen, Cherie M Kuzmiak, Bo Kyoung Seo, Yeonhee Lee, and Dag Pavic. A comparative study of mobile electronic data entry systems for clinical trials data collection. *International journal of medical informatics*, 75(10):722–729, 2006.

[58] Ko Ko Lwin and Yuji Murayama. Web-based gis system for real-time field data collection using personal mobile phone. *Journal of Geographic Information System*, 3(04):382, 2011.

[59] AP Pakhare, S Bali, and G Kalra. Use of mobile phones as research instrument for data collection. *Indian Journal of Community Health*, 25(2):95–98, 2013.

[60] Jilian A Sacks, Elizabeth Zehe, Cindil Redick, Alhoussaine Bah, Kai Cowger, Mamady Camara, Aboubacar Diallo, Abdel Nasser Iro Gigo, Ranu S Dhillon, and Anne Liu. Introduction of mobile health tools to support ebola surveillance and contact tracing in guinea. *Global Health: Science and Practice*, 3(4):646–659, 2015.

[61] Angela Crandall. Kenyan farmers' use of cell phones: Calling preferred over sms. *Proceedings of M4D 2012 28-29 February 2012 New Delhi, India*, 28(29):119, 2012.

[62] Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, and Tapan S Parikh. A comparative study of speech and dialed input voice interfaces in rural india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 51–54. ACM, 2009.

[63] Cristina Delogu, Andrea Di Carlo, Paolo Rotundi, and Danilo Sartori. A comparison between dtmf and asr ivr services through objective and subjective evaluation. In *Interactive Voice Technology for Telecommunications Applications, 1998. IVTTA'98. Proceedings. 1998 IEEE 4th Workshop*, pages 145–150. IEEE, 1998.

[64] Kwan Min Lee and Jennifer Lai. Speech versus touch: A comparative study of the use of speech and dtmf keypad for navigation. *International Journal of Human-Computer Interaction*, 19(3):343–360, 2005.

[65] Madelaine Plauche, Udhyakumar Nallasamy, Joyojeet Pal, Chuck Wooters, and Divya Ramachandran. Speech recognition for illiterate access to information and technology. In *Information and Communication Technologies and Development, 2006. ICTD'06. International Conference on*, pages 83–92. IEEE, 2006.

[66] Indrani Medhi, SN Gautama, and Kentaro Toyama. A comparison of mobile money-transfer uis for non-literate and semi-literate users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1741–1750. ACM, 2009.

[67] Jacob AC Badenhorst and M Davel. Data requirements for speaker independent acoustic models. 2008.

[68] Etienne Barnard, Marelie Davel, and Charl Van Heerden. Asr corpus design for resource-scarce languages. ISCA, 2009.

[69] Tanja Schultz, Martin Westphal, and Alex Waibel. The globalphone project: Multilingual lvcsr with janus-3. In *Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop*, pages 20–27. Citeseer, 1997.

[70] Tanja Schultz and Alex Waibel. Language independent and language adaptive large vocabulary speech recognition. In *Fifth International Conference on Spoken Language Processing*, 1998.

[71] Tanja Schultz and Alex Waibel. Fast bootstrapping of lvcsr systems with multilingual phoneme sets. In *Fifth European Conference on Speech Communication and Technology*, 1997.

[72] Tanja Schultz and Alex Waibel. Adaptation of pronunciation dictionaries for recognition of unseen languages. In *Proc. SPIIRAS International Workshop on Speech and Computer, St. Petersburg*, pages 207–210. Citeseer, 1998.

[73] Andrei Constantinescu and Gerard Chollet. On cross-language experiments and data-driven units for alisp (automatic language independent speech processing). In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 606–613. IEEE, 1997.

[74] Gérard Chollet, Jan Černockỳ, Andrei Constantinescu, Sabine Deligne, and Frederic Bimbot. Toward alisp: A proposal for automatic language independent speech processing. In *Computational Models of Speech Pattern Processing*, pages 375–388. Springer, 1999.

[75] Nishanth Ulhas Nair and TV Sreenivas. Joint decoding of multiple speech patterns for robust speech recognition. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 93–98. IEEE, 2007.

[76] Dhananjay Bansal, Nishanth Nair, Rita Singh, and Bhiksha Raj. A joint decoding algorithm for multiple-example-based addition of words to a pronunciation lexicon. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4293–4296. IEEE, 2009.

[77] Edmondo Trentin and Marco Gori. A survey of hybrid ann/hmm models for automatic speech recognition. *Neurocomputing*, 37(1-4):91–126, 2001.

[78] Xuedong Huang, Fileno Alleva, Hsiao-Wuen Hon, Mei-Yuh Hwang, Kai-Fu Lee, and Ronald Rosenfeld. The sphinx-ii speech recognition system: an overview. *Computer Speech & Language*, 7(2):137–148, 1993.

[79] Jonathan G Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 347–354. IEEE, 1997.

[80] Rita Singh, Bhiksha Raj, and Richard M Stern. Automatic generation of subword units for speech recognition systems. *IEEE Transactions on Speech and Audio Processing*, 10(2):89–99, 2002.

[81] Torbjørn Svendsen. Pronunciation modeling for speech technology. In *Signal Processing and Communications, 2004. SPCOM'04. 2004 International Conference on*, pages 11–16. IEEE, 2004.

[82] Jahanzeb Sherwani, Rahul Tongia, Roni Rosenfeld, Nosheen Ali, Yousuf Memon, Mehtab Karim, and Gregory Pappas. Healthline: Towards speech-based access to health information by semi-literate users. 2004.

[83] Jahanzeb Sherwani. *Speech interfaces for information access by low literate users*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, May 2009. Available as Technical Report CMU-CS-09-131.

[84] Charl Van Heerden, Etienne Barnard, and Marelie Davel. Basic speech recognition for spoken dialogues. 2009.

[85] S Young, P Woodland, G Evermann, and M Gales. The htk toolkit 3.4. 1. *Cambridge Univ. Eng. Dept. CUED*, 2013.

[86] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book. *Cambridge university engineering department*, 3:175, 2002.

[87] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. Sphinx-4: A flexible open source framework for speech recognition. 2004.

[88] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[89] Thomas Kisler, Florian Schiel, and Han Sloetjes. Signal processing via web services: the use case webmaus. In *Digital Humanities Conference 2012*, 2012.

[90] Christian Gaida, Patrick Lange, Rico Petrick, Patrick Proba, Ahmed Malatawy, and David Suendermann-Oeft. Comparing open-source speech recognition toolkits. *Tech. Rep., DHBW Stuttgart*, 2014.

[91] lex4all/lex4all: pronunciation lexicons for any low-resource language. https://github.com/lex4all/lex4all. (Accessed on 11/23/2018).

[92] Audacity ® — free, open source, cross-platform audio software for multi-track recording and editing. https://www.audacityteam.org/. (Accessed on 11/19/2018).

[93] Anne-Marie Beukes and Marne Pienaar. Identities in extended afrikaans speech communities. 2014.

[94] South africa - languages — ethnologue. https://www.ethnologue.com/country/ZA/languages. (Accessed on 05/28/2018).

[95] Assibi Apatewon Amidu. Kiswahili: People, language, literature and lingua franca. *Nordic Journal of African Studies*, 4(1):104–123, 1995.

[96] Zimbabwe - tribes and people. https://www.geni.com/projects/Zimbabwe-Tribes-and-People/23140. (Accessed on 10/02/2018).

[97] Sabine Zerbian and Manfred Krifka. Quantification across bantu languages. *Quantification: A cross-linguistic perspective*, 64:383–414, 2008.

[98] Microsoft word - census in brief ros updates 28 oct 2012.doc. http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf. (Accessed on 10/02/2018).

[99] South africa's people — south african government. https://www.gov.za/about-sa/south-africas-people. (Accessed on 01/21/2019).

[100] Patrick Biernacki and Dan Waldorf. Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research*, 10(2):141–163, 1981.

[101] Moto e3 smartphone — lenovo uk. https://www.lenovo.com/gb/en/smart-devices/smartphones-and-watches/moto/smartphones/Moto-E3/p/PMIPMIF11MW. (Accessed on 02/10/2019).

[102] Voice recorder – apps on google play. https://play.google.com/store/apps/details?id=com.media.bestrecorder.audiorecorder&hl=en_ZA. (Accessed on 02/10/2019).

[103] WH Abdulla, NK Kasabov, and Dunedin-New Zealand. Improving speech recognition performance through gender separation. *changes*, 9:10, 2001.

[104] R: The r project for statistical computing. https://www.r-project.org/. (Accessed on 03/19/2018).

[105] seaborn: statistical data visualization — seaborn 0.8.1 documentation. https://seaborn.pydata.org/. (Accessed on 03/19/2018).

[106] Welcome to python.org. https://www.python.org/. (Accessed on 03/19/2018).

[107] Peter Q Pfordresher and Steven Brown. Enhanced production and perception of musical pitch in tone language speakers. *Attention, perception, & psychophysics*, 71(6):1385–1398, 2009.

[108] Jackson L Lee and Stephen Matthews. When french becomes tonal: Prosodic transfer from l1 cantonese and l2 english. In *PRONUNCIATION IN SECOND LANGUAGE LEARNING AND TEACHING CONFERENCE (ISSN 2380-9566)*, page 63, 2014.

[109] Yuh-Shiow Lee, Douglas A Vakoch, and Lee H Wurm. Tone perception in cantonese and mandarin: A cross-linguistic comparison. *Journal of Psycholinguistic Research*, 25(5):527–542, 1996.

[110] Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. The cmu sphinx-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, volume 1, pages 2–5, 2003.

[111] Download microsoft speech platform - runtime (version 11) from official microsoft download center. https://www.microsoft.com/en-us/download/details.aspx?id=27225. (Accessed on 02/08/2019).

Appendix A

Ethics clearance

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Faculty of Science
University of Cape Town
RONDEBOSCH 7701 South Africa
E-mail: timm.hoffman@uct.ac.za
Telephone: + 27 21 650 5551

17 January 2017

Mr Nick Kayokwa Chibuye
Department of Computer Science

Using question-specific vocabularies with SALAAM to support speech Data Collection

Dear Mr Nick Kayokwa Chibuye

I am pleased to inform you that the Faculty of Science Research Ethics Committee has approved the above-named application for research ethics clearance, subject to the conditions listed below.

- Implement the measures described in your application to ensure that the process of your research is ethically sound; and
- Uphold ethical principles throughout all stages of the research, responding appropriately to unanticipated issues: please contact me if you need advice on ethical issues that arise.

Your approval code is: **FSREC 077** – 2016

I wish you success in your research.

Yours sincerely

Signature removed to avoid exposure online

Prof Timm Hoffman
Chair: Faculty of Science Research Ethics Committee

Cc: Dr. Brian DeRenzi (Supervisor)

Appendix B

Target language vocabularies

seSotho

1 Bohobe

2 Dijo

3 Ditholwana

4 Makotomane

5 Diphoofolo

6 Nakong ea dipula

7 Mokopu

8 E

9 Manyolo

10 Kwae

11 Avocado

12 Dilamuni

13 Metsi

14 Temo

15 Nyalothe

16 Mabele

17 Ditamati

18 Ditapole

19 Letswai

20 Seshebo

21 Tlhapi

22 Boemo ba Lehodimo

23 Dinawa

24 Dierekisi

25 Lebese

26 Kgoho

27 Nama

28 Podi

29 Kolobe

30 Mahe

31 Mango

32 Peo

33 Lema

34 Tshimo

35 Mabopo

36 Pineapple

37 Dinotsi

38 Makgea

39 Ho lema

40 Re lemme

41 Ho kotula

42 Ho lema

43 Likgomo

44 Ho hlaola

45 Pula

46 Noka

47 Meroho

48 Rice

49 Letata

50 Papa

51 Dianyanese

52 Di-mushroom

53 Poone

54 Kh’abeche

55 Che

56 Ole ea lijo

57 Madi

58 Batswali

59 Moputso

60 Meriana

61 Thutu

62 Mme

63 Sehwai

64 Hoseng

65 Mmaraka

66 Difate

67 ‘Ngoe

68 Peli

69 Tharo

70 Nne

71 Hlano

72 Tshelela

73 Supa

74 Robeli

75 Robong

76 Leshome

77 Sontaha

78 Mantaha

79 Labobeli

80 Laboraro

81 Labone

82 Labohlano

83 Moqebelo

84 Pherekgong

85 Hlakola

86 Hlakubele

87 Mmesa

88 Motsheanong

89 Phuptjane

90 Phupu

91 Phato

92 Loetse

93 Mphalane

94 Pudungwana

95 Tshitwe

96 Ha nngoe

97 Ha beli

98 Ha raro

99 Ha nne

100 Ha hlano

Chishona

1 Chingwa

2 Chikafu

3 Michero

4 Nzungu

5 Zviphuyo

6 Zhizha

7 Nhanga

8 Ehe

9 Muphudze

10 Fodya

11 Kotapeya

12 Maranjisi

13 Mvura

14 Kurima

15 Mapfunde

16 Mbambaira

17 Madomasi

18 Mbatatisi

19 Munyu

20 Muriwo

21 Hove

22 Mamiriro ekunze

23 Bhochisi

24 Mukaka

25 Huku

26 Nyama

27 Mbudzi

28 Nguruve

29 Mazai

30 Gaka

31 Mbeu

32 Geja

33 Munda

34 Muhomba

35 Nanazi

36 Nyuchi

37 Uchi

38 Kudyara

39 Takadyara

40 Kukohwa

41 Kurima

42 Mombe

43 Kusakura

44 Mvura

45 Rwizi

46 Chikoro

47 Muriwo

48 Mupunga

49 Dhadha

50 Sadza

51 Hanyanisi

52 Howa

53 Chibage

54 Mufarinya

55 Kabheji

56 Aiwa

57 Mafuta ekubikisa

58 Ropa

59 Vabereki

60 Mushonga

61 Dzidzo

62 Amai

63 Varimi

64 Makuseni

65 Musika

66 Miti

67 Poshi

68 Piri

69 Tatu

70 China

71 Shanu

72 Tanhatu

73 Nomwe

74 Sere

75 Pfumbamwe

76 Gumi

77 Svondo

78 Muvharo

79 Chipiri

80 Chitatu

81 China

82 Chishanu

83 Mugovera

84 Ndira

85 Kukadzi

86 Kurume

87 Kubvumbi

88 Chivabvu

89 Chikumi

90 Chikunguru

91 Nyamavhuuhu

92 Gunyana

93 Gumiguru

94 Mbudzi

95 Zvita

96 Kamwe

97 Kaviri

98 Katatu

99 Kunokwana Kana

100 Kashanu

Kiswahili

1 Mkate

2 Chakula

3 Matunda

4 Njugu

5 Wanyama

6 Msimu wa mvua

7 Mlenge

8 Ndio

9 Mbolea

10 Kiraiko

11 Parachichi

12 Chungwa

13 Maji

14 Kilimo

15 Wimbi

16 Mtama

17 Nyanya

18 Viazi

19 Chumvi

20 Mboga

21 Samaki

22 Hali ya hewa

23 Maharagwe

24 Maziwa

25 Kuku

26 Nyama

27 Mbuzi

28 Nguruwe

29 Mayai

30 Embe

31 Mbegu

32 Palilia

33 Kiwanja

34 Nanasi

35 Nyuki

36 Uki

37 Kupanda

38 Tulipanda

39 Kuvuna

40 Ukulima

41 Ng’ombe

42 Kupalilia

43 Mvua

44 Mto

45 Shule

46 Mboga

47 Mchele

48 Bata

49 Ugali

50 Vitunguu

51 Uyoga

52 Mahindi

53 Mhogo

54 Mboga

55 Hapana

56 Mafuta ya kupika

57 Damu

58 Wazazi

59 Mshahara

60 Dawa

61 Elimu

62 Mama

63 Mkulima

64 Asubuhi

65 Soko

66 Miti

67 Moja

68 Mbili

69 Tatu

70 Nne

71 Tano

72 Sita

73 Saba

74 Nane

75 Tisa

76 Kumi

77 Jumapili

78 Jumatatu

79 Jumane

80 Jumatano

81 Alhamisi

82 Ijumaa

83 Jumamosi

84 Januari

85 Februari

86 Machi

87 Aprili

88 Mei

89 Juni

90 Julai

91 Agosti

92 Septemba

93 Octoba

94 Novemba

95 Decemba

96 Mara moja

97 Mara mbili

98 Mara tatu

99 Mara nne

100 Mara tano

Afrikaans

1 Brood

2 Kos

3 Vrugte

4 Grondbone

5 Vee

6 Reënseisoen

7 Pampoen

8 Ja

9 Kunsmis

10 Tabak

11 Avokadopeer

12 Lemoene

13 Water

14 Landbou

15 Komkommer

16 Blaarslaai

17 Tamaties

18 Aartappels

19 Sout

20 Smoor

21 Vis

22 Weer

23 Bone

24 Ertjies

25 Melk

26 Hoender

27 Vleis

28 Bok

29 Vark

30 Eiers

31 Waatlemoen

32 Saad

33 Ploeg

34 Veld

35 Rante

36 Pynappel

37 Bye

38 Heuning

39 Om te plant

40 Ons het geplant

41 Om te oes

42 Om te boer

43 Beeste

44 Skoffel

45 Reën

46 Rivier

47 Groente

48 Rys

49 Eend

50 Uie

51 Sampioene

52 Mielies

53 Druiwe

54 Kool

55 Nee

56 Kookolie

57 Bloed

58 Ouers

59 Salaris

60 Medisyne

61 Onderwys

62 Moeder

63 Boere

64 Oggend

65 Mark

66 Bome

67 Een

68 Twee

69 Drie

70 Vier

71 Vyf

72 Ses

73 Sewe

74 Agt

75 Nege

76 Tien

77 Sondag

78 Maandag

79 Dinsdag

80 Woensdag

81 Donderdag

82 Vrydag

83 Saterdag

84 Januarie

85 Februarie

86 Maart

87 April

88 Mei

89 Junie

90 Julie

91 Augustus

92 September

93 Oktober

94 November

95 Desember

96 Een keer

97 Twee keer

98 Drie keer

99 Vier keer

100 Vyf keer

English

1 Bread

2 Food

3 Fruits

4 Groundnuts

5 Livestock

6 Rainy season

7 Pumpkin

8 Yes

9 Fertiliser

10 Tobacco

11 Avocado

12 Oranges

13 Water

14 Agriculture

15 Millet

16 Sorghum

17 Tomatoes

18 Potatoes

19 Salt

20 Relish

21 Fish

22 Weather

23 Beans

24 Peas

25 Milk

26 Chicken

27 Meat

28 Goat

29 Pig

30 Eggs

31 Mango

32 Seed

33 Plough

34 A field

35 Ridges

36 Pineapple

37 Bees

38 Honey

39 To plant

40 We planted

41 To harvest

42 To farm

43 Cattle

44 Weeding

45 Rain

46 A river/river

47 School

48 Vegetables

49 Rice

50 Duck

51 Onions

52 Mushrooms

53 Maize

54 Cassava

55 No

56 Cooking oil

57 Blood

58 Parents

59 Salary

60 Medicine

61 Education

62 Mother

63 Farmers

64 Morning

65 Market

66 Trees

67 One

68 Two

69 Three

70 Four

71 Five

72 Six

73 Seven

74 Eight

75 Nine

76 Ten

77 Sunday

78 Monday

79 Tuesday

80 Wednesday

81 Thursday

82 Friday

83 Saturday

84 January

85 February

86 March

87 April

88 May

89 June

90 July

91 August

92 September

93 October

94 November

95 December

96 Once

97 Twice

98 Three times

99 Four times

100 Five times