UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

CS/IT Honours Project Final Paper 2022

Title: Building a Gold Standard Corpus for Automatic Speech Recognition for isiXhosa

Author: Mosamat Sabiha Shaikh

Project Abbreviation: SABCTXT

Supervisor(s): Dr Hussein Suleman

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	0
Theoretical Analysis	*0*	*25*	0
Experiment Design and Execution	*0*	*20*	15
System Development and Implementation	*0*	*20*	15
Results, Findings and Conclusions	*10*	*20*	20
Aim Formulation and Background Work	*10*	*15*	10
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	
Total marks		**80**	

Building a Gold Standard Corpus for Automatic Speech Recognition for isiXhosa

Addressing the Low-Resource Challenge in South African Languages

Mosamat Sabiha Shaikh
Computer Science
University of Cape Town
shkmos004@myuct.ac.za

ABSTRACT

South Africa is a linguistically diverse country with 11 official languages. Nine of the languages are termed low-resource languages due to there being not enough electronic documents available for them. The exceptions are English and Afrikaans. Natural language processing techniques require electronic documents to extract content for processing to enable computers to understand, interpret and generate spoken languages. The goal of this paper is to assess the feasibility of using crowdsourcing to construct a gold-standard ASR corpus for isiXhosa. It involved collecting audio data, transcribing it with isiXhosa speakers, and applying several statistical methods to evaluate the transcriptions' quality and consistency. The corpus, while not without limitations due to small sample sizes and potential variations, represents a significant step towards addressing the scarcity of electronic linguistic resources in low-resource African languages.

CCS CONCEPTS

• natural language processing • corpus linguistics • low-resource languages • gold standard corpus • data collection • data annotation •

KEYWORDS

corpus, crowdsourcing, low-resource language

1 INTRODUCTION

The preservation and exploration of linguistic diversity are essential components of understanding and respecting cultural heritage. The linguistic diversity of South Africa, with its eleven official languages, presents a unique challenge for natural language processing (NLP). There exists a glaring disparity in the availability of electronic textual resources for all languages except English and Afrikaans [1][2][3]. This limitation hinders the development of language technologies, particularly in the context of low-resource languages. IsiXhosa is a prime example of that category and the language chosen for this project. It is an indigenous language in the Western Cape in South Africa, where the research was conducted, and is also the second most spoken South African language with around 16% of the population speaking it [4].

The South African Centre for Digital Language Resources (SADILAR) has developed speech technology tools that are accessible to the public [6]. The project as a whole aims to investigate various existing techniques to determine if language and speech recognition tools attained from SADILAR can be used to produce accurate and appropriate transcripts of audio in isiXhosa. The transcripts produced form a text corpus which will be compared to a gold standard Automatic Speech Recognition (ASR) corpus to validate it. The success of the project would mean that text corpora can be accurately developed using standard speech tools and models, which is essential for natural language processing as it is a primary source of data.

The focus of this project is the development of the gold standard ASR corpus. A corpus is essentially a collection of written texts or audio recordings of a language that is processed to learn the behaviours of that language [7]. Corpora exist in various forms. A text corpus is a collection of written texts that are used for various natural language processing tasks[27]. On the other hand, an ASR corpus is a collection of speech recordings of spoken language along with their corresponding transcriptions [28]. Gold standard corpora are manually annotated collections of audio used as dependable sources of information regarding languages [8]. They are necessary for the training and meaningful evaluation of algorithms.

This paper describes the process of developing the corpus. It includes participant recruiting, audio data collecting, transcription, and statistical analysis. The findings shed light on transcription consistency, linguistic pattern compliance, and vocabulary diversity in the developed corpus, while also addressing inherent limitations. The conclusion underlines the importance of the research, while the future directions section discusses opportunities for expanding the impact and enriching isiXhosa linguistic resources.

2 BACKGROUND AND RELATED WORK

Corpora are crucial for NLP and are available in a variety of formats such as monolingual and parallel collections [9][10]. A monolingual corpus is made up of text or speech data in only one language whereas parallel corpora are collections of connected texts or speech from two or more languages. Each sentence or text in one language is paired with its comparable translation in another language in a parallel corpus. The corpus in conversation is a monolingual corpus. Benchmarking NLP algorithms is done using Gold Standard Corpora (GSC), which are manually annotated datasets [11]. These datasets serve as credible standards against which the performance and accuracy of NLP algorithms may be systematically examined, giving a solid foundation for assessing their efficacy and developing their capabilities.

There has been various work done in corpus development in South Africa and elsewhere. The NCHLT broadband corpus and the Lwazi corpus, both of which concentrate on multilingual spoken language resources, are examples of South Africa's corpus development initiatives [12][13]. Crowdsourcing is the technique of sourcing needed services, ideas, or content from several contributors, particularly from the internet community, as opposed to from traditional employers or suppliers [14]. Crowdsourcing is an emerging collaborative approach that can be used for the acquisition of annotated corpora and a wide range of other linguistic resources [15]. Academics have investigated cutting-edge techniques that rely on crowdsourcing and gamification to create corpora for languages [7][16]. Data collection also makes use of specialized technologies and platforms like Microsoft's Universal Human Relevance System (UHRS) and Amazon's Mechanical Turk (MTurk) [17][18].

Crowdsourcing emerges as a cost-effective approach to corpus development [15][19][20]. Project definition, data preparation, project execution, and data review and aggregation are all steps in its process. Crowdsourcing has several benefits, including the potential to involve many contributors and insights into human behaviour [7][21]. Planning is necessary, nevertheless, because of issues like quality control and fluctuating contributor motivation [16][20]. Additionally, there are necessary considerations for legal and ethical issues, such as contributor privacy and appropriate acknowledgement [15].

Due to their extensive social networks and technological know-how, Hughes et al. [21] successfully recruited university students for data collection via crowdsourcing. Amazon does not engage South African workers for crowdsourcing assignments and hence the usage of MTurk is not possible despite it being a popular crowdsourcing platform for low-resource languages [19][22][23]. For a natural language experiment to be evaluated properly, the gold standard ASR corpus must be of high quality and accuracy. Various quality assessment techniques and analytic methods such as Zipf's law, token-to-type ratio, Heap's law, and inter-transcriber agreement are used to assess the quality of corpora [24][25]. Packham and Suleman note that monetary incentives played a significant role in motivating crowdsourcing participation [16]. In their study, they highlight how the inclusion of monetary incentives is a powerful motivator that encourages people to actively participate in crowdsourcing in South Africa.

3 PROBLEM STATEMENT

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

4 RESEARCH QUESTION

Can crowdsourcing via an online web application effectively produce a high-quality gold standard ASR corpus for isiXhosa, a low-resource South African language, from audio data collected from casual conversations and broadcast news? A gold standard ASR corpus consists of manually transcribed speech data of high quality. Crowdsourcing is the process of obtaining a service from a group of people, typically via the Internet. In this case, the service is the transcription of audio from fluent isiXhosa speakers. The audio segments are a combination of casual conversations and broadcast news.

5 METHODS

This section describes several critical components that work together to ensure a rigorous and complete approach to producing a gold-standard ASR corpus for low-resource languages. This section describes the participant recruitment process, audio data gathering means, audio transcription process, and statistical analysis tools used to assess quality.

5.1 Participants Recruitment

The target audience for the project was isiXhosa-speaking students from the University of Cape Town (UCT). This is because students at universities have extensive social networks and technological knowledge. First and foremost, we acquired ethics clearance from the UCT ethics committee (clearance code - **FSREC 051-2023**) to maintain the highest ethical standards. The project entailed minimal risks, and no identifying or sensitive data was required for the experiment itself.

Participants were recruited through the distribution of mass emails, which were thoughtfully sent out by course convenors. These emails were sent to first-year Computer Science students as well as students enrolled in courses offered by the Department of African Languages. Additionally, digital and physical posters were created and distributed within the university eco-system. This approach ensured that a diverse and large group of potential participants was reached. The participants had to meet the strict criteria of being fluent in isiXhosa to ensure the language corpus's quality and trustworthiness while increasing its utility for diverse natural language processing activities and linguistic studies. A compensation amount of R50 was offered for each task completed. A task is either taking part in the audio collection process or completing an audio transcription. Participants could do either or both tasks. Two informed consent forms were drafted for each of the tasks requiring participants. See Appendix 1 for the email, poster, and consent forms.

5.2 Audio Data Collection

The audio data collection process involved obtaining two distinct sets of audio segments. One set originated from news segments, characterized by their formal language and absence of background noise, and was referred to as "structured audio." The second set was gathered from the recruited participants, characterized by casual conversation and ambient background noise, and was termed "unstructured audio". The casual conversations contain slang, code-switching, and dialect differences. These two sets of audio types were deliberately chosen to capture diverse aspects of the language across varying contexts.

5.2.1 Structured Audio

The South African Broadcasting Cooperation (SABC), the government-run national broadcaster, is required to have programmes in all the South African languages [5]. The SABC is required to "promote the development and use of all official languages" and "ensure that all South Africans have access to information and programming in their language". This makes it content-heavy and thus where the structured audio was sourced from.

The use of SABC news segments within the research project was carried out in strict accordance with the parameters set forth by copyright laws. Rule 4 of SABC's terms and conditions stipulate that "The User and/or Subscriber is only permitted to use the service for personal and non-commercial use. The User and /or Subscriber is duly granted a non-transferable and non-exclusive right to access our services and content"[5]. The research study followed the ethical and legal incorporation of these resources by aligning the use of SABC's videos with the restrictions of copyright regulations, further contributing to the project's integrity and credibility. The SABC lindaba channel on YouTube hosts lindaba zesiXhosa daily news segments. Videos were downloaded from the channel using online third-party video download tools. The default Windows video editor tool was used to trim and split them into 10-minute segments.

5.2.2 Unstructured Audio

The unstructured audio data was collected from a portion of the recruited participants. The participants were divided into groups of 2 or 3 based on when they would be available. The participants were instructed to assemble in front of the Sarah Baartman Hall in UCT. The location was chosen because of its convenient location and its accessibility and recognizability for students. The environment is uncontrolled, introducing various noise factors such as unstable weather and external conversations in the background. This was suitable and desirable for the task at hand.

The collection process began with the group being eased into the task by a short introduction and answering any questions or concerns they may have. The procedure of the collection process was explained. The goal was to collect 40 minutes of audio in total from four individual 10-minute topic-specific conversations. The topics chosen had general themes but were broad, making it possible for individuals who did not know each other before this meeting to engage in lengthy and in-depth conversations. The topics were sports, food, education, and entertainment. There were bullet points for each topic to help facilitate the conversation and ensure that it went on for the desired duration. A sample of this is shown in Appendix 2.

To include diversity in the corpus, one group was given a range of topics and instructed to converse for the entire 40-minute duration. There are many benefits to diversifying a conversational corpus. It mimics actual conversations by recording a range of subjects, leading to more natural language use and adaptive models. By lowering biases and boosting the resilience of language models, diverse datasets improve system generalization. They result in more realistic and natural interactions that are advantageous for both research and application areas needing extensive context handling.

5.3 Audio Transcription

The audio segments were transcribed on an open-source online crowdsourcing web application called Turkle[26]. Turkle was deployed and hosted on the Internet using the Amazon Web Service (AWS) EC2 cloud service. AWS was also used to host the audio files. A pilot test was performed to ensure that the web application worked as expected before it was deployed to real users.

5.3.1 Turkle Web Application

Turkle is a Django-based web application clone of MTurk[1]. MTurk is a crowdsourcing marketplace that enables individuals and businesses to outsource small tasks or "Human Intelligence Tasks" (HITs) to a distributed workforce. Using MTurk in a South African context is not feasible due to cost and payment issues thus making Turkle a suitable alternative. Turkle offers a way for people to manage their crowdsourcing platform using features that are similar to those provided by MTurk.

It is a flexible platform with many essential features intended to improve its usability and effectiveness. It allows multiple workers to be assigned to each task in a project. This allows for one audio to be transcribed by more than one person. As a result, transcription accuracy can be assessed. Turkle includes an intuitive admin GUI that makes managing users, groups, projects, and batches of tasks simple to streamline administrative operations.

Administrators are equipped to manage many areas of the platform with ease due to this user-friendly interface. An HTML file is uploaded to form the interface of the transcription task, which is complemented by a CSV file with links to the audio files. The audio files can be on any online hosting platform and listed in the CSV file as downloadable links. The annotations can be downloaded as CSV files. Figures 1 and 2 illustrate the administration interface for project and batch management respectively.

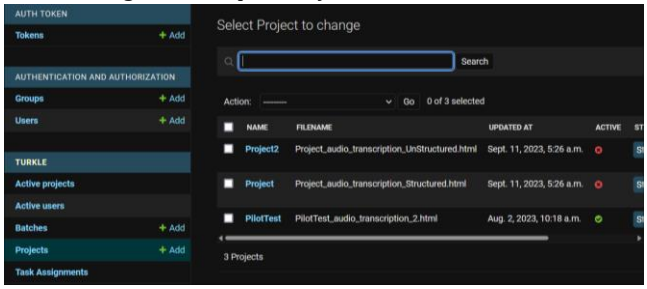

Figure 1: Administrator Project Management Interface

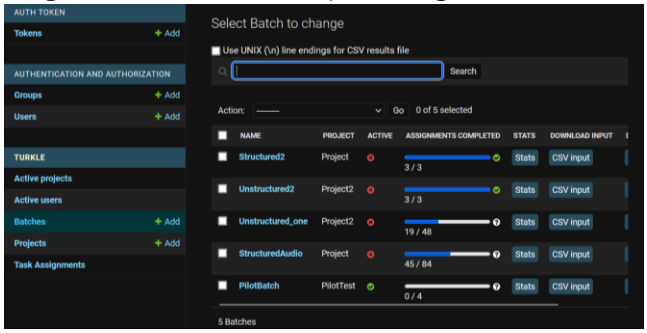

Figure 2: Administrator Batch Management Interface

[1] https://www.mturk.com/

5.3.2 Amazon Web Service

The web application was hosted on Amazon EC2 and ran with *Gunicorn*[2], *Nginx*[3] and *Supervisord*[4]. The web application was configured to provide users with a dependable and seamless experience. It was set up on an Amazon EC2 instance to accomplish this, utilizing the capacity of the cloud to manage multiple workloads. The Python Web Server Gateway Interface (WSGI) HTTP server *Gunicorn*, which can handle multiple concurrent requests well, was added to the application's runtime environment.

Nginx, a web server and reverse proxy, was used to expand capabilities by effectively managing incoming client requests. Additionally, *Supervisord*, a process control system, was used to closely monitor and oversee the program's operations. *Supervisord* made sure the application kept functioning dependably and automatically restarted any crashed processes.

Additionally, the AWS S3 buckets were used as the designated hosting platform for the audio files collected. A public bucket was created for both the structured and unstructured audio segments. Using these AWS buckets, a safe and scalable infrastructure for storing and accessing the gathered audio files was constructed. This decision enabled effective management and smooth distribution of the audio data.

5.3.3 Pilot Test

A pilot test was performed to test the stability of the web application and evaluate its performance, functionality, and usability before deploying it to real users. Four participants were chosen for the limited-scale trial of the web application. They were registered with usernames and passwords and a test project was created with a batch of two audio files that each had to be transcribed four times.

The participants were given their user information and instructed that their task was to transcribe the given audio segments. They were then given time to explore the application and complete the task set for them. The participants used either the Mac, Windows, or Linux operating system to allow for a comprehensive testing of the application. Each participant was informally interviewed afterwards for feedback. The overall result was positive with the application described as working "smoothly" and "as expected". The interface was deemed "straightforward" and "intuitive". Some suggestions were given for the user interface, but usability and performance were positive.

[2] https://gunicorn.org/
[3] https://www.nginx.com/
[4] http://supervisord.org/running.html#running-supervisord

5.3.4 Transcription Task

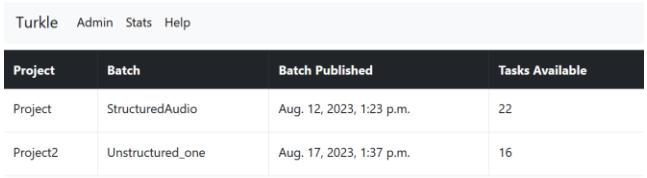

Figure 3: Turkle Interface Displaying Tasks

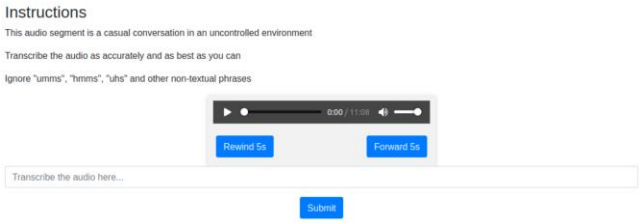

Figure 4: Turkle Interface for Transcription

Publishing a project on Turkle required the uploading of an HTML file. The default audio transcription file was used as a base and built on to make it more intuitive and specific to this project. Suggestions were taken from the pilot test to improve the user interface. Two projects were created, each highlighting the type of audio being transcribed. Batches are CSV files containing links to the audio files that are being hosted on the AWS S3 bucket. Separate CSV files containing links to the structured and unstructured audio files were uploaded to the site. Figures 3 and 4 show what the audio transcription system looks like.

Individual usernames and passwords were generated for the participants. Their username was set to be their UCT student number, and the password was obtained from an online password-generating website. An email was sent out to each participant detailing their authentication information and instructions on how to access the web application and how to complete the transcriptions. Each project was set so that each audio segment had to be transcribed by three individual participants.

5.4 Statistical Analysis

Various techniques were used in the statistical analysis phase to assess the validity and reliability of the audio transcriptions that eventually formed the corpus. To assess accuracy, the inter-transcriber similarity between transcriptions of the same audio segment was calculated. Zipf's law involves comparing word frequency distributions and was used to evaluate corpus reliability. This analytical approach helps in discovering abnormalities and data quality issues in language corpora. Token-to-type ratio, which indicates vocabulary diversity, was used to assess the corpus's richness and potential data quality issues.

5.4.1 Pre-Preprocessing Data

Data preprocessing for the corpus construction involved a series of transformative operations that convert raw transcribed text into a structured format appropriate for linguistic analysis. The first step was to remove transcriptions that were translations. Some of the participants misunderstood the assignment and translated the segments to English instead of transcribing them in isiXhosa. A Python file was then used to extract the results from a CSV file and build the corpus after processing the transcriptions. To begin, the transcription text was split into individual words or tokens, laying the foundation for further processing. Converting the words into lowercase then ensures consistency in how words are handled across different cases. The code then excludes English words through the *words.words()* list from *nltk* to maintain a monolingual corpus.

5.4.2 Inter-Transcriber Similarity

The Levenshtein distance, also known as edit distance, is a statistic used to compare two sequences of characters [30], often strings. It counts the number of single-character modifications (insertions, deletions, or replacements) required to transform one string into another. Language corpora frequently include a wide range of language variations and dialects thus the Levenshtein distance can be used to quantify the linguistic similarity or dissimilarity between different dialects or languages of the same text. Most studies of transcription consistency have focused on inter-transcriber agreement [31].

A Python file was developed to calculate the similarity pairs between transcripts of the same audio using Levenshtein distance. The code then compares the inter-transcriber similarity score of all the audio transcriptions as pairs for the same audio segment. For example, if there are 3 transcriptions then 1 and 2, 1 and 3, and 2 and 3 are compared. The average is then calculated to obtain an overall similarity view. This is useful for quality assurance or further language analysis.

5.4.3 Zipf's Law

Zipf's Law is an empirical law that describes a certain form of frequency distribution found in a variety of natural language and non-language events[29]. The formula for Zipf's Law is:

$$f(r) = \frac{C}{r^s}$$

Where:

- $f(r)$ is the frequency of the word/item at rank r.
- C is a constant.
- s is the exponent that characterizes the distribution, often close to 1.

In this formula, as r (the rank) increases, the frequency $f(r)$ decreases, and this decrease is inversely proportionally to r^s. This is the mathematical representation of the rank-frequency relationship described by Zipf's Law. Calculating Zipf's Law involves looking at the frequency distribution of items in a dataset, ranking them by frequency, and visualising the data on a log-log scale. A Python file was created to do the calculations and plot the graph. The word frequencies in a text corpus are examined by this code using libraries like *matplotlib.pyplot* and *Counter*. By using this method, it is possible to determine whether Zipf's Law's basic principles are followed by the word frequency distribution of the corpus.

5.4.5 Token-Type Ratio

The Type-Token is a linguistic statistic used to assess the lexical richness or diversity of a text or corpus[31]. It is a quantitative measurement that contrasts the number of distinctive word types to the overall word count (tokens) in a particular text. It uses a simple formula:

$$Type - Token\ Ratio = \frac{Number\ of\ Types}{Number\ of\ Tokens}$$

A high type-token ratio denotes high lexical variety, whereas a low type-token ratio denotes the reverse. The range consists of a theoretical value of 0 (unlimited repetition of a single type) and a value of 1 (the total absence of repetition found in concordance). The type-token ratio is frequently used in linguistic analysis to compare the lexical diversity or richness of various texts or to monitor vocabulary evolution through time[31]. It can shed light on the richness and diversity of linguistic usage.

Some research uses a "token-type" ratio rather than the more commonly used "type-token" ratio. The number of tokens is divided by the number of kinds in these experiments. The results are reported in a range, with a token-type ratio of 1 indicating the greatest degree of variance conceivable and higher ratios indicating lower degrees of variation. [25] [32]. The Token-To-Type Ratio(TTR henceforth) was used for the simple purpose of ease of comparison. A simple Python file was written for calculation.

6 RESULTS AND DISCUSSION

We outline the findings of our research in this section, such as details about the information gathered, how it was analyzed, and the subsequent deductions drawn as a result. The effects of inter-transcriber similarity, adherence to Zipf's Law, and token-to-type ratio are also examined. Each of these phenomena provides important insights into the linguistic patterns and makeup of the isiXhosa language corpus.

6.1 Corpus Statistics

A total of 21 participants, consisting of both male and female students, volunteered to take part in the project. Out of these participants, 13 volunteered to take part in the audio collection portion and 17 volunteered to take part in the transcription portion. In total, 10 participants were active contributors to the transcription portion. Two of the participants did not respond to follow-up emails.

A total of 28 10-minute audio segments were sourced from the SABC lindaba channel resulting in 84 segments to transcribe (each of the 28 should be transcribed by three independent people). For unstructured audio, the goal was to collect the same number of audio segments. This would have been achieved by having seven groups converse for 40 minutes each. However, due to insufficient participants, only four 40-minute audio segments were collected. This resulted in having 48 10-minute unstructured audio segments to transcribe.

Out of the 84 transcribable structured audio segments, 48 were completed. From those, some were translated to English due to the participant misinterpreting the task. Consequently, the statistical analysis was conducted on a slightly smaller batch of 45 transcriptions. A corpus of size 26794 words was created for structured audio. Out of the 48 transcribable unstructured audio segments, 22 were completed. The issue of translating to English instead of transcribing posed an issue here as well, resulting in 18 valid transcriptions.

The reason for the very low level of transcriptions for the unstructured audio segments was due to the unstructured audio segments that had been uploaded later in the project because they had to be collected first. Participants were requested to start on the unstructured audio transcriptions but there was little enthusiasm and effort as compared to the structured audio transcriptions. Consequently, a corpus of size 12764 words was developed. Table 1 shows the statistical summary of the two corpora. See Appendix 3 for a longer list of the most and least appeared words alongside their appearance number.

Table 1: Collection's Summary and Comparisons

Description	Structured Corpus	Unstructured Corpus
Audio Segments	84	45
Transcriptions	45	22
Total Words	26794	12764
Distinct Words	9837	5165
Most Appeared Word	Ukuba	Uba
Least Appeared Word	Minist	Iyandi

6.2 Structured Audio

This section provides the results of inter-transcriber similarity, Zipf's law, and type-token ratio analysis for the structured audio corpus.

6.2.1 Inter-Transcriber Similarity

Table 2: Audio ID and Inter-Transcriber Similarity Comparison for Structured Audio

Audio ID	Inter-Transcriber Similarity / %	No. of Transcriptions	Standard Deviation
1	62.27	3	20.94
2	68.33	2	0.00
3	81.02	2	0.00
4	71.08	2	0.00
6	99.43	2	0.00
7	64.74	3	20.90
8	85.65	3	13.86
9	95.45	3	2.84
10	31.11	3	15.00
11	87.19	3	11.58
13	54.73	2	0.00
14	74.40	2	0.00
15	15.09	2	0.00
17	75.96	2	0.00
21	28.31	2	0.00

The inter-transcriber similarity score in Table 2 shows how well different transcribers agreed on each audio recording. In cases where there were only 2 transcriptions, only one inter-transcriber similarity score could be calculated (similarity score between the two transcriptions). Consequently, the standard deviation for those segments was 0. The standard deviation for the inter-transcriber similarities between the three transcriptions shows little difference between them. Audio segments that had only one transcription were not included as there are no transcriptions to compare it against. See Appendix 4 for the detailed transcription similarity scores.

These results provide insightful information about the accuracy and reliability of the transcriptions. Notably, the scores between various audio recordings vary widely. Audios 6, 9, and 11 exhibit exceptionally high inter-transcriber similarity scores of 99.43%, 95.45%, and 87.19%, respectively. These high ratings indicate a high level of agreement among transcribers, implying that these transcriptions are likely to be reliable and accurate.

Some of the audio segments in the corpus have a more moderate level of consistency. This category includes audios 3, 8, and 17, with inter-transcriber similarity scores ranging from above 75%. While not as consistent as the previously described audios, they nonetheless show a significant degree of agreement among transcribers. As a result, these transcriptions can be regarded as high-quality and dependable for corpus growth.

A subset of audio segments have lower inter-transcriber similarity scores, indicating potential consistency and accuracy problems. The ratings for audios 1, 2, 4, 7, 13, and 14 range from 54.73% to 68.33%, indicating moderate to low agreement between transcribers. These disparities indicate the presence of inconsistencies or flaws in the transcriptions, emphasising the importance of additional examination and refinement to improve the quality of these transcripts.

Finally, the corpus contains audios with extremely low inter-transcriber similarity scores, such as Audios 10, 15, and 21, which have a score of less than 32%. These scores indicate that there are considerable differences across the transcriptions, implying that these specific audios may require close review and extensive revision. The quality of these transcriptions appears to be noticeably inferior to that of others in the corpus. News reports use a formal linguistic style, potentially causing comprehension challenges, even for native isiXhosa speakers. Those particular segments containing less common formal words could explain the extremely low scores achieved.

6.2.2 Zipf's Law

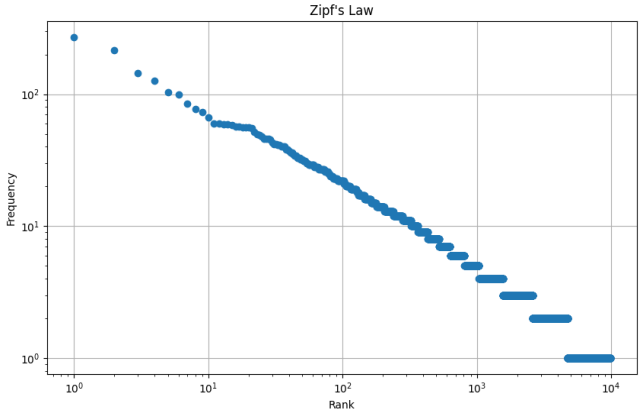

Figure 5: Zipf's Law Distribution: Frequency vs. Rank for Structured Audio

Zipf's law simply states that a word's frequency is inversely related to its rank. This indicates that the most common term has a rank of one, the second most common word has a rank of two, and so on. According to Figure 5, the frequency of the language corpus drops as its rank rises. The slope of the line in the graph is not perfectly -1, indicating that the corpus does not exactly obey Zipf's law. However, this is justified because certain words are more prevalent than expected, while others are less common. The slope being near the -1 gradient implies that the corpus closely follows Zipf's law.

6.2.3 Token-To-Type- Ratio

A TTR of 2.70 was calculated for the corpus which has a size of 26794 words. This implies a balanced yet moderate level of vocabulary diversity. This indicates that a mixture of common and uncommon terms can be found in the corpus. Although the corpus is not overly repetitious, it nevertheless lacks variety. This could be viewed as an expected ratio for news reporting because it follows a balance between using everyday language to assure comprehension and occasionally using more sophisticated vocabulary to raise the level of communication.

6.3 Unstructured Audio

This section provides the results of inter-transcriber similarity, Zipf's law, and type-token ratio analysis for the unstructured audio corpus.

6.3.1 Inter-Transcriber Similarity

Table 3: Audio ID and Inter-Transcriber Similarity Comparison for Unstructured Audio

Audio ID	Inter-Transcriber Similarity / %	No. of Transcriptions	Standard Deviation
4	55.15	2	0.00
5	72.25	2	0.00
11	70.05	2	0.00
12	70.05	2	0.00

The information in Table 3 illustrates an average degree of agreement amongst transcribers for the unstructured audio segments. There was only one inter-transcriber similarity score for two transcriptions explaining why the standard deviation equalled zero. None of the transcriptions had more than two transcriptions as a result the standard deviation is zero. Audio segments that had only one transcription were not included as there are no transcriptions to compare it against. See Appendix 4 for the detailed transcription similarity scores.

The similarity scores of Audios 5, 11, and 12 display moderately high inter-transcriber similarity scores around 70%. This suggests reasonable agreement among transcribers. Considering the informal and noisy context of the unstructured audio segments, these transcripts can be considered relatively reliable for corpus development. The highest similarity scores being less than that of the structured audio segments suggests that transcribing casual talks with background noise can be a more difficult task.

The inter-transcriber similarity score for Audio 4 is relatively low compared to the other audio segments. This indicates less agreement among transcribers. The differences in similarity could be explained by variables such as audio clarity, content complexity, adherence to rules, and transcriber expertise. To improve the dependability and quality of these transcripts, extensive examination and correction may be required.

6.3.2 Zipf's Law

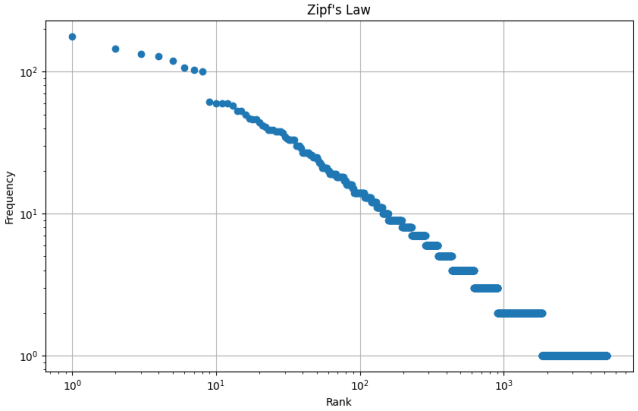

Figure 6: Zipf's Law Distribution: Frequency vs. Rank for Unstructured Audio

Figure 6 illustrates that the frequency of words in the isiXhosa corpus follows Zipf's law. This is consistent with findings for other natural language corpora. However, the graph does deviate from a straight line at the lower ranks. A reason for the deviation could be that the unstructured corpus is not large enough. Zipf's law assumes that all terms in the corpus have an equal chance of being used. This assumption, however, may not be valid for small corpora. Some words may be more common than others in a small corpus simply because they appear more frequently in the text. Casual conversations often contain more repetition and are less structured than formal speech. This could result in a higher percentage of uncommon terms and function words in the corpus, which could also account for the divergence from Zipf's law. The graph's general adherence to Zipf's rule shows that the corpus is sufficiently large to capture the overall frequency of terms used in isiXhosa discussions. It can be deduced that despite the corpus not being complete, it is large enough to be statistically significant.

6.3.3 Token-to-Type Ratio

A TTR of 2.5 for a corpus of 12764 words indicates moderate vocabulary diversity. This is because the corpus comprises a mixture of common and uncommon terms. The studied corpus has a rich and varied vocabulary, with a rather wide range of unique words compared to the total number of words. There is a fair amount of repetition found in the corpus. The unstructured nature and presence of background noise can have an impact on the corpus's lexical variety. This is because it may be challenging to understand what is being said due to background noise. As a result, the listener could be unable to distinguish between various words, which can result in the repetition of words.

6.4 Discussion

The average of the average inter-transcriber agreement for the structured audio corpus is 62.25% and the standard deviation is 25.60. The unstructured audio corpus produces a similar result with 66.88% and 6.8362 as the mean and standard deviation respectively. Munyaradzi et al. [24] found that "a high degree of consistency correlates with a high degree of accuracy". It can thus be argued that while there are variations in the quality of the corpora, as a whole it provides better than average reliability and accuracy. It may be necessary to pay more attention to and revise segments with lower inter-transcriber similarity ratings to improve the overall quality. The corpora can also be improved for their intended use in research or language analysis by considering elements like the audio's context, the transcribers' level of experience, and the presence of potential transcription problems.

Figure 5 and Figure 6 demonstrate that both the corpora conform to Zipf's law. This implies that the corpora follow the linguistic patterns seen in natural languages. The high-frequency pattern is followed by frequently used words like articles, conjunctions, and prepositions while the low-frequency pattern is followed by less frequently used words. Commonly used words like articles, conjunctions, and prepositions follow the high-frequency pattern, while less common words adhere to the low-frequency trend. This pattern is observed in numerous language corpora and cuts across linguistic boundaries, supporting its universal application [25][33][34].

Marquard et al. [32] used web crawling to develop an isiXhosa corpus. The resultant Zipf's law graph had a heavy-tailed distribution. They argue that due to isiXhosa being an agglutinative language, Zipf's law may not be a good predictor of the statistics. IsiXhosa is an agglutinative language, which means that it employs several prefixes and suffixes to change the core meaning of a root word[35]. The subject, objects, and verbs can all be combined into a single word [32]. However, the overall correspondence to Zipf's law suggests that crowdsourcing transcriptions via fluent and native speakers may be a more reliable method of corpus development. A corpus for the Bantu languages Citumbuka and Chichewa was developed by Chavula et al.[36]. Both produced a Zipfian plot that conforms to Zipf's law. This further supports the validity of the produced corpus as isiXhosa also belongs to the Bantu family.

The corpus developed by Marquard et al. [32] using web crawling produced a TTR of 4.81 without duplicate sentences and 11.81 with duplicate sentences at a corpus size of 800 014 words. Considering the size difference, the developed corpora can be said to align with it. Higher lexical variation is expected of isiXhosa with it being an agglutinating language. This contrasts with the findings of Mustafa et al. [25] where a corpus for Arabic and English obtained a TTR of 26.71 and 27.17 respectively. Arabic and English are not agglutinative languages.

6.5 Limitations and Improvements

The statistical analysis used on the corpora developed has numerous limitations that should be considered when evaluating the results. These restrictions shed light on potential sources of variability that may have an impact on the analysis's reliability and breadth. The first restriction is the presence of unstructured audio in the structured corpus. The SABC news segments cut to interviews where the audience starts speaking informal isiXhosa, English, isiZulu etc. These segments, linguistically distinct from the primary isiXhosa focus, could introduce noise into the analysis, potentially affecting the accuracy of language-specific metrics and insights. Additional work had to be done to filter out noise from the audio segments.

A constraint specific to the unstructured corpus is the gathering technique. The usage of a phone microphone in an organized setting with predetermined topics for discussion introduces an element of artificiality. This controlled arrangement may not fully represent the genuine dynamics of spontaneous dialogues that individuals may have in their daily contacts. Additionally, it is important to recognize that the participants did not know each other before the interactions. This aspect may change the type of interactions, hence influencing the authenticity and depth of the talks. Diversifying the data collection method could be a potential solution. Conversations could be recorded in real-life settings, such as recording classroom discussions or workplace meetings.

A major restriction for both corpora is the variation in spellings among various transcriptions. Similar phrases spelt differently might cause errors in frequency estimations and linguistic patterns. This variation might make important linguistic observations difficult to see and possibly explain, for example, the significant difference between the lower-bound and upper-bound inter-transcriber similarity percentages. The depth and generalizability of the analysis are further restricted by the constraint of a small number of transcriptions. The transcription pool could be expanded to have diverse sources and contributors for quality control. Clearer transcription guidelines can be provided. This can also solve the problem where participants translated the segments to English instead of transcribing.

Finally, a small sample size might not accurately represent the range of language usage and linguistic patterns, which could result in biased or unreliable results. It's important to acknowledge that not all transcriptions could be finished due

to time constraints, which made it more difficult to conduct a thorough and reliable statistical analysis. The timeframe of the study must be expanded to tackle this issue. Recruiting more participants and acquiring more transcriptions will allow the creation of a more accurate and reliable corpus. These limitations highlight the importance of careful interpretation of the analysis results. While useful insights can be obtained, the limited sample size, controlled environment, and participant unfamiliarity need a nuanced evaluation of the findings and their relevance to real-world scenarios.

7 CONCLUSIONS

This study addressed the pressing issue of the lack of electronic linguistic resources in low-resource African languages, particularly isiXhosa in South Africa. Through the application of crowdsourcing via the Turkle web application, a gold standard ASR corpus in isiXhosa was developed for two types of audio segments. A corpus of size 26794 and 12764 was developed for structured audio and unstructured audio respectively.

The results shed light on various aspects of the corpus, drawing attention to both its strengths and limitations. The inter-transcriber similarity analysis revealed varying degrees of agreement in transcriptions, offering insights into the consistency of the data. The adherence of the corpus to Zipf's Law provided evidence of linguistic patterns inherent in natural languages, while the token-to-type ratio highlighted a balanced yet moderate vocabulary diversity. However, the study faced limitations such as the presence of unstructured audio in the structured audio, variations in spellings of the transcriptions, and most importantly, a small number of transcriptions. These limitations emphasize the need for cautious interpretation of the findings and their applicability beyond the study's scope.

In conclusion, while there were problems and constraints, this study revealed the potential of using crowdsourcing to produce a gold-standard ASR corpus for a low-resource South African language. The study contributes to the larger area of linguistic research and aids the development of computational language processing systems by providing significant insights into linguistic patterns and word richness in isiXhosa. This work serves as a platform for future research and collaborations focused on alleviating linguistic resource shortages in African languages.

8 FUTURE WORK

Moving forward, there are several interesting options for extending and expanding the impact of this research. First and foremost, increasing the size and diversity of the corpus is a top priority. A more comprehensive portrayal of isiXhosa language dynamics can be produced by adding a greater range of topics, conversation types, and participant demographics. Efforts should also be made to increase transcription consistency possibly through clearer instructions, continued training, and better quality assurance procedures.

The potential of using the gold standard corpus for sophisticated language studies and applications is promising. Deeper insights into isiXhosa use can be gained by investigating sentiment analysis, language evolution, and syntactic patterns. The corpus can be used for machine translation, speech recognition, and natural language processing. Collaborative efforts involving local communities, linguists, and language enthusiasts can improve authenticity and relevance while also promoting a sense of ownership and cultural connection. This can be further expanded from one language to several of the other South African languages that share linguistic characteristics.

ACKNOWLEDGMENTS

I extend my gratitude to my supervisor and co-supervisor for their guidance and support, my dedicated group partners for their collaboration, our generous sponsor for making this project possible, and the welcoming Xhosa community for sharing their knowledge. Your contributions have been vital to our success, and I am thankful for your support.

REFERENCES

[1] J. A. Nel, V. H. Valchev, S. Rothmann, F. J. R. Vijver, D. Meiring, and G. P. Bruin, "Exploring the Personality Structure in the 11 Languages of South Africa," in *Journal of Personality* 80, vol. 4, pp. 915-948, 2012. DOI: https://doi.org/10.1111/j.1467-6494.2011.00751.x

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

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

[4] Science Portals & Science Gateways. Retrieved August 13, 2023 from https://www.sdsc.edu/services/hpc/science_gateways.html

[5] SABC – Official Website – South African Broadcasting Corporation. Retrieved August 26, 2023 from https://www.sabc.co.za/sabc/

[6] SADiLaR. Retrieved August 26, 2023 from https://sadilar.org/index.php/en/

[7] N. Nowshin, Z. S. Ritu, and S. Ismail, "A Crowd-Source Based Corpus on Bangla to English Translation," in *21st International Conference of Computer and Information Technology (ICCIT),* pp. 1-5 2018. DOI: https://doi.org/10.1109/iccitechn.2018.8631947

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

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

[10] Corpus types: monolingual, parallel, multilingual… | Sketch Engine. *Sketch Engine*. Retrieved March 13, 2023 from https://www.sketchengine.eu/corpora-and-languages/corpus-types/

[11] L. Wissler, M. Almashraee, D. Monett, A. Paschke, "The Gold Standard in Corpus Annotation," in *Proceedings of the IEEE GSC*, Passau, Germany, vol. 21, pp. 26–27, June 2014.

[12] E. Barnard, M. H. Davel, C. J. V. Heerden, F. D. Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," In *4th Workshop on Spoken Language Technologies for Under-resourced Languages,* SLTU 2014*,* ISCA, St. Petersburg, Russia, pp. 194–200, May 2014.

[13] E. Barnard, M. Davel, and C. v. Heerden, "ASR corpus design for resource-scarce languages," in *Interspeech,* pp. 2847-2850, 2019. DOI: https://doi.org/10.21437/interspeech.2009-727

[14] Merriam-Webster: America. Retrieved August 26, 2023 from https://www.merriam-webster.com/

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

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

[17] Amazon Mechanical Turk. *Amazon Mechanical Turk*. Retrieved March 13, 2023 from https://www.mturk.com/

[18] UHRS. *UHRS*. Retrieved March 13, 2023 from https://prod.uhrs.playmsn.com/

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

[20] W. Y. Wang, D. Bohus, E. Kamar, and E. Horvitz, "Crowdsourcing the acquisition of natural language corpora: Methods and observations," in *IEEE Spoken Language Technology Workshop (SLT)*, pp. 73-78, 2012. DOI: https://doi.org/10.1109/slt.2012.6424200

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

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

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

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

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

[26] hltcoe/turkle: Django-based clone of Amazon. Retrieved August 26, 2023 from https://github.com/hltcoe/turkle

[27] M. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank," in *Comput. Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[28] M. A. Haidar and M. Rezagholizadeh, ''Fine-tuning of pretrained end-to-end speech recognition with generative adversarial networks,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP*), Toronto, ON, Canada, pp. 6204–6208, June 2021. DOI: https://10.1109/ICASSP39728.2021.9413703

[29] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," in *Psychonomic Bulletin & Review,* vol. 21, no 5, pp. 1112-1130, 2014. DOI: https://doi.org/10.3758/s13423-014-0585-6

[30] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions aud Reversals," in *Soviet Physics Doklady*, vol. 10, pp. 707, 1966.

[31] B. Eisen and H. G. Tillman, "Consistency of judgments in manual labeling of phonetic segments: The distinction between clear and unclear cases," in *Proceedings of ICSLP*, Banff/Canada, pp. 871-874, 1992.

[31] D. Thomas, "Type-token ratios in one teacher's classroom talk: An investigation of lexical complexity," United Kingdom: University of Birmingham, 2005.

[31] B. Richards, "Type-token ratios: what do they really tell us?, "in *Journal of Child Languages*, vol. 14, no 14, pp. 201-209, 1987. DOI: https://doi.org/10.1017/s0305000900012885

[32] C. Marquard and H. Suleman," Focused Crawling for Automated IsiXhosa Corpus Building," in *Proceedings of SAICSIT 2023: South African Institute of Computer Scientists and Information Technologists*, Muldersdrift, South Africa, 1878, pp 19-31, 2023

[33] N. Hatzigeorgiu, G. Mikros, and G. Carayannis, "Word Length, Word Frequencies and Zipf's Law in the Greek Language," in *Journal of Quantitative Linguistics,* vol. 8, no. 3, pp. 175-185, 2001, DOI: https://doi.org/10.1076/jqul.8.3.175.4096

[34] H. Xiao, "On the applicability of Zipf's law in Chinese word frequency distribution," in *Journal of Chinese Language and Computing,* vol. 18, no. 1, pp. 33-46, 2008.

[35] M. Pascoe and M. Smouse, "Masithethe: Speech and language development and difficulties in isiXhosa," in *South African Medical Journal,* vol. 102, no. 6, pp. 469, 2012. DOI: https://doi.org/10.7196/samj.5554

[36] C. Chavula and H. Suleman, ``Assessing the impact of vocabulary similarity on multilingual information retrieval for bantu languages,'' *in Proc. 8th Annu. Meeting Forum Inf. Retr. Eval*., pp. 16-23, Dec. 2016.

APPENDICES

Appendix 1 – Recruitment Documents

Appendix 1.1 Email

Dear students,

I hope this email finds you well. My name is Mosamat Sabiha Shaikh, and I am a Computer Science Honours student at the University of Cape Town. I am reaching out to invite you to participate in an exciting language research project focused on building a gold standard corpus in isiXhosa, a critically important language in our community.

Before I proceed, I want to assure you that I have obtained ethical clearance from our university's research committee to conduct this study (clearance code - **FSREC 051-2023)**. Your privacy and well-being are of utmost importance to us, and we will strictly adhere to the guidelines and regulations set forth by the committee.

The project is divided into two parts:

1. Unstructured audio collection: The purpose of this section is to collect unstructured audio data, which refers to casual conversations spoken in an uncontrolled environment. The audio data collected will be used to develop a gold standard corpus that can be used in various areas of natural language processing.

   We are inviting all native/fluent isiXhosa speakers to participate in a group conversation with two other individuals. You will engage in a casual conversation for a total of 40 minutes. Each conversation will be divided into four topics, with each topic discussed for approximately 10 minutes. The objective is to simulate a natural conversational setting, allowing us to capture a wide range of linguistic features and patterns. You will receive a payment of R50 for the completion of the task.

   Your participation in this study would be immensely valuable in helping us build a comprehensive gold standard corpus for isiXhosa. By contributing your voice and insights, you will be making a significant contribution to the study and advancement of this rich and vibrant language.

2. Transcription of audio: Alternatively, if you prefer not to engage in group conversations, you can participate in the transcription aspect of the project. The purpose of this section is to develop a corpus which contains transcriptions of audio in isiXhosa. As a transcription participant, you will be responsible for transcribing 10-minute audio recordings, each 10-minute audio recording being a task, in isiXhosa. A task should take around 45 minutes to an hour to complete for an average person, and you will receive R50 as compensation for each

completed task.

You are welcome to participate in either or both parts of the project. To participate in the transcription portion you will need to meet the following criteria:

1. Be fluent in isiXhosa: As the task involves transcribing audio in isiXhosa, fluency in the language is essential.
2. Have basic technology skills: You should have access to a computer or a device with internet access.
3. Attention to detail: Accurate transcription is crucial, so attention to detail is highly valued.

Please note that your participation is entirely voluntary, and you have the right to withdraw at any time without providing a reason. If you choose to participate, your identity and personal information will be kept strictly confidential. All data collected will be anonymized and stored securely, and only the research team will have access to the audio recordings.

If you meet the stated criteria and are interested in participating or have any questions or concerns, please contact me at shkmos004@myuct.ac.za. Please do inform me which portion you would be interested in. I would be more than happy to provide you with additional information and address any queries you may have.

Your involvement in this project would be immensely appreciated, and we sincerely hope you will consider taking part. Together, we can make a significant impact on the study of the isiXhosa language and contribute to its preservation and understanding. Thank you for your time, and we look forward to hearing from you soon.

Warm regards,
Mosamat Sabiha Shaikh
University of Cape Town

Appendix 1.2 Poster

Appendix 1.3 Audio Transcription Consent Form

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF CAPE TOWN
PRIVATE BAG X3
RONDEBOSCH 7701
SOUTH AFRICA

RESEARCHER/S: Mosamat Sabiha Shaikh
TELEPHONE: +27-60-346 6797
E-MAIL: Shkmos004@myuct.ac.za
URL: dept@cs.uct.ac.za

Informed Voluntary Consent to Participate in Research Study

SABC2TXT: Gold Standard Corpus development

Invitation to participate, and benefits: You are invited to participate in a research study conducted with a student from the University of Cape Town. The study aim is to develop gold standard corpus in isiXhosa for use in natural language processing. I believe that your experience would be a valuable source of information, and hope that by participating you may gain useful knowledge.

Procedures: During this study, you will be asked to transcribe audio in isiXhosa on an online crowdsourcing platform. 1 segment should take approximately 1 hour to transcribe for the average person.

Risks: There are no potentially harmful risks related to your participation in this study.

Feedback: You will receive feedback about the results of this research via email if you wish.

Compensation: You will receive R50 compensation for the completion of each transcription.

Disclaimer/Withdrawal: Your participation is completely voluntary; you may refuse to participate, and you may withdraw at any time without having to state a reason and without any prejudice or penalty against you. Should you choose to withdraw, the researcher commits not to use any of the information you have provided without your signed consent. Note that the researcher may also withdraw you from the study at any time.

Confidentiality: All information collected in this study will be kept private in that you will not be identified by name or by affiliation to an institution. Confidentiality and anonymity will be maintained as pseudonyms will be used.

What signing this form means: By signing this consent form, you agree to participate in this research study. The aim, procedures to be used, as well as the potential risks and benefits of your participation have been explained verbally to you in detail, using this form. Refusal to participate in or withdrawal from this study at any time will have no effect on you in any way. You are free to contact me, to ask questions or request further information, at any time during this research.

I agree to participate in this research (tick one box) ☐ Yes ☐ No _____ (Initials)

_____	_____	_____
Name of Participant	Signature of Participant	Date
_____	_____	_____
Name of Researcher	Signature of Researcher	Date

Appendix 1.3 Audio Collection Consent Form

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF CAPE TOWN
PRIVATE BAG X3
RONDEBOSCH 7701
SOUTH AFRICA

RESEARCHER/S: Mosamat Sabiha Shaikh
TELEPHONE: +27-60-346 6797
E-MAIL: Shkmos004@myuct.ac.za
URL: dept@cs.uct.ac.za

Informed Voluntary Consent to Participate in Research Study

SABC2TXT: Gold Standard Corpus development

Invitation to participate, and benefits: You are invited to participate in a research study conducted with a student from the University of Cape Town. The study aim is to develop gold standard corpus in isiXhosa for use in natural language processing. I believe that your experience would be a valuable source of information, and hope that by participating you may gain useful knowledge.

Procedures: During this study, you will be asked to have 4 casual conversations (topics will be provided) in isiXhosa with 2 other participants for approximately 10 minutes each.

Recording: We will take an audio recording as part of the study. These recordings will be used to develop the corpus. If you object to this, please indicate below.

Risks: There are no potentially harmful risks related to your participation in this study.

Feedback: You will receive feedback about the results of this research via email if you wish.

Compensation: You will receive R50 compensation for completing and participating in the conversation session.

Disclaimer/Withdrawal: Your participation is completely voluntary; you may refuse to participate, and you may withdraw at any time without having to state a reason and without any prejudice or penalty against you. Should you choose to withdraw, the researcher commits not to use any of the information you have provided without your signed consent. Note that the researcher may also withdraw you from the study at any time.

Confidentiality: All information collected in this study will be kept private in that you will not be identified by name or by affiliation to an institution. Confidentiality and anonymity will be maintained as pseudonyms will be used.

What signing this form means: By signing this consent form, you agree to participate in this research study. The aim, procedures to be used, as well as the potential risks and benefits of your participation have been explained verbally to you in detail, using this form. Refusal to participate in or withdrawal from this study at any time will have no effect on you in any way. You are free to contact me, to ask questions or request further information, at any time during this research.

I agree to participate in this research (tick one box) ☐ Yes ☐ No _____ (Initials)
I agree to be audio-recorded ☐ Yes ☐ No _____ (Initials)
I agree to the use of anonymized audio recordings in a crowdsourcing platform
☐ Yes ☐ No _____ (Initials)

_____	_____	_____
Name of Participant	Signature of Participant	Date

_____	_____	_____
Name of Researcher	Signature of Researcher	Date

Appendix 2 – Audio Collection Topics

Topic: Sports

1. **Sports:**
    - "Do you enjoy watching or playing sports?"
        - Expanding: Share your own preferences, whether you're a fan of watching sports, playing them, or both. Discuss your favourite sports and why you find them appealing.
    - "Have you ever been part of a sports team?"
        - Expanding: Talk about your experiences being part of a sports team, the camaraderie, and lessons learned. Ask about their own experiences playing on a team or any memorable sports-related moments.
    - "What's your favourite sport to watch, and why?"
        - Expanding: Explain your favourite sport, what you enjoy about watching it, and any memorable games you've seen. Ask about their favourite sport and reasons for their preference.
2. **UCT Sports**
3. **Sports Heroes:**
    - "Do you have a favourite sports player or athlete?"
        - Expanding: Share your admiration for a particular athlete and what qualities you find inspiring. Ask about their favourite athletes and any reasons they look up to them.
    - "Have you ever met a sports celebrity or attended a live game?"
        - Expanding: Share any encounters you've had with sports celebrities or your experiences attending live sporting events. Ask about their own experiences with meeting athletes or attending games.
4. **Favourite Sports Moments:**
    - "What's the most memorable sports moment you've witnessed?"
        - Expanding: Describe your own unforgettable sports moment and why it left a lasting impression. Encourage them to share their own memorable moments and the emotions tied to them.
    - "Is there a sports event you'd love to witness in person?"
        - Expanding: Talk about a sports event you've always wanted to attend and why. Ask about their dream sports event to attend and the reasons behind their choice.
5. **Team Rivalries:**
    - "Do you have a favourite sports team?"
        - Expanding: Talk about your favourite sports team and your loyalty to them. Inquire about their favourite team and whether they have any special memories associated with them.
    - "Have you ever experienced a friendly rivalry between fans of different teams?"
        - Expanding: Share any light-hearted rivalries you've experienced and how they've added excitement to your sports enjoyment. Ask about their experiences with team rivalries and any funny anecdotes.
6. **Sports and Life Lessons:**
    - "Have you learned any valuable life lessons from playing sports?"
        - Expanding: Share any life lessons you've gained from your sports experiences, such as teamwork, discipline, or perseverance. Ask about any lessons they've learned from playing sports and how they've applied them to other areas of life.
7. **Uncommon or Unique Sports:**
    - "Are there any unusual or lesser-known sports you find interesting?"
        - Expanding: Talk about any unique sports you've come across and why they caught your attention. Ask about any lesser-known sports they've encountered or participated in.

Topic: Food

1. **Food Preferences:**
    - "What type of cuisine do you enjoy the most?"
        - Expanding: Share your favourite type of cuisine and what specific dishes you love. Ask about their culinary preferences and any memorable dining experiences they've had.
    - "Do you have any dietary restrictions or food allergies?"
        - Expanding: Discuss any dietary preferences or restrictions you have and how you manage them. Inquire about their dietary needs and any creative ways they've adapted recipes.
2. **UCT Food/Shops/In Res?**
3. **Cooking and Culinary Adventures:**

- "Do you like cooking? What's your signature dish?"
  - Expanding: Talk about your cooking experiences, favorite recipes, and any culinary experiments you've tried. Ask about their cooking style, signature dish, and any cooking achievements they're proud of.
- "Have you ever taken a cooking class or tried making a challenging recipe?"
  - Expanding: Share your experiences with cooking classes or attempting difficult recipes. Ask about any cooking challenges they've taken on and the results.

4. **Favourite Food Memories:**
   - "What's your most cherished food-related memory?"
     - Expanding: Share a special food memory from your past, whether it's a family gathering or a memorable meal during travel. Encourage them to share their own cherished food memories and the emotions attached to them.
   - "Is there a particular dish that reminds you of your childhood?"
     - Expanding: Describe a dish that brings back fond childhood memories for you. Ask about their own nostalgic food favorites and any stories related to those dishes.
5. **Culinary Explorations:**
   - "Are you adventurous when it comes to trying new foods?"
     - Expanding: Share a story about trying a new and unfamiliar food, whether you loved it or found it challenging. Ask about their experiences with trying diverse cuisines and any unique foods they've encountered.
   - "Have you ever tried a local delicacy while traveling?"
     - Expanding: Talk about any local delicacies you've tasted during your travels and the cultural significance behind them. Inquire about their own experiences trying regional foods in different parts of the world.
6. **Food and Culture:**
   - "How do you think food reflects a culture's identity?"
     - Expanding: Discuss your thoughts on how food is intertwined with culture and traditions. Ask about their perspective on the role of food in expressing cultural identity and fostering connections.
   - "Are there any cultural dishes you find particularly intriguing?"
     - Expanding: Talk about a cultural dish that captivates you and the reasons behind your fascination. Ask about their own interests in cultural cuisine and any dishes they've found intriguing.
7. **Favourite Food Destinations:**
   - "Do you have a favourite food destination you've visited?"
     - Expanding: Share your favourite food-related travel experience and the culinary delights you discovered. Ask about their own favourite food destinations and the memorable meals they've had while traveling.

Topic: Education

1. **Educational Journey:**
   - "Tell me about your educational background. Where did you study?"
     - Expanding: Share your educational path, including the schools you attended and any notable experiences. Ask about their own educational journey, their alma mater, and any memorable moments from their time in school.
   - "What subjects or fields of study have you been most interested in?"
     - Expanding: Discuss your own academic interests and how they've evolved. Ask about their preferred subjects and any reasons behind their fascination with those areas.
2. **University Life:**
   - "What was your major?"
     - Expanding: Talk about your university experience, your major, and any extracurricular activities you were involved in. Ask about their university journey, major, and any standout experiences they had.
   - "How do you think university life differs from high school?"
     - Expanding: Share your insights on the differences between high school and university life, including the level of independence and academic rigor. Ask about their observations and how they navigated the transition.
3. **Favourite Educational Memories:**
   - "Do you have a favourite memory from your school or university days?"
     - Expanding: Share a cherished memory from your educational journey, whether it's a heartwarming moment or a significant achievement. Encourage them to share their own memorable experiences from school or university.
   - "Have you had a particularly influential teacher or professor?"
     - Expanding: Talk about a teacher or professor who left a lasting impact on your education and personal growth. Inquire about their own influential educators and the lessons they imparted.

4. **Education and Future Goals:**
    - "Did your education influence your career choices?"
        - Expanding: Discuss how your educational background impacted your career decisions and opportunities. Ask about their own experiences in choosing career paths based on their education.
    - "Are you considering further education or advanced degrees?"
        - Expanding: Talk about any plans or aspirations for pursuing higher education. Inquire about their thoughts on further studies and whether they're considering advanced degrees.
5. **Education System Discussions:**
    - "What do you think about the current education system?"
        - Expanding: Share your thoughts on the strengths and shortcomings of the education system. Encourage them to discuss their perspective on the education system and any improvements they would suggest.
    - "If you could change one thing about the education system, what would it be?"
        - Expanding: Talk about a specific aspect of the education system you'd like to see improved and your reasons for the change. Ask about their own ideas for positive changes in education.

Topic: Entertainment

1. **Favourite Books:**
    - "Do you have a favourite book that left a lasting impact on you?"
        - Expanding: Share a book that has deeply influenced you and why it holds significance. Ask about their favourite book and what resonated with them in that particular story.
    - "Are there any books you could read over and over again?"
        - Expanding: Talk about a book you find so captivating that you can read it multiple times. Inquire about their own timeless favourites and what draws them back to those books.
2. **Movie Preferences:**
    - "Do you enjoy watching movies? What's your favourite genre?"
        - Expanding: Discuss your movie preferences and the types of films you're most drawn to. Ask about their favourite movie genre and any specific films within that genre they've enjoyed.
    - "Is there a movie that you can quote lines from?"
        - Expanding: Share a movie that you know so well you can quote lines from it. Inquire about their own memorable movie quotes and the context behind them.
3. **Book-to-Movie Adaptations:**
    - "Do you think book-to-movie adaptations usually live up to the source material?"
        - Expanding: Share your thoughts on the challenges of adapting books into movies and instances where it was successful or fell short. Ask about their opinions on adaptations and any memorable examples.
    - "What's your favourite book-to-movie adaptation?"
        - Expanding: Talk about a book-to-movie adaptation you believe captured the essence of the original story. Ask about their favorite adaptations and what made them stand out.
4. **Unforgettable Stories:**
    - "Is there a book or movie that you find yourself recommending to others often?"
        - Expanding: Share a book or movie that you frequently recommend and why it's so impactful. Ask about their go-to recommendations and the reasons behind their choices.
    - "Have you ever read a book or watched a movie that changed your perspective on something?"
        - Expanding: Discuss a book or movie that shifted your viewpoint or deepened your understanding of a topic. Ask about their own transformative reading or viewing experiences.
5. **Characters**
    - "Do you have a favourite fictional character?"
        - Expanding: Talk about a fictional character you admire and the qualities that resonate with you. Inquire about their favorite characters and what draws them to those particular personalities.
6. **Movie Magic:**
    - "What's the last movie you watched that really captivated your attention?"
        - Expanding: Discuss a recent movie that held your interest and why it stood out. Ask about the last movie they watched that left a lasting impression.
    - "Is there a classic movie you think everyone should watch?"
        - Expanding: Share a classic movie recommendation and the reasons you believe it's a must-see. Ask about their own classic movie favorites and what makes them timeless.

Appendix 3 – Corpora Raw Data

Appendix 3.1 Word Statistic for the Structured Audio

Table 4: Most Appeared Word

Word	Appearance
Ukuba	270
Ke	214
Abantu	145
Kodwa	127
Emva	104

Table 5: least Appeared Word

Word	Appearance
Minist	1
Appreci	1
Collapse	1
Ukuhambisa	1
Iyasilela	1

Appendix 3.2 Word Statistic for the Unstructured Audio

Table 6: Most Appeared Word

Word	Appearance
Uba	177
Mna	145
Ngoku	133
Xa	128
Kodwa	119

Table 7: least Appeared Word

Word	Appearance
Iyandi	1
Nawo	1
Amantombazana	1
Kungacingwa	1
Kungazubakho	1

Appendix 4 – Inter-Transcriber Similarity

Appendix 4.1 Full Similarity Score List for Structured Audio Segments

Audio URL: 1
Similarity between Transcript 1 and Transcript 2: 0.4893
Similarity between Transcript 1 and Transcript 3: 0.5059
Similarity between Transcript 2 and Transcript 3: 0.8729

Audio URL: 2
Similarity between Transcript 1 and Transcript 2: 0.6833

Audio URL: 3
Similarity between Transcript 1 and Transcript 2: 0.8102

Audio URL: 4
Similarity between Transcript 1 and Transcript 2: 0.7108

Audio URL: 6
Similarity between Transcript 1 and Transcript 2: 0.9943

Audio URL: 7
Similarity between Transcript 1 and Transcript 2: 0.8626
Similarity between Transcript 1 and Transcript 3: 0.5762
Similarity between Transcript 2 and Transcript 3: 0.5033

Audio URL: 8
Similarity between Transcript 1 and Transcript 2: 0.9996
Similarity between Transcript 1 and Transcript 3: 0.7848
Similarity between Transcript 2 and Transcript 3: 0.7852

Audio URL: 9
Similarity between Transcript 1 and Transcript 2: 1.0000
Similarity between Transcript 1 and Transcript 3: 0.9318
Similarity between Transcript 2 and Transcript 3: 0.9318

Audio URL: 10
Similarity between Transcript 1 and Transcript 2: 0.2422
Similarity between Transcript 1 and Transcript 3: 0.2607
Similarity between Transcript 2 and Transcript 3: 0.4304

Audio URL: 11
Similarity between Transcript 1 and Transcript 2: 0.9955
Similarity between Transcript 1 and Transcript 3: 0.8094
Similarity between Transcript 2 and Transcript 3: 0.8109

Audio URL: 13
Similarity between Transcript 1 and Transcript 2: 0.5473

Audio URL: 14
Similarity between Transcript 1 and Transcript 2: 0.7440

Audio URL: 15
Similarity between Transcript 1 and Transcript 2: 0.1509

Audio URL: 16
Similarity between Transcript 1 and Transcript 2: 0.4178

Audio URL: 17
Similarity between Transcript 1 and Transcript 2: 0.7596

Audio URL: 21
Similarity between Transcript 1 and Transcript 2: 0.2831

Appendix 4.2 Full Similarity Score List for Unstructured Audio Segments
Audio: 4
Similarity between Transcript 1 and Transcript 2: 0.5515

Audio: 5
Similarity between Transcript 1 and Transcript 2: 0.7225

Audio URL: 11
Similarity between Transcript 1 and Transcript 2: 0.7005

Audio URL: 12
Similarity between Transcript 1 and Transcript 2: 0.7005