UNIVERSITY OF THE WITWATERSRAND

MSC IN COMPUTER SCIENCE

Tone Labelling Algorithm for Sesotho

Author:
Mpho Raborife

Supervisors:
Prof. S. Ewert and Prof. S. Zerbian

September 16, 2011

Declaration Of Authorship

I, Mpho Ivy Raborife, declare that the dissertation entitled Tone Labelling Algorithm for Sesotho and the work presented in the dissertation are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a MSc degree at the University of the Witwatersrand;
- where any part of this research report has previously been submitted for a degree or any other qualification at the University of the Witwatersrand or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the research is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Signed: .....

Abstract

Studies have shown that text-to-speech systems need detailed prosodic models of a language in order to ideally sound natural to native speakers of the language. A text-to-speech system developed for Sesotho needs to have tone implemented in it since Sesotho is a tonal language which uses pitch variations to distinguish lexical and/or grammatical meaning.

In order to implement tone for a language such as Sesotho, it is necessary for a tone modeling algorithm to receive as input the tone labels of the syllables of a word. This allows the algorithm to predict the appropriate intonation of the word. The aim of our study is to improve a basic tone labeling algorithm that predicts tone labels using three Sesotho tonal rules. The application of this algorithm is restricted to polysyllabic verb stems. The research study involves implementing an extended tone labeling algorithm that implements four additional Sesotho tonal rules and extends its application to all the other parts of speech.

The results of our study show that the extended tone labeling algorithm significantly improves the basic algorithm by increasing the number of matched tone labels. Furthermore, our study provides the basic step to tone modeling for languages such as Sesotho which do not mark tone labels in orthography.

Acknowledgements

I owe my deepest gratitude to my supervisors, Prof. Sigrid Ewert and Prof. Sabine Zerbian, whose help, stimulating suggestions and encouragement helped me in the conduct of the research study and writing of this research report.

I would like to express my gratitude to all those who assisted me in completing this research study. I would like to thank the Human Language Technology group at the Meraka Institute (Council of Scientific and Industrial Research) for providing me with the financial assistance in the form of a Master studentship. Most specifically, the text-to-speech team for making me feel at home and for their ideas and suggestions during the conduct of the research study. Furthermore, I would also like to thank the external examiners for their useful comments.

Lastly, I offer my regards and blessings to all those who supported me in any respect during the conduct of the research study.

Preface

Two papers based on some of the results of this research have been presented at conferences. Their details are as follows:

- Mpho Raborife, Sigrid Ewert and Sabine Zerbian. Tone Labelling Algorithm for Sesotho. *Proceedings of the Postgraduate Symposium of the Annual Conference of the South African Institute for Computer Scientists and Information Technologists*. Bela Bela, South Africa. 2010.
- Mpho Raborife, Sigrid Ewert and Sabine Zerbian. An African Solution for an African Problem: A step towards perfection. *Proceedings of the Twenty-First Annual Symposium of the Pattern Recognition Association of South Africa*. Stellenbosch, South Africa. 2010.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Introduction

Sesotho is a Southern Bantu language belonging to the Niger-Congo language family. It is closely related to three languages in the Sotho-Tswana language group, that is Setswana, the Northern Sotho languages, and Silozi (spoken in Zambia). Bantu languages are spoken largely in central Africa, eastern Africa, and southern Africa.

Bantu languages are tonal languages in that they use pitch to distinguish between meaning as opposed to stress-accent languages such as English. One might liken tone to the relative loudness given to a syllable and stress to the emphasis given to a syllable in a word or a word in a phrase. Stress in a language such as English may be phonetic, that is, noticeable to the listener but not affecting meaning, or it may be used to distinguish meaning.

Text-to-speech (TTS) systems convert text input to a speech waveform that ideally sounds natural to native speakers of the target language. TTS systems require detailed prosodic models of a language such as tone for Bantu languages in order to sound natural and interesting. Prosody refers to the rhythm, stress, and intonation of connected speech.

The research study aimed to improve an algorithm developed by Raborife [2009] which predicts tone labels based on three Sesotho tonal rules, namely high tone spread (HTS1), iterative high tone spread (HTS2) and grammatical tone insertion (GTI). We intended to improve this algorithm by implementing four other tonal rules as well as extending the application of the tonal rules to all other parts of speech since the application of Raborife [2009]'s algorithm is restricted to verb stems. The four tonal rules are the right branch delinking rule (RBD), left branch delinking rule (LBD), specifier high tone delinking rule (SHTD) and the finality restriction rule (FR).

We refer to the algorithm developed by Raborife [2009] as the basic algorithm and our current version of the algorithm as the extended algorithm.

1.2 Theoretical Basis

The basic algorithm as well as the extended algorithm were formulated on the basis of Sesotho tonal rules as described by Khoali [1991][1]. These tonal rules apply within certain phonological domains as discussed in Section 2.3. The phonological domains within which the tonal rules are implemented by both the basic and the extended algorithms are the clitic phrase, phonological word and the phonological phrase. These domains make use of the theory of the Prosodic Hierarchy as proposed by Nespor and

[1] Kunene [1974] also investigates the tonology of Sesotho. However, we could not access his thesis thus we only used Khoali [1991].

Vogel [1986]. The prosodic hierarchy describes the phonological domains to which the phonological rules are restricted.

1.3 Motivation for the Research Study

Extensive research has been dedicated to tone modeling for East-Asian languages such as Chinese [Wang *et al.* 1996; Lee *et al.* 2002; Li *et al.* 2004; Ni *et al.* 2008]. Recently, studies have focused on tone modeling for African tone languages such as Yoruba [Odejobi *et al.* 2006; Dtunji *et al.* 2007; Odejobi *et al.* 2008]. For Bantu languages, we only know of one published attempt, namely by Louw *et al.* [2005], to model tone for an isiZulu TTS system. Not having tone implemented into such systems compromises the perceived naturalness as well as the intelligibility of the system. This illustrates the need for research based on tone modeling for this group of languages.

To model tone into TTS systems developed for Yoruba, the tone labels on the syllables are assigned appropriate pitch values using methods such as fuzzy logic based rules and decision trees. In Yoruba, tone labels are marked in orthographic writing in contrast to Southern Bantu languages. Considering that Southern Bantu languages do not mark tone in orthography, it is difficult to assign appropriate pitch values to syllables as it is not known which of the syllables are high or low-toned.

It is clear that a solution which allows for the equal treatment and advancement of Southern Bantu languages, with respect to other tonal languages in terms of speech systems developed for such languages, is needed. Louw *et al.* [2005] suggest that such a solution might be an algorithm that predicts which of the syllables of a word are to be marked for a high or a low tone and the appropriate intonation of the word. They argue that with such an algorithm, a TTS system would produce more "natural" tones.

Raborife [2009] focused on how such an algorithm can be implemented. Raborife [2009] found that such an algorithm would need to use linguistically-defined tonal rules to make the tone mark (label) predictions. That is, the tonal rules are described in the literature of the language in question and the algorithm implements them.

1.4 Scope and Limitations of the Research Study

Prosody modeling is one of the fundamental components in building TTS systems. This aspect is handled by the prosody prediction module which involves generating pitch, stress, duration and the likes as specified in the language which is being modeled. This module is responsible for making the systems ideally sound as natural as human speech. Sesotho is a tonal language and the prosody prediction module in a TTS system developed for this language would have to model tone.

To generate audible tone patterns, the prosody prediction module needs to be served as input the individual tone labels for each syllable in a Sesotho sentence. Once it has this information, then the appropriate pitch values can be assigned to the syllables.

In Sesotho, unlike a language such as Yoruba, tone labels are not marked in orthography. This makes it difficult to investigate the relationship between the tone marks and the actual pitch values. Once the tone labels on the syllables are known, the relationship between the labels and the actual pitch values can be investigated. It is therefore clear that the first step to tone modeling for TTS systems developed for tonal languages, such as Sesotho, which do not mark tone in orthography, involves being able to effectively predict which of the syllables in a word have either a high or a low tone.

Our study was centered around improving an algorithm developed by Raborife [2009] which makes use of linguistic tonal rules to predict the tone marks on the syllables, thereby tackling the first aspect of Louw *et al.* [2005]'s suggestion which states that to successfully model tone into Bantu languages such

as isiZulu and Sesotho, we need an algorithm that can predict which of the syllables in a word have a high or a low tone.

The second aspect suggested by Louw *et al.* [2005] involves the development of an algorithm which can predict the overall intonation of the word. This aspect involves mapping the tone marks on the syllable to actual pitch values. Our study did not investigate this and as a result we cannot measure how tone affects the perceived naturalness of a TTS system.

1.5 Research Aim and Hypothesis

Our study aimed to improve a basic algorithm that uses three linguistically-defined Sesotho tonal rules to make tone label predictions on the syllables of Sesotho words. The basic algorithm predicts the tone labels of a syllable based on the underlying tones, as described in a Sesotho dictionary by Du Plessis *et al.* [1974], which marks tone as well as the tense, mood and aspect of the verb stems. This algorithm restricts its applications to polysyllabic verb stems. Monosyllabic verb stems have a complex tonal behaviour which differs from that of polysyllabic verb stems. Applying the tonal rules to monosyllabic verb stems would have exceeded the scope of the study.

The main objective of our study was to improve the basic algorithm by implementing four other tonal rules, namely the right branch delinking rule, the left branch delinking rule, the finality restriction principle and the specifier high tone delinking rule. Furthermore, we extended the application of the tonal rules to other word classes. We formulated our research hypothesis as follows:

> The extended algorithm improves the basic algorithm by increasing the number of matched tone labels.

Matched tone labels are the labels predicted by either one of the algorithms which are exactly like their transcribed counterpart. The sentences on which we tested the algorithms were recorded and the tone labels on the syllables of the words in the sentences were transcribed as described in the next section. The transcriptions were used in analyzing the algorithms and testing the hypothesis.

1.6 Research Method

1.6.1 Corpus Compilation and Algorithm Improvement

Forty-five Sesotho sentences were chosen to be used as data in such a way that they show occurrences of the tenses that are relevant for the application of both algorithms. These sentences were recorded by three native Sesotho speakers. Only one speaker's recording was necessary for transcription since the tonal rules implemented by both algorithms are not restricted to speakers of a particular dialect [Khoali 1991].

The recorded speech of the chosen speaker was then transcribed by three independent labelers with different backgrounds in tone studies. The transcriptions included tone labels and were compared across all three labelers. In the cases where there were disagreements amongst the labelers with respect to the tone labels, the final tone label was based on the tone label that was agreed upon by at least two labelers.

The extended algorithm was designed to improve the basic algorithm as follows:

- Implements four other tonal rules and their domains of application.
- Extends the application of the tonal rules to other parts of speech.

The application of the algorithm developed by Raborife [2009] is restricted to verb stems and their preceding clitics. In our study, we extend the application of the extended algorithm to include all the other parts of speech such as nouns and adjectives.

- Treats each verb stem independently according to its tense when applying the tonal rules to verb stems.

  In Sesotho, a sentence can have more than one verb stem with the verb stems having different tenses. In such a sentence, the algorithm by Raborife [2009] applies the grammatical tone insertion rule followed by the iterative high tone spread rule to all the verb stems in the sentence if one of the verb stems satisfies the morphological contexts which allow for the application of these rules. Otherwise, the algorithm applies high tone spread rule within the clitic phrase boundaries that the verb stems are in. This led to mismatched tone labels in the analysis by Raborife [2009] prompting the suggestion that in a case where there is more than one verb stem in a sentence, each verb stem should be treated independently according to its tense.

The basic and the extended algorithms were tested on the 45 Sesotho sentences independently of each other.

1.6.2 Analysis

The transcribed tone labels of the actual recordings of the 45 sentences were used in analyzing the improvement of the extended algorithm from the basic one. We compared the transcriptions of the tone labels to the tone labels produced by the basic algorithm. Furthermore, the transcribed tone labels were compared to the tone labels output by the extended algorithm. The basic and the extended algorithm produced as output, the tone labels as predicted by the tonal rules they implemented.

Since our hypothesis assumed that the extended algorithm was an improvement on the basic algorithm, we measured improvement as the number of matched tone labels produced during our comparisons. If the number of matched tone labels between the basic algorithm and the transcriptions was less than the number of matched tone labels between the extended algorithm and the transcriptions, then our hypothesis was validated.

In addition, the transcribed tone labels were compared to the following:

- The tone labels on the syllables all marked as low.
- The underlying tone labels on the syllables as described in a Sesotho dictionary by Du Plessis *et al.* [1974].

1.7 Results

The results of our research study validated our research hypothesis. That is, the number of matched tone labels between the extended algorithm and the transcriptions was greater than the number of matched tone labels between the basic algorithm and the transcriptions. In addition, to statistically verify our results, we performed a t-test on the results we got from comparing the output of the extended algorithm with the transcription as well as the results from comparing the output produced by the basic algorithm with the transcriptions.

A t-test compares one variable between two groups. In our study, the variable is the tone label of each of the syllables in the 45 Sesotho sentences on which the two algorithms were tested. In our study, we used this test to analyze the experimental effect of improving the basic algorithm by implementing

the extended algorithm. The results of this test showed that our results are statistically significant, that is, the extended algorithm significantly improves the basic one.

The results of our analysis are summarized in Table 1.1.

Compare Transcribed Tone Labels with:	**Percentage of Matched Tone Labels (out of 565)**
Unmarked Tones	60.2%
Underlying Tones	64.7%
Basic Algorithm	64.6%
Extended Algorithm	72.4%

Table 1.1: Summary of results

1.8 Contribution of the Research Study

Accurate prosody prediction methods are essential in the development of TTS systems. Such predictions affect the naturalness of the synthesized speech produced by the system.

Our study involved predicting tone marks on the syllables of a word using linguistically-defined tonal rules. These tone marks can then be used to predict the pitch values using statistical means. However, this aspect is beyond the scope of our research study.

Our research study provides a first step to modeling tone for Sesotho. This step provides a solution that brings our local languages such as Sesotho up to par with other tonal languages in the development of TTS systems. Furthermore, it also provides a solution for implementing tone in TTS systems that is relevant to local Bantu languages.

1.9 Structure of the Document

- Chapter 2 provides the background literature which is the basis of the proposed research study. This chapter first presents the prosodic hierarchy as well as the tonal rules of Sesotho and their domain of application. Previous studies on tone in Sesotho and another related language are discussed in this chapter.
- Chapter 3 presents and discusses the research method that was used to conduct the research study. The context of the research is described. The research question and hypothesis are presented and motivation is given for them and their appropriateness to the study.
- Chapter 4 provides a detailed description of how the sentences which we used as input data were compiled and transcribed.
- Chapter 5 presents a description of how the data was analyzed and evaluated. Furthermore, the results of the research study are outlined in this chapter.
- Chapter 6 presents a summary of the research study as well as the major points presented in this document. Furthermore, it discusses suggestions for further research.
- Appendix A contains the glosses that are used in this report.
- Appendix B contains the 45 Sesotho sentences chosen for the testing of the algorithm.

- Appendix C contains the 9 minimal pairs that were used to choose the recorded speech for transcriptions.
- Appendix D presents the information sheet handed to the respondents of our research, the consent form as well as the language profile sheet which was used to verify the primary language of the speaker.
- Appendix E contains the transcribed sentences by the three labelers.
- Appendix F contains the final tone labels from the transcribers as determined by the majority vote.
- Appendix G contains output of the basic algorithm.
- Appendix H contains output of the extended algorithm.
- Appendix I presents the t-test table used in our statistical analysis.

Chapter 2

Related Background and Research

2.1 Introduction

"As information technology becomes increasingly pervasive in society, there is a strong desire to support local languages through technology" [Zerbian and Barnard 2008, p.248]. For speech technology, TTS systems are useful human language technology systems for use in illiterate communities with typical applications in health and education information dissemination [Gibbon *et al.* 2006]. Usually TTS system development environments adapt an existing voice in one language to a voice in another language [Gibbon *et al.* 2006]. This approach is only feasible when languages are prosodically and phonemically similar[1], such as Sesotho and Setswana, but is problematic when languages are not similar, such as Sesotho and English, since the prosodic information of the language will not be correctly implemented in the TTS system [Gibbon *et al.* 2006].

The main objectives in developing text-to-speech systems are for the synthesized speech to be intelligible and to sound as natural as human speech. In order for text-to-speech systems developed for Bantu languages to sound natural, tone needs to be implemented onto them. To implement tone, we need a speech corpus that has tone labels annotated onto it. A corpus is a structured set of texts usually used in hypothesis testing, checking occurrences or validating linguistic rules.

For Southern Bantu languages, it is difficult to annotate tone markings onto a speech corpus since in orthography, tone information is not included. Our study was concerned with improving an algorithm that predicts tone labels (which are either high or low) on the syllables of the words in a speech corpus. To make such predictions, the basic as well as the extended algorithm uses Sesotho tonal rules which are described by Khoali [1991].

This chapter provides some linguistic background on the tonal rules and a prior study that involves developing a tone label prediction algorithm.

2.2 The Tonal System of Sesotho

Sesotho is a tonal language; it uses pitch variations to convey grammatical or lexical meaning. According to Doke and Mofokeng [1957], the exact pitch of a tone in a word varies from speaker to speaker; it is the relation of the pitch of one syllable to that of the next that is constant with all speakers. Sesotho has only two tonal levels, namely the high tone (H) and the low tone (L)[2].

Sesotho is classified as a grammatical tone language with a restricted tonal system, that is, a system where not every syllable must be encoded for tone in the lexicon [Demuth 1995]. In other words, it is not

[1] 'Prosodically and phonemically similar' languages are languages that have the same speech melody.

[2] á: the diacritic on the vowel *a* is used to indicate high tone, low tones are not marked.

necessary to show low tone underlyingly but it is important to show high tone underlyingly as illustrated in Example 2.1 [Demuth 1995].

(2.1) ho-bóna - "to see"

Tone bearing units that end up with no tone specification at the surface are filled in with a rule of default low tone insertion [Demuth 1995]. That is, syllables which are not marked by any tone label end up with a low tone by default. According to Demuth [1995], if a verb stem is lexically marked for high tone, a high tone is predicted to be associated with the first syllable of the verb stem as exemplified in 2.1 (from Demuth [1995, p.18]) whereas if a verb stem is a low-toned verb stem, it will surface with a low tone on the first syllable as exemplified in 2.2 (from Demuth [1995, p.18]).

(2.2) ho-batla - "to want"

Tone in Sesotho may be used to differentiate meaning on a lexical level as shown in 2.3. The word *bona* has different tonal patterns in Example 2.3. It is this difference in tonal patterns that affects the meaning of the word.

(2.3) bóna - "see"
boná - "they"

Tone in Sesotho may also be used to show grammatical relationships. For instance, first person singular subject markers can only be distinguished from similar third person forms by means of tone [Doke and Mofokeng 1957]. The following example from Doke and Mofokeng [1957, p.39] illustrates this grammatical relationship:

(2.4) Ké motho - "It is a person"
Ke motho - "I am a person"

It is, however, important to note that not all words composed of identical phonemes[3] can be distinguished by tone as some are also identical in tone. For such words, only the context in which they are used clarifies the meaning. The following, from Doke and Mofokeng [1957, p.38] are examples of such words:

(2.5) likoto
"knob-kerries" or "pieces"

(2.6) hlolo
"hare" or "creation"

In sum, it is only the high tones that are explicitly specified on the syllables in the speaker's mental lexicon, and low tones appear when a syllable is tonally under-specified. It is also important to note that nasal sounds in Sesotho (*m, n, ng* and *ny*) form syllables on their own when they are not immediately followed by a vowel such as the first *n* in *monna ("man").* In that instance the nasal sounds can carry tone (*móna*). Nasal sounds are sounds that are pronounced with breath escaping mainly through the nose.

According to Zerbian [2006, p.147], "the behaviour of high tones is one of the most fundamental phenomena of Bantu tonology"[4]. For instance, high tones in Sesotho do not only surface on the syllables which they are associated with underlyingly, but also on succeeding syllables as discussed in Section 2.4. This behaviour does not occur at arbitrary places in the sentence but within certain phonological (prosodic) constituents which we discuss in Section 2.3.

First, however, we present the phonological constituents within which tonal rules apply. Then, we discuss the tonal rules with reference to the phonological constituents.

[3] A phoneme is a segmental unit of sound employed to form meaningful contrasts between utterances. An example of a phoneme is the /p/ sound in the words pot and spot.

[4] The study of forms and uses of tone in a language.

2.3 The Prosodic Hierarchy

The prosodic hierarchy describes a series of increasingly smaller regions of prosodic constituents as shown in Figure 2.1 [Nespor and Vogel 1986]. The constituents of the prosodic hierarchy are domains within which phonological rules apply. When a phonological (or tonal) rule is restricted within a certain domain, it may only apply if both the segments triggering the application of the rule and the segments undergoing the change are all included within that domain.

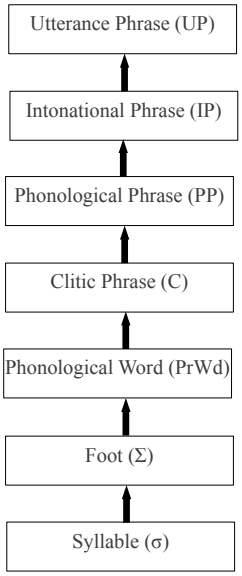

Figure 2.1: Prosodic hierarchy

The phonological constituents are important for our study since some tonal rules in Sesotho, such as the high tone spread rule and the iterative high tone spread rule as discussed in Section 2.4.1, are restricted within certain phonological constituents. If tonal rules, such as the high tone spread rule which involves the spreading of a high tone from the syllable it is associated with underlyingly to the adjacent syllable, were to apply throughout a sentence, one would expect that the high tone on the monosyllabic verb stem *ja* would spread to the first syllable of the noun stem *nama* as exemplified in 2.7[5].

(2.7) *(ke já)ϖ náma
1SG.SM eat meat
"I eat meat"

However, this is not the case. Hence an asterisk is inserted to show that the tonal behaviour of the high tones in the sentence is incorrect. In Example 2.7 there is no other explanation as to why the high

[5]Numbers indicate the grammatical person (for example, 1 indicates first person) and underlying high tones are underlined. Between the Sesotho sentence and its English translation is a gloss that describes the grammatical categories of the words in the sentence. A gloss is a brief notation of the meaning of the word (refer to Appendix A for a list of glosses that are used in the document). ϖ refers to the phonological word domain which is discussed in detail in Section 2.3.1.

tone on the verb stem does not spread to the first syllable of the noun stem *na* other than to refer to the prosodic constituent which in this case is the phonological word as discussed in Section 2.3.1.

The phonological word, clitic phrase and the phonological phrase are the only prosodic constituents that are of relevance in our study since the tonal rules which form the basis of this study are restricted within these three constituents. Therefore, we only discuss these three constituents.

2.3.1 The Phonological Word

The phonological word (ϖ) is a domain for phonological processes such as HTS1 and HTS2 in Sesotho as discussed in Section 2.4.1.

This domain in Shona, a Bantu language spoken primarily in Zimbabwe, consists of a full word with all the function words preceding it as shown in Example 2.8 (from [Myers 1987, p.123-124]). Full words are verbs, nouns, adverbs, determiners, adjectives and pronouns in Shona [Myers 1987]. Function words are prepositions, class prefixes, inflectional prefixes and auxiliary verbs [Myers 1987].

(2.8) (né-rwenyú)ϖ (rúnako)ϖ
"with-your beauty"

In Sesotho, this domain is a bit more complex. The domain of the phonological word in Sesotho is different for monosyllabic verb stems and polysyllabic verb stems.

A monosyllabic verb stem and the prefix to the left of the verb stem make up one phonological word domain as shown in Example 2.9 (from Khoali [1991, p.216]) where the verb stem *ja* is in the same phonological word domain as the object prefix *ée*.

(2.9) ó ya (ée já)ϖ
3SG.SM T/A OM eat
"S/he eats it"

Polysyllabic verb stems form a phonological word domain on their own as exemplified in 2.10 (from [Khoali 1991, p.224]). In Example 2.10, the verb stem *tsamaya* is in a separate phonological word domain from the future marker *tlá*.

(2.10) ó tlá (tsamaya)ϖ
3SG.SM FUT go
"S/he will go"

According to Khoali [1991], in Sesotho there are particles whose high tones spread to the next syllable as a consequence of the high tone spread rule (HTS1) and those that do not. A particle whose high tone spreads as a result of the high tone spread rule forms a phonological word with the noun class prefix of the noun stem following it, as illustrated in Example 2.11 (from [Khoali 1991, p.126]). In Example 2.11, the particle *ho* belongs to the same phonological word domain as the noun prefix *mo*.

(2.11) (hó mó)ϖ-(rena)ϖ
PREP NP1-king
"to the king"

A particle whose high tone does not spread forms its own phonological word domain as exemplified in 2.12 (from [Khoali 1991, p.126]), where the particle *ke* belongs to a different phonological word from the noun class prefix of the noun stem following it.

(2.12) (ké)ϖ (morena)ϖ
1SG.SM king
"It is a king"

In sum, content words in Sesotho form their own phonological word domain. There is a special case for noun stems where the following holds: if the noun stem is preceded by a particle that spreads by the high tone spread rule, the class marker of the noun stem and that particle form a single ϖ-domain as exemplified in 2.11. Content words in Sesotho include nouns, verbs, adjectives and adverbs.

2.3.2 The Clitic Phrase

The clitic phrase (C) is a domain for phonological processes such as HTS1 in Sesotho as discussed in Section 2.4.1. In Sesotho the clitic phrase consists of the verb stem and all the clitics (prefixes) to the left of the verb stem [Khoali 1991]. A clitic is a word that cannot stand on its own and has to attach to another word (host), such as the English clitic *s*, as in *"Mpho's coming"*.

Example 2.13 (from Khoali [1991, p.216]) presents a polysyllabic verb stem (*jésa*) forming its own phonological word and forming a clitic phrase with all the prefixes to the left of it. This illustrates that the clitic phrase is a higher domain than the phonological word and that the phonological word is contained entirely within the clitic phrase.

(2.13) (ó yá é (jésa)ϖ)C
3SG.SM T/A OM feed
"He feeds it"

In Example 2.14 (from [Khoali 1991, p.292]), the verb stem (*paqáma*) is not preceded by any clitics hence the phonological word is equivalent to the clitic phrase (which in this case is equal to a morphological word, which is the verb stem).

(2.14) ((paqáma)ϖ)C
lie down
"lie down!"

The clitic phrase domain is only necessary for verb stems and other parts of speech such as nouns and adjectives do not make reference to this domain [Khoali 1991]. This is shown in Example 2.15 where the noun *pódí* is not contained within any clitic phrase and the verb stem *rátá* is in the same clitic phrase domain as the subject marker that precedes it.

(2.15) (ke (rátá)ϖ)C (pódí)ϖ (hahólo)ϖ
1SG.SM like goat ADV.much
"I like a goat too much"

The clitic phrase domain in Sesotho is made up of a verb stem and all the function words preceding that verb stem. In Section 2.4.1, we discuss the application of HTS1 outside of the phonological word, that is, within the clitic phrase.

2.3.3 The Phonological Phrase

The phonological phrase (ϕ) is a domain for phonological processes such as the finality restriction rule in Sesotho, discussed in Section 2.4.2. In Sesotho, the domain includes the phrasal head (for a noun phrase, the phrasal head would be a noun) and whatever precedes it within the same phrase [Nespor and Vogel 1986].

Khoali [1991] uses the concept of Chomsky-adjoined phrases to define the boundaries of the phonological phrase domain. Chomsky-adjoined phrases are phrases that are not needed for their heads to "make sense". Post-nominal modifiers such as adjectives are regarded as Chomsky-adjoined since they

are not needed for the noun to "make sense". Post-nominal modifiers are optional phrases or words placed after nouns whose removal from the sentences does not render the sentences ungrammatical as shown in Example 2.16 (from Khoali [1991, p.67]). In Example 2.16(a), the adjectival phrase *é kgóló* can be removed without affecting the grammaticality of the sentence as illustrated in Example 2.16(b).

(2.16) (a) Pódi **é kgóló** é já mabelé
"A big goat eats sorghum"

(b) Pódí é já mabelé
"A goat eats sorghum"

According to Khoali [1991], Chomsky-adjoined phrases are in different phonological phrase domains than their heads. We will refer to Chomsky-adjoined phrases as modifiers in our study.

The right edge of a phonological phrase domain is inserted between a head and its modifier as exemplified in 2.17[6].

(2.17) (a) Ke ráta)ϕ haholo ho tsamaya
1SG like ADV.much INF go
"I like to travel a lot"

(b) Pódi)ϕ é kgóló é já mabelé
goat SC9 ADJ.big SC9 eat sorghum
"A big goat eats sorghum"

In Example 2.17(a) (from Khoali [1991, p.196]), the verb stem *ráta* is followed by an adverb *haholo*. Adverbs are modifiers since they are not needed for the verb stem to make sense – removing them does not make the sentence ungrammatical. Hence, in Example 2.17(a), a right edge of a phonological phrase domain is inserted immediately after the verb stem.

In Example 2.17(b) (from Khoali [1991, p.67]) the noun *pódí* is followed by an adjectival phrase *é kgóló*. As discussed above, this phrase is a post-nominal modifier, thus in Example 2.17(b), a right edge of a phonological phrase domain is inserted immediately after the noun before the adjectival phrase.

The right edge of the phonological phrase is also inserted on the last word on the right edge of a sentence. This sentence can be either simple, compound or complex as exemplified in 2.18[7] (from Khoali [1991, p.57]).

(2.18) (a) Thabo ó rátá pódi)ϕ
Thabo SC1 like goat
"Thabo likes a goat"

(b) Thabo ó rátá pódí mmé ó bátlá bá jé yoná féla)ϕ
Thabo SC1 like goat CONJ SC1 OC2 eat PRN ADV.only
"Thabo likes goat and wants them to eat it only"

(c) Thabo ó bátlá pódí hore ré tló e hlába)ϕ
Thabo SC1 wants goat that 2SG FUT OC9 kill
"Thabo wants the goat so that we slaughter it"

According to Khoali [1991], constituents which are to the left of the head of a phrase such as topicalized nouns, as exemplified in 2.19 (from Khoali [1991, p.66]), are within the same phrase which they

[6] In Example 2.17(a), the original English translation for the sentence is given in Khoali [1991] as "I like to go a lot".

[7] In Example 2.18(b), the English translation from the original text is given as "Thabo likes mutton and wants them to eat it only".

specify. Topicalization refers to placing the topic at the beginning of the sentence. In Example 2.19, the noun stems *pó̲di* and *phó̲folo* are in the same phonological phrase domain even though the second noun stem *phó̲folo* specifies the first noun stem *podi*.

(2.19) P̲ódí phó̲fóló̲ é̲ jé̲lé mabelé̲)ϕ
Goat animal SC9 eat.PERF NP6.sorghum
"goat, the animal has grazed sorghum"

The left edge of the phonological phrase boundary is inserted at the following positions by default:

- On the leftmost edge of a sentence. This sentence can be either compound, simple or complex as exemplified in 2.20 (from [Khoali 1991, p.57]).

  (2.20) (Thabo ó̲ rá̲tá pó̲di mmé̲ ó̲ bá̲tlá bá̲ jé̲ yoná̲ fé̲la)ϕ
  Thabo SC1 like goat CONJ SC1 OC2 eat PRN ADV.only
  "Thabo likes goat and wants them to eat it only"

- Whenever there is a right edge, a left edge is inserted immediately after the right edge as exemplified in 2.21 (from [Khoali 1991, p.43]). In Example 2.21, the left edge is inserted immediately after the right edge (after the noun stem *lekala*).

  (2.21) (Ke na le lekala)ϕ (le letle)ϕ
  1SG have NP5.branch SC5 nice
  "I have a nice branch"

In this section, we have shown that the right edge of the phonological phrase domain is between a phrasal head and its modifier as well as at the end of a sentence. By default, the left edges of a phonological phrase are inserted at the beginning of a sentence as well as immediately before a right edge.

2.4 The Phonological Rules

This section discusses the tonal rules that are implemented in the extended algorithm.

Section 2.4.1 discusses the three Sesotho rules which formed the basis of the study by Raborife [2009], namely HTS1, HTS2 and GTI. We also discuss the domains of application of the tonal rules with reference to the prosodic constituents discussed in the previous section.

Section 2.4.2 discusses the four tonal rules that have been added with the aim of improving the basic algorithm. These rules are also discussed with reference to their domains of application.

2.4.1 Tonal Rules Implemented in the Previous Study

In this section we discuss the three tonal rules which are implemented in the basic algorithm.

High Tone Spread

High tone spread (also referred to as High Tone Doubling in some literature [Demuth 1995]) involves the spreading of an underlying high lexical tone to the adjacent syllable as exemplified in 2.22 –2.24 from Demuth [1995, p.13].

Example 2.22 shows a high tone on the first syllable of the high-toned verb stem *ré̲* spreading to the second syllable of the verb stem *ké*. We can only account for this behaviour by referring to the high tone spread rule.

(2.22) ke ré̲kéla...
1SG.SM buy
"I am buying for.."

Example 2.23 shows a high tone on the subject marker ó̲ spreading to the tense and aspect marker *á*.

(2.23) ó̲ á lema
3SG.SM T/A plough
"S/he is ploughing"

In Example 2.24, the high tone from the subject marker ó̲ spreads to the first syllable of the verb stem *lé*. The verb stem *lema* surfaces as a high-toned verb stem in Example 2.24 and a low-toned verb stem in Example 2.23. We can thus conclude that this verb stem is a low-toned verb stem and in Example 2.24 it acquires the high tone on the first syllable due to HTS1.

(2.24) ó̲ léma...
3SG.SM plough
"S/he ploughs..."

According to Khoali [1991], HTS1 applies within the domain of the phonological word as exemplified in 2.25 (from Demuth [1995, p.13]) and also within the domain of the clitic phrase as shown in Examples 2.26–2.27 (from Demuth [1995, p.13]).

In Example 2.25, the initial high tone of the verb stem spreads to the next syllable since the polysyllabic verb stem forms a phonological word as defined by Khoali [1991].

(2.25) (ke (ré̲kéla)ϖ)C...
1SG.SM buy...
"I buy for..."

In Example 2.26, the high tone in the subject marker ó̲ spreads to the first syllable of the verb stem *le* since the subject marker and the verb stem are within the same clitic phrase. Examples 2.25–2.26 serve as evidence that HTS1 applies within the clitic phrase since the high tone on the subject marker spreads to the first syllable of the verb stem.

(2.26) (ó̲ (léma)ϖ)C...
3SG.SM plough
"S/he ploughs..."

The tone in the tense and aspect marker *a* surfaces as a high tone in Example 2.27 because the high tone on the subject marker ó̲ spreads to the tense and aspect marker. This behaviour is expected by the high tone spread since the subject marker and the tense and aspect marker are within the same clitic phrase.

(2.27) (ó̲ á (lema)ϖ)C
3SG.SM T/A plough
"S/he is ploughing"

In this section, we show the application of the high tone spread rule within the phonological word as well as the clitic phrase domain.

Grammatical Tone Insertion

The grammatical tone insertion rule associates the second syllable of a verb stem with a high tone as exemplified in 2.28 [Khoali 1991].

(2.28) (ha bá (batlé)ϖ)C batho
NEG 3PL.SM want people
"They do not want people"

According to Khoali [1991] it occurs in the following contexts:

- The imperative mood - this mood is discussed in detail in this section.
- The negative verb form - forms that indicate a negative of the verb.
- The participial narrative past tense - indicates a sequence of past events which are related by a speaker and is characterized by a sequence of verbs.
- The habitual verb form - expresses an action that happens often (habit).
- The perfect tense - used to express action that has been completed.
- The subjunctive mood - expresses an action that might occur.

In this document we discuss the imperative, the other contexts described above work accordingly. For the morphological formations of the above-mentioned contexts, the interested reader may refer to Doke and Mofokeng [1957, p.184–240].

The imperative The imperative mood expresses an instruction, command or a politely strong request such as *thola!* (keep quiet!) in Sesotho. In Sesotho, this mood always surfaces with a high tone on the second syllable as presented in Example 2.29 from [Khoali 1991, p.291]. Example 2.29 shows a high tone surfacing on the second syllable of an underlyingly low-toned verb stem. This high tone is referred to as the grammatical tone.

(2.29) batlá
want
"want!"

The verb stem *batla* surfaces without a high tone on its second syllable in a declarative sentence as shown in Example 2.30. However, in the imperative, the verb stem *batla* has a high tone surfacing on the second syllable as exemplified in 2.29. In Example 2.30, the high tone on the first syllable of the verb stem *bátla* is acquired by the high tone spread rule.

(2.30) (ó (bátla)ϖ)C nama
3SG.SM want meat
"S/he wants meat"

Examples 2.29 and 2.30 illustrate that the grammatical tone insertion rule occurs in the imperative mood. There is no need to refer to prosodic constituents in the application of the grammatical tone insertion rule since this rule applies exclusively to verb stems in specific morphological contexts.

Iterative High Tone Spread

The iterative high tone spread rule involves the spreading of a grammatical high tone iteratively within the phonological word domain as shown in the examples below.

In Example 2.31, a grammatical high tone surfaces on the second syllable of the low-toned imperative verb stem *kgarákgárámétsáng* as discussed above, this high tone then spreads iteratively until the end of the verb stem. The verb stem in Example 2.31 is polysyllabic and since we know that polysyllabic verb stems in Sesotho form their own phonological word, we can conclude that the high tone in Example 2.31 spreads iteratively within the phonological word domain.

(2.31) ((kgarákgárámétsáng)ϖ)C [Khoali 1991, p.293]
push push.PL
"push" "push"

Example 2.32 presents a grammatical high tone surfacing on the second syllable of a low-toned verb stem *kgarámétsé* in the negative form. Since we know that a grammatical high tone surfaces on verb stems in the negative form, this grammatical high tone spreads iteratively within the phonological word domain.

(2.32) (ha ké (kgarámétsé)ϖ)C motho [Khoali 1991, p.248]
NEG 1SG.SM push a person
"I do not push a person"

Example 2.33 shows a high-toned verb stem *rékísé* that surfaces with a high tone on the second syllable and the final syllable. We know that lexical tones spread only to the adjacent syllable (by the high tone spread rule) and not iteratively. We can conclude that the high tone on the second syllable is a grammatical high tone and that it spreads to the following syllables until the end of the phonological word domain within which the verb stem is.

(2.33) (ha ké (rékísé)ϖ)C nama [Khoali 1991, p.248]
NEG 1SG.SM sell meat
"I do not sell meat"

We can thus conclude that HTS2 is ordered after the grammatical tone insertion rule since for a grammatical high tone to spread iteratively it has to be first associated with the second syllable of the verb stem in the contexts mentioned in the section above. The data presented in this section also shows that HTS2 applies strictly within the phonological word domain.

The analysis of the basic algorithm was restricted to polysyllabic verb stems as opposed to monosyllabic verb stems due to the complex tonal behaviour of monosyllabic verb stems. For instance, the high tone of the object prefix does not spread onto the adjacent monosyllabic verb stem as expected by the high tone spread rule [Khoali 1991]. Therefore, we also exclude monosyllabic verb stems from our study.

2.4.2 Tonal Rules Implemented in this Study

In this section we discuss the four tonal rules which we implemented in order to improve the algorithm developed by Raborife [2009]. These tonal rules are the right branch delinking rule, left branch delinking rule, specifier high tone delinking rule and the finality restriction rule.

Right Branch Delinking Rule

The right branch delinking rule disassociates the right branch of a multiply-linked high tone if, and only if, there is a high tone immediately after the target [Khoali 1991]. A multiply-linked high tone is a high tone that is linked to more than one syllable after the application of a tone spread rule such as HTS1, where the high tone of the source syllable will be linked to both the source syllable and the target syllable.

Consider a scenario whereby HTS1 spreads the underlying high tone from the syllable with which it is associated (say σ_i) to the adjacent syllable (say σ_{i+1}). If the syllable σ_{i+2} has a high tone, then the high tone which has been spread to σ_{i+1} will be "delinked" and σ_{i+1} will have a low tone as described by the right branch delinking rule.

This is shown in Example 2.34 (from Raborife [2009, p.43]) where the present tense marker *a* surfaces with a low tone instead of a high tone as we would expect after the application of HTS1 from the subject marker *bá* to the present tense marker *a*. The application of the high tone spread rule results in the high tone of the subject marker being multiply-linked to both the subject marker and the present tense marker. However, the present tense marker is followed by a high-toned verb stem *ága*, thus the high tone is delinked from the present tense marker.

(2.34) **Bá a ága** banna a mótse **bá a ága**.
3PL.SM PRS build men AGR village 3PL.SM PRS build
"The village men are building"

Example 2.35 (from Khoali [1991, p.143]) shows the application of this rule within the phonological word domain. In Example 2.35, the high tone on the second syllable *be* of the noun *lebelete* spreads to *le*, the third syllable of the noun. The high tone on the third syllable of the noun is delinked since it is linked to the second and third syllable of the noun and the last syllable is high-toned.

(2.35) (lebéleté)ϖ
NP5.untamed_animal
"untamed animal"

Example 2.36 (from Khoali [1991, p.225]) shows the application of this rule within the clitic phrase domain. In Example 2.36, the past tense marker *ne* surfaces with a low tone. This marker is underlying low-toned and since it is preceded by a high-toned subject marker *bá*, by the high tone spread rule, the high tone on the subject marker spreads to the past tense marker. This leads to the high tone of the subject marker being multiply-linked to both the subject marker and the past tense marker. However, the past tense marker is followed by a high-toned subject marker *bá*. This allows for the application of the right branch delinking rule which delinks the high tone on the past tense marker. Thus, the past tense marker surfaces with a low tone.

(2.36) (Bá ne bá (tsamaya)ϖ)C
3PL.SM PST 3PL.SM walk
"They were walking"

The evidence provided in this section shows that the right branch delinking rule is a ϖ-domain as well as C-domain span rule.

Left Branch Delinking Rule

The left branch delinking rule delinks the left branch of a multiply-linked high tone if it is preceded by a high tone [Khoali 1991].

For instance, HTS1 spreads the high tone from one syllable with which it is associated (say σ_i) to the adjacent syllable (say σ_{i+1}). If the syllable σ_{i-1} has a high tone, by the left branch delinking rule, the high tone on σ_i will be delinked. Thus, σ_i will have a low tone.

In Example 2.37 (from [Khoali 1991, p.254]), the high tone on the second syllable of the verb stem *kgurukgurumetse* surfaces with a low tone. This syllable is associated with a high tone by GTI since the verb stem is in the negative form. This high tone then spreads by HTS2 until the penultimate syllable. The high tone on the second syllable of the verb stem is delinked by the left branch delinking rule because the first syllable is high-toned.

(2.37) ha ké kgúrukgúrúmétsé
NEG 1SG.SM cover.REDUPL.NEG
"I don't cover a bit"

This rule is a phonological word domain span rule as exemplified in 2.38. The verb stem in Example 2.37 is a reduplicated form of the verb stem in Example 2.38. Reduplication refers to the repetition of a root or stem of a word. Therefore, the explanation of the application of LBD in Example 2.37 holds for Example 2.38. In Example 2.38, all the syllables involved in the application of this rule are within the same verb stem. In Sesotho, polysyllabic verb stems form their own phonological word.

(2.38) ha ké (kgúrumétsé)ϖ
NEG 1SG.SM cover.NEG
"I don't cover a bit"

Specifier High Tone Delinking Rule

The specifier high tone delinking rule delinks a high tone on an object concord or reflexive prefix if they are not within the same phonological word as the stem [Khoali 1991]. This rule only delinks the high tone on an object (or reflexive) concord after the high tone spread rule has applied. It delinks the high tone on a reflexive or object concord regardless of whether it is multiply-linked or not. This is a major difference of this rule from the left branch delinking since for the left branch delinking rule to apply, the high tone on the source syllable should be multiply-linked to the source and the target syllable.

Example 2.39 (from [Khoali 1991, p.274]) shows an underlyingly high-toned object prefix *mo* surfacing with a low tone. The high tone on the object prefix spreads by HTS1 to the first syllable of the verb stem *kgarameditse*. The high tone on the object prefix then delinks since it is not in the same phonological word domain as the verb stem.

(2.39) ó mo kgárámédítsé
3SG.SM OC1 push
"He pushed him/her"

Finality Restriction Rule

The finality restriction rule exempts syllables at the end of a phonological phrase from the application of tonal rules [Khoali 1991]. In Example 2.40 (from [Khoali 1991, p.38]), the high tone on the first syllable of the verb stem *reka* does not spread to the second syllable as expected by the high tone spread. This high tone is blocked by the finality restriction rule.

(2.40) ((ho (réka)ϖ)C)ϕ
INF buy
"to buy"

In Example 2.41, the second syllable of the verb stem *ré̲ka* surfaces with a high tone. This is due to the fact that in this example, the second syllable of the verb stem is not at the end of the phonological phrase domain but the application of HTS1 is blocked within the noun stem *bú̲ka*. As expected by HTS1, the second syllable of the noun stem *ka* should surface with a high tone as the first syllable of the noun stem *bú̲* has a high tone. Nevertheless, the target syllable *ka* is at the end of the phonological phrase domain, hence by the finality restriction rule, this syllable surfaces with a low tone.

(2.41) ((ho (ré̲ká)ϖ)C (bú̲ka)ϖ)ϕ
INF buy NP.book
"to buy a book"

In Northern Sotho and Sesotho, underlying high tones are allowed to surface in phrase-final positions [Khoali 1991; Zerbian 2007] as exemplified in 2.42 (from Khoali [1991, p.48]). In Example 2.42, the high tone on the noun *lehé̲* is at the end of the phonological phrase domain. The high tone on the final syllable of the noun stem is retained and not delinked because it is underlying.

(2.42) (ó̲ na lé̲ (leh é̲)ϖ)ϕ
3SG.SM have egg
"He has an egg"

2.4.3 Summary

The two preceding sections explain in detail the tonal rules that formed the basis of the study by Raborife [2009] as well as the tonal rules that are implemented in the extended algorithm. Some of these tonal rules need other tonal rules to apply before they can apply. This phenomenon is called feeding. If rule A feeds from rule B, then rule B needs to apply in order to create a new context in which rule A can apply. If rule B does not apply then rule A cannot apply.

An example of a feeding relationship is between the grammatical tone insertion rule and the iterative high tone spread rule. As shown in Example 2.43, a grammatical tone surfaces on the second syllable *rá* of the imperative verb stem *kgarákgárámétsáng*. This grammatical high tone then spreads by the iterative high tone spread until the penultimate syllable *tsá*. The grammatical high tone does not spread to the final syllable of the verb stem *ng*, since this syllable is at the end of the phonological phrase domain. Thus, by the finality restriction rule, it is exempt from the application of the tonal rules.

(2.43) (((kgarákgárámétsáng)ϖ)C)ϕ
push push.PL
"push" "push"

Example 2.43 also shows that the finality restriction "feeds" from the iterative high tone spread rule. If the iterative high tone spread rule did not apply in Example 2.43, then the finality restriction rule would not apply on the final syllable of the verb stem which is at the end of the phonological phrase domain. This example shows that the iterative high tone spread rule "feeds" from the grammatical tone insertion rule and in turn, the iterative high tone spread rule creates a context in which the finality restriction rule applies.

The left branch delinking rule, the right branch delinking rule and the finality restriction rule all "feed" from the high tone spread rule as discussed in the explanation of these tonal rules in Section 2.4.2. The high tone spread rule has to apply before they can apply. Khoali [1991] argues the following rule order for the tonal rules that have been added to the algorithm implemented by Raborife [2009].

1. High tone spread

2. Finality restriction

3. Specifier high tone delinking

4. Right branch delinking

5. Left branch delinking

The basic algorithm predicts tone labels on the syllables of a word using linguistically-defined tonal rules. This algorithm is the first step and the basis of tone modeling in TTS systems developed for Bantu languages such as Sesotho. In the following section we discuss the basic structure of TTS systems and the importance of tone modeling in these systems.

2.5 Text-to-Speech Systems

Text to Speech systems are built to convert given text to a speech waveform that ideally sounds natural to native speakers of the target language. These systems are useful in illiterate communities, such as rural areas in South Africa, where accessing information is difficult. Such systems are essential in the developing world, where there are many such communities and where the importance of information dissemination by voice is often greater than in literate communities.

The main challenge in developing text-to-speech systems is to build speech systems that sound as natural as human speech. According to Zerbian and Barnard [2008], for TTS systems developed for tonal languages, reliable tone and intonation models are of great importance since the intelligibility of the system depends heavily on them.

TTS systems usually have three modules:

- The text processing module
- The prosody prediction module
- The speech production module

The text processing module converts the textual input into a format suitable for phonetization[8] [Louw 2008]. This module is responsible for text normalization by converting abbreviations, numbers and symbols into their full word forms.

The prosody prediction module generates tone, stress, duration and the likes. Some prosodic features such as word stress in English can be modeled implicitly from recorded data, but in other cases, like Sesotho tone, explicit models must be created [Black 2006]. This is due to the complexity of the behaviour of tone in Bantu languages, thus making it difficult to model tone based solely on statistical methods [Louw *et al.* 2005; Kuun *et al.* 2005; Govender *et al.* 2005b; Govender 2006]. The processes that occur in the text processing module and the prosody prediction module are collectively referred to as *Natural Language Processing* (NLP).

The speech production module accepts as input the data generated in the NLP Stage and converts it to synthetic speech. This stage is referred to as the *Digital Speech Processing* (DSP) stage. Louw [2008] classifies the functional blocks of a TTS system as illustrated in Figure 2.2.

[8]Process of representing sounds by phonetic symbols, for example, the phonetic symbol for the sound $< th >$ (”thigh” in English) is θ in the IPA (International Phonetic Alphabet).

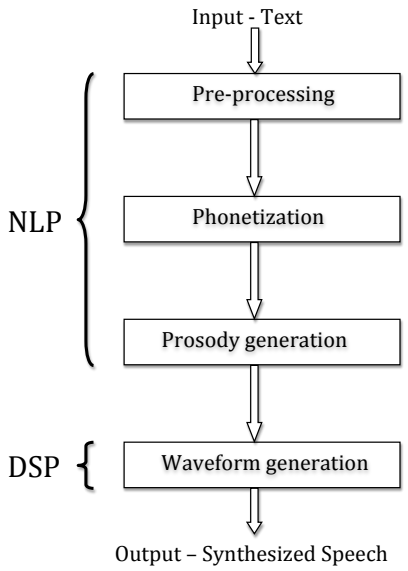

Figure 2.2: Functional blocks of a TTS system

2.6 Tone Modeling Attempt in a Bantu Language TTS System

2.6.1 Introduction

Although in recent years there have been studies focusing on tone modeling for African tonal languages such as Ibibio (spoken in Nigeria) and Yoruba, we only know of one published attempt to implement the tone in a Bantu language TTS system, namely the isiZulu TTS system as described in Louw *et al.* [2005]. In this section we discuss this attempt at tone modeling.

2.6.2 An Attempt at Tone Implementation in an isiZulu TTS System

To model tone into their TTS system, Louw *et al.* [2005] recorded isiZulu sentences using native speakers of the language. They then analyzed intonation patterns pronounced by an isiZulu speaker producing continuous sentences using a speech analysis software. Louw *et al.* [2005] expected that accurate production of appropriate pitch contours would be an important element of an intelligible system.

After analyzing the most noticeable intonation patterns produced in the sentences by the speaker, it became clear that tone production in natural isiZulu speech is less predictable from syntactic considerations than in a language such as Ibibio [Louw *et al.* 2005].

According to Louw *et al.* [2005], in their study, linguistic features such as syllabic stress (prominence of a syllable) and boundary tones[9] were not explicitly conveyed. This validates the statement by Govender [2006] that at a closer inspection, the instinctive notion that tone is expressed in the fundamental frequency of an utterance does not hold.

[9]Boundary tones are tones that occur on the edge-most syllable of a domain.

Fundamental frequency (F0) is the pitch of speech. According to Zerbian and Barnard [2008], fundamental frequency is considered a correlate for intonation and tone, both in studies on tone languages such as the Southern Bantu languages and in studies on stress languages such as English. Furthermore, in the acoustic studies available on Bantu tone, F0 variations have been taken as the sole indicator for tone [Zerbian and Barnard 2008].

According to Louw *et al.* [2005], accurate prediction of the marking of tones in the syllables was beyond their capabilities, and the consensus of both isiZulu speakers and system developers was that treating all words as "unmarked" would produce more understandable speech. Thus currently, the isiZulu TTS system developed by Louw *et al.* [2005] only has the sentence-penultimate falling tone (HL-tone) implemented into it.

Louw *et al.* [2005] suggest that in order to model tone, an algorithm that predicts which of the syllables of a word are to be marked for a high tone and also to predict the appropriate intonation of the words is needed.

2.7 Previous Study on a Tone Mark Prediction Algorithm

Raborife [2009] validated the suggestion by Louw *et al.* [2005] that an algorithm which predicts the tone labels on the syllables is necessary for tone modeling in a language such as isiZulu and Sesotho. The common denominator of these two languages is that they do not mark tone in orthography.

The aim of the study by Raborife [2009] was to develop a tone prediction algorithm. The algorithm by Raborife [2009] used tonal rules to make the predictions regarding the tone of the syllables. The research study by Raborife [2009] involved the development and description of a linguistically-motivated algorithm for three tonal rules of Sesotho, namely, high tone spread, iterative high tone spread and the grammatical tone insertion rule. The application of this algorithm was restricted to the clitic phrase domain containing polysyllabic verb stems.

2.7.1 Text Corpus Compilation

Fifteen Northern Sotho sentences were chosen from a large corpus of Northern Sotho data [Raborife 2009]. The 15 Northern Sotho sentences were chosen in such a way that they show the following occurrences [Raborife 2009]:

- The present principal tense.
- The present participial tense.
- The perfect tense.
- The imperative mood.

These tenses are relevant for the application of the basic algorithm. Some sentences which have long verb stems were selected as they allow the tonal rules to show their full consequence even though the preferred verb stem size in Bantu languages is in general two syllables [Raborife 2009].

The basic algorithm was formulated on the basis of the literature on tone rules available for Sesotho [Khoali 1991]. However, the sentences for testing and evaluation were taken from a corpus of Northern Sotho [Raborife 2009]. Although it would have been preferable to use sentences either from a corpus of Sesotho or to extract tone rules from the literature on Northern Sotho, this was not possible for the following reasons [Raborife 2009]:

- The reason for choosing Northern Sotho sentences was that prerecorded and adequately prepared sentences were not available in Sesotho. The prerecorded sentences were needed for evaluating our algorithm. Collecting and preparing these sentences for Sesotho would have gone beyond the time frame of the research study.

- A Northern Sotho literature could have been used as the basis of our research. However, the only source for tone rules in Northern Sotho is by Lombard [1976], a doctoral thesis written in Afrikaans. Thus it is not accessible to non-Afrikaans speakers. Having it translated would have gone beyond the time frame allocated to the research study.

A previous study has shown that Northern Sotho and Sesotho do not differ from each other in terms of the tonal rules investigated in our research study [Zerbian and Khoali 2009]. Thus, using Northern Sotho sentences does not have any negative impact on the results of the research study by Raborife [2009].

The recordings of the fifteen sentences were transcribed by three independent labelers. The labelers had different background and experience with respect to tone in Bantu languages and were all sensitive to tonal issues [Raborife 2009]. The labeled sentences were then compared across all three labelers and the labelers reached a unanimous agreement on 72.3% of the cases (196 out of 271 tone-bearing syllables). A final transcription was generated based on majority votes in the case of unanimous labels [Raborife 2009].

2.7.2 An Explanation of the Basic Algorithm

The algorithm takes as input a sentence with the underlying high tones on the syllables, grammatical categories of the words, and the tense, mood and aspect of the verbs tagged.

The algorithm assumes that there are no monosyllabic verb stems in the sentence since monosyllabic verb stems were not the focus of our research study due to their complex tonal behaviour. If there are any monosyllabic verb stems in the sentence, the basic algorithm does not implement the tonal rules on the verb stem. The algorithm inserts the prosodic boundaries around the verb stem.

The algorithm starts by inserting the prosodic boundaries (the clitic phrase boundary and the phonological word boundary). The algorithm inserts the prosodic domains as follows:

1. Inserts the clitic phrase boundary from the verb stem to the left until it finds a subject marker.

2. Inserts the phonological word boundary around the verb stem within the clitic phrase.

The insertion of the prosodic boundaries needs to precede the application of the rules since HTS1 and HTS2 are sensitive to the phonological word and the clitic phrase domains.

If there is a monosyllabic verb stem in the sentence, the algorithm inserts the clitic phrase boundary just as it does for polysyllabic verb stems. The phonological word boundary includes the verb stem and the agreement marker that precedes the verb stem as discussed in Khoali [1991] in contrast to polysyllabic verb stems which form their own phonological word domains.

Then, the algorithm checks the morphological context of the verb stem, that is, the tense, mood and aspect (TMA). If the verb falls within one of the morphological contexts necessary for the application of the grammatical high tone insertion rule, the algorithm applies the grammatical high tone insertion rule, followed by the iterative high tone spread rule. Otherwise, the algorithm applies the high tone spread rule.

The algorithm applies the most specific rule first, that is, the grammatical tone insertion rule. The reason for this rule ordering is that the grammatical tone insertion rule occurs in specific contexts; if none of its contexts is met, the algorithm applies the high tone spread rule.

In the case in which there is more than one verb stem in the sentence, the algorithm checks if any of the verb stems provides the context that allows for the application of the grammatical tone insertion rule. If yes, then the algorithm applies the grammatical tone insertion rule followed by the iterative high tone spread rule to all the verb stems in the sentence. Otherwise, the algorithm applies the high tone spread rule within the clitic phrase boundaries within which each verb stem is.

Our algorithm is shown in Algorithm 1.

Algorithm 1 Basic tone labeling algorithm

procedure TONE(S) ▷ S is the input sentence
  $Locate\ V$ ▷ Find the all the verb stems in the sentence.
  $Insert\ the\ C-phrase\ boundary\ from\ V\ to\ the\ left\ until\ we\ find\ SM\ or\ AGR$
  $number\ of\ \sigma\ in\ V \leftarrow \alpha$
  if $\alpha = 1$ **then** ▷ Inserts the phonological word boundary.
    $Insert\ the\ \varpi-boundary\ from\ V\ till\ preceding\ AGR$
  else
    $Insert\ the\ \varpi-boundary\ around\ V$
  end if
  for all V **do**
    $Check\ TMA$
    if $TMA = PERF\ or\ NEG\ or\ HAB\ or\ PST\ or\ IMP\ or\ SUBJ$ **then**
      for all V **do**
        $tone\ \sigma_2\ verb\ stem = H$ ▷ Grammatical Tone Insertion Rule
        $Spread\ the\ H\ until\ the\ end\ of\ the\ \varpi-boundary\ around\ V$ ▷ HTS2
      end for
      $break$
    else
      for all $C-phrases$ **do**
        $Within\ the\ C-phrase\ find\ a\ H-toned\ \sigma$
        $\sigma_i \leftarrow \sigma$ ▷ The high-toned syllable
        for all $\sigma_i \in C-phrase$ **do**
          if $\sigma_{i+1} = L-tone$ **then** ▷ σ_{i+1} is the adjacent syllable
            $\sigma_{i+1} = H-tone$
          end if
        end for
      end for
      $break$
    end if
  end for
  $Return\ S$
end procedure

Trace of the Basic Algorithm on a Sentence

Table 2.1 shows the trace of the algorithm on the input sentence in Example 2.44. Since Example 2.44

is in the negative form, we expect the basic algorithm to apply the grammatical high tone insertion rule and the iterative high tone spread rule.

(2.44) Input: S = ha ké rékise nama [Khoali 1991, p.248]
NEG 1SG.SM sell meat
"I do not sell meat"

Line in Code	**Output**
Line 2	V = *rékise* ▷ Locating the verb stem in the sentence
Line 3	(ha ké rékise)C nama
Line 4	(ha ké (rékise)ϖ)C nama
Line 5	TMA = NEG
Line 6	TMA = NEG ▷ If-statement is true
Line 7	*rékíse*
Line 8	*rékísé*
Line 8	*rékísé*
Line 18	S = ha ké rékísé nama

Table 2.1: Execution of the basic algorithm

Our input sentence, *S* is "ha ké rékise nama". The algorithm first finds the verb stem *rekise* in the sentence. It then inserts the clitic phrase boundary followed by the phonological word boundary around the verb stem. The sentence has the negative form morpheme *se* and the verb stem has the negative suffix (*e*), which illustrates that the sentence is in the negative *(TMA = NEG)*. The sentence is in the negative causing the basic algorithm to insert a grammatical high tone on the second syllable of the high-toned verb stem (rékise) and then spread it iteratively to the end of the verb stem. The algorithm then outputs the sentence *S* = (ha ké (rékísé)ϖ)C nama.

2.7.3 Analysis and Results

Analysis

The application of the basic algorithm was restricted to the clitic phrase domain, thus, the comparisons between the output by the basic algorithm and its transcribed counterparts were restricted to the clitic phrase domain. The transcriptions of the sentences were then compared to the output produced by the basic algorithm. Furthermore, the transcriptions were compared to sentences with unmarked tone labels and sentences with underlying tone labels independently within the clitic phrase domain [Raborife 2009].

Results

The results of the study by Raborife [2009] showed that in order to predict the tone labels on the syllables, we need an algorithm that makes use of linguistic theories, specifically linguistically-defined tonal rules. The results from this study also showed that implementing the three tonal rules is not sufficient for tone labeling in Sesotho and have shown that there are three other tonal rules which feed from HTS1 that needed to be implemented. These tonal rules are the left branch delinking rule, the right branch delinking rule and the finality restriction rule. These three rules are implemented by the extended algorithm.

Table 2.2 presents the results of the study by Raborife [2009] which showed that using underlying tones is better than having no tone information on the syllables. Furthermore, implementing the three tonal rules resulted in an increase in the number of matched tone labels.

Compare Transcribed Tone Labels with:	Number of Matched Tone Labels (out of 118)
Unmarked Tones	37
Underlying Tones	79
Basic Algorithm	100

Table 2.2: Summary of results from the previous study on a tone label prediction algorithm

2.7.4 Limitations of the Basic Algorithm

The analysis of the basic algorithm revealed the following limitations:

- There are three Sesotho tonal rules not implemented by the basic algorithm that need to be implemented. These rules are the left branch delinking rule, the right branch delinking rule and the finality restriction rule.
- In a sentence where there is more than one verb stem, the basic algorithm checks if any of the verb stems are in the context that allows for the application of the grammatical tone insertion rule. If there is such a verb stem, the algorithm applies the grammatical tone insertion rule followed by the iterative high tone spread rule to all the verb stems in the sentence. Otherwise, the algorithm applies the high tone spread rule within the clitic phrase boundaries that the verb stems are in. This leads to the algorithm applying incorrect tonal rules to verb stems which have contexts that do not allow the application of the grammatical tone insertion rule.
- The implementation of the algorithm restricts its application to the clitic phrase domain.

2.8 Conclusion

The aim of this research study was to improve a basic tone label prediction algorithm that uses three Sesotho tonal rules to make the predictions [Raborife 2009]. These rules are discussed in detail in Section 2.4.1. The extended algorithm implemented four other tonal rules as discussed in Section 2.4.2. The extended algorithm therefore implements a total of seven Sesotho tonal rules.

The tonal rules that are implemented by both the basic and extended algorithms are restricted within certain phonological domains, with the exception of the grammatical tone insertion rule. These rules are restricted to three phonological domains, namely the phonological word, clitic phrase and the phonological phrase. The tonal rules as well as their domains of application are described in Khoali [1991].

A tone label prediction algorithm implements the tonal rules within the domains to which the rules are restricted. This chapter has provided background literature on the tonal rules as well as their domains of application. We also discussed the basic algorithm which we aimed to improve in our research. The next chapter discusses the methodology that was followed to improve the basic algorithm.

Chapter 3

Research Methodology

3.1 Introduction

The previous chapter provided an overview and basis of our research study. This research study sought to improve on some of the limitations of the basic algorithm which predicts tone labels based on three Sesotho tonal rules.

This chapter provides a framework on which our research is based. We also present our research hypothesis. Furthermore, the research method employed in our study is presented and its appropriateness to the research is discussed.

3.2 Research Context

Text-to-speech systems require reliable tone and intonation models since the intelligibility of these systems depends heavily on them [Zerbian and Barnard 2008]. However, not much research has been dedicated to developing tone models for Bantu languages. Louw *et al.* [2005] attempted to implement tone into their TTS system, however, in their attempt, it was found that treating all tones as "unmarked" would produce more understandable speech by the system. Louw *et al.* [2005] suggested that for TTS systems to produce more "naturally" perceived tones, we need an algorithm that can predict the tone labels of the syllables and the appropriate intonation of the word. Zerbian and Barnard [2010] discussed the steps that are necessary in order to develop such an algorithm focusing on the Sotho-Tswana languages as follows:

1. Morphological analysis - This is important for part-of-speech tagging as well as for the tagging of the tense, mood and aspect of the verb stems. This process is important because some tonal rules, such as the grammatical tone insertion rule, only apply to words of certain grammatical categories [Zerbian and Barnard 2010].

2. Pronunciation dictionaries - These are important for the assignment of tone patterns for noun, verb and adjective stems as well as for grammatical morphemes such as subject concord as well as tense, mood and aspect markers [Zerbian and Barnard 2010].

3. Algorithm for prosodic phrase boundaries - Tonal rules in the Sotho-Tswana languages do not just apply anywhere in a sentence where their phonological requirements are met. They are constrained within certain domains. These domains are referred to as the prosodic (phonological) domains. It is important that the insertion of the domains precedes the application of the rules since some of the tonal rules are sensitive to these domains [Raborife 2009].

4. Application of the tonal rules - Once Steps 1 – 3 above are implemented, the tonal rules can apply as described in the literature. Khoali [1991] would serve as the literature specifying the tonal rules and their applications for Sesotho.

A research study by Raborife [2009] developed the basic algorithm that implements Steps 3 and 4 of the above hierarchy, using three Sesotho tonal rules as a basis for the study. The basic algorithm has some limitations which we aimed to improve in order to increase the number of matched tone labels between the transcribed tone labels and the tone labels predicted by the basic algorithm. A matched tone label is a label predicted by an algorithm which is exactly like its transcribed counterpart.

3.3 Research Aims

Our research study sought to improve a linguistically-motivated algorithm that describes three Sesotho tonal rules. Our study involved the following:

- Implementing four other tonal rules, namely, the right branch delinking rule, the left branch delinking rule, the specifier high tone delinking rule as well as the finality restriction rule.
- Extending the application of the tonal rules to other parts of speech such as adverbs and adjectives.
- Implementing the tonal rules on each verb stem independently in a case where there is more than one verb stem in a sentence.
- Testing the algorithm on a larger corpus of data. Our corpus is made up of 45 Sesotho sentences whereas in Raborife [2009]'s study, only 15 Northern Sotho sentences were used.

3.4 Research Hypothesis

Louw *et al.* [2005] suggest a tone label prediction as well as an intonation prediction algorithm to model tone into TTS systems developed for languages such as isiZulu and Sesotho. These languages do not mark tone in orthography. According to our knowledge, there is only one algorithm that has been developed to predict tone labels on the syllables for the Sotho-Tswana languages [Raborife 2009]. The basic algorithm developed by Raborife [2009] restricts its application to the clitic phrase domain.

Our aim was to improve the basic algorithm by, amongst other things, implementing four additional Sesotho tonal rules. Furthermore, we extended the application of the tonal rules to all the other parts of speech. Hence, our research hypothesis was formulated as follows:

> The extended algorithm improves the basic algorithm by increasing the number of matched tone labels.

The research study conducted by Raborife [2009] showed that failure to implement the right branch delinking, left branch delinking and finality restriction rules, which "feed" from the high tone spread rule, leads to some mismatches in the tone labels. The right branch delinking rule accounted for 1.69% and the left branch delinking rule for 0.85% out of 15.3% of the mismatches found in the study by Raborife [2009]. The finality restriction rule accounted for 7.63% of the mismatched syllables found in the study conducted by Raborife [2009]. The specifier high tone delinking rule did not account for any of the mismatches found in Raborife [2009]'s study since object markers were not included in their data.

3.5 Research Method

Figure 3.1 provides an overview of the processes involved in our study and also outlines the steps described in Zerbian and Barnard [2010]. The text in bold indicates the output of each process.

The first two processes correspond to Steps 1–2 of Zerbian and Barnard [2010]'s framework. Since we did not have access to tone-marked pronunciation dictionaries or a morphological analyzer, these two steps were processed manually.

In the final process, the high tone spread occurs within the phonological word domain for the noun stem as well as the clitic phrase domain for the verb stem and its preceding clitics. The last two processes correspond to Steps 3–4 of Zerbian and Barnard [2010]'s framework. The following sections describe in detail the manner in which we implement these processes.

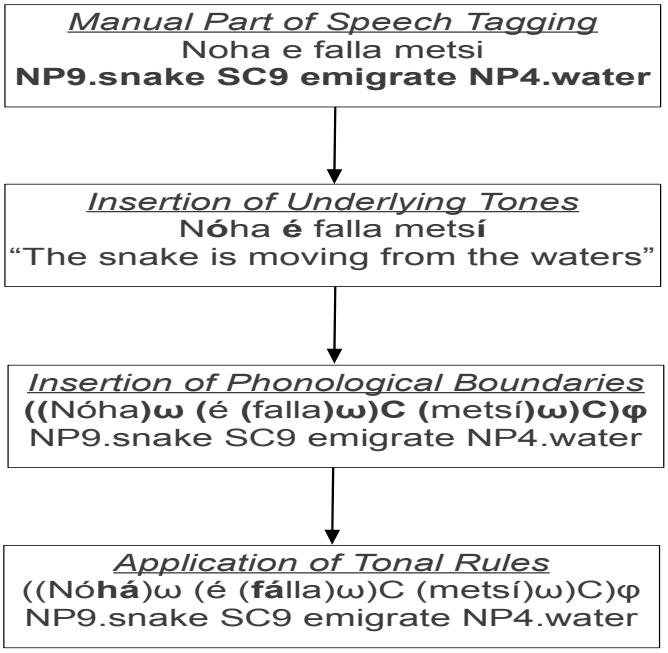

Figure 3.1: Approach taken in our study

3.5.1 Preparations of Input Data

The first two steps shown in Figure 3.1 represent the necessary processes we took to prepare the sentences which were used to test both algorithms. This includes tagging the parts of speech of the words in the sentences as well as marking the underlying tone labels on the syllables. In this section, we discuss these two processes in detail.

Part of Speech (POS) Tagging

The parts of speech are tagged manually. POS tagging was important in our study since the grammatical

tone insertion rule applies only to words of certain grammatical categories. We tagged the POS category of a word using the dictionary by Du Plessis *et al.* [1974]. Also, when inserting phonological boundaries, the extended algorithm uses grammatical information as a cue, as discussed in Section 3.6.

Faasz *et al.* [2009] describe an automatic POS tagger developed for Northern Sotho. One of the challenges encountered in their study is the complexity of content words in Northern Sotho. For instance, adding the suffix *-ng* to the noun stem *toropo* ("town"), forms *toropong* ("in/at/to town") [Faasz *et al.* 2009]. Such derivatives show up as unknown in their POS tagger.

Another challenge Faasz *et al.* [2009] have come across in their study is the ambiguity of certain function words. Function words include demonstratives, pronouns and conjunctions. In Northern Sotho, as well as in Sesotho, function words are highly ambiguous. For instance, the morpheme *-a-* can mean nine different things in Northern Sotho: it could be a subject concord of noun class 1 or 6, an object concord of class 6, a possessive concord of class 6, a question particle, a hortative particle or a verbal morpheme indicating past or present tense [Faasz *et al.* 2009]. According to Faasz *et al.* [2009], it is difficult for an automatic POS tagger to disambiguate such morphemes since many of them appear in the context of other function words. Thus, using local context would not necessarily disambiguate.

For the purposes of our research, it was not sufficient to tag just the POS but we also needed information on whether a particle spreads or not or if a certain word precedes a modifier. In the latter case the extended algorithm would insert the right edge of the ϕ-domain (phonological phrase) before such a word when inserting the phonological domains as discussed in Section 3.6.2. In the case of derivational suffixes such as *-ng* in *toropo-ng*, a morphological analyzer would separate the *-ng* suffix from the noun stem. A POS tagger would tag the word as a noun stem, which is incorrect because the addition of the suffix *-ng* has changed the grammatical category of the word to an adverb (modifier). A more sensible approach for us in our study would be to have access to a refined POS tagger, which is both a morphological analyzer as well as a POS tagger. According to our knowledge, currently, there is no refined POS tagger that has been developed for the Sotho-Tswana languages. Appendix B shows how a refined automatic POS tagger would tag the words for the purposes of the implementation of the extended algorithm on the data.

Labeling of Underlying Tones

According to Schadeberg [1981], there are only two Sesotho dictionaries that have tone marks (Du Plessis *et al.* [1974], Mabille *et al.* [1961]). In our study, we used Du Plessis *et al.* [1974] to mark the underlying tones since we could not access Mabille *et al.* [1961].

Du Plessis *et al.* [1974] show the consequence of the high tone spread in its entries as exemplified in 3.1.

(3.1) rékéla
"buy for"

In Example 3.1, the high tone from the first syllable of the verb stem *re* spreads to the second syllable of the verb stem *ke*. We know that the high tone on the second syllable of the verb stem is not underlying as it would have been deleted by the Meeussens rule. The Meeussens rule deletes the second high tone of two adjacent underlying high tones [Khoali 1991]. In our tagging, we did not include the high tones which are a consequence of the high tone spread since this rule forms part of our analysis and is implemented by the algorithm by Raborife [2009].

Du Plessis *et al.* [1974] marks the tones of the noun stem *kunutu* as *kúnútú*. We can assume that the high tone on *nú* is a consequence of the high tone spread rule and is therefore not underlying. It is delinked by the right branch delinking rule. We only tag underlying tones thus we assume the following marking in our data.

(3.2) kúnutú
"secret"

In the data, the noun stem *tebogo* which is not a Sesotho word but a Sepedi word is included. This does not compromise the results of our study since the two languages are closely related and its Sesotho counterpart is *teboho*. We use the dictionary by Ziervogel and Mokgokong [1971] to mark its underlying tones.

Some of the words we used were not in Du Plessis *et al.* [1974]. In such cases, other sources of literature were consulted. Below we give the source of tone patterns of these words:

- *tse*
  - According to Doke and Mofokeng [1957] relative concords have a high tone. Thus, we tagged this morpheme with a high tone.
- *ha*
  - Corresponding *fa* has a high tone in Setswana. It is not in Du Plessis *et al.* [1974] but in Khoali [1991] it is marked as a high tone. We use Khoali's transcription for this morpheme [Khoali 1991].
- *ile*
  - The tone pattern for this morpheme is not in Khoali [1991] and Du Plessis *et al.* [1974]. We use the Sepedi dictionary by Ziervogel and Mokgokong [1971]; according to them, this morpheme has a low-low (LL) tone pattern.
- *wa*
  - Possessive concords have a high tone in Tswana [Creissels *et al.* 1997]. This is not in Du Plessis *et al.* [1974], but the possessive concord in Sesotho *ya* has a low tone. Acoustic data for this morpheme shows it to be relatively high. We assume a high tone for this morpheme in our study.
- *lona*
  - Doke and Mofokeng [1957] also classifies the tone pattern of absolute pronouns as low-high (LH) with the final *-na* having a high tone (H). This classification for absolute pronouns was also assumed in our tagging because there are discrepancies between Du Plessis *et al.* [1974] and Mabille *et al.* [1961] regarding the tone pattern of *lona*. According to Du Plessis *et al.* [1974], it has an LL tone pattern and according to Doke and Mofokeng [1957] *lona* has an LH tone pattern. Doke and Mofokeng [1957]'s tone classification is supported by acoustic data across the three recorded speakers where this morpheme has an LH pattern.

The next section explains the last two steps presented in Figure 3.1, which were in essence the crux of our research study, the actual algorithm.

3.5.2 The Algorithm

The last two steps shown in Figure 3.1 refer directly to the extended algorithm. These steps were the actual focus of our study which implements the algorithm. The algorithm assumes that there are no monosyllabic verb stems in the data, since monosyllabic verb stems have a complex tonal behaviour

which exceeds the scope of our study. The algorithm first inserts the phonological domains within which the tonal rules apply. Then, the algorithm applies the tonal rules accordingly. Section 3.6 discusses the actual implementation of the extended algorithm. In this section, we present an overview as well as the pseudo-code of the extended algorithm.

Insertion of Phonological Boundaries

There are three phonological domains within which the tonal rules apply, namely, the phonological word, the clitic phrase and the phonological phrase. The phonological domains are inserted as described in the previous chapter, specifically, Section 2.3.

The phonological word (ϖ) The extended algorithm inserts the ϖ-domain by "looking" at the POS category of each word in a sentence. If the word is a content word (except for nouns), the algorithm inserts the phonological word domain around the content word. Content words include nouns, verbs, adjectives and adverbs. If the content word is a noun, the algorithm checks if the word preceding it is a particle that spreads by the high tone spread rule (HTS1). If it is, the algorithm inserts the ϖ-domain around the particle and the first syllable of the noun stem. Otherwise, the algorithm inserts the phonological word domain around the noun stem.

The clitic phrase (C) Recall that the clitic phrase domain is only motivated for verb stems. Therefore, for each verb stem in a sentence, the C-boundary is inserted from the verb stem to the left until a subject marker is encountered. That is, the right edge of the C-domain is inserted immediately after a verb stem and the left edge is inserted before the subject marker that precedes the verb stem.

The phonological phrase (ϕ) The right edge of a phonological phrase domain is inserted between a head and its modifier as well as after the last word on the right edge of a sentence. The head of a phrase defines the category of the phrase (for example, the head of a verb phrase is a verb). The algorithm then inserts the left edge immediately after a right edge as well as before the first word of a sentence. The sentence can be either compound, simple or complex.

Application of the Tonal Rules

The extended algorithm locates the verb stem in a sentence. It then checks the TMA of the verb stem. If the TMA of the verb stem falls in the category that allows for the application of GTI, the algorithm applies GTI followed by the HTS2. HTS2 is a strictly phonological word domain span rule. If not, the algorithm applies HTS1. The high tone spread rule (HTS1) is a clitic phrase domain span rule.

The extended algorithm searches for a clitic phrase in a sentence, "looks" for the verb stem and checks its context. If the context allows for the application of the GTI, the algorithm applies GTI followed by HTS2. Else, it applies HTS1 and then RBD (or LBD) if necessary. The algorithm searches for the next clitic phrase in the sentence and does the same for all the clitic phrases in a sentence.

Once all the clitic phrases have been found and the tonal rules applied in them, the algorithm searches for the phonological word domains and applies the HTS1 followed by RBD and LBD accordingly within this domain.

Within each phonological phrase domain, the algorithm checks the tone on the last syllable of the domain. If the syllable is a surface high tone, the algorithm changes the tone on that syllable to a low tone due to the finality restriction rule (FR).

After the application of HTS1, the algorithm applies the specifier high tone delinking (SHTD) rule by changing the tone on all object concords to low.

A pseudo-code of the extended algorithm is shown in Algorithm 2.

Algorithm 2 Extended tone labeling algorithm

```
1:  procedure TONE_LABELING(S)                          ▷ S is the input sentence
2:      Insert C boundary
3:      Insert ϖ boundary
4:      Insert ϕ boundary
5:      α → syllable
6:      n → number of syllables in phrase
7:      GTI → PERF or NEG or HAB or PST or IMP or SUBJ
8:      for all C − phrases do
9:          Apply HTS1
10:     end for
11:     for all P − word do
12:         Check POS
13:         if POS = V then
14:             Check TMA
15:             if TMA ∈ GTI then
16:                 Apply GTI
17:                 Apply HTS2
18:             else
19:                 Apply HTS1, RBD OR LBD
20:             end if
21:         else
22:             Apply HTS1, RBD OR LBD
23:         end if
24:     end for
25:     for all P − phrase do
26:         if α_n = H − tone then                      ▷ Underlying tone
27:             α_n → L − tone
28:         end if
29:     end for
30:     Apply SHTD
31:     Return S
32: end procedure
```

Trace of the Extended Algorithm on a Sentence

Table 3.1 shows the trace of the extended algorithm on the input sentence in Example 3.3.

(3.3) Input: S = ó tla tsamaya [Khoali 1991, p.224]
SC1 FUT go
"S/he will go"

Our input sentence, *S* is "ó tla tsamaya". The extended algorithm first inserts the prosodic boundaries as follows:

Line Code	Output
Line 2	(ó tla tsamaya)C
Line 3	((ó tla)ϖ tsamaya)ϖ)C
Line 4	((ó tla)ϖ (tsamaya)ϖ)C)ϕ
Line 9	ó tlá tsamaya
Line 22	ó tlá
Line 19	tsamaya
Line 30	S = ó tlá tsamaya

Table 3.1: Execution of the extended algorithm

- The clitic phrase domain includes the verb stem and all the clitics preceding it.
- There are two phonological word domains.
  - The first phonological word domain includes the subject concord and the future tense marker.
  - The second phonological word domain includes the verb stem.
- The phonological phrase domain includes the whole sentence.

The sentence has a future tense morpheme *tla* which disallows the application of the GTI and HTS2 rules on the verb stem. In the clitic phrase, the algorithm applies the HTS1 rule in which the high tone on the subject concord *ó* spreads to the future tense marker *tla*. The algorithm then goes through each phonological word, however, no contexts are met for the application of HTS1. The algorithm then moves to the phonological phrase domain and checks the last syllable in the domain. The last syllable in the phonological phrase domain is low-toned, thus, FR does not apply in this case. The SHTD does not apply because there are no object concords in the input sentence. The algorithm then outputs the sentence S = ó tlá tsamaya.

In the next section, we discuss the actual implementation of the algorithm within the context of the architecture of *Speect* (Speech synthesis with extensible architecture). *Speect* is a multilingual TTS system developed by the Meraka Institute [Louw 2008].

3.6 Implementation of the Algorithm

3.6.1 Input Processing

We used the back-end of the *Speect* system to implement the extended algorithm. The extended algorithm used the overall architecture of this system. ***Speect*** revolves around a globally accessible data structure known as the *utterance structure*. The *utterance structure* is represented internally as a *Heterogeneous Relation Graph* (HRG) which consists of a set of relations where each relation contains some items [Louw 2008]. The relations represent structures such as words, syllables or phonemes and the items are the content of these structures [Louw 2008]. The SyllableStructure relation in Figure 3.2 can define which word (*item*) is a parent of a specific syllable (*item*) and which syllables are next or previous chronologically. Items consist of a bundle of features and a set of named links to nodes in relation, for instance a syllable (*item*) can have a tone feature which can be set to either high or low [Taylor *et al.*

1998]. Figure 3.2 shows an example representation of an utterance structure using an HRG with five relations (namely, Tokens, Words, Syllables, SyllableStructure, Segments) and their items.

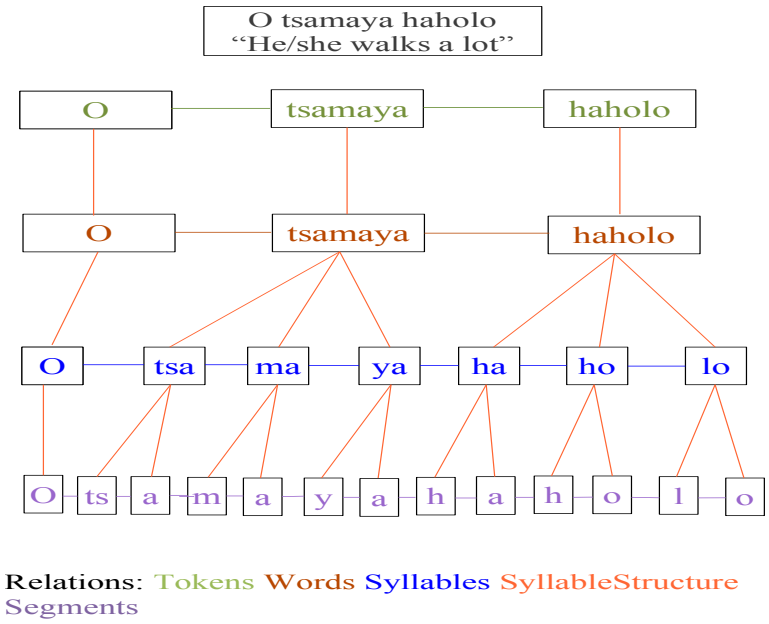

Figure 3.2: An example representation of an utterance structure using a heterogeneous relation graph

The extended algorithm receives as input the sentence from the console as typed in by the user. The sentence received as input is then processed by the Speect text processing module. This module creates a SyllableStruture "Relation" which contains the sequences of the word as entered by the user with each word having the following attributes:

- grammatical category (part of speech),
- phonemes,
- syllables with underlying tones associated to them, and
- Boolean flag on each syllable which are all reset to False.

The above mentioned information is obtained by Speect's text-processing module from a dictionary which has lexical entries with each entry having these attributes. In the following section, we discuss how we use the SyllableStructure relation to create our phonological domains.

In sum, the SyllableStructure relation has the words (*items*) with each word having the POS feature. The word is a parent of one or more syllables which have tone features. Furthermore, this relation defines which syllables are next or previous chronologically.

In our implementation, each phonological domain is coded as a method. Furthermore, each rule is coded as a method and traverses the phonological domain within which it is restricted.

3.6.2 Inserting Prosodic Domains

Clitic Phrase Domain Once the input sentence has been processed and the SyllableStructure relation of the sentence defined, the **main** method calls the **clitic** method (*clitic phrase*). A method is a subroutine that is associated with either a class or an object. This method creates a new relation called CPhrase (clitic phrase). It then "indexes" the SyllableStructure relation of the sentence. An iterator (*iter1*) is used to traverse the elements in the SyllableStructure relation. It starts with the first item in the relation and checks the POS feature of the item. If the POS is a verb stem, it is inserted as an item with its POS feature in the *CPhrase* relation. Then, the syllables — together with their features such as the tone labels — of the verb stem are stored in a temporary list (*list1*). A separate iterator (*iter2*) is then used to backtrack from the verb stem searching for all the clitics preceding the verb stem, storing their tone labels as it encounters them in a separate temporary list (*list2*) until it hits a content word. Since the clitics are inserted in a temporary list as they are encountered, the clitic which is immediately before the verb stem is placed first in the list.

(3.4) ó a téna...
H L HL...
"he/she is putting on..."

In Example 3.4, *list1* would have "LH" and *list2* would have "LH". However the correct order of *list2* should be "HL". So the order of the tone labels in *list2* is reversed. The list containing the tone labels of the clitics preceding the verb stem (*list2*) and the list containing the tone labels of the verb stem (*list1*) are then merged together. The elements of the new list are added as daughters of the verb stem item in the CPhrase relation. The iterator *iter1* then searches for the next verb stem in the sentence and the above-mentioned process is repeated for the verb stem. This process is done iteratively for all the verb stems in a sentences. Once all the verb stems in the sentence have been traversed, control is transferred back to the **main** method.

Phonological Word Domain Once control is transferred back to the **main** method by the **clitic** method, the **main** method calls the **pwrd** method (*phonological word*). This method creates a new relation called PWord (*phonological word*). It then "indexes" the SyllableStructure relation of the sentence. An iterator (*iter1*) is used to traverse the elements in the SyllableStructure relation. It starts with the first item in the relation, checks the POS feature of the item and does the following:

1. If the POS of the item is an element of the list of content words, except for nouns, it is inserted as an item with its POS feature in the PWord relation.

2. If the POS of the item is a noun, it is inserted as an item with its POS feature in the *PWord* relation. A separate iterator (*iter2*) is then used to backtrack to the item that precedes the noun.

    - If the POS feature of the item indexed by *iter2* is PTS[1], the daughter of *iter2* and the first daughter of *iter1*[2] are inserted as daughters of the newly-inserted item. A new item is then inserted into the PWord relation and the remaining daughters[3] of the item indexed by *iter1* are inserted as daughters of this item.
    - Else, the item is inserted into the PWord relation together with all of its daughters.

[1] particle that spreads
[2] which is the noun class prefix
[3] root of the noun

3. If the POS of the item is a FUT, a new item is created in the PWord relation and the item indexed by *iter1* added as a daughter to the new item. A separate iterator (*iter2*) is then used to backtrack to the item that precedes *iter1* in the Word relation. The item indexed by *iter2* is added as a daughter to the item that already has the FUT as a daughter in the PWord relation.

iter1 then moves to the next item in the Syllable structure and the same process is followed. This process repeats until all items in the SyllableStructure relation have been traversed. Control is then transferred to the **main** method.

Phonological Phrase Domain Once control is transferred back to the **main** method by the **pwrd** method, the **main** method calls the **pphrase** method (*phonological phrase*). This method creates a new relation called PPhrase (*phonological phrase*) and a new item for the relation. It then "indexes" the SyllableStructure relation of the sentence. This method iteratively (using an iterator, say *iter1*) traverses the items of the SyllableStructure relation. At each iteration, the POS of the item is checked if it is a NAGR or VAGR[4]. As *iter1* traverses the items (checking their POS features) in a SyllableStructure relation, these items are inserted into a single PPhrase item until an item with a modifier POS is encountered which is either tagged as NAGR or VAGR (or ADJ/ADV).

If a modifier is encountered, the **pphrase** method calls the **pphrasemod** method that inserts the phonological phrase boundary around modifiers. This method retrieves the PPhrase and the SyllableStructure relation to get the last item indexed by the iterator. It then creates a new item in the relation and uses an iterator to traverse the item that follows the item passed to it in the SyllableStructure relation, inserting them as daughter to the newly-inserted item until it reaches another modifier or the end of a sentence. If another modifier is encountered, the method breaks out of the loop and the **pphrasemod** method is called again.

3.6.3 Applying the Tonal Rules

The basic algorithm already implements three tonal rules. The application of these tonal rules is restricted to polysyllabic verb stems. The application of the tonal rules is now extended to other parts of speech except for monosyllabic verb stems. The tone feature of each syllable is assigned a flag which is set to False for all the syllables in the beginning.

High Tone Spread incl. RBD and LBD within Clitic Phrase Once all the prosodic domains are inserted, there are two HTS methods, the one for verb stems and the other for all the other parts of speech. The **main** method first calls the **HTS_verbs** method. In this method, the clitic phrase is traversed. The **HTS_verbs** method calls the **HTS** method. The **HTS** method does the following for all the items in the CPHRASE structure: It checks the first daughter (syllable) of the item. If it is a high tone, the tone feature of the following syllable, say β, is changed to high if it is low and flagged as True. If β already has a high tone, the iterator moves to β, checks the following syllable, say α, and does the same for it as it did for β. After the application of the HTS1, the method checks the syllable preceding the source syllable. If there is one and it has a high tone, the tone of the source syllable is set to low. The method also checks the tone feature of the syllable following the target syllable, which is the syllable that received a high tone due to the application of HTS1. If it is high, the tone feature of the target syllable is reset to low.

High Tone Spread incl. RBD and LBD within Phonological Word Once control is transferred to the **main** method, it calls the **HTS_all** method. This method applies the HTS1 rule to all the parts of speech.

[4]These are POS tags for the modifiers in the dictionary.

The iterator checks all the items in the PWord relation. For each item, it checks the POS category of the word.

- If the POS category of the word is a verb, then the iterator checks the TMA category of the item against the list of TMA categories that allow for the application of the grammatical tone insertion rule.
  - If the POS category of the word is in that list, this method calls the **GTI** method and passes the item and the utterance as parameters for this method. The **GTI** method gets as input the item (word and all its attributes) as well as the whole utterance. The **GTI** method gets the second syllable of the item (second daughter in the structure) and changes the tone feature to high and sets the flag value of the syllable to True. The **GTI** method then calls another method, **HTS2** and passes the item and the utterance as parameters to the **HTS2** method. This method then accesses the item as it is in the SylStructure relation. It then gets the second daughter (syllable) of the item and sets the tone feature of this syllable to high and the flag value to True. This method then calls the **HTS2** method, passing the verb stem as a parameter to this method. The **HTS2** method accesses the item as it is in the SylStructure relation and accesses the third daughter (if it exists) and any other daughters following the third daughter, changing the tone features to high and the flag values to True.
  - Else, the **HTS_all** method calls the **HTS** method, passing the item as a parameter to the method. The **HTS** method then applies the high tone spread rule to this item as discussed below.
- Else, the **HTS_all** method calls the **HTS** method, passing the item as a parameter to the method. The **HTS** method then applies the high tone spread rule to this item as follows:
  If it is a high tone, the tone feature of the following syllable, say β, is changed to high if it is low and flagged as True. Otherwise, if β already has a high tone, the iterator moves to β, it checks the following syllable, say α, and does the same for it as it did for β. After the application of the **HTS** method, the method checks the syllable preceding the source syllable. If one exists and has a high tone, the tone of the source syllable is set to low. The method also checks the tone feature of the syllable following the target syllable. If it is high, the tone feature of the target syllable is reset to low.

Finality Restriction Once control is transferred to the **main** method by the **HTS_all** method, the **main** method calls the **FR** method. This method applies the finality restriction rule. This method accesses the PPHRASE structure. For each item in the PPHRASE relation, this method gets the last daughter of the item. It then accesses the last daughter of the item (this_daughter) as it is in the SylStructure relation. In the SylStructure relation, this_daughter is listed as an item (this_item) and has its syllables as its daughters. The FR method then accesses the last daughter of this_item. It then checks the flag value of the last daughter - if the flag value is True, it means that the value has been changed and is therefore not underlying. Then, it checks the tone feature of this daughter - if it is high, it is reset to low.

Specifier High Tone Delinking When control is transferred to the **main** method by the **FR** method, the **main** method calls the **SHTD** method. This function checks for the object markers in the sentence. If there is an object marker, the tone feature is set to low.

3.7 Conclusion

Our study involved the improvement of a linguistically-motivated algorithm developed by Raborife [2009]. To improve this algorithm, we did the following:

- Implemented four other Sesotho tonal rules as described by Khoali [1991].
- Extended the application of the tonal rules to other parts of speech.
- Treated each verb stem independently according to its own grammatical context.

The Python implementation of the extended algorithm used the back-end of Speect, a multilingual TTS system. This algorithm defines the phonological domains within which the tonal rules apply. Then, it applies the tonal rules accordingly.

In the next chapter, we discuss how the sentences which we used to test the extended algorithm were compiled, recorded and transcribed. We also discuss the significance of the transcriptions.

Chapter 4

Corpus Compilation

4.1 Introduction

The extended algorithm aimed to improve the basic algorithm by implementing four Sesotho tonal rules in addition to the three Sesotho tonal rules implemented by the basic algorithm. The previous chapter presented the research methodology employed to improve the basic algorithm.

This chapter provides a detailed description of the 45 Sesotho sentences which were used to test the algorithm. Since the output produced by the algorithm on the sentences is compared to recorded speech, we discuss the recording process as well as the transcription process in this chapter.

4.2 Text Corpus Compilation

Forty-five Sesotho sentences were chosen from a large corpus of Sesotho data which includes folk-tales and poetry (see Appendix B)[1]. The sentences were chosen to include the following occurrences:

- The present principal tense.
- The past tense.
- The perfect tense.
- The imperative mood.
- The future tense.

These tenses are relevant for the application of both the basic and extended algorithm. Proper nouns and/or loan words are not included in the corpus since their tone patterns are not available in the literature [Zerbian and Barnard 2010; Du Plessis *et al.* 1974]. Verb forms showing the Potential Mood (using *ka*) have been excluded since the tonology of these forms is subject to considerable dialectal variation [Creissels *et al.* 1997; Cole and Mokaila 1962; Lombard 1976]. Furthermore, there are no monosyllabic verb stems in our data given the fact that they have complex tonal characteristics [Khoali 1991; Zerbian 2009] and have therefore been excluded from the algorithm developed by Raborife [2009] as well as its extended version.

Since we did not have pre-recorded Sesotho sentences for the transcriptions, the 45 sentences were recorded by three native Sesotho speakers as discussed in Section 4.3. The transcriptions are needed for

[1]The Human Language Technology group at the Meraka Institute has a corpus which has Sesotho sentences taken from poetry and folk-tales as well as the Government Gazette. The sentences were selected from this corpus.

the evaluation phase of our research study where the transcriptions of these sentences (which include tone labels) are compared to the output produced by both the basic and extended algorithm as discussed in Chapter 5.

In the next section, we discuss the recording process as well as the protocol used for the recording session.

4.3 Recording Session

The research study involved recording native Sesotho speakers (respondents). These respondents were not remunerated for the recordings and the recordings can be used for further research studies. The respondents were made aware of this aspect.

Three Sesotho speakers recorded the sentences, with each speaker recording the 45 sentences. In addition to the 45 sentences, the respondents also recorded eight Sesotho tonal minimal pairs (see Appendix C). These minimal pairs were used to select the respondent with the most pronounced tones as discussed in Section 4.4. This session took approximately three and a half hours for all speakers and consisted of a number of stages which are explained in the subsequent sections.

4.3.1 Introduction

This phase involved formal introductions between the interviewer and the respondent. The respondent was given the research information sheet (see Appendix D.1) which explains the purpose of this research study. The respondent together with the interviewer read the sheet and the interviewer made sure that the respondent understood the contents of the document clearly.

4.3.2 Confirmation of Consent

Once the respondent understood the recording process as well as the purpose of the research study, he or she made a final decision as to whether he or she was willing to participate in the recording session. The interviewer emphasized that participation was completely voluntary and that the data gathered in the recording would remain confidential, but could be used for further research.

The respondent was also informed that pseudonyms were used to store the data. Based on this information, if the respondent chose to continue with the recording session, he/she was given a formal consent form to sign (see Appendix D.2).

The consenting respondent was also required to fill out a language profile questionnaire (see Appendix D.3). This would serve as proof that according to our knowledge, the respondent is a native Sesotho speaker.

4.3.3 Recording Guidelines

In this phase of the research study, the interviewer explained the purpose of the recording and its contribution to the research study. The interviewer also explained and loaded the following instructions onto a computer screen.

- Silently read and make sure you understand the English meaning of the sentence given below the Sesotho sentence.
- Only say the Sesotho sentence, not its English meaning.
- Please speak in a natural voice and tone.

- You can record the sentence again if you are not happy with the first recording.
- Please press enter only after you have finished recording the sentence.

The respondent was then asked to read out the instructions above. The instructions were recorded and played back to the respondent to make sure that the microphone was placed at an appropriate angle. Once the respondent understood the instructions, the interviewer proceeded to the final phase of the session as described in the following section.

4.3.4 Recording the Sentences

This is the final phase of the session. In this phase, the recorder was switched on and a presentation was then loaded onto a computer screen. The respondent was then asked to read out the sentences as they appeared on the slides. The sentences did not have any tone information on them. Furthermore, they did have their English meanings below them.

The tonal minimal pairs were presented to the respondent before the 45 sentences. Each minimal pair was on a separate slide with the English meaning of the word/phrase below. The English meaning was to guide the respondent as to which tone pattern was to be pronounced for that phrase/word. Tone labels were not marked on the minimal pairs, nor on the 45 Sesotho sentences.

Once all the minimal pairs were recorded, each sentence (in the compiled corpus) was placed on a separate slide. If the participant mispronounced any of the words in a sentence or spoke unnaturally, they were asked to repeat the sentence. Once all the 45 sentences were recorded, the recorder was switched off and the respondent was thanked for their participation.

For the purposes of our research study, only one respondent was needed for transcriptions. In the next section, we provide a reason for using only one respondent as well as for how the respondent was chosen from the three recorded in the phase we discuss in this section.

4.4 Selection Process

To evaluate the extended algorithm, we compared its output (the 45 selected sentences described in Section 4.2, including the tone marks predicted by the extended algorithm) to the speech recorded by a native Sesotho speaker. This speech needed to be encoded in a way that allows for this comparison, that is, it needed to have tone marks (either high or low). Since the extended algorithm output the tone marks on the syllables, to label the tone marks on the recorded speech, we transcribed the speech to orthography[2] which included the tone marks on the syllables.

It was not necessary to transcribe the recordings made by all three recorded respondents because the tonal rules implemented by the extended algorithm are not restricted to speakers of a specific Sesotho dialect. We used only one respondent's recording for the transcriptions. The respondent we chose for transcriptions had to have produced the most "clearest" tones. By "clearest", we mean the most pronounced and thus better perceived tones for the transcribers (labelers). We needed clearly pronounced tones for the transcription process since the labelers used mostly perception in addition to acoustic means to transcribe the speech as discussed in Section 4.5.1.

To choose the speaker with the "clearest" tones in their speech from the three respondents, two processes of elimination were employed.

- Eight tonal minimal pairs were used to determine the speaker's ability to distinguish tone.

[2]The alphabetic spelling system used by the language.

- The other process involved perceptually and acoustically verifying the clarity of the tones the speaker produced in their speech.

Tonal minimal pairs are two words or phrases which differ only in tone and have a distinct meaning. We used them to test if the speaker can distinguish tone. We named our respondents as TK, MM and NM. In this process of elimination, respondent NM could not perceptually identify the tonal distinction in the following minimal pair:

(4.1) Le já dijó
"you eat food"

Lé já dijó
"he/she eats food"

This minimal pair shows the changing of tone affecting the subject marker *le*. In the first sentence, *le* refers to a 2^{nd} person plural whereas in the second sentence it refers to a 3^{rd} person singular. We thus excluded respondent NM's recording from the transcriptions.

Respondent TK produced clear tones and spoke in a natural voice compared to respondent MM, who spoke in an unnatural voice and exaggerated the tone and words. In some instances, speaker MM produced each word without taking the context into consideration. This is problematic since the tonal rules are restricted within certain phonological boundaries and some of the rules in specific morphological contexts. Also, a word in isolation might be tonally ambiguous and one would need the context to pronounce the appropriate tone pattern as discussed in Section 3.5.1.

Respondent TK's recording was transcribed by three independent labelers, each having some experience in tonal studies. The transcriptions are in orthography and include tone labels. In the next section, we discuss how the labelers transcribed the speech by TK.

4.5 Transcription of the Sentences

4.5.1 Transcription Process

The transcription of prosody is widely used in linguistic research as well as in the development and testing of speech systems [Syrdal and McGory 2000]. It is used in understanding linguistic variables and their relation to the interpretation of human speech. In our study, three independent expert labelers transcribed tone as perceived from TK's speech. We used these transcriptions to evaluate the extended algorithm. The transcriptions represent spoken language in written form including tone labels.

Perceptual tests are the most widely used methods for evaluating TTS systems. In these tests, native speakers of a language in question are required to grade the synthesized speech with respect to naturalness. This is because synthesized speech produced by TTS systems needs to ideally sound as natural as human speech.

We used transcriptions to evaluate the extended algorithm since they represent Sesotho speech as pronounced by a native Sesotho speaker. Given that the extended algorithm is developed for the purpose of tone modeling for a Sesotho TTS system, it is crucial in our study that the tone labels predicted by the extended algorithm match those in the transcriptions. We did not employ any statistical methods to predict the tone labels as such methods are not based on perception and thus cannot give us an indication of how well the extended algorithm can fare on a perception test. However, we employed statistical methods to establish the reliability of the transcriptions as discussed in Section 4.5.2.

Initially, we aimed to use 12 labelers to transcribe TK's recording. These labelers were to be divided into two groups, six being experts in tonal studies and the other six having no experience in tonal studies.

In each of the two groups, three labelers were to be native Sesotho speakers and the other three were to be speakers of non-tonal languages. However, in the expert group, only three labelers were found whereas in the non-expert group, all six labelers were found. Table 4.1 summarizes the initial criteria for the labelers.

Group	Native	Non-native
Expert	1	2
Non-expert	3	3

Table 4.1: Labeler criteria

We found that the labelers in the non-expert group had difficulty transcribing the tone labels. In the non-expert group, even native speakers had difficulty transcribing the labels. It is not straightforward to transcribe tone and one usually needs some experience with the transcription system. Also, native speakers of tonal languages are usually not aware of their language's tonal system as this is not taught at schools. Since there were more labelers in the non-expert group than in the expert group, they would have outweighed the expert labelers in the majority votes used to choose the tone labels in the syllables in cases of disagreements between the labelers as discussed later in this section. Thus, we chose to use the three expert labelers to transcribe the tone labels. Although it would have been preferable to have more expert transcribers, it was not easily achievable since there are not many experts in this area of study.

The transcribers had different backgrounds with respect to tonal studies, with two of them being linguists and the other being a speech engineer. Furthermore, two of the transcribers were speakers of Germanic languages (linguist (A) and speech engineer (B)) and the other was a native Sesotho speaker (linguist (C)). The labelers who are non-native Sesotho speakers developed a framework that introduces an algorithm which derives word-level tone assignments for the Sotho-Tswana languages [Zerbian and Barnard 2010]. It is this framework on which the study by Raborife [2009] is based. The labeler who is a native Sesotho speaker implemented this framework by developing the basic algorithm [Raborife 2009]. Appendix E shows the transcriptions of the labelers.

To evaluate the improvement made by the extended algorithm on the basic algorithm, we used the labels transcribed by our individual labelers. The labelers were not given a training manual, thus they relied on perception and acoustic means to label the tone marks. The individual transcriptions were compared across the three labelers, finding an agreement of 55.9% (316 out of 565 tone-bearing syllables). In the case of disagreements between the labelers, the final tone label was decided by a majority vote, that is, it was based on the tone label that two labelers agreed on. The final tone labels are shown in Appendix F.

The agreement between each pair of transcribers out of 565 tone-bearing syllables is shown in Table 4.2.

Transcriber	Matched Syllables	Agreement Percentage
A and B	375	66.4%
A and C	374	66.5%
B and C	400	70.8%

Table 4.2: Agreement with respect to each pair of transcribers

The labelers used Praat to label the tone marks on the syllables whose tones they could not audibly perceive. Praat is a speech analysis software that has a pitch tracking algorithm which estimates pitch

from a speech waveform [Boersma 2001]. It produces a pitch contour that estimates pitch, which the labelers observed when marking the tone labels. In perceptual analysis, the transcribers relied on hearing, interpreting and understanding the tone patterns in the speech. However, perception is subjective and varies from one transcriber to the other. This serves as a disadvantage in our study as discussed in the following section.

According to Govender *et al.* [2005a], there are two most widely used pitch tracking algorithms, namely, Praat and Yin. Govender *et al.* [2005a] compare these algorithms in extracting pitch for isiZulu. In this study, it has been found that Praat performs slightly better than Yin. According to Govender [2006], for the Nguni languages, Praat is more accurate and noise-resistant than Yin, which has no post-processing on the input speech. Thus, our transcribers used Praat for the acoustic analysis aspect on the test data. An example of the extracted pitch contour (in blue; if you have no colour, it is the faint line between 0Hz and 5000Hz across the spectrum) by Praat is shown in Figure 4.1.

Even though the Praat pitch tracking algorithm has been found to be accurate, it does not fare well with unvoiced sounds as shown in Figure 4.1 (the unvoiced sound *f* is highlighted and has no visible pitch track) [Govender *et al.* 2007]. In speech we have voiced and unvoiced sounds. When producing voiced sounds the vocal folds vibrate and when unvoiced sounds are produced the vocal folds are at rest. According to Govender *et al.* [2007], unvoiced segments usually do not have a defined pitch contour (recall the unvoiced sound *f*). Unvoiced segments were ignored when assigning the tone labels to the syllable. A syllable always contains at least one voiced segment.

In the study by Raborife [2009], the labelers reached unanimous agreement on the tone labels in 72.3% of the cases. The recordings used in Raborife [2009] were for speech synthesis purposes, thus the speaker was asked to produce a rather monotonous speech melody. The higher agreement in that study can be attributed to this speech melody. The sentences chosen in this study were fairly complex and spoken in a natural, fast voice making it difficult to track pitch changes.

4.5.2 Inter-Transcriber Reliability

As discussed in Section 4.4, three labelers transcribed the 45 recorded sentences independently. They reached unanimous agreement on the tone labels in 55.9% of the cases. A reason for such a low agreement between the transcribers might lie in the fact that perception is subjective and that the tone transcribers did not have a training manual. Two of our labelers were speakers of non-tonal languages. Studies have shown that speakers of non-tonal languages are less sensitive to the perception of tone but are not necessarily tone "deaf" [Halle *et al.* 2004; Gandour and Harshman 1978]. In these studies, the speakers of the non-tonal languages were naive and had no experience in tonal studies. In our study, all the labelers were sensitive to tone. Furthermore, they had different backgrounds and experiences regarding tone in Bantu languages.

Our study relied heavily on the transcriptions since they formed a central part in evaluating the extended algorithm. Thus, it was important to statistically verify the transcriptions since there was a fairly low agreement between our transcribers. To determine the reliability of the transcriptions, we used the kappa (κ) coefficient which is a statistical measure used for analyzing the reliability of agreement between a number of labelers (raters) during observer reliability studies [Fleiss 1971; Brennan and Prediger 1981; Randolph 2005]. This statistical measure was more reliable than a simple percentage agreement calculation since it took into account the agreement occurring by chance between the labelers.

Fleiss' kappa is a well-known multirater kappa which assumes that raters are forced to assign a certain number of cases to each category, hence, it is referred to as a fixed-marginal kappa [Brennan and Prediger 1981]. For instance, in our study, it assumes that the labelers are restricted in the number of syllables (cases) they can assign to each of the tone labels (categories). However, this is not the case as the labelers are free to assign any number of the syllables to each of the two tone labels. We thus make

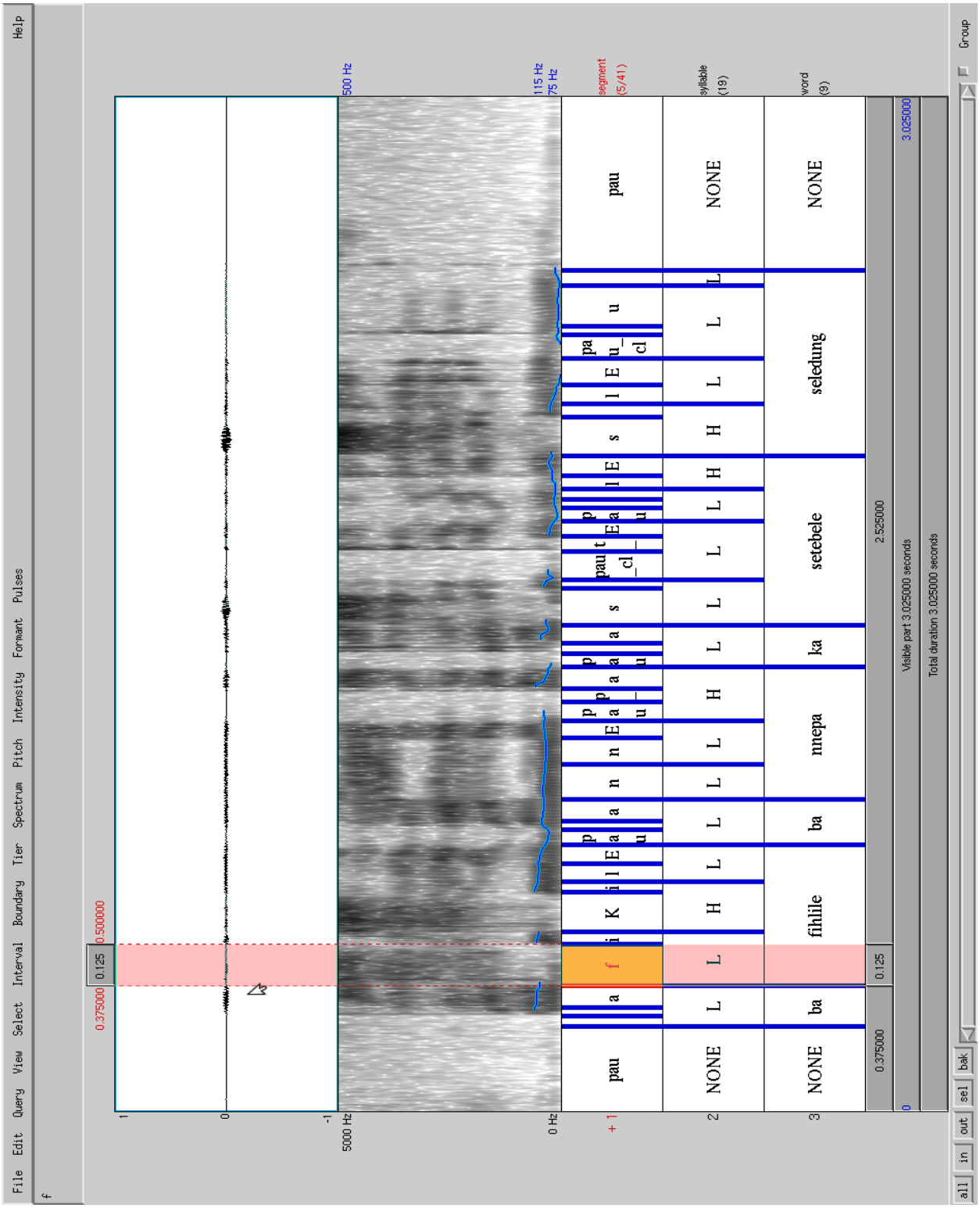

Figure 4.1: Extracted signal from a Sesotho speech recording

use of a free-marginal kappa which assumes that raters are not restricted to assign a specific number of cases to each of the categories as proposed by Brennan and Prediger [1981].

Randolph [2005] defines the free-marginal kappa, κ, as follows:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{4.2}$$

$\bar{P}$ is the observed percentage agreement and $\bar{P}_e$ is the percentage of agreement expected by chance alone.

According to Fleiss [1971], the factor $1 - \bar{P}_e$ gives us the measure of agreement that is attainable above what would be predicted by the labelers by chance and $\bar{P} - \bar{P}_e$ is the degree of agreement actually achieved above chance. If κ = 1, then there is complete agreement between the labelers. If $\kappa \leq 0$, then there is no agreement among the labelers other than what would be expected by chance, as if the raters had simply guessed every rating.

We used an online kappa calculator which allows us to enter all the necessary variables and then calculates the κ-value [Randolph 2008]. The free-marginal κ-value between three labelers, assigning 565 syllables to two categories as calculated by Randolph [2008], is as follows: $\kappa = 0.41$.

The positive value of kappa indicates that the agreement is better than what would have been expected by chance. To interpret this value we use Table 4.3 as described by Landis and Koch [1977]. According to the classification by Landis and Koch [1977], the labelers in our study seem to be in moderate agreement on the tone labels they have transcribed, since $\kappa = 0.41$. Recall that a perfect agreement would have a kappa of 1, and chance agreement would have a kappa of 0 or less. Thus we can use our transcriptions to evaluate the extended algorithm and test our hypothesis.

κ	**Interpretation**
≤ 0	Poor Agreement
$0.0 - 0.20$	Slight Agreement
$0.21 - 0.40$	Fair Agreement
$0.41 - 0.60$	Moderate Agreement
$0.61 - 0.80$	Substantial Agreement
$0.81 - 1.00$	Almost Perfect Agreement

Table 4.3: Interpretation of κ values

4.6 Conclusion

Forty-five Sesotho sentences were chosen to test the extended algorithm. These sentences were chosen in such a way that they represent the contexts (tenses) that are necessary for the application of this algorithm. They were recorded by three native Sesotho speakers (respondents) but only one respondent's recording was transcribed.

The recording was transcribed by three labelers, each with experience in tonal studies. The transcriptions were in orthography and included tone labels. They were based on acoustic means and/or perception. We have shown that there is moderate agreement between the transcribers on the labels they have transcribed. This allowed us to use the transcriptions to evaluate the extended algorithm and to determine if it was an improvement on the basic algorithm.

In the following section, we describe the manner in which we used the transcriptions to evaluate the extended algorithm. We also present the results to answer whether or not the extended algorithm is an improvement on the basic algorithm.

Chapter 5

Analysis and Results

5.1 Introduction

To analyze the extended algorithm, we compared the output produced by the algorithm against the transcriptions. As discussed in the previous section, the transcriptions were labeled by three independent transcribers. The transcriptions are based on sentences recorded by a native Sesotho speaker and include tone labels.

Since the aim of our study was to improve the basic algorithm, we needed to know if it had been improved by testing it on the same data that the extended algorithm was tested on. We measured improvement as an increase in the number of matched tone labels between the extended algorithm and the transcriptions as compared to the number of matched tone labels between the output by the basic algorithm and the transcriptions. Moreover, in cases of mismatches between the extended algorithm output and the transcriptions, we attempted to provide plausible reasons for the mismatches based on the observations we made on the data.

This chapter provides a detailed description of the analysis of our research study. The observations that emerged in the research study as well as the results of the research study are also presented.

5.2 Analysis

5.2.1 Introduction

Our analysis consisted of the following two phases:

- Comparing the tone labels predicted by the basic algorithm to the tone labels on the transcriptions.
- Comparing the tone labels predicted by the extended algorithm to the tone labels on the transcriptions.

Our analysis involved comparing the tone labels of the syllables which are within the phonological phrase domain, as opposed to Raborife [2009], whose analysis was restricted to the clitic phrase domain. Since the application of the basic algorithm is restricted to the clitic phrase domain, which in Sesotho consists of a verb stem and the prefixes preceding it, the tone labels of the syllables that are outside this domain were left with their underlying tones as defined by Du Plessis *et al.* [1974] when comparing the tone labels predicted by the basic algorithm to their transcribed counterparts. These syllables were also included in our analysis.

Example 5.1 shows a sentence from our test data which shows the transcribed sentence in TR and the output by the basic algorithm in BA. In this example, the basic algorithm applies the high tone spread

rule (HTS1) within the clitic phrase whereby the high tone on the first syllable of the verb stem *kganna* spreads to the second syllable. In Example 5.1, there are five mismatches of which the two that are within the clitic phrase are due to the discrepancies in the underlying tones between the transcription and the basic algorithm. In TR, within the clitic phrase, it seems that the underlying high tone is on the second syllable of the verb stem which spreads to the last syllable of the verb stem by HTS1 whereas in BA the underlying high tone is on the first syllable of the verb stem and spreads to the second syllable of the verb stem by HTS1. The other three mismatches are outside the clitic phrase domain and thus exceed the scope of application of the basic algorithm. Therefore, we will not discuss the other three mismatches.

(5.1) TR: Modísána ó kgaṅná dinku tsé sekete.
BA: Modisana (ó̲ (kgá̲ṅna)ϖ)C dinku tse sekete
"The shepherd is driving a thousand sheep"

Example 5.2 shows a sentence from our test data that illustrates the application of the extended algorithm. In this example, there are three mismatched syllables which are attributed to the realization of underlying tones in the transcriptions. The first two mismatches are within the clitic phrase domain and have been discussed above. The third mismatch is on the last syllable of the noun *dinku* on which the transcribed tone label is low whereas according to Du Plessis *et al.* [1974] this syllable is underlyingly high-toned. EA is the output of the extended algorithm and TR is the transcribed sentence.

(5.2) TR: Modísána ó kgaṅná dinku tsé sekete.
EA: ((Mo̲dísána)ϖ (ó̲ (kgá̲ṅna)ϖ)C (dinkú̲)ϖ)ϕ (tsé (sekete)ϖ)ϕ
"The shepherd is driving a thousand sheep"

Our hypothesis is that the extended algorithm is an improvement on the basic algorithm. If the number of matched tone labels between the tone labels predicted by the extended algorithm and the transcribed tone labels is greater than the number of matched tone labels between the tone labels predicted by the basic algorithm and the tone labels transcribed by the labelers, then we have validated our hypothesis and justified the need for the extended algorithm.

To test our research hypothesis, each sentence – with the tone labels as predicted by the algorithm in that phase – was compared to its transcribed counterpart. Comparing the output of the basic algorithm with the transcriptions yielded 365 matched tone labels whereas comparing the output of the extended algorithm (on the same 45 sentences on which the basic algorithm was tested) yielded 409 matched tone labels out of 565. Table 5.1 shows a summary of the results within each phase.

Phase	**Percentage of Matched Tone Labels**
Basic Algorithm	64.6%
Extended Algorithm	72.4%

Table 5.1: Number of matched tone labels within each phase

Since our study was concerned with implementing the extended algorithm, our analysis focused on the mismatches found when comparing the tone labels predicted by the extended algorithm to the transcribed tone labels. In the next section, we provide a detailed analysis of the second phase which involved comparing the output of the extended algorithm to the transcriptions.

5.2.2 Analysis

There are 565 tone-bearing syllables in the 45 sentences. We used these 565 syllables in the analysis

of our research study. In this section, we provide a detailed description of the mismatches between the output by the extended algorithm and the transcriptions.

In analyzing the extended algorithm, we grouped the sentences according to the similarities found in their mismatches. That is, sentences whose mismatches occurred under similar contexts were placed together in one group. Our analysis yielded six groupings of which the first one had perfect matches on all syllables in the sentence that was being compared.

Table 5.2 shows the five groups of sentences with mismatched syllables. We also present the number of mismatched tone labels in each group. In total, we have 156 mismatched syllables out of 565 syllables.

Group	Number of Mismatched Tone Labels (out of 156))
B	3
C	22
D	53
E	11
F	67

Table 5.2: Number of mismatched tone labels within each group

Group B

The mismatches found in Group B are due to underlying high tones at the end of the phonological phrase being realized as low tones in the transcriptions. These syllables are not deleted by the finality restriction rule since they are not a consequence of a tone spread rule but are underlying on the syllables with which they are associated. Khoali [1991] describes a rule that can account for this phenomenon. This is the mid-tone insertion rule.

By the mid-tone insertion rule, a high tone on the ϕ-domain (phonological phrase) final syllable preceding a falling tone (HL-tone) is realized as a mid-tone [Khoali 1991]. In our transcriptions, the transcribers used only two tone levels, high or low; the mid-tone is not transcribed in the labels. Our transcribers transcribed the syllables which meet the phonological requirements for the application of the mid-tone insertion rule as low-toned. The extended algorithm does not implement the mid-tone insertion rule. Thus, syllables which meet the phonological requirements that are necessary for the application of this rule are left as high-toned. This rule is exemplified in 5.3, shown below.

In Example 5.3[1] (see Appendix H, sentence 19), the underlying high tone on the final syllable of the pronoun *lona* surfaces as a low tone. This syllable is at the end of the phonological phrase domain and the two syllables preceding it, *yá* and *lo*, have high and low tones respectively (HL); together they are a falling tone. By the mid-tone insertion rule, this syllable is realized as a mid-tone. However, our labelers transcribed a low tone. This rule accounted for 1.9% of the mismatches.

(5.3) ((Re tla (hl̲ó᷄mpha)ϖ)C (meetlo)ϖ)ϕ (yá̲ (lonä)ϖ)ϕ
1PL.SC FUT respect NP4.tradition PC4 PRN
"We will respect your traditions"

Group C

In Group C, the transcriptions showed a pattern whereby an underlying high tone spreads to the succeeding syllable as expected by the high tone spread rule. This high tone on the target syllable (say β) then

[1]The diacritic ä indicates a mid-tone.

spreads further onto the syllable following the target syllable (say α). Then, the high tone is deleted on its source syllable. Observations from our data show this rule to be a clitic phrase domain-span rule. Furthermore, the clitic phrase within which this rule applies should not be the rightmost edge domain in the sentence. That is, the clitic phrase within which this rule applies should not be the last clitic phrase within the phonological phrase domain it is in.

Example 5.4[2] (see Appendix F, sentence 2) shows the subject marker á in the second clitic phrase which is underlyingly high-toned according to Du Plessis *et al.* [1974] surfacing with the low tone. The first syllable of the verb stem *fetoha* is underlyingly low-toned as described by Du Plessis *et al.* [1974] but surfaces with a high tone that is a result of the high tone spread rule between the subject marker *a* and the syllable *fé*. The syllable *tó* following *fé* also surfaces with a high tone. According to Du Plessis *et al.* [1974], this syllable is underlying low-toned and since *fé* is also underlyingly low-toned, the high tone on *tó* is not a result of the high tone spread rule. It seems to have been associated with the subject marker *a*. The high tone on the subject marker (within the second clitic phrase) spread to the two syllables following it before delinking on the subject marker. This phenomenon accounts for 14.1% of the mismatches.

(5.4) ((Matsha)ϖ a (tshóloha)ϖ)C)ϕ (((a (fétóha)ϖ)C (maw<u>a</u>tle)ϖ)ϕ
NP6.lake SC6 pour SC6 turn NP6.sea
"The lakes are overpouring, they are turning into seas"

Group D

In Group D, there were discrepancies in the syllables regarding their underlying tones as described by Du Plessis *et al.* [1974] and their transcribed counterparts. In other words, a syllable which is marked as underlyingly high-toned (or low-toned) in Du Plessis *et al.* [1974] is transcribed as low-toned (or high-toned) by the labelers.

Example 5.5 (from [Khoali 1991, p.126]) shows an instance whereby the preposition *ho* is underlyingly high-toned (as described by Du Plessis *et al.* [1974]) in EA and is low-toned in TR. In EA, the high tone on the preposition spreads to the first syllable of the noun *morena* as expected by the high tone spread rule. In TR, since the preposition is low-toned, the first syllable noun is also low-toned since the phonological requirements of the high tone spread rule are not met. When comparing TR and EA, we get two mismatched syllables. The first mismatched syllable (*ho*) is due to the difference in the underlying tones between TR and EA. The second mismatch is a consequence of the first mismatch for which the high tone spread rule does not apply in TR due to its phonological conditions not being met. In EA, the high tone spread rule applies since its phonological conditions are met.

(5.5) TR:ho morena
EA: (h<u>ó</u> mó)ϖ-(rena)ϖ
PREP NP1-king
"to the king"

The mismatches between the output of the extended algorithm and the transcribed sentences are attributed to the disagreements between the labelers in the transcriptions. Recall that our transcriber agreement is 55.9% with a kappa value of 0.41, thus illustrating the reliability of our transcriptions. Although we have shown that our transcriptions are reliable and can be used to evaluate both the basic and extended algorithms as well as test our hypothesis, we have to acknowledge that some of the agreement between our transcribers is solely by chance. Therefore, we attribute the mismatches in this group to the

[2]Please note that although the transcriptions do not have phonological domains, they are inserted in this example to aid the reader.

chance agreement between our labelers, thus accounting for the discrepancies between the literature and the transcriptions. It is also possible that the literature is not accurate, or that our speaker has different tone assignments to the relatively old literature.

In Example 5.6 (see Appendix F and H, sentence 1), TR shows the first syllable of the noun stem *noha* having a low tone. According to Du Plessis *et al.* [1974], this syllable is underlyingly high-toned as shown in EA. Since in the transcriptions it is low-toned, the syllable following it, *ha*, surfaces with a low tone in TR. In EA, the syllable *ha* surfaces with a high tone as expected by the high tone spread rule.

(5.6) TR: Noha é fálla metsí
EA: ((Nóhá)ϖ (é (fálla)ϖ)C (metsí)ϖ)ϕ
"The snake is migrating from its waters"

Other instances of mismatches that can be attributed to the transcriptions occur when the phonological requirements of a tonal rule are met in the transcriptions but the tonal rule fails to apply. In Example 5.7, TR shows the first syllable of the verb stem *bala* surfacing with a low tone even though it is preceded by a high-toned subject marker *ba*. We know that *ba* is underlyingly high-toned in TR since it is not preceded by another high-toned syllable. In EA, the first syllable of the verb stem *bala* surfaces with a high tone as expected by the high tone spread rule since it is preceded by a high-toned subject marker *ba* within the same clitic phrase domain.

(5.7) TR: Barwétsána bá bala dinkgo
EA: ((Barwétsána)ϖ (bá (bála)C)ϕ (dinkgo)ϖ)ϕ
"The girls are counting the clay pots"

This group accounts for 34% of the mismatches.

Group E

The transcriptions in Group E showed trisyllabic noun stems at the end of the phonological phrase domain surfacing with high tones on their first syllables and low tones on all the other succeeding syllables. A trysyllabic noun stem is a noun stem with three syllables, for example *di.po.tso*[3] "questions".

Example 5.8 shows a trisyllabic noun stem *dipotso* with an underlying high tone on its second syllable in *po* on EA as described by Du Plessis *et al.* [1974]. In TR, this syllable surfaces with a low tone and the first syllable, the class marker *di*, surfaces with a high tone, even though as discussed in Khoali [1991], all class markers are underlyingly low-toned.

(5.8) TR: Ó tla útlwá bá bótsá dípotso
EA: ((Ó tla (útlwá)ϖ)C (bá (bótsá)ϖ)C (dipótso)ϖ)ϕ
"He/She (you) will hear them asking questions"

This group accounts for 7.1% of the mismatches.

Group F

The mismatches grouped in F showed peculiar patterns which are neither represented in the groups discussed above nor can they be accounted for by any of the tonal rules discussed in Khoali [1991]. These mismatches are open for further research based on the linguistic characteristics of the behaviour of tone in Sesotho. This group accounts for 42.9% of the mismatches found in our analysis.

[3]The full stops in the word represent syllable boundaries, that is, they separate the syllables.

5.2.3 Problems Encountered in Our Analysis

In Sesotho, function words are highly ambiguous. A single word such as *o* can either mean "he/she"(ó) or "you" (o) and this semantic distinction is affected by tone. Orthography does not include tone marks and when we recorded the sentences, respondents were not given the English meaning of the sentences. Thus, we have no means to discern to which meaning the respondent is referring. The word *o* accounts for 0.53% of the syllables in our study.

One solution to this problem is to use both instances of *o* in our analysis. That is, compare *o* to its transcribed counterpart when it is both high and low-toned as shown in Example 5.9. However, in all the cases where this word has been transcribed, the labelers transcribed it as a high tone indicating that the respondent was referring to *he/she*. In our analysis, we assumed that this morpheme is underlyingly high-toned.

(5.9) (a) TR: Ó fúmané monna wá sébele.
EA[4]: ((Ó (fúmáné)ϖ)C (monná)ϖ)ϕ ((wá sé)ϖ(bele)ϖ)ϕ
"He/she has found a man of substance"

(b) TR: Ó fúmané monna wá sébele.
EA[5]: ((O (fúmáné)ϖ)C (monná)ϖ)ϕ ((wá sé)ϖ(bele)ϖ)ϕ
"You have found a man of substance"

5.3 Additional Analysis

In the study by Louw *et al.* [2005], the consensus between the TTS system developers and the isiZulu speakers who evaluated the system was that all the syllables should be treated as unmarked as it would allow the system to produce understandable speech. In our study, we also compared the transcriptions to unmarked syllables. As discussed in Chapter 2, in Sesotho, unspecified (unmarked) syllables received a low tone by the default low tone insertion rule [Demuth 1995]. We compared the transcriptions to the sentences with all the syllables assigned low tones. These sentences were the same sentences used to test the basic and extended algorithms. Comparing the sentences on which all syllables were marked as low-toned with the transcriptions yielded 340 matched tone labels out of 565 tone labels.

In addition, we also assumed a case whereby a TTS system only realized underlying tones. For instance, consider a case in which the consensus between a particular TTS system developer and native speakers of the target language is that the tone labels on the syllables of the words be realized as the underlying tone labels, as described in a tone dictionary of the target language, without applying any tonal rules. The target language is the language for which the TTS system is developed. In our case, we compared the 45 sentences with only the underlying tone information marked on them to their transcribed counterparts. In our comparison, we compared the tone labels on the syllables of the words in the sentences. This approach gave us 366 matched tone labels out of 565 matched tone labels. Table 5.3 provides a summary of our additional analysis.

The results of our analysis showed that applying the basic algorithm to the selected sentences yields slightly worse results than using only underlying tones, a margin of 0.1%. One of the reasons for this is that the basic algorithm restricts its application to the clitic phrase domain. In our analysis, all the syllables that are outside the clitic phrase domain were left with their underlying tones.

Our sentences were chosen to present the contexts that are necessary for the basic algorithm as well as the extended algorithm to apply. The basic algorithm only applies three tonal rules, two of which (tone spread rules) provide the phonological contexts for the other tonal rules implemented by the extended

[4]When *o* is underlyingly high-toned.

[5]When *o* is underlyingly low-toned.

Phase	Percentage of Matched Tone Labels
Unmarked Tones	60.2%
Underlying Tones	64.7%

Table 5.3: Summary of the additional analysis

algorithm. The mismatches between the basic algorithm and the transcriptions occur because the two tonal rules which are implemented by this algorithm create contexts which allow other tonal rules to apply that are not implemented by the basic algorithm.

5.4 Results

In our study, we tested a single research hypothesis which assumes that the extended algorithm is an improvement on the basic algorithm. Our research hypothesis was formulated as follows:

> The extended algorithm improves the basic algorithm by increasing the number of matched tone labels.

To validate our hypothesis, we tested both algorithms as follows:

1. The first phase included analyzing the improvement made by the extended algorithm on the basic algorithm when both algorithms are restricted to the clitic phrase domain.

2. The second phase included analyzing the improvement made by the extended algorithm on the basic algorithm when there is no restriction in the domain of application on which the algorithms can apply.

In both tests, there were two phases. The first phase was to determine if the number of matched tone labels between the extended algorithm and the transcriptions is greater than the number of matched tone labels between the basic algorithm and the transcriptions. The second phase involved statistically verifying the significance of the results found in the first phase.

To statistically verify the results in each phase, we employed an unpaired two-sample t-test. A two-sample t-test is a statistical test which assesses if there is a significant difference between two groups of dependent samples [Boslaugh and Watters 2008]. It is mostly used when the variances of two normal distributions are unknown and the sample size is relatively small. The t-value is calculated as follows:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{5.10}$$

where $\mu_N = Np$ and $\sigma_N^2 = Np(1-p)$, $p\,\epsilon[0,1]$ is the number of matched tone labels in a particular phase of evaluation and N_p is the total number of tone labels.

The t-value has a distribution with $N_1 + N_2 - 2$ degrees of freedom (df) that advances the normal distribution. Degrees of freedom are the number of values that are free to vary. When the t-test is employed, the first thing to be done is to calculate the t-value.

To test significance, we need to set a significance level (α). In statistics, a result of probability greater or equal to 0.05 is generally considered to be significant: In other words, a distribution is statistically significant if the likelihood that it has come about by chance is below 0.05 (5%).

A t-test may be one-tailed or two-tailed. In a one-tailed t-test, the differences between the groups are not only explained but are also specified in the direction in which they exist. For instance, here we are saying that we expect children in Brazil to play better football than children in South Africa. A two-tailed t-test declares the difference between the groups but does not specify the direction in which it exists. In our study, we used the one-tailed t-test since we correlated the number of tonal rules an algorithm applies and its restricted domain of application in a sentence to the number of matched tone labels.

We then compare the t-value to the appropriate critical t-value in the t-test table (See Appendix I). The critical t-value is calculated by aligning $\alpha = 0.05$ with $df = N_1 + N_2 - 2 = 565 + 565 - 2 = 1128 \approx \infty$. This gives us a critical t-value of 1.645. If $t \geq 1.645$, then the extended algorithm is significantly better than the basic algorithm.

In the following section, we discuss the two phases described above in detail.

5.4.1 Restriction in the Application of Both Algorithms

In this phase of our testing, we tested the basic algorithm on the 45 selected sentences. The extended algorithm was also tested on the same data set. Our data was chosen in such a way as to represent all the contexts necessary for the application of both algorithms. The syllables that are outside the clitic phrase domain were excluded in the analysis. This allows the basic algorithm a fair chance to perform to its maximum ability and ensures that there is no bias against it. The total number of syllables used in this analysis is 327.

We then evaluated the basic algorithm by comparing its predicted tone labels to the tone labels in the transcriptions within the clitic phrase domain. The extended algorithm was also evaluated in the same way. That is, the tone labels predicted by it were compared to their transcribed counterparts within the clitic phrase domain. The first evaluation, involving the transcriptions and the basic algorithm, yielded 194 matched tone labels. The second evaluation, between the transcriptions and the extended algorithm, yielded 226 matched tone labels. The extended algorithm predicted 32 more matched labels than the basic algorithm.

In our study, we measured improvement by the number of matched tone labels produced during our evaluations. Recall that for our hypothesis to be validated, the number of matched tone labels between the basic algorithm and the transcriptions has to be less than the number of matched tone labels between the extended algorithm and the transcriptions. Therefore, we have validated our algorithm since the evaluation of the extended algorithm yields 32 more matched tone labels than the evaluation of the basic algorithm.

To statistically verify the results in this phase, we perform the t-test as follows:

$$t = \frac{\mu_{EA} - \mu_{BA}}{\sqrt{\sigma_{EA}^2 + \sigma_{BA}^2}} = \frac{226 - 194}{\sqrt{69.9 + 79.1}} = 2.62 \begin{cases} \mu_{EA} = 327 \times 0.69 & p = \frac{226}{327} \\ \mu_{BA} = 327 \times 0.59 & p = \frac{194}{327} \\ \sigma_{EA}^2 = 225.6 \times 0.31 & 1 - p = 0.31 \\ \sigma_{BA}^2 = 192.9 \times 0.41 & 1 - p = 0.41 \end{cases} \tag{5.11}$$

As calculated above, $t = 2.62 > 1.645$, thus statistically verifying our results. This validates our research hypothesis which states that the extended algorithm is an improvement on the basic algorithm.

5.4.2 Non-restriction in the Application of Both Algorithms

The previous section shows that the extended algorithm is an improvement on the basic algorithm when both algorithms are in a restricted domain. The extended algorithm, however, does not restrict its domain

of application to a particular subset of a sentence. In this section, we look at a scenario whereby both algorithms were tested to their maximum potential.

Both algorithms were tested on the 45 sentences independently of each other. We then compared the tone labels predicted by both algorithms independently to their transcribed counterparts. In evaluating the basic algorithm, the syllables outside the clitic phrase domain retained their underlying tones as defined in Du Plessis *et al.* [1974].

Comparing the tone labels predicted by the extended algorithm, across the whole sentence in the 45 sentences, to their transcribed counterparts gave us 409 matched tone labels. Comparing the tone labels predicted by the basic algorithm, across the whole sentence in the 45 sentences[6], to their transcribed counterparts yielded 365 matched tone labels. The extended algorithm predicted 44 more matched labels than the basic algorithm in this phase.

We employ the t-test in order to statistically verify the results in this phase.:

$$t = \frac{\mu_{EA} - \mu_{BA}}{\sqrt{\sigma^2_{EA} + \sigma^2_{BA}}} = \frac{409 - 365}{\sqrt{113.9 + 128.6}} = 2.82 \begin{cases} \mu_{EA} = 565 \times 0.72 & p = \frac{409}{565} \\ \mu_{BA} = 565 \times 0.65 & p = \frac{365}{565} \\ \sigma^2_{EA} = 406.8 \times 0.28 & 1 - p = 0.28 \\ \sigma^2_{BA} = 367.3 \times 0.35 & 1 - p = 0.35 \end{cases} \tag{5.12}$$

As calculated above, $t = 2.82 > 1.645$, thus statistically verifying our results. The method of analysis employed in this phase also validates our research hypothesis.

5.5 Conclusion

The basic and extended algorithms were tested on the same data set (45 selected sentences) independently of each other. The tone labels predicted by both algorithms were then compared independently to the transcribed tone labels. The results of the comparisons were then compared to each other in order to evaluate both algorithms. The results arising from the comparison between the basic algorithm and the transcriptions were compared to the results arising from the comparison between the extended algorithm and the transcriptions. Our evaluation shows that the extended algorithm is a significant improvement on the basic algorithm, thus validating our research hypothesis. This chapter has presented the analysis and results of the research study.

The next chapter provides a summary of our research study. The limitations of this research are discussed as well as suggestions for further research in this area of study.

[6]The syllables outside the clitic phrase domain have their underlying tones.

Chapter 6

Conclusion

6.1 Introduction

In the previous chapter, we showed that the extended algorithm is an improvement on the basic algorithm. The extended algorithm is the first step to tone modeling for languages such as Sesotho, which do not have any tone information in the orthography.

In this chapter, we provide a summary of this research study, including the tonal rules on which the basic and extended algorithms are based. Furthermore, we introduce ideas for further research which would complete the tone modeling process for Sesotho. These ideas include developing an algorithm that can map appropriate pitch values to the tone labels predicted by our algorithm as well as conducting a perception test which can test whether implementing tone into TTS systems developed for Bantu languages improves the perceived naturalness of the speech produced by the system.

6.2 Overview and Basis of the Research Study

Bantu languages are tonal languages; they use tone to distinguish between lexical or grammatical meaning. Thus, tone is an important aspect of prosody for Bantu languages. TTS systems need detailed prosodic models of a language in order to sound natural and interesting. Hence, it is important for Bantu languages that tone is implemented for speech technologies developed for this group of languages.

Tone in Bantu languages is assigned to each syllable. To implement tone into a TTS system developed for a Bantu language, such as Sesotho, it would have to be known which tone label a syllable should be assigned. This is the basis of tone modeling for Bantu languages such as Sesotho as shown by Raborife [2009] and isiZulu as suggested by Louw *et al.* [2005]. Hence, our study involves improving an algorithm which makes tone label predictions on the syllables in a Sesotho sentence.

Raborife [2009] has shown that such an algorithm should use linguistic theories, particularly, phonological theories to make such predictions. Raborife [2009] developed an algorithm based on three Sesotho tonal rules as described by Khoali [1991]. The application of these tonal rules is restricted within certain phonological constituents. The three Sesotho tonal rules, namely the high tone spread rule, the grammatical tone insertion rule and the iterative high tone spread rule, which form the basis of Raborife [2009]'s study are defined as follows:

- The high tone spread rule involves the spread of a high tone from the syllable with which it is associated to the adjacent syllable. This rule applies strictly within the phonological word domain and the clitic phrase domain.

- The grammatical tone insertion rule associates the second syllable of a verb stem with a high tone. This rule is morphologically governed; it occurs in certain morphological contexts.
- The iterative high tone spread rule involves the spreading of a grammatical high tone until the end of the phonological word that contains the verb stem. This rule is ordered after the grammatical tone insertion rule.

Our study focused on implementing four other Sesotho tonal rules, namely right branch delinking, left branch delinking, specifier high tone delinking and finality restriction, to the algorithm developed by Raborife [2009]. These rules are also restricted within certain phonological domains, and most importantly "feed" directly from the high tone spread rule. They are defined by Khoali [1991] as follows:

- The right branch delinking rule delinks the right branch of a multiply-linked high tone if, and only if, there is a high tone immediately after the target. This rules "feeds" directly from the high tone spread rule and thus also applies within the clitic phrase domain as well as the phonological word domain.
- The left branch delinking rule delinks the left branch of a multiply-linked high tone if it is preceded by a high tone. This rule is a phonological word domain as well as clitic phrase domain-span rule and also "feeds" directly from the high tone spread rule.
- The specifier high tone delinking rule disassociates a high tone on an object or reflexive prefix if the prefix is not within the same phonological word as the stem.
- The finality restriction principle exempts syllables at the end of a phonological phrase from the application of tonal rules.

Both algorithms were tested on 45 Sesotho sentences that were chosen in such a way that they represented all the contexts necessary for the application of the seven tonal rules. The sentences were recorded by three Sesotho respondents. One respondent's speech was transcribed by three independent labelers, with an agreement of 55.9% being achieved. The transcriptions included tone labels which we used in our analysis.

The tone labels predicted by the basic algorithm as well as the extended algorithm were compared to their transcribed counterparts independently of each other. In addition, the transcribed tone labels were compared to their following counterparts:

- Sentences with no tone information, that is, all the tone labels in the sentences are marked as low-toned.
- Sentences with underlying tone information as described in Du Plessis *et al.* [1974].

6.3 Summary of Findings

The aim of the research study was to improve an algorithm that uses Sesotho tonal rules to predict the tone labels of the syllables in an input sentence. To make such predictions, underlying tone information on each syllable was tagged as described in Du Plessis *et al.* [1974]. This algorithm also required grammatical information of each word in an input sentence to be tagged.

The research hypothesis which we tested in our study is formulated as follows:

> The extended algorithm improves the basic algorithm by increasing the number of matched tone labels.

To validate this hypothesis, we tested the algorithms in two ways:

- Restricting the application of both algorithms to the clitic phrase domain.
- Allowing the algorithms to perform to their full potential.

In both tests, we compared the tone labels output by the basic algorithm to their transcribed counterparts, and compared the tone labels output by the extended algorithm to their transcribed counterparts. We found that in both tests, the number of matched tone labels between the extended algorithm and transcriptions exceeded the number of matched tone labels between the basic algorithm and the transcriptions. The results were statistically verified using a t-test. This test allowed us to show that there is a real difference between the algorithms. The results of both t-tests, for each testing phase, validated our hypothesis.

6.4 Suggestions for Further Research

This research study provides the first and most basic step to tone modeling for languages such as Sesotho. Such languages do not mark tone information in orthography thus making it difficult to model tone into TTS systems developed for them. A tone modeling algorithm requires the tone marks on the syllables in order to assign appropriate pitch values on the syllables, thus predicting the overall intonation of the word as discussed in Odejobi *et al.* [2008] for Yoruba. Yoruba is a tonal language spoken in Nigeria which marks tone information in orthography.

The study by Odejobi *et al.* [2008] focuses on predicting the overall intonation of a word using decision trees which are based on the tone labels. This involves predicting, for instance, the pitch value to be assigned to a high tone which is preceded by another high tone. Such an algorithm can now be developed for Sesotho since we have a tone label prediction developed for this language. The next step in tone modeling for Sesotho (and other languages that are similar with respect to tone) is to develop an algorithm that can assign appropriate pitch values to the syllables allowing the prediction of the overall intonation of the word.

Currently, our algorithm is not fully integrated into a TTS system. We only use the back end of Speect. Until our algorithm is fully integrated into a TTS system, we cannot verify the extent to which implementing tone in TTS systems developed for Bantu languages can affect the perceived naturalness of the system. Once an algorithm that can assign pitch values to the tone labels predicted by the extended algorithm has been developed, future research can focus on conducting a perception test. A perception test will allow native speakers of the target language to grade the synthesized speech. This is important since the synthesized speech has to sound natural to the native speakers of the language.

6.5 Conclusion

Our main aim was to improve the basic algorithm developed by Raborife [2009], which predicts the tone labels on the syllables using three Sesotho tonal rules. We have improved this algorithm by implementing its extended version (referred to as the extended algorithm), which does the following:

- Implements four additional Sesotho tonal rules.
- Applies the tonal rules to all grammatical categories.
- Treats all verb stems that occur in a single sentence independently according to their own contexts.

Our research study verifies the suggestion by Louw *et al.* [2005] that to implement tone into TTS systems for languages such as isiZulu, as well as the Sotho-Tswana languages as shown by Raborife [2009], we need an algorithm that predicts the tone marks of the syllables as well as the appropriate intonation of the words. The extended algorithm deals with the former, the latter can only be dealt with once the relationship between the tone labels and the fundamental frequency is known for Sesotho [Van Niekerk and Barnard 2010].

Appendix A

Glosses

Abbreviation	Full Form
AGR	Agreement Marker
CA	Case Assigner
FUT	Future Tense
HAB	Habitual form
IMP	Imperative form
INF	Infinitive
INS	Instrument
NAGR	Noun Agreement Marker
NEG	Negative Form
OM	Object Marker
OC	Object Concord
PERF	Perfect tense
PL	Plural
PREP	Preposition
PROG	Progressive Marker
PRS	Present Tense
PST	Past Tense
PTS	Particle
REDUPL	Reduplication
SG	Singular
OC	Object Concord
SM	Subject Marker
SUBJ	Subjunctive form
T/A	Tense and Aspect
TMA	Tense, Mood and Aspect
V	Verb Stem
VAGR	Verb Agreement Marker
σ	Syllable

Table A.1: Glosses

Appendix B

Corpus

For each sentence, there are four rows:

- The first row is the sentence.
- The second row is the morphological analysis of the sentence.
- The third row represents how the morphological analyses are tagged in the algorithm, refined POS tagging.
- The fourth row is the English translation.

1. Noha e falla metsi
   NP9.snake SC9 emigrate NP4.water
   N AGR VPRS N
   “The snake is migrating from its waters”

2. Matsha a tsholoha, a fetoha mawatle
   NP6.lake SC6 pour SC6 turn NP6.sea
   N AGR VPRS VAGR VPRS N
   “The lakes are overpouring, they are turning into seas”

3. Modisana o kganna dinku tse sekete
   NP1.shepherd SC1 drive NP10.sheep RC7 NP7.thousand
   N AGR VPRS N NAGR N
   “The shepherd is driving a thousand sheep”

4. Malwetse a ipepesa kgafetsa
   NP6.illness SC6 reveal ADV.often
   N AGR VPRS ADV
   “Illnesses are revealing themselves often”

5. Metsi a dinoka a phorosela butle
   NP4.water PC4 NP10.river SC4 flow ADV.slow
   N AGR N/MOD AGR VPRS ADV
   “The water from the rivers is flowing slowly”

6. Barwetsana ba bala dinkgo
   NP2.girl SC2 count NP10.clay pot

N AGR VPRS N
"The girls are counting the clay pots"

7. Metsi a lelemela, a phalla
   NP4.water SC4 glide SC4 flow
   N AGR VPRS VAGR VPRS
   "The water is gliding, it is flowing"

8. Ke a utlwa ke a tshaba
   1SG.SC PRS hear 1SG.SC PRS fear
   AGR AGR VPRS SC VAGR VPRS
   "I hear, I fear"literal meaning
   "I fear the worst"figurative meaning

9. O ile a bona botshabeho a kgotsa
   SC1 PST SC1 see NP14.dreadfulness SC1 take refuge
   AGR PST AGR V N NAGR V
   "He/she (you) saw dreadfulness and sought refuge"

10. Nonyana di ile tsa kgutsa
    NP10.bird SC10 PST SC10 V
    N AGR PST AGR V
    "The birds became silent"

11. O ne a ba sheba
    SC1 PST SC1 OC2 look
    AGR PST AGR AGR V
    "He(you) looked at them"

12. Ba ne ba tumme
    SC2 PST SC2 known.SUBJ
    AGR PST AGR VSUBJ
    "They were well-known"

13. O ne a pesotse dimpa
    SC1 PST SC1 stick.PERF NP10.stomach
    AGR PST AGR VPERF N
    "He/she stuck out his/her stomach"

14. Leshodu le ile la baleha ha le bona mapolesa
    NP5.thief SC5 PST SC5 run CONJ.when SC5 see NP6.police
    N AGR PST AGR V CONJ AGR V N
    "The thief ran away when he/she saw the police"

15. Mosadi o ne a hebihebisana le monna wa hae
    NP1.woman SC1 PST SC1 argue PREP.with NP1.man PC1 PRN
    N AGR PST AGR V PTS N NAGR PRN
    "The woman was arguing with her husband"

16. Ba ile ba mo tshwara, mme ba mo neela morena
    SC2 PST SC2 OC1 capture CONJ SC2 OC1 surrender NP1.king

AGR PST AGR AGR V CONJ AGR AGR V N
"They captured him and surrendered him to the king"

17. Ke ile ka ema ka leba hae ke swabile
    1SG.SC PST SC1 stand SC1 head ADV.home 1SG.SC to be sad.PERF
    AGR PST VP AGR V ADV AGR VPERF
    "I stood up, headed home sadly"

18. Le tla mo fumana a phomotse
    2PL.SC FUT OC1 find SC1 rest.PERF
    AGR FUT AGR VPRS AGR VPERF
    "You will find him/her resting"

19. Re tla hlompha meetlo ya lona
    1PL.SC FUT respect NP4.tradition PC4 PRN
    AGR FUT V N NAGR PRN
    "We will respect your traditions"

20. Ke tla o ruta ho phela le batho
    1SG.SC FUT OC2 teach INF live PREP.with NP2.people
    AGR FUT AGR VPRS AGR VPRS PTS N
    "I will teach you how to live with people" figurative meaning
    "I will teach you a lesson" literal meaning

21. O tla le rekela dijo tse monate
    SC1 FUT OC2 buy NP8.food RC8 NP3.nice
    AGR FUT AGR VPRS N NAGR N
    "He/she will buy you good food"

22. O tla utlwa ba botsa dipotso
    SC1 FUT hear SC2 ask NP10.question
    AGR FUT VPRS AGR VPRS N
    "He/She (you) will hear them asking questions"

23. Sefate se tla thiba lehaha
    NP7.tree SC7 FUT hide NP5.cave
    N AGR FUT VPRS N
    "The tree will hide the cave"

24. Ke tla hlokomela ba tswileng dikotsi
    1SG.SC.SC FUT take.care SC2 out.REL NP10.danger
    AGR FUT VPRS AGR V N
    "I will take care of those who are in danger"

25. Qelang thuso ya moruti wa lona
    Ask.IMP NP9.help PC9 NP1.teacher PC1 PRN
    VIMP N NAGR N NAGR PRN
    "Ask your teacher for help"

26. Thea tsebe, o mamele morwetsana
    listen NP9.ear SC1 hear.SUBJ NP1.girl

VIMP N AGR VSUBJ N
"Listen attentively, girl"

27. Bua le pelo ya hao
speak PREP.with NP9.heart PC9 PRN
VIMP PTS N NAGR PRN
"Speak to you heart"

28. Mme, ntate, tsohang le re thuse
NP1a.mother NP1a.father wake_up.IMP SC.2PL OC.1PL help.SUBJ
N N VIMP AGR AGR VSUBJ
"Mother, father, wake up and help us"

29. Basotho, nkineleng matsoho metsing
NP2.sotho_people dip.IMP NP6.hand NP4.water.LOC
N VIMP N ADV
"The Sotho people, dip your hands in the water for me" literal meaning
"The Sotho people, please forgive me" figurative meaning

30. Binelang Modimo pina ya tebogo
Sing.IMP NP1.God NP9.song PC9 NP9.thank
VIMP N N NAGR N
"Sing songs of thanks to the Lord"

31. Tshwarelang ba le sitetsweng
Forgive.IMP SC2 OC.2PL wrong.REL
VIMP AGR AGR V
"Forgive those who have wronged you"

32. Jwalo ka naledi morena re tataise
like a NP9.star NP1.king OC.1PL guide.SUBJ
ADJ AGR N N AGR VSUBJ
"Like a star, King guide us"

33. Seroki se rokile morena
NP7.praise poet SC7 praise.PERF NP1.king
N AGR VPERF N
"The praise poet has praised the king"

34. O fumane monna wa sebele
SC1 find.PERF NP1.man PC1 NP7.substance
AGR VPERF N PTS N
"He/she has found a man of substance"

35. Ke o bapisitse moratuwa, ke hlotswe
1SG.SC OC1 confirm.PERF NP1.loved one 1SG.SC fail.PERF
AGR AGR VPERF N AGR VPERF
"I have confirmed my love for you, I have failed"

36. Banna ba ruletse dintlo tsa bona
NP2.man SC2 thatch.PERF NP10.house RC10 PRN

N AGR VPERF N NAGR PRN
"The men have thatched their houses"

37. Monna o kgokoloditse matlakala
NP1.man SC1 gather.PERF NP6.dirt
N AGR VPERF N
"The man has gathered the dirt"

38. Lepolesa le phenyekollotse makunutu kaofela
NP5.police SC5 uncover.PERF NP6.secret all
N AGR VPERF N ADJ
"The policeman has uncovered all the secrets"

39. Ke labalabetse tokoloho
1SG.SC yearn.PERF NP9.freedom
AGR VPERF N
"I have yearned for freedom"

40. Ke tla palama maqhubu, ke phalle
1SG.SC FUT climb NP6.wave 1SG.SC flow.SUBJ
AGR FUT V N AGR VSUBJ
"I will climb waves and flow"

41. Ba fihlile ba nnepa ka setebele seledung
SC2 arrive.PERF SC2 hit PREP.with NP7.fist NP7.chin.LOC
AGR VPERF AGR V VAGR N ADV
"They arrived and hit me with a fist on the chin"

42. Dintja di ile tsa bohola ha di utlwa mashodu a atamela
NP10.dog SC10 PST SC10 bark when SC10 hear NP6.thief SC6 approach
N AGR PST AGR V AGR AGR V N AGR V
"The dogs barked when they heard the thieves approaching"

43. Ke a tshepa hore o tla o rata
1SG.SC PRS hope that SC1 FUT SC1 like
AGR AGR V CONJ AGR FUT AGR V
"I hope she/he will like you"

44. Ba ile ba kgutla ba fumana pitsa e tletse
SC2 PST SC2 return SC2 find NP9.pot SC9 fill up.PERF
AGR PST AGR V AGR V N NAGR VPERF
"They came back and found the pots full"

45. Le tla binela morena sefela hobane o entse dintho tse makatsang
2PL.SC FUT sing NP1.king NP5.song CONJ SC1 do.PERF NP10.thing RC10 amaze.REL
AGR FUT VIMP N N CONJ AGR VPERF N VAGR V
"You will sing to the king for he has done amazing things"

Appendix C

Minimal Pairs

The following minimal pairs are from Doke and Mofokeng [1957].

Minimal pairs in isolation

(C.1) (a) ho báká
"to repent"

(b) ho baka
"to cause"

(C.2) (a) ho téná
"to put on"

(b) ho tena
"to disgust"

Minimal pair verb phrases

(C.3) (a) o já dijó
" you eat food "

(b) ó já dijó
"he eats food"

(C.4) (a) le já dijó
"you eat food"

(b) lé já dijó
"he (the thief) eats food"

(C.5) (a) bá a sébá
"they are doing mischief"

(b) bá a seba
"they are gossiping"

Short sentences differing in tone

(C.6) (a) Ke ngwaná wá háo
"I am your child"

(b) Ké ngwaná wá háo
"He/she/it is your child"

(C.7) (a) O mobé
"You are ugly"

(b) Ó mobé
"'He/she is ugly"

(C.8) (a) Ke motho
"I am a person"

(b) Ké motho
"It is a person"

Appendix D

Recording Process

D.1 Research Information Sheet

The University of the Witwatersrand
School of Computer Science
Private Bag 3
Wits 2050
South Africa

(+27)12 717 4303
(+27)78 403 2288

Date:

INFORMATION SHEET FOR THE PARTICIPANT

As part of my Masters degree in Computer Science, I am conducting a research study involving Sesotho speakers and how they speak. I am particularly interested in the pitch variations in their speech and how the variations can be implemented in text-to-speech systems developed for Sesotho. This involves improving an existing algorithm that predicts these variations and testing it on 45 Sesotho sentences.

For this reason I would like to record a native Sesotho speaker. I would like to record the 45 Sesotho sentences that will be used to test the algorithm. The recordings are going to be transcribed. I will analyse particular features that appear on the transcriptions against output of the algorithm.

Participation in the research study is **STRICTLY VOLUNTARY**. If you do choose to participate, the data recorded will be **STRICTLY CONFIDENTIAL** and I will be the only one with access to this information. The data will be kept securely and will be used for academic purposes only.

I have approached you because my research requires me to record people who are native Sesotho speakers. I would be very grateful if you would agree to take part. Your assistance and time in this research is greatly appreciated. Please feel free to ask any questions during the course of the session.

If you have any queries about the study, please feel free to contact me at the address above or by email on raborifem@cs.wits.ac.za.

Kind Regards,

Mpho Raborife

D.2 Consent Form

The University of the Witwatersrand
School of Computer Science

Consent Form

Project title: Recording of speech for MSc Dissertation

1. I have read and had explained to me by Mpho Raborife the Information Sheet relating to this study.
2. I have had explained to me the purpose of the study and what will be required of me, and any questions have been answered to my satisfaction. I agree to the arrangements described in the Information Sheet in so far as they relate to my participation.
3. I understand that my participation is entirely voluntary and that I have the right to not participate in this study.
4. I have received a copy of this Consent Form and of the accompanying Information Sheet.

Name:
Signed:
Date:

D.3 Language Profile

The University of the Witwatersrand
School of Computer Science

Language Profile Questionnaire

1. What is your first language?
   -
2. What is your second language?
   -
3. Are you bilingual?
   -
4. What other languages do you speak?
   -
5. What language do you primarily speak?
   -
6. Where were you born (Country/Province/City)?
   -

7. Where were you raised (Country/Province/City)?
    -
8. Where do you currently live (Country/Province/City)?
    -

Appendix E

Transcriptions by Three Labelers

1. A: Noha é fálla metsí.
   B: Nohá e falla metsí.
   C: Noha é fálla metsí.

2. A: Matshá á tshóloha, a fétóha mawátle.
   B: Matshá a tshóloha, a fétóha mawátle.
   C: Matshá a tsholoha, á fétoha mawátle.

3. A: Modísána ó kgańńa dinkú tsé sekéte.
   B: Modísaná o kganńa dinku tsé sekete.
   C: Modísána ó kgańńa dinku tsé sekete.

4. A: Malwétsé á ipépésa kgáfetsa.
   B: Malwétsé a ipepesa kgáfetsa.
   C: Malwétsé a ipépesa kgáfetsa.

5. A: Metsí á dínóká á phórósela bútle.
   B: Metsí a dínoka a phorosela bútle.
   C:Metsí á dínoka á phórósela butle.

6. A: Barwétsána bá bala dinkgó.
   B: Barwétsana bá bala dinkgó.
   C: Barwétsána bá bala dinkgó.

7. A: Metsí á lélemela á phállá.
   B: Metsí a lelemela á phalla.
   C: Metsí á lélemela á phálla.

8. A: Ke a útlwá ke a tshaba.
   B: Ke a útlwa ke a tshaba.
   C: Ke a útlwá ke a tshaba.

9. A: Ó ilé á boná bótshabého á kgótsa.
   B: O ilé a bona botshabého a kgotsa.
   C: Ó ilé a boná botshabeho a kgótsa.

10. A: Nonyana dí íle tsá kgútsa.
    B: Nonyana dí ile tsa kgútsa.
    C: Nonyana dí ile tsá kgútsa.

11. A: Ó ne á ba sheba.
    B: Ó ne a ba sheba.
    C: Ó ne a ba sheba.

12. A: Bá ne bá tuḿmé.
    B: Bá ne bá tumme.
    C: Bá ne bá tumme.

13. A: Ó ne á pésotsé dímpá.
    B: Ó ne a pésotsé diḿpa.
    C: Ó ne a pésotsé dimpa.

14. A: Leshodu lé íle lá balého há lé boná mápolésa.
    B: Leshodu lé íle la balého ha le boná mapolesa.
    C: Leshódu le ílé la baleha ha le bona mapolesa.

15. A: Mosádí ó ne a hébíhébisana lé móńna wá haé.
    B: Mosádí o ne a hébíhebisana lé mónna wa hae.
    C: Mosádí o ne a hébíhébisana le monna wa hae.

16. A: Bá íle bá mó tshwará mmé bá mo neélá moréna.
    B: Bá íle ba mo tshwará mme bá mo neéla moréna.
    C: Bá ílé bá mo tshwará ḿme bá mo neela moréna.

17. A: Ke íle ká emá ka lebá haé ké swabíle.
    B: Ke ilé ka émá ka lebá haé ke swabíle.
    C: Ke ilé ká ema ka lebá haé ke swabíle.

18. A: Lé tla mo fúmána á phómotse.
    B: Lé tlá mo fumána á phómotse.
    C: Le tla mó fumana á phómotse.

19. A: Re tla hlóḿphá meetlo yá lona.
    B: Re tla hlómphá méetlo yá lona.
    C: Ré tla hloḿphá meetlo yá lona.

20. A: Ke tla ho rútá hó phela le bátho.
    B: Ké tla ho rútá hó phela le bátho.
    C: Ke tla ho rutá hó phéla le batho.

21. A: Ó tlá le reélá dijó tsé monáte.
    B: Ó tlá le rékéla dijó tsé mónate.
    C: Ó tlá lé rekelá dijo tsé monate.

22. A: Ó tla útlwá bá bótsá dípotso.
    B: Ó tla útlwa bá bótsá dípotso.
    C: O tla útlwá bá bótsa dipotso.

23. A: Sefaté sé tlá thibá léhaha.
    B: Sefáté se tlá thíbá léhaha.
    C: Sefáté se tla thiba lehaha.

24. A: Ke tla hlókómela bá tswileńg díkotsi.
    B: Ké tla hlókómela bá tswiléng díkotsi.
    C: Ke tla hlókóméla bá tswileng díkotsi.

25. A: Qéláńg thúso yá morúti wá lona.
    B: Qéláńg thuso ya morúti wá lona.
    C: Qéláńg thúso ya morútí wa lóna.

26. A: Théá tsebé ó mámelé morwétsána.
    B: Theá tsebé o maméle mórwétsana.
    C: Théa tsebé ó mamele morwétsana.

27. A: Búa lé pelo yá haó.
    B: Búa lé pelo ya háo.
    C: Búa le pelo yá háo.

28. A: Mmé, ntaté, tsoháng lé re thúse.
    B: Mmé, ntaté, tsoháng le re thúse.
    C: Mmé, ntaté, tsóháńg le ré thuse.

29. A: Basóthó, nkínéleng matsóhó metsíng.
    B: Baśotho, nkíneleng matsóhó metsing.
    C: Basótho, nkínéléńg matsóhó metsing.

30. A: Bínélang Modímó pina yá tébogo.
    B: Bínelaáng Modímó pina yá tebogo.
    C: Bínéláng Modimó pína ya tébogo.

31. A: Tshwárélang bá le sitétswéng.
    B: Tshwareláng bá le sitetswéng.
    C: Tshwáreláng bá lé sitétswéńg.

32. A: Jwaló ká naledí, morena ré tátaise.
    B: Jwaló ka naledí, morena ré tátaise.
    C: Jwaló ka naledí, morena ré tatáise.

33. A: Serókí sé rokílé mórena.
    B: Serókí se rokíle mórena.
    C: Serókí se rokíle morena.

34. A: Ó fúmane mońna wá sébele.
    B: Ó fúmáné monna wá sébele.
    C: Ó fúmane monna wá sébele.

35. A: Ke o bápísitse morátúwa, ké hlotswé.
    B: Ké o bápísitse mórátuwa, ke hlotswe.
    C: Ke o bápísitse moratúwá, ke hlotswe.

36. A: Bańná bá rúlétsé díntló tsá bona.
    B: Bańná ba rúletse díntlo tsá bona.
    C: Banná bá rúletse díntlo tsa bona.

37. A: Moṅná ó kgokólóditsé mátlakala.
    B: Moṅna o kgokolodítse mátlakala.
    C: Monná o kgókolodítsé mátlakala.

38. A: Lepólésa lé phenyékóllotsé makunutú káofélá.
    B: Lepolésá lé phenyékollotsé makunútu kaoféla.
    C: Lepolésá lé phenyékóĺĺotsé makunútú kaofela.

39. A: Ke lábálabetsé tókoloho.
    B: Ke lábálabetsé tókolohó.
    C: Ke lábálábétse tókóloho.

40. A: Ke tla palámá maqhubú ké phállé.
    B: Ké tla páláma maqhúbu ke phálle.
    C: Ke tla pálámá maqhubu ke phalle.

41. A: Bá fíhlíle bá nnépa ká sétebelé séledúng.
    B: Bá fíhlíle ba nnépa ká setebelé seledúng.
    C: Ba fihlíle ba nnepá ka setebelé séledung.

42. A: Dintjá dí ílé tsá bohóla há dí boná máshodu a átámela.
    B: Dintjá di ilé tsa bohóla há di bona mashodu á átámela.
    C: Dintjá dí ile tsá bohóla ha di boná mashodu a átámela.

43. A: Ke a tshépá hore ó tlá o ráta.
    B: Ké a tshépa hore ó tlá o rata.
    C: Ke a tshépa hore ó tlá o rata.

44. A: Bá íle bá kgutla bá fúmáná pitsá é tletsé.
    B: Ba ílé ba kgútla bá fúmáná pitsá e tletse.
    C: Ba ile bá kgutla bá fúmána pitsá é tletse.

45. A: Le tla bínélá mórena sefela hobáné ó éntsé dinthó tsé makátsaṅg.
    B: Le tla bínélá mórena sefela hóbáne ó éṅtsé diṅthó tsé makatsáng.
    C: Le tla binélá morena sefela hobáne ó entse dinthó tsé makatsang.

Appendix F

Transcriptions Based on a Majority Vote

1. Noha é fálla metsí.
2. Matshá a tshóloha, a fétóha mawátle.
3. Modísána ó kgańńa dinku tsé sekete.
4. Malwétsé a ipépesa kgáfetsa.
5. Metsí á dínoka á phórósela bútle.
6. Barwétsána bá bala dinkgó .
7. Metsí á lélemela, á phálla.
8. Ke a útlwá ke a tshaba.
9. Ó ilé a boná botshabého a kgótsa.
10. Nonyana dí ile tsá kgútsa.
11. Ó ne a ba sheba.
12. Bá ne bá tumme.
13. Ó ne a pésotsé dimpa.
14. Leshodu lé íle la balého ha le boná mapolesa.
15. Mosádí o ne a hébíhébisana lé mónna wa hae.
16. Bá íle bá mo tshwará, mme bá mo neéla moréna.
17. Ke ilé ká emá ka lebá háe ke swabíle.
18. Lé tla mo fumána á phómotse.
19. Re tla hlómphá meetlo yá lona.
20. Ke tla o rútá hó phela le bátho.
21. Ó tlá le rekélá dijó tsé monate.

22. Ó tla útlwá bá bótsá dípotso.
23. Sefáté se tlá thibá léhaha.
24. Ke tla hlókómela bá tswileng díkotsi.
25. Qéláńg thúso ya morúti wá lona.
26. Théá tsebé, ó mamele morwétsana
27. Búa lé pelo yá háo.
28. Mmé, ntaté, tsoháng le re thúse.
29. Basótho, nkínéleng matsóhó metsing.
30. Bínéláng Modímó pina yá tébogo.
31. Tshwáreláng bá le sitétswéng.
32. Jwaló ka naledí morena ré tátaise.
33. Serókí se rokíle mórena.
34. Ó fúmané monna wá sébele.
35. Ke o bápísitse morátúwa, ke hlotswe.
36. Bańná bá rúletse díntlo tsá bona.
37. Mońná o kgokolodítsé mátlakala.
38. Lepolésá lé phenyékóllotsé makunútú kaoféla.
39. Ke lábálabetsé tókoloho.
40. Ke tla pálámá maqhubu, ke phálle.
41. Bá fíhlíle ba nnépa ká setebelé séledúng.
42. Dintjá dí ilé tsá bohóla há di utlwá mashodu a átámela.
43. Ke a tshépa hore ó tlá o rata.
44. Ba íle bá kgutla bá fúmáná pitsá é tletse.
45. Le tla bínélá mórena sefela hobáne ó éntsé dinthó tsé makatsang.

Appendix G

Output by Basic Algorithm

1. Noha (é (fálla)ϖ)C metsi.
2. Matsha (á (tshóloha)ϖ)C (á (fétoha)ϖ)C mawatle.
3. Modisana (ó (kgáńna)ϖ)C dinku tse sekete.
4. Malwetse (á (ípépésa)ϖ)C kgafetsa.
5. Metsi a dinoka (á (phórosela)ϖ)C butle.
6. Barwetsana (bá (bála)ϖ)C dinkgo.
7. Metsi (á (lélemela)ϖ)C (á (phálla)ϖ)C.
8. (Ke a (útlwa)ϖ)C (ke a (tshábá)ϖ)C.
9. (Ó íle á (bóná)ϖ)C botshabeho (á (kgótsá)ϖ)C.
10. Nonyana (dí íle tsá (kgútsá)ϖ)C.
11. (Ó né á bá (shéba)ϖ)C.
12. (Bá né bá (túmme)ϖ)C.
13. (Ó ne á (pesótsé)ϖ)C dimpa.
14. Leshodu (lé íle lá (báleha)ϖ)C (há lé (bóná)ϖ)C mapolesa.
15. Mosadi (ó né á (hébihebisana)ϖ)C le monna wa hae
16. (Bá íle bá mó (tshwára)ϖ)C mme (bá mó (nééla)ϖ)C morena.
17. (Ke ile ka (émá)ϖ)C (ka (lébá)ϖ)C hae (ké (swabílé)ϖ)C.
18. (Le tla mó (fúmáná)ϖ)C (á (phomótsé)ϖ)C.
19. (Re tla (hlómṕha)ϖ)C meetlo ya lona.
20. (Ke tla o (rútá)ϖ)C (ho (phela)ϖ)C le batho.
21. (Ó tlá le (rékéla)ϖ)C dijo tse monate.

22. (Ó tlá (útlwá)ϖ)C (bá (bótsá)ϖ)C dipotso.
23. Sefate (sé tlá (thíbá)ϖ)C lehaha.
24. (Ke tla (hlókómela)ϖ)C (bá (tswíleng)ϖ)C dikotsi
25. ((Qélάńg)ϖ)C thuso ya moruti wa lona.
26. ((Théá)ϖ)C tsebe, (ó (máméle)ϖ)C morwetsana
27. ((Búá)ϖ)C le pelo ya hao.
28. Mme ntate ((tsóháńg)ϖ)C (lé ré (thúsé)ϖ)C.
29. Basotho ((nkínéléńg)ϖ)C matsoho metsing.
30. ((Bínélάńg)ϖ)C Modimo pina ya tebogo.
31. ((Tshwárélάńg)ϖ)C (bá lé (sítétswéńg)ϖ)C.
32. Jwalo ka naledi morena (ré (tátáísé)ϖ)C.
33. Seroki (sé (rókíle)ϖ)C morena.
34. (Ó (fúmáné)ϖ)C monna wa sebele.
35. (Ke o (bapísítsé)ϖ)C moratuwa (ke (hlótswé)ϖ)C.
36. Banna (bá (rúlétsé)ϖ)C dintlo tsa bona.
37. Monna (ó (kgokólódítsé)ϖ)C matlakala.
38. Lepolesa (lé (phényékóĺlótsé)ϖ)C makunutu kaofela.
39. (Ke (labálábétsé)ϖ)C tokoloho.
40. (Ke tla (páláma)ϖ maqhubu (ké (phálle)ϖ)C.
41. (Bá (fíhlile)ϖ)C (bá (ńnepa)ϖ)C ka setebele seledung.
42. Dintja (dí íle tsá (bóhóla)ϖ)C (há dí (útlwa)ϖ)C mashodu (á (átamela)ϖ)C.
43. (Ke a (tshépá)ϖ)C hore (ó tla o (rátá)ϖ)C.
44. (Bá íle bá (kgútla)ϖ)C (bá (fúmáná)ϖ)C pitsa (é (tlétsé)ϖ)C.
45. (Le tla (bínélá)ϖ)C morena sefela hobane (ó (eńtsé)ϖ)C dintho tse makatsang

Appendix H

Output by Extended Algorithm

1. ((Nóhá)ϖ (é (fálla)ϖ)C (metsí)ϖ)ϕ.
2. ((Matshá)ϖ (á (tshóloha)ϖ)C)ϕ ((á (fétoha)ϖ)C (mawatle)ϖ)ϕ.
3. ((Modísána)ϖ (ó (kgáńna)ϖ)C (dinkú)ϖ)ϕ (tsé (sekete)ϖ)ϕ.
4. ((Malwétsé)ϖ (á í(pépésa)ϖ)C)ϕ ((kgafetsa)ϖ)ϕ.
5. ((Metsí)ϖ á (dinóká)ϖ (á (phórosela)ϖ)C)ϕ ((bútle)ϖ)ϕ.
6. ((Barwétsána)ϖ (bá (bála)ϖ)C (dinkgo)ϖ)ϕ.
7. ((Metsí)ϖ (á (lélemela)ϖ)C)ϕ ((á (phálla)ϖ)C)ϕ.
8. ((Ke a (útlwá)ϖ)C (ke a (tshába)ϖ)C)ϕ.
9. ((Ó íle á (boná)ϖ)C (botshábého)ϖ (á (kgótsa)ϖ)C)ϕ.
10. ((Nonyana)ϖ (dí íle tsá (kgútsa)ϖ)C)ϕ.
11. ((Ó ne á ba (shéba)ϖ)C)ϕ.
12. ((Bá ne bá (túmme)ϖ)C)ϕ.
13. ((Ó ne á (pésotse)ϖ)C (dimpa)ϖ)ϕ.
14. ((Leshodu)ϖ (lé íle lá (báleha)ϖ)C (há lé (boná)ϖ)C (mapolésa)ϖ)ϕ.
15. ((Mosádi)ϖ (ó ne á (hébihebisana)ϖ)C (lé mó)ϖ(nná)ϖ)ϕ (wá (háe)ϖ)ϕ.
16. ((Bá íle bá mo (tshwára)ϖ)C (mmé)ϖ (bá mo (nééla)ϖ)C (morena)ϖ)ϕ.
17. ((Ke ile ka (émá)ϖ)C (ka (léba)ϖ)C)ϕ ((háé)ϖ (ké (swábíle)ϖ)C)ϕ.
18. (((Lé tlá)ϖ mo (fúmána)ϖ)C (á (phómótse)ϖ)C)ϕ.
19. (((Re tla)ϖ (hlómṕha)ϖ)C (meetlo)ϖ)ϕ (yá (loná)ϖ)ϕ.
20. (((Ke tla)ϖ ó (rútá)ϖ)C (ho (phela)ϖ)C (lé bá)ϖtho)ϕ.
21. (((Ó tlá)ϖ le (rékéla)ϖ)C (dijó)ϖ)ϕ (tsé (monáte)ϖ)ϕ.

22. (((Ó tlá)ϖ (útlwá)ϖ)C (bá (bótsá)ϖ)C (dipótso)ϖ)ϕ.

23. ((Sefáté)ϖ ((sé tlá)ϖ (thíbá)ϖ)C (leháha)ϖ)ϕ.

24. (((Ke tla)ϖ (hlókómela)ϖ)C (bá (tswíleng)ϖ)C (dikótsi)ϖ)ϕ.

25. ((Qéláng)ϖ (thúso)ϖ)ϕ (yá (morúti)ϖ)ϕ (wá (loná)ϖ)ϕ.

26. ((Théá)ϖ (tsebé)ϖ (ó (mámélé)ϖ)C (morwétsána)ϖ)ϕ.

27. ((Búa)ϖ lé (peló)ϖ)ϕ (yá (háo)ϖ)ϕ.

28. ((Mmé)ϖ (ntaté)ϖ (tsóháng)ϖ (lé ré (thúse)ϖ)ϕ.

29. ((Basotho)ϖ (nkínéléng)ϖ (matsóho)ϖ)ϕ ((metsíng)ϖ)ϕ.

30. ((Bínéláng)ϖ (Modímó)ϖ (pína)ϖ yá (tebogo)ϖ)ϕ.

31. ((Tshwáréláng)ϖ (bá lé (sítétsweng)ϖ)ϕ

32. ((Jwáló)ϖ ká (nalédí)ϖ (morena)ϖ (ré (tátáíse)ϖ)ϕ.

33. ((Seróki)ϖ (sé (rókílé)ϖ)C (morena)ϖ)ϕ.

34. ((Ó (fúmáné)ϖ)C (monná)ϖ)ϕ ((wá sé)ϖ(bele)ϖ)ϕ.

35. ((Ke o (bapísítsé)ϖ)C (morátúwa)ϖ (ke (hlótswe)ϖ)C)ϕ.

36. ((Banná)ϖ (bá (rúlétsé)ϖ)C (dintlo)ϖ)ϕ (tsá (boná)ϖ)ϕ.

37. ((Monná)ϖ (ó (kgokólódítsé)ϖ)C (matlakala)ϖ)ϕ.

38. ((Lepolésá)ϖ (lé (phényékóĺlótsé)ϖ)C (makú nútú)ϖ (káófela)ϖ)ϕ.

39. ((Ke (labálábétsé)ϖ)C (tokoloho)ϖ)ϕ.

40. ((Ke tla (páláma)ϖ)C (maqhubú)ϖ (ké (phálle)ϖ)C)ϕ.

41. ((Bá (fíhlílé)ϖ)C (bá (nnépa)ϖ)C ká (setebele)ϖ)ϕ ((seledung)ϖ)ϕ.

42. ((Dintjá)ϖ (dí íle tsá (bóhóla)ϖ)C (há dí (útlwá)ϖ)C (mashodu)ϖ (á (átamela)ϖ)C)ϕ.

43. ((Ke a (tshépa)ϖ)C)ϕ ((hore)ϖ (ó tlá o (ráta)ϖ)C)ϕ.

44. ((Bá íle bá (kgútla)ϖ)C (bá (fúmána)ϖ)C (pitsá)ϖ (é (tlétse)ϖ)C)ϕ.

45. (((Le tla)ϖ (bínélá)ϖ)C (morena)ϖ (sefela)ϖ (hobáné)ϖ (ó (éńtsé)ϖ)C (dinthó)ϖ tsé (makatsang)ϖ)ϕ.

Appendix I

Table of Significance

df	0.10	0.05	0.025	0.01	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.313
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.782
8	1.397	1.860	2.306	2.896	3.355	4.499
9	1.383	1.833	2.262	2.821	3.250	4.296
10	1.372	1.812	2.228	2.764	3.169	4.143
11	1.363	1.796	2.201	2.718	3.106	4.024
12	1.356	1.782	2.179	2.681	3.055	3.929
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
70	1.294	1.667	1.994	2.381	2.648	3.211
80	1.292	1.664	1.990	2.374	2.639	3.195
90	1.291	1.662	1.987	2.368	2.632	3.183
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.090

Table I.1: t-test table

References

[Black 2006] A.W. Black. Multilingual speech synthesis. In T. Schultz and K. Kirchhoff, editors, *Multilingual Speech Processing*, pages 207–231. Elsevier, USA, 2006.

[Boersma 2001] P. Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5:341–345, 2001.

[Boslaugh and Watters 2008] S. Boslaugh and P. A. Watters. *Statistics in a nutshell*. O'Reilly Media, Inc., USA, 1st edition, 2008.

[Brennan and Prediger 1981] R. L. Brennan and D. J. Prediger. Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41(1):687–699, 1981.

[Cole and Mokaila 1962] D.T. Cole and D.M Mokaila. *A course in Tswana*. Georgetown University, Washington, USA, 1st edition, 1962.

[Creissels *et al.* 1997] D. Creissels, A.M. Chebane, and H.M. Nkhwa. *Tonal morphology of the Setswana verb*. Munich, Lincom, Europe, 1st edition, 1997.

[Demuth 1995] K. Demuth. Problems in the acquisition of tonal systems. In J. Archibald, editor, *The Acquisition of Non-linear Phonology*, pages 111–134. Lawrence Erlbaum Associates, Hillsdale, New York, 1995.

[Doke and Mofokeng 1957] C.M Doke and S.M. Mofokeng. *Textbook of Southern Sotho grammar*. Longmans Green and Co, London, 1st edition, 1957.

[Dtunji *et al.* 2007] A.D. Dtunji, , S. Wonga, and A.J. Beaumonta. A fuzzy decision tree-based duration model for standard Yoruba text-to-speech synthesis. *Computer Speech and Language*, 21:325–349, 2007.

[Du Plessis *et al.* 1974] J.A. Du Plessis, J.G. Gildenhuys, and J.J Moiloa. *Tweetalige woordeboek Afrikaans-Suid-Sotho / Bukantswe ya maleme-pedi Sesotho-Seafrikanse*. Via Afrika Beperk, Cape Town, 1st edition, 1974.

[Faasz *et al.* 2009] G. Faasz, U. Heid, E. Taljard, and D.J. Prinsloo. Part-of-speech tagging of Northern Sotho: disambiguating polysemous function words. In *EACL 2009 Workshop on Language Technologies for African Languages*, pages 38–44, Athens, Greece, 2009.

[Fleiss 1971] J.L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 5:378–382, 1971.

[Gandour and Harshman 1978] J. T. Gandour and R. A. Harshman. Crosslanguage differences in tone perception: a multidimensional scaling investigation. *Language and Speech*, 21:1–33, 1978.

[Gibbon *et al.* 2006] D. Gibbon, E. Urua, and M. Ekpenyong. Problems and solutions in African tone language text-to-speech. In *ISCA Workshop on Multilingual Speech and Language Processing*, pages 1–14, Stellenbosch, South Africa, 2006.

[Govender *et al.* 2005a] N. Govender, E. Barnard, and M. Davel. Developing intonation corpora for isiXhosa and isiZulu. In *Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa*, pages 147–151, Cape Town, South Africa, 2005.

[Govender *et al.* 2005b] N. Govender, E. Barnard, and M. Davel. Fundamental frequency and tone in isiZulu: initial experiments. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 1417–1420, Lisbon, Portugal, 2005.

[Govender *et al.* 2007] N. Govender, E. Barnard, and M. Davel. Pitch Modelling for the Nguni Languages. *South African Computer Journal*, 38:28–39, 2007.

[Govender 2006] N. Govender. Intonation modelling for the Nguni languages. Master Thesis, University of Pretoria, 2006.

[Halle *et al.* 2004] P.A. Halle, Y. Chang, and C.T. Best. Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, 32:395–421, 2004.

[Khoali 1991] B. Khoali. *A Sesotho tonal grammar*. PhD Thesis, University of Illinois, 1991.

[Kunene 1974] D.P. Kunene. *The sound system of Southern Sotho*. PhD Thesis, University of Cape Town, 1974.

[Kuun *et al.* 2005] C. Kuun, V. Zimu, E. Barnard, and M. Davel. Statistical investigations into isiZulu intonation. In *Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa*, pages 111–115, Cape Town, South Africa, 2005.

[Landis and Koch 1977] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[Lee *et al.* 2002] T. Lee, G. Kochanski, C. Shih, and Y. Li. Modeling tones in continuous Cantonese speech. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pages 2401–2404, Colorado, USA, 2002.

[Li *et al.* 2004] Y. Li, T. Lee, and Y. Qian. Analysis and modeling of F0 contours for Cantonese text-to-speech. *ACM Transactions on Asian Language Information Processing*, 3(3):169–180, 2004.

[Lombard 1976] D.P. Lombard. *Aspekte van toon in Noord-Sotho*. PhD Thesis, University of South Africa, 1976.

[Louw *et al.* 2005] J.A Louw, M. Davel, and E. Barnard. A general-purpose isiZulu speech synthesiser. *South African Journal of African Languages*, 25(2):92–100, 2005.

[Louw 2008] J. Louw. Speect: a multilingual text-to-speech system. In *Proceedings of the 9th Annual Symposium of the Pattern Recognition Association of South Africa*, pages 165–168, Cape Town, South Africa, 2008.

[Mabille *et al.* 1961] A. Mabille, H. Dieterlen, and R.A. Paroz. *Southern Sotho-English dictionary*. Morija: Morija Sesutu book depot, Lesotho, 8th. rev. and enlarged ed edition, 1961.

[Myers 1987] S. Myers. Tone and the structure of words in Shona. PhD Thesis, University of Massachusetts, 1987.

[Nespor and Vogel 1986] M. Nespor and I. Vogel. *Prosodic phonology*. Foris Publications, Dordrecht, 1st edition, 1986.

[Ni *et al.* 2008] J. Ni, S. Sakai, T. Shimizu, and S. Nakamura. Prosody modeling from tone to intonation in Chinese using a functional F0 model. In *Proceedings of the 2nd International Symposium on Universal Communication*, pages 397–404, Osaka, Japan, 2008.

[Odejobi *et al.* 2008] O.A. Odejobi, S. Wonga, and A.J. Beaumonta. A modular holistic approach to prosody modelling for standard Yoruba speech synthesis. *Computer Speech and Language*, 22:39–68, 2008.

[Odejobi *et al.* 2006] O.A. Odejobi, A.J. Beaumonta, and S. Wonga. Intonation contour realisation for Standard Yoruba text-to-speech synthesis: a fuzzy computational approach. *Computer Speech and Language*, 20:563–588, 2006.

[Raborife 2009] M. Raborife. The implementation of Sesotho tonal rules in a text-to-speech system. Honours Research Report, School of Computer Science, University of the Witwatersrand, 2009.

[Randolph 2005] J. J. Randolph. Free-marginal multirater kappa: an alternative to Fleiss' fixed-marginal multirater kappa. In *Joensuu University Learning and Instruction Symposium*, Joensuu, Finland, 2005.

[Randolph 2008] J.J. Randolph. Online kappa calculator, 2008.

[Schadeberg 1981] T.C. Schadeberg. Tone in South African Bantu Dictionaries. *Journal of African Languages and Linguistics*, 3:175–180, 1981.

[Syrdal and McGory 2000] A.K. Syrdal and J. McGory. Inter-transcriber reliability of ToBI prosodic labeling. In *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, 2000.

[Taylor *et al.* 1998] P. Taylor, A. Black, and R. Caley. The architecture of the Festival speech synthesis system. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pages 147–151, Australia, 1998.

[Van Niekerk and Barnard 2010] D.R. Van Niekerk and E. Barnard. An intonation model for TTS in Sepedi. In *INTERSPEECH*, Makuhari, Japan, 2010.

[Wang *et al.* 1996] R. Wang, Q. Liu, and D. Tang. A new Chinese text-to-speech system with high naturalness. In *Proceedings of the 4th International Conference on Spoken Language Processing*, pages 1441–1444, Philadelphia, USA, 1996.

[Zerbian and Barnard 2008] S. Zerbian and E. Barnard. Phonetics of intonation in South African Bantu languages. *Southern African Linguistics and Applied Language Studies*, 26(2):235–254, 2008.

[Zerbian and Barnard 2010] S. Zerbian and E. Barnard. Word-level prosody in Sotho-Tswana. In *Proceedings of Speech Prosody*, Chicago, USA, 2010.

[Zerbian and Khoali 2009] S. Zerbian and B. Khoali. Extensions necessary for tone modeling in Southern and Northern Sotho, 2009. Project progress report for the NHN-project Phonetics for Advanced Speech Technology, Meraka Institute and University of the Witwatersrand, Johannesburg, South Africa.

[Zerbian 2006] S. Zerbian. High tone spread in the Sotho verb. In *Proceedings of the 35th Annual Conference on African Linguistics*, pages 147–157, Somerville, 2006.

[Zerbian 2007] S. Zerbian. Phonological phrasing in Northern Sotho (Bantu). *The Linguistic Review*, 24(2-3):233–262, 2007.

[Zerbian 2009] S. Zerbian. Segmental and suprasegmental properties of monosyllables in Sotho/Tswana. In *International Conference "Monosyllables - from phonology to typology"*, pages 111–115, University of Bremen, Germany, 2009.

[Ziervogel and Mokgokong 1971] D. Ziervogel and P. C. Mokgokong. *Comprehensive Northern Sotho dictionary*. Van Schaik, South Africa, 1st edition, 1971.