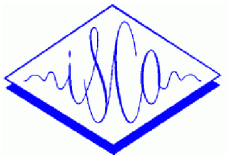
ISCA Archive
http://www.isca-speech.org/archive



Third Workshop on Spoken Language
Technologies for Under-resourced Languages
Cape Town, South Africa
May 7-9, 2012

EMPIRICAL MEASUREMENTS ON A SESOTHO TONE LABELING ALGORITHM

Mpho Raborife[1], *Sabine Zerbian*[2] *and Sigrid Ewert*[3]

Meraka Institute[1], Potsdam University[2] and University of the Witwatersrand[3]
Human Language Technology Group[1], Department of Linguistics[2] and School of Computer Science[3]
mraborife@csir.co.za[1], sabine.zerbian@uni-potsdam.de[2] and sigrid.ewert@wits.ac.za[3]

ABSTRACT

This article discusses the empirical assessments employed on two versions of a Sesotho tone labeling algorithm. This algorithm uses linguistically-defined Sesotho tonal rules to predict the tone labels on the syllables of Sesotho words. The two versions differed in the number of tonal rules that they employ as well the lexical categories that the tone rules apply to. Both versions were tested on the same input text and we employed t-tests to prove that one version is an improvement on the other.

Index Terms— Sesotho, Tone, Algorithm

1. INTRODUCTION

Sesotho is a Southern Bantu language spoken primarily in South Africa and Lesotho. It is an official language in both countries with approximately 4.2 million native speakers in South Africa and 1.7 million native speakers in Lesotho [1].

Bantu languages are tonal languages, using two-level tones, namely high (H) and low (L). They use tone to distinguish meaning. Tone in Bantu languages may be used to differentiate meaning on a lexical level as shown in Example 1. The word *bona* has different tonal patterns in this example. It is this difference in tonal patterns that affects the meaning of the word.

(1) bóna - “see”
boná - “they”

Tone 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 [2]. The following example from [2, p.39] illustrates this grammatical relationship:

(2) Ké motho - “It is a person”
Ke motho - “I am a person”

Our study was concerned with improving a Sesotho algorithm that predicts the tone labels (which are either high or low) on the syllables of the words in a speech corpus. To make such predictions, the tone labeling algorithm which is discussed in this article uses linguistically-defined Sesotho tonal rules which are described by [3]. This article discusses the empirical measurements done on two versions of the same algorithm to demonstrate the improvement of one algorithm on the other. We will refer to these algorithms as the basic algorithm (BA) and the extended algorithm (EA).

This paper is structured as follows: Section 2 discusses the motivation of the research study. Section 3 provides a detailed description of the two versions of the algorithm. Section 4 discusses the aims and objectives of our study. In Section 5 we discuss the input text collection as well as the measurements done on the algorithm. Section 6 presents the results based on the empirical measurements. Section 7 presents a summary of the research study as well as the major points presented in this paper.

2. MOTIVATION

As stated in [4, p.248], “As information technology becomes increasingly pervasive in society, there is a strong desire to support local languages through technology”. For speech technology, text-to-speech (TTS) systems are useful human language technology systems for use in illiterate communities with typical applications in health and education information dissemination [5].

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 sound as natural as human speech as possible. 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, the appropriate pitch values can be assigned to the syllables.

There are two main objectives in developing text-to-speech systems:

- Intelligibility.
- Naturalness.

Southern Bantu languages have fewer tonal minimal pairs than East-Asian languages. Furthermore, in conversation, these tonal minimal pairs are hardly ever used in isolation thus speakers use context to disambiguate the meaning [6]. Therefore, tone does not have much impact on the intelligibility of the synthesized speech. However intelligibility is not the only objective in the development of TTS systems, it is also important that the synthesized speech sounds natural to native speakers of the target language. In order to achieve this, these systems have to remain true to the prosodic system of the target language.

In order for a TTS system developed for a language such as Sesotho to sound natural, tone needs to be implemented in it since it is an important prosodic feature in this language. To do this, one needs input text that has tone labels annotated onto it. For Southern Bantu languages, it is difficult to annotate tone markings onto a speech corpus since in orthography, tone information is not included. However, surface tone can be predicted from underlying lexical tone in conjunction with morphological information and prosodic constituency. It is therefore clear that the first step to tone modeling for TTS systems developed for Southern Bantu involves being able to effectively predict which of the syllables in a word have a high or a low tone.

In the next section, we provide the descriptions of the basic and extended algorithm.

3. BACKGROUND

3.1. The Basic Algorithm

The basic algorithm is the first tone labeling algorithm which was developed for Sesotho [7]. This algorithm predicts the surface tones of the syllables of a word based on the underlying tones as specified in the lexicon, the grammatical categories of the words, and the tense, mood and aspect of the verb stems [7]. To make these predictions, the algorithm uses linguistic rules which are documented in [3]. One of the rules implemented by this algorithm is the high tone spread rule. This rule involves the spreading of an underlying high lexical tone to the adjacent syllable as exemplified in 3 ([8, p.13]). Example 3 shows a high tone on the first syllable of the high-toned verb stem *ré* 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.

(3) ke ré-ké-la…
1SG.SM buy
"I buy for…"

The basic algorithm implements two other tonal rules, namely the grammatical high tone spread rule and the iterative high tone spread rule. The application of these rules is restricted to polysyllabic verb stems and their preceding clitics. This domain of application is referred to as the clitic phrase domain. The algorithm provided a basis for the development of tone labeling algorithm for languages such as Sesotho which is a step towards tone modeling for the Southern Bantu languages. However, it has some limitations on which the extended algorithm aims to improve as discussed in Section 3.2.

3.2. The Extended Algorithm

The extended algorithm was designed to improve the basic algorithm as follows [9]:

- Implements four other tonal rules and their domains of application.

  The additional rules implemented by the extended algorithm are the right branch delinking rule, the left branch delinking rule, the finality restriction principle and the specifier high tone delinking rule. This algorithm implements a total of seven Sesotho tonal rules.

- Extends the application of the tonal rules to other parts of speech.

  The basic algorithm restricts its application to verb stems and their preceding clitics. The application of the extended algorithm includes 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 basic algorithm applies its tonal rules according to the context of the first verb stem in a sentence. The analysis by [7] showed that in such cases, each verb stem should be treated independently according to its own tense.

In the next section, we discuss the main objective in implementing the extended algorithm.

4. AIMS AND OBJECTIVES

The main objective of our study was to improve the basic algorithm by implementing four other tonal rules and extending 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 algorithm which are exactly like their transcribed counterparts. 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.

In the next section, we discuss the empirical measurements done for both versions of the algorithm.

5. EMPIRICAL MEASUREMENTS

5.1. Input Text Compilation

In this section we provide a brief overview on how the input text used to test the basic and extended algorithm was compiled. The article by [10] provides a more detailed account on this phase of our study, as well as on the recording and transcription process.

To demonstrate the improvement the extended algorithm makes on the basic algorithm, both algorithms were tested on 45 Sesotho sentences (with 565 tone bearing syllables) independently of each other. Proper nouns and/or loan words were not included in the corpus since their tone patterns are not available in the literature [11, 12]. Verb forms showing the Potential Mood (using *-ka*) have been excluded, since the tonology of these forms is subject to considerable dialectical variation [13, 14, 15]. Furthermore, there are no monosyllabic verb stems in our data, given that they have complex tonal characteristics [3, 16].

Since we did not have pre-recorded Sesotho sentences for the transcriptions, the 45 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. 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. They reached unanimous agreement on the tone labels in 55.9% of the cases.

5.2. Analysis

The transcribed tone labels of the actual recordings of the 45 sentences were used in determining the improvement of the extended algorithm on the basic one. 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.

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.

6. RESULTS

To validate the hypothesis (as stated in Section 4), we tested both algorithms as follows:

1. In the first phase we analyzed the improvement made by the extended algorithm on the basic algorithm when both algorithms are restricted to the clitic phrase domain.
2. In the second phase we analyzed the improvement made by the extended algorithm on the basic algorithm when there is no restriction in the domain of application.

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 [17]. 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:

$$t = \frac{\mu_{\mathrm{EA}} - \mu_{\mathrm{BA}}}{\sqrt{\sigma^2_{\mathrm{EA}} + \sigma^2_{\mathrm{BA}}}} \ ,$$

where, for $\mathrm{alg} \in \{\mathrm{EA}, \mathrm{BA}\}$,

- $p_{\mathrm{alg}} \in [0, 1]$ is the probability that the algorithm alg will produce a matched label,
- N_{alg} is the total number of tone labels used by algorithm alg,
- $\mu_{\mathrm{alg}} = N_{\mathrm{alg}} \times p_{\mathrm{alg}}$, and
- $\sigma^2_{\mathrm{alg}} = N_{\mathrm{alg}} \times p_{\mathrm{alg}}(1 - p_{\mathrm{alg}})$.

The t-value has a distribution with $N_{\mathrm{BA}} + N_{\mathrm{EA}} - 2$ 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. 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 (cf. [17]). The critical t-value is calculated by aligning $\alpha = 0.05$ with $df = N_{\mathrm{BA}} + N_{\mathrm{EA}} - 2 = 565 + 565 - 2 = 1128 \approx \infty$. 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.

6.1. Restriction in the Application of Both Algorithms

In this phase of our testing, both algorithms were tested on the 45 selected sentences. 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.

The output tone labels by the basic and extended algorithm were then compared to their transcribed counterparts independently of each other within the clitic phrase domain. The evaluation involving the transcriptions and the basic algorithm, yielded 194 matched tone labels. The evaluation comparing the transcriptions and the extended algorithm yielded 226 matched tone labels. The extended algorithm thus predicted 32 more matched labels than the basic algorithm.

In this phase of evaluation we have validated our hypothesis since the number of matched tone labels between the basic algorithm and the transcriptions is less than the number of matched tone labels between the extended algorithm and the transcriptions.

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

Firstly

- $p_{\mathrm{EA}} = \frac{226}{327} = 0.69$, $\mu_{\mathrm{EA}} = 327 \times 0.69 = 225.6$ and $\sigma^2_{\mathrm{EA}} = 225.6 \times 0.31 = 69.9$, and
- $p_{\mathrm{BA}} = \frac{194}{327} = 0.59$, $\mu_{\mathrm{BA}} = 327 \times 0.59 = 192.9$ and $\sigma^2_{\mathrm{BA}} = 192.9 \times 0.41 = 79.1$.

Then

$$t = \frac{\mu_{\mathrm{EA}} - \mu_{\mathrm{BA}}}{\sqrt{\sigma^2_{\mathrm{EA}} + \sigma^2_{\mathrm{BA}}}} = \frac{226 - 194}{\sqrt{69.9 + 79.1}} = 2.62$$

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.

6.2. No 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 where 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 to their transcribed counterparts. In evaluating the basic algorithm, the syllables outside the clitic phrase domain retained their underlying tones as defined in [12]. The total number of syllables used in this analysis is 565.

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[1], to their transcribed counterparts yielded 365 matched tone labels. The extended algorithm thus 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. Firstly

- $p_{\mathrm{EA}} = \frac{409}{565} = 0.72$, $\mu_{\mathrm{EA}} = 565 \times 0.72 = 406.8$ and $\sigma^2_{\mathrm{EA}} = 406.8 \times 0.28 = 113.9$, and
- $p_{\mathrm{BA}} = \frac{365}{565} = 0.65$, $\mu_{\mathrm{BA}} = 565 \times 0.65 = 367.3$ and $\sigma^2_{\mathrm{BA}} = 367.3 \times 0.35 = 128.6$.

Then

$$t = \frac{\mu_{\mathrm{EA}} - \mu_{\mathrm{BA}}}{\sqrt{\sigma^2_{\mathrm{EA}} + \sigma^2_{\mathrm{BA}}}} = \frac{409 - 365}{\sqrt{113.9 + 128.6}} = 2.82$$

As calculated above, $t = 2.82 > 1.645$, thus statistically verifying our results.

6.3. Discussion

In this section we discuss the mismatches between transcriptions and the extended algorithm in the final phase of evaluation.

The Sesotho tonal system is described in detail by [3]. In this dissertation, ten Sesotho tonal rules are discussed. The extended algorithm implements seven of these rules. In our analysis we found that underlying high tones at the

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

end of a certain domain were realized as low tones in the transcriptions. [3] describes a rule that can account for this phenomenon. This is the mid-tone insertion rule.

A large number of the mismatches cannot be accounted for by any of the tonal rules discussed in [3]. The transcribed tone labels show some tonal behaviour which can be explained by tonal rules that we believe are not yet investigated or/and documented. These mismatches are open for further research based on the linguistic characteristics of the behaviour of tone in Sesotho.

Although we have shown that our transcriptions are reliable (cf. [10]) 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. In this instance, there are discrepancies between the lexical tones described in [12] and their transcribed counterparts. Since the extended algorithm uses lexical tone to predict the surface tone labels, if there already discrepancies in the lexical tones then the surface tones will be affected. Thus leading to mismatches in our evaluations.

It is also possible that the literature is not accurate, or that our speaker has different tone assignments to the relatively old literature.

6.4. Future Work

The evaluation of the basic and extended algorithms relied heavily on the transcriptions. In our study, we used only one speaker for the transcription process. It would have been ideal to transcribe a larger set of speech. However, manual tone labeling is a time-consuming task because tone transcription is not a straightforward task and one usually needs some experience with the transcription system.

It would have been preferable to have more expert transcribers, however, this was not easily achievable since there are not many experts in this area of study. Furthermore, being a native Sesotho speaker does not qualify one as an expert transcribers since native speakers of tonal languages are usually not aware of their language's tonal system as this is not taught at schools. Thus, in future, more research should be dedicated to adopting acoustic measures such as the fundamental frequency to perform such a classification [18].

This research study provides the first and most basic step to tone modeling for languages such as Sesotho. The next step is to predict the overall intonation of a word. 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.

7. CONCLUSION

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. In this article, we provide empirical means to evaluate tone labeling algorithms. This empirical measures can be used to verify the importance of tone labeling algorithms developed for other tonal languages.

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.

8. REFERENCES

[1] M.P. Lewis, *Ethnologue: Languages of the World*, SIL International, Dallas, Texas, 16th edition, 2009.

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

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

[4] S. Zerbian and E. Barnard, "Phonetics of intonation in South African Bantu languages," *Southern African Linguistics and Applied Language Studies*, vol. 26, no. 2, pp. 235–254, 2008.

[5] 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*, Stellenbosch, South Africa, 2006, pp. 1–14.

[6] J.C. Roux, "On the perception and production of tone in Xhosa," *South African Journal of African Languages*, vol. 15, pp. 196–204, 1995.

[7] 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.

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

[9] M. Raborife, "Tone labelling algorithm for Sesotho," Masters dissertation, School of Computer Science, University of the Witwatersrand, 2011.

[10] M. Raborife, S. Zerbian, and S. Ewert, "Developing a corpus to verify the performance of a tone labelling algorithm," in *Proceedings of the 22nd Annual Symposium of the Pattern Recognition Association of South Africa*, Vanderbijlpark, South Africa, 2011, pp. 126–131.

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

[12] 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.

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

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

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

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

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

[18] L. Mohasi, H. Mixdorff, and T. Niesler, "An acoustic analysis of tone in Sesotho," in *17th International Congress of Phonetic Sciences*, Hong Kong, China, 2011, pp. 1402–1405.