Distinctive transitional probabilities across words and morphemes in Sesotho*

Andrei Anghelescu
University of British Columbia

Abstract: This project examines a corpus of child-directed Sesotho speech in order to gauge how adequate transitional probability (TP) between units of sound is for predicting word boundaries and word-internal morpheme boundaries. The predictor is the likelihood of some phonological unit, a segment or syllable, given either the preceding or following unit, *i.e.* the (forward or backward) transitional probability of those two units. I conclude that transitional probability between adjacent segments and between adjacent syllables is not useful as a predictor of boundaries in Sesotho.

Keywords: phonological learning, Sesotho, transitional probability

1 Introduction

Several models of word segmentation in first language acquisition are based on using the probability of phonological sequences as a predictor. For example, Saffran et al. (1996a) argue that children can segment words out of the speech stream based on the likelihood of a syllable or segment sequence. If transitional probability is a reliable way for children who are acquiring their first language to segment the speech stream, then there should be an observable trend in real language data such that better than chance predictions can be made about boundary-hood based on TP; this has been demonstrated using computational modeling for some languages by Daland and Pierrehumbert (2011).

The situation modeled by this examination is that of an infant attempting to posit boundaries between units of speech sounds. An infant learner has access to a wide variety of stimuli that are relevant in this task; however, in this model we abstract away from meaning and just examine what contribution can be made by simply paying attention to the distribution of sounds (specifically, syllables and segments) in the speech stream.

Sesotho is a Bantu language spoken by about 6 million people in Lesotho and South Africa (Lewis et al. 2013). It was chosen as the language of investigation because of the simple syllable inventory, high ratio of morphemes per word and availability of relevant data. The large number of polymorphemic words facilitates investigation of not only word boundaries, but also word-internal morpheme boundaries.[1] The particular corpus used in this paper, Katherine Demuth's Sesotho CHILDES corpus (Demuth 1992), is ideal for morphological and phonological analysis as it contains both morpheme glosses and a phonemictranscription.

This project contributes to two areas that have not been developed in phonological learning. Because Sesotho has many poly-morphemic words, this project examines not only the traditional word boundary identification, but also morpheme boundary identification. Since some experimental

*Thank you to Kathleen Currie-Hall, Carla Hudson Kam, Douglas Pulleyblank, and the Language and Learning Lab group.

Contact info: aanghelescu@gmail.com

[1] In this paper, I will refer to word-internal morpheme boundaries simply as morpheme boundaries. Though word boundaries are implicitly also morpheme boundaries, I use the term morpheme boundary exclusively to refer to the edges of word-internal morphemes.

In UBC Qualifying Papers 3 (2013–2015),
University of British Columbia Working Papers in Linguistics,
Andrei Anghelescu, Joel Dunham, and Natalie Weber (eds.), 2016.

work has demonstrated that both forward and backward transitional probability are available to learners in parsing (Pelucchi et al. 2009; Saffran et al. 1996b), this project examines both measures as predictors of boundary-hood.

The paper is organized as follows. In the next section I present background information on transitional probabilities (Section 2.1), how they have been used in artificial language learning tasks (Section 2.2), and an overview of Sesotho (Section 2.3). Section 2 closes with a summary of the predictions made about the Sesotho corpus (Section 2.4). In Section 3, I describe the corpus used in this study (Section 3.1), and report the findings of the study (Section 3.2). In Section 4, I discuss the results, and in Section 5 I offer concluding remarks.

2 Background

Children who are learning their first language must learn which sound sequences co-occur with which meanings and map those two elements together as a 'word'. Correctly identifying where a unit of meaning begins and ends in the speech stream is part of this challenge. Children must posit boundaries in long strings of inflectional morphemes which are part of the same word; this raises the question of how informative TPs are in identifying boundaries found between words (referred to as word boundaries in this paper) and the boundaries found between the subparts of a word (referred to as morpheme boundaries in this paper). This paper is concerned with the question of whether children could use information about the sound patterns in natural language to predict word boundaries and word-internal morpheme boundaries; this approach does not take into account any of the meaning associated with these units.

The rest of this section is organized as follows: Section 2.1 provides an explanation of how TPs could be used in language learning and how they are computed. Section 2.2 provides an example of a laboratory experiment which demonstrates that infants are sensitive to differences in TP. Section 2.3 provides information about the corpus, what is included in it, the segments and syllables observed in it, and the distribution of TPs in it.

2.1 Statistical Learning

One possible approach to boundary identification comes from using transitional probability between units (such as segments or syllables) as a way to determine where boundaries should be placed (Harris 1955, 1967; Saffran et al. 1996b). It has been claimed that identifying a TP as low relative to some other TP or set of TPs is a viable way to guess between which syllables a boundary should be placed. Specifically, segments with low transitional probability have a high likelihood to occur across a word boundary (Harris 1955). This idea follows from the observation that sequences of segments in a language like English are not equally distributed within words compared to across words. Specifically, there is a much smaller possible set of sequences within a word than across two words. This means that sequences across words are harder to predict; these sequences have lower probability. This approach to statistical learning has yielded positive results for some languages (Daland 2009; Daland and Zuraw 2013); however, it is not clear that TPs will relate to boundaries in the same way for all languages.

The transitional probability (TP) between two units X and Y is the probability of observing one given the other. Mathematically, it is equivalent to conditional probability: $p(Y|X)$, which can be read as 'the probability of Y given X'. In other words, $p(Y|X)$ is the forward transitional probability of XY. The backward transition probability of a sequence XY is equivalent to the conditional probability $p(X|Y)$, the probability of X given Y.

(1) a. Forward transitional probability
$$p(Y|X) = \frac{p(XY)}{p(X)}$$
b. Backward transitional probability
$$p(X|Y) = \frac{p(XY)}{p(Y)}$$

In this study, the sequence XY stands for a sequence of adjacent segments, such as [t͡ɬ, a], or a sequence of adjacent syllables, such as [t͡ɬa, t^{h}o]. As illustration, consider the computation of TP for two segments. To compute the transitional probability of two segments, [t͡ɬ, a], we first count the number of times the sequence 't͡ɬa' occurs in the corpus in order to find the probability of the sequence:

(2) $$p(\text{t͡ɬa}) = \frac{\text{\# of t͡ɬa bigrams}}{\text{total \# of bigrams}}$$

We then count how often 't͡ɬ' occurs as the first member of a bigram and how often 'a' occurs as the second member of a bigram and compute their probabilities so that we can later find both the forward and backward TP of the sequence [t͡ɬ, a]:

(3) a. $$p(\text{t͡ɬ\_}) = \frac{\text{\# of t͡ɬ\_ bigrams}}{\text{total \# of bigrams}}$$

b. $$p(\text{\_a}) = \frac{\text{\# of \_a bigrams}}{\text{total \# of bigrams}}$$

Finally, we can plug in the values for $p(\text{t͡ɬa})$, $p(\text{t͡ɬ\_})$, and $p(\text{\_a})$ into the equations in (1):

(4) a. Forward transitional probability of [t͡ɬ, a]
$$p(\text{a}|\text{t͡ɬ}) = \frac{p(\text{t͡ɬa})}{p(\text{t͡ɬ\_})}$$
b. Backward transitional probability of [t͡ɬ, a]
$$p(\text{t͡ɬ}|\text{a}) = \frac{p(\text{t͡ɬa})}{p(\text{\_a})}$$

2.2 Artificial Language Learning

Following the logic laid out above, Saffran et al. (1996a) asked if infants could use distinctions in TP to segment words out of a sequence of syllables. While most computational approaches to morphological segmentation use phonemes or phones as the basic unit of processing, there is evidence that suggests infants first perceive syllables (Bertoncini 1981; Bijeljac-Babic et al. 1993). Therefore, there is some merit to investigating this claim using syllables as basic units of discernibility.

Saffran et al. (1996a) constructed a lexicon of four three-syllable words and then concatenated these words into a speech stream composed such that the forward transitional probabilities between any two syllables within a word are much higher than between any two words (Forward TP between syllables within a word = 1.0; Forward TP between syllables across a word boundary = 0.33). Consider the example below, constructed with the same principles.

(5) Sample Words
to.ki.bu gi.ko.ba
go.pi.la ti.po.lu

(6) Sample speech stream
toki**bugiko**bagopilatipolutokibugopilatipolutoki**bugiko**bagopilagikobatokibugopilatipo
lugikobatipolugikobatipolugopilatipolutokibugopilatipolutokibugopilatipolutokibugopi
lagikobatipolutokibugopilagikobatipolugikobatipolugikobatipolutoki**bugiko**bagopilatipo
lugikobatokibugopilatoki**bugiko**bagopilatokibu

In the sample above, the sequence **bugiko**, composed of syllables from two words in the lexicon (to.ki.**bu#gi.ko**.ba), is referred to as a part-word.

Saffran et al. (1996a) found that infants listened longer to part-words (like **bugiko**) than to the statistically defined words after being exposed to a speech stream. If the part-words count as different from the words, then the infant must be sensitive to the TPs.

Both forward and backward TPs have been shown to be relevant. Pelucchi et al. (2009) demonstrate that children are able to discern between words and part-words only based on backward transitional probability. As Pelucchi et al. (2009) discuss, backward transitional probability is more informative than forward transitional probability in several situations. For instance, when considering TPs between morphemes in languages where determiners precede nouns and there is gender agreement between determiners and nouns (*e.g.* Spanish), the noun is an excellent predictor of the determiner since it predicts the gender.

With respect to Sesotho, backwards transitional probability could be more informative due to the templatic morphology of the language. Both nouns and verbs are constructed with monosyllabic prefixes which occur before roots. Because these prefixes are drawn from a small class of concoordial morphemes, they are more predictable from the perspective of the root, which is drawn from a large open class, than the root is from the perspective of the prefix. This means that backward TP will be higher for these sequences than forward TP would be.

Since higher TP relative to other bigrams is associated with a bigram existing within a single domain, such as within a morpheme, the sequences that contain a root and prefix will be less distinct from sequences that occur within a single morpheme in terms of backward TP. Therefore, the forward TP of such sequences will be less similar to the TP of sequences within a word (but more similar to across-word sequences). Under the assumption that there is a larger gap between the TPs of across-word sequences and across-morpheme sequences than the TPs of across-morpheme and within-morpheme sequences, then the forward TP will be the most distinct predictor of within word morpheme boundaries.

Because many prefixes are monosyllabic, it should be the case that syllabic bigrams are useful in predicting morpheme boundaries. However, there are reasons to believe that segmental bigrams will fare better; these are discussed in Section 2.3, below.

2.3 Sesotho

Sesotho is spoken in South Africa and Lesotho by about six million people. It has many features found in other Bantu languages, such as extensive concordial morphology which appears on both verbs and nouns.

All of the data used in this investigation is drawn from Demuth (1992)'s corpus. The corpus was accessed through the CHILDES database. This corpus contains ninety-eight transcribed hours of interaction focused on four children acquiring Sesotho in Lesotho. In addition to the target learners, the database has transcribed and glossed utterances from adults. The utterances of adults are selected for use in this study as child directed speech. Adults include parents, grandparents and teen-aged siblings. Concretely, a filter was used to select only utterances produced by speakers with the following tags: 'Adult', 'Grandmother', 'Mother', 'Father', 'Uncle', 'Teenager'. Playmates and target children are excluded on the assumption that their productions are not fully adult-like. It is assumed that this is representative of the input that a first language learner of Sesotho might have. The resulting corpus had 16,941 utterances containing a total of 199,551 segments of 41 types; 188,005 segmental bigrams; 130,050 syllables of 188 types; and 94,994 syllabic bigrams.

The corpus is transcribed into Sesotho orthography, which is largely phonemic. The Sesotho orthography does not encode the full range of vowel distinctions. This is unfortunate because vowel harmony is entirely obscured in the orthographic rendering of Sesotho as [±ATR] is collapsed into a single symbol (*i.e.*, orthographic <e> maps to [e] and [ɛ], <o> maps to [o] and [o], <i> maps to [i] and [ɪ], and <u> maps to [u] and [ʊ]); see Section 4 for a discussion of how vowel harmony and morphological structure could be relevant to the predictive power of TPs. Similarly, the transcription does not encode tone. Finally, vowel length is not represented in the orthography, but is largely predictable in Sesotho. Specifically, penultimate vowels generally become lengthened. While this could be coded back into the data, the regularity of penultimate lengthening is not known and therefore left unmarked.[2]

Table 1 lays out the phone inventory of Sesotho (following Doke and Mofokeng (1957) and Demuth (1983) for phonemes) which includes all surface segments found in the corpus.

Sesotho has five possible syllable shapes: CV, V, N, C^wV, and L (Doke and Mofokeng 1957). Most syllables have a consonant as the onset and a vowel as the nucleus: CV. Syllables may or may not have an onset. Onset consonants may be labialized. No syllable has a coda. Nasals and liquids can act as nuclei, but syllables with nasal or liquid nuclei cannot have onsets. The inventory of syllable shapes is shown below along with examples illustrating each syllable type in an utterance. The counts and percentages are reported for the utterances included in this analysis; the corpus analysed here contained 130,050 syllables.

(7) Sesotho syllable shapes (with frequency and count)

Type	% of total syllables	Count
CV	65.3	84,860
V	20.1	27,283
N (syllabic nasal)	10.9	14,208
C^wV	2.3	3,052
L (syllabic liquid)	0.5	647

[2] See Zerbian (2007) for a study on the contexts in which penultimate lengthening applies.

a.

e.re	m.p^{h}e	n.t^{h}o	e.na
V.CV	N.CV	N.CV	V.CV

‘Say: give me this thing’

b.

t͡ɬi.sa	k^{w}a.no
CV.CV	C^{w}V.CV

‘Bring it here’

c.

o.se.a.i.lo.!e.te.l.la	a.re.fe.e.la	l.lo
V.CV.V.V.CV.CV.CV.L.CV	V.CV.CV.V.CV	L.CV

’He is going to end up just saying *llo*’

Table 1: Sesotho phone inventory

(a) Consonant inventory

	Labial	Alveolar	Lateral	Postalveolar	Velar	Glottal
Click				! ᵑ! !^{h}		
Stop	p b p^{h}	t d t^{h}			k g k^{h}	
Nasal	m	n	ɲ	ŋ		
Fricatives	f v	s z	ɬ	ʃ ʒ	x	h
Affricates		t͡s t͡s^{h}	t͡ɬ	t͡ʃ t͡ʃ^{h}	k͡x	
Approximant	w		l	j		
Trill					ʀ	

(b) Vowel inventory

	Front	Mid	Back
High	i		u
Mid	e		o
Low		a	

Consider the example below; because the vowel, [ó], occurs word initially, we know that it is its own syllable. Since Sesotho does not allow codas, the segment [t͡ɬ] is parsed as the onset of the syllable [t͡ɬá]. The next syllable, [mo] readily fits the template of CV syllables, as do [ré], [ké] and [la].[3]

(8) Sesotho Verb Demuth (2007)

ó-	t͡ɬá-	mo-	rék	-él	-a
1SM-	FUT-	1OM-	buy	-BEN	-IN

‘He will buy food for someone.’

[3] The morpheme glosses in the example are as follows: 1SM: Noun class 1 subject marker, FUT: future tense marker, 1OM: noun class 1 object marker,BEN: benefactive marker, IN: indicative mood marker.

The example above is a single word of Sesotho. It contains six morphemes: the first person subject marker (SM1) [ó-], the future tense marker (FUT) [t͡ɬá-], the first person object marker (OM1) [mo-], the verb root meaning to buy [rék], the benefactive mood marker (BEN) [-él] and the final vowel [-a]. Of particular note is the mismatch between morpheme boundaries and syllable boundaries. Consider the two representations of (8) below.

(9) Morpheme break: [o-t͡ɬa-mo-rek-el-a]
Syllable break: [o.t͡ɬa.mo.re.ke.la]

The first three morpheme boundaries, [o-t͡ɬa-mo-] coincide with syllable boundaries. However, the last two morpheme boundaries, [rek-el-a] do not. Morpheme boundaries like the last three are undetectable to a learner that considers syllables as the basic unit of analysis.

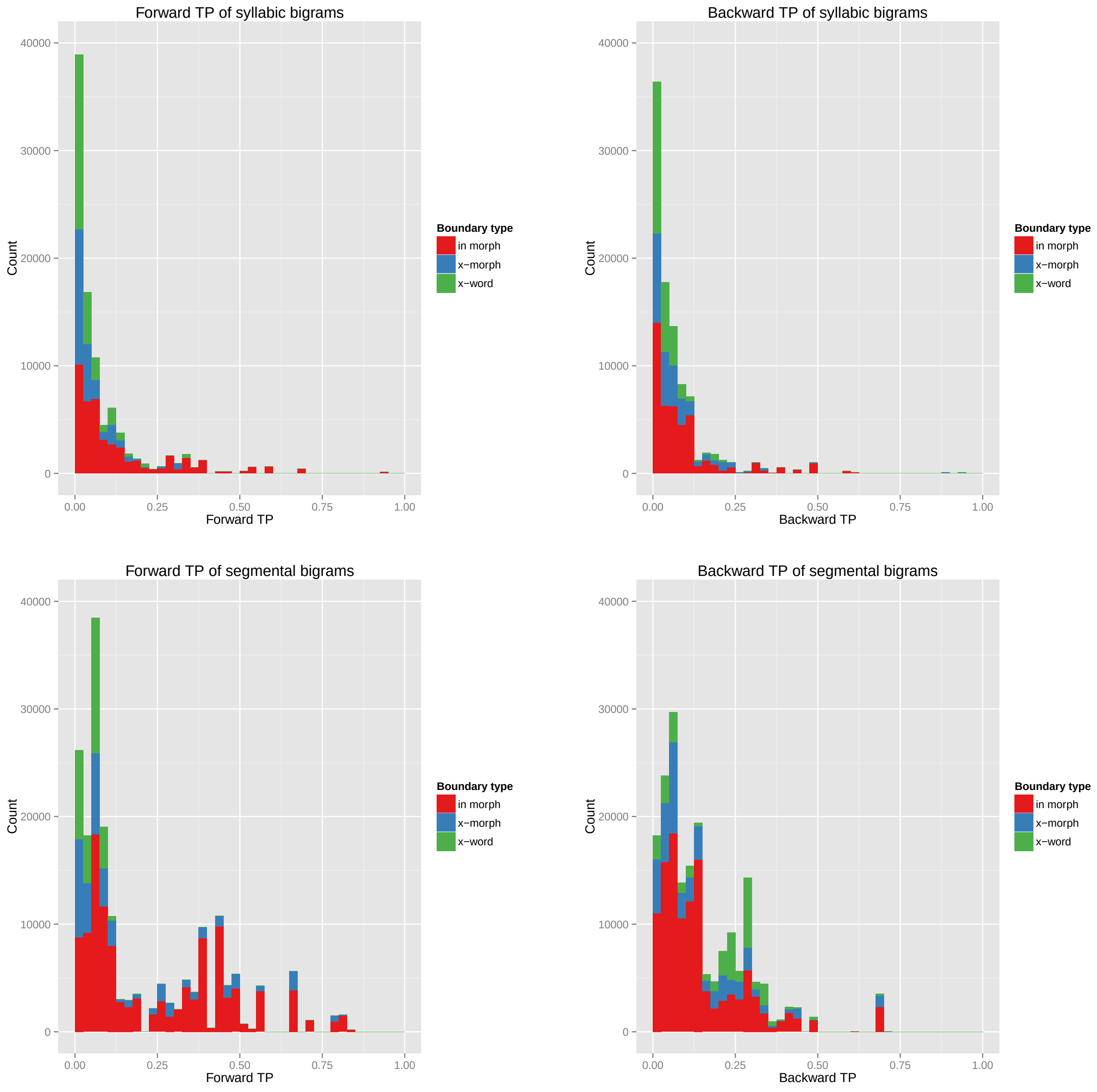
The example above is drawn from a class of cases which cannot be correctly predicted by using syllables as the basic unit of analysis. There are many CVC verb roots, such as 'rek' which combine with morphemes that are simply a vowel to form two syllables. However, since the morpheme boundary always exists between an onset and its coda, it cannot be found by examining the TP between two syllables.

Out of 158,380 total morpheme boundaries, 67,821 coincide with syllable boundaries. This means that 67,821 morpheme boundaries can be detected using a syllabic parse as counted in the corpus used for this study. The morpheme boundaries which cannot be detected include the boundaries following CVC- verb roots as well as any other morphemes which do not align with syllable boundaries.

In contrast, it is assumed that all word boundaries are syllable boundaries.[4] Therefore, the syllabic bigram model has the potential to correctly predict the existence of every word boundary. The segmental bigram model can do better at predicting morpheme boundaries because is has access to every possible morpheme boundary.

The distribution of transitional probabilities by the boundary over which they occur is shown in the four panels of the figure below. These graphs display the counts for each bigram token sorted by its transitional probability and coded by the boundary over which it occurs; they are displayed as stacked histograms where the height of a column indicates the total number of bigrams which have that TP. In these figures, 'x-morph' stands for a token of a bigram which occurs over a morpheme boundary, 'x-word' stands for a token of a bigram which occurs over a word boundary, and 'in morph' stands for a token of a bigram which does not occur over any boundary.

[4] Since Sesotho lacks syllables with codas, it is not possible for a coda to be syllabified as an onset. Liquids and nasals which act as nuclei could become the onset of otherwise onsetless syllables across a word boundary; however, this process is not attested in the literature regarding Sesotho phonology.


Forward TP of syllabic bigrams

Backward TP of syllabic bigrams

Forward TP of segmental bigrams

Backward TP of segmental bigrams

Figure 1: Histograms of forward and backward TP for syllabic bigrams and segmental bigrams

The following charts give the average TP for each category for the unit type; recall that the range for TP is between 0 and 1.

(10) a. Average TP by boundary type for syllabic bigrams

	Forward TP	Backward TP
In morph	0.133	0.090
X-morph	0.045	0.070
X-word	0.040	0.039

b. Average TP by boundary type for segmental bigrams

	Forward TP	Backward TP
In morph	0.244	0.142
X-morph	0.172	0.145
X-word	0.050	0.200

In general, most bigrams have low transitional probability. In the two charts above, none of the average TPs are above .244, which is in the lower quarter of the possible range. Bigrams that contain two units within a morpheme have the widest distribution of TPs. Considering syllabic bigrams, there is not a clear distinction between the TPs of bigrams that contain morpheme boundaries and the TPs of bigrams that contain word boundaries, the average TP chart clearly supports this for forward TP. Considering segmental bigrams, bigrams that occur across a morpheme have slightly higher maximum TPs than bigrams that occur across word boundaries; however, the distributions largely overlap.

These observations run counter to our expectations. Word-internal sequences were predicted to have higher TPs than both cross-morpheme and cross-word bigrams. Cross-morpheme bigrams were predicted to have higher TPs than cross-word bigrams. The second expectation is arguably met by the segmental bigrams. Note that our prediction about the backward TP of cross-morpheme bigrams being more similar to the TP of within morpheme bigrams compared to forward TP is borne out by the average TPs.

2.4 Prediction Summary

In the preceding section we have seen that there are several facts about Sesotho which lead us to predict that different directionality and units will be more or less useful in deriving boundaries from transitional probability. This section summarizes these predictions.

Because syllable boundaries do not always align with morpheme boundaries, we predict that segmental bigrams will be more useful in predicting within-word morpheme boundaries. Because of the prefixing inflectional morphology of Sesotho, we predict that backwards transitional probability will be more useful in predicting within-word morpheme boundaries.

The null hypothesis is that transitional probability has no predictive value for the presence of either word or morpheme boundaries. In other words, TPs are as informative as randomly deciding if a bigram contains a boundary or not. Therefore, under the null hypothesis, we do not expect differences to arise from unit of analysis either. Furthermore, there should not be any difference between the predictive value of forward and backward TP.

3 Method and Results

This study considers the predictions a learner could make about the existence of word and morpheme boundaries from observing the transitional probability between units of sound, specifically the transitional probabilities between adjacent segments and the transitional probabilities between syllables. By focusing on the data we are able to determine how accurate and precise an idealized learner could be in their predictions about boundaries. In the following two subsections I lay out how the evaluation models were constructed and then analyze the results of the models.

3.1 Data

The models described in this section fall into two broad classes that define the unit of analysis: syllabic bigrams and segmental bigrams. Though both models share the same basic components, the syllabic models also rely on syllabification of the Sesotho corpus.

The Sesotho corpus includes morphological segmentation, but does not include syllabification. Each utterance is represented by a string of letters each of which uniquely represents a single phone of Sesotho or one of the two boundaries under consideration. For instance, the segment [p^h] is encoded as ‘P’, while the segment [p] is encoded as ‘p’. Recall from (7) that Sesotho has five syllable shapes: CV, V, N, C^wV, and L. The corpus was automatically syllabified as follows: beginning from left and moving rightward through each utterance, the first segment was checked. If the first segment was a single vowel, it was syllabified as a V syllable. If the first segment was a nasal or liquid, the following segment was checked. If the following segment was another consonant, the nasal or liquid was syllabified as a syllabic segment; if the following segment was a vowel the sequence was syllabified as NV or LV. It is not expected that a nasal or liquid can act as a labialized onset. The syllabification algorithm includes a check in case such sequences are found; however, there were none. If the first segment was a consonant other than a nasal or liquid, the second segment was checked. If the second segment was a ‘w’, the next segment was checked. If that segment was a vowel, all three segments were syllabified as a C^wV syllable. Finally, if the first segment was a consonant and was not followed by a ‘w’ but instead followed by a vowel, the sequence was syllabified as a CV syllable. The syllabification algorithm then went on to the next unsyllabified segment, repeating the process described above until no more segments remained in the utterance.

For the segmental bigram models, the database was scanned with a two segment window in order to extract the bigrams.[5] Every bigram was recorded as well as the bigram context it occurred in: within a morpheme (not over any boundary), over a morpheme boundary, or over a word boundary. The same process was applied to the syllabified corpus, however the bigrams were composed of two syllables. The resulting dataset is the testing data for the experiment; it contained each bigram token along with the forward, and backward TP as well as coding for the presence or absence of word and morpheme boundaries.

The examples below illustrate a sample utterance from the corpus along with the computed transitional probabilities. Under the utterance in (11a), are the syllabic bigrams it contains; under the utterance in (11b) the segmental bigrams it contains. Under the bigrams are their forward and backward transitional probabilities. The presence of a morpheme boundary is indicated by a dash, ‘-’; the presence of a word boundary is indicated by a hash mark, ‘#’; bigrams are given in square brackets with a comma separating the two units of the bigram. Boundaries that occur within a syllable are not shown.

[5] Labialized onsets were treated as two segments, a consonant followed by a glide.

(11) a. u-r-e#u-lo-rek-a-ŋ

Syllabic bigrams:	[u, -re]	[re, #u]	[u, -lo]	[lo, -re]	[re, ka]	[ka, -ŋ]
Forward TP:	0.036	0.082	0.001	0.006	0.015	0.013
Backward TP:	0.121	0.024	0.005	0.006	0.009	0.005

b. u-r-e#u-lo-rek-a-ŋ

Segmental bigrams:	[u, -r]	[r,-e]	[e, #u]	[u, -l]	[l, o]	[o, -r]	[r, e]
Forward TP:	0.049	0.673	0.046	0.060	0.15	0.035	0.673
Backward TP:	0.138	0.060	0.188	0.051	0.09	0.197	0.060
Segmental bigrams:	. . .	[e, k]	[k, -a]	[a, -ŋ]			
Forward TP:		0.070	0.383	0.101			
Backward TP:		0.233	0.116	0.682			

The segmental bigram [r, e] is repeated twice in the utterance. Because TP is computed over the whole corpus, the measure for any instance of a bigram will have the same TP as another instance even though some instances occur over a boundary and some do not.

In order to model the relationship between transitional probability and segmentation, a set of categorizers was created. The categorizers are simple predictive algorithms that return categorical predictions about some variable based on an input variable. In this case, the algorithm attempts to make predictions about a boundary (presence of a morpheme boundary, presence of a word boundary) using transitional probability (forward TP, backward TP).

Creating the categorizer was done by using logistic regression in the R Software package and glm() function (R Core Team 2014). This function fits generalized linear models. The linear model maps a TP to the probability of a boundary. The models were given a list of every bigram, that bigram's (forward or backward) TP, and if it occurred over a (word or morpheme) boundary. The model created a function which maps TP to the probability of a boundary. None of the models had statistically significant fits, indicating that the relationship between the predictor and variable was not meaningful.

After creating linear models for all eight relationships, the bigram data was fed back into the model in order to obtain the probability that each bigram would occur over a boundary. These probabilities were converted into categorical responses; if the probability for a bigram having a boundary was higher than or equal to 0.50, the model was considered to have predicted a boundary. If the probability for a bigram having a boundary was lower than 0.50, the model was considered to have predicted no boundary. These categorical results were then compared to the actual boundaries for each bigram.

In summary, eight models were created: four models which use syllable bigrams to predict boundaries and four models which use segmental bigrams. Of the four models for each unit, two models use forward transitional probability and two use backward transitional probability; of the models which use forward transitional probability, one predicts word boundaries and one predicts morpheme boundaries, likewise for the two models which use backward transitional probability. None of these models had statistically significant relationships between TP and the existence of a boundary. These eight models are summarized in the chart below. The model names can be read as '**Y** *predicted from* **X**'; for example, '**Morph** $\sim$ **Fwd TP**' is a model that 'predicts **morpheme boundaries** from **forward transitional probability**'. The next section summarizes the performance of each model.

(12) Model summary

	Syllabic bigrams	Segmental bigrams
Forward	Morph $\sim$ Syll Fwd TP	Morph $\sim$ Seg Fwd TP
	Word $\sim$ Syll Fwd TP	Word $\sim$ Seg Fwd TP
Backward	Morph $\sim$ Syll Bkwd TP	Morph $\sim$ Seg Bkwd TP
	Word $\sim$ Syll Bkwd TP	Word $\sim$ Seg Bkwd TP

3.2 Results

In this section I present three measures for understanding the models: precision, recall and F-score (Manning et al. 2009). These measures are common to signal analysis of any type and have been used extensively to report on learning algorithms.

Precision and recall are measured in terms of hits, misses, correct rejection and false positives. *Hits* are correct predictions made by the categorizer in cases where there was a boundary and the categorizer reported it; *misses* are incorrect predictions in cases where there was a boundary and the categorizer reported none. From the perspective of sequences that do not contain a boundary, a *correct rejection* is a case when there is no boundary and the categorizer predicts none; likewise, a *false positive* is a case when there is no boundary and the categorizer predicts there to be one.

The table below illustrates the predictions of a categorizer compared to the actual data and the classification of the response.

(13) Sample categorizer results

Prediction	Actual	Classification
Morpheme Boundary	Morpheme Boundary	Hit
Morpheme Boundary	No Boundary	False Positive
No Boundary	No Boundary	Correct Rejection
No Boundary	Morpheme Boundary	Miss

Precision measures how likely the categorizer is to identify actual boundaries as opposed to non-boundaries; it is computed by dividing the number of hits by the sum of the hits and false positives (*i.e.* the number of correctly predicted boundaries over the total predicted boundaries):

(14) $$\text{Precision} = \frac{\text{\# of hits}}{\text{\# of hits + \# of false positives}}$$

Recall measures how likely the categorizer is to identify all actual boundaries; it is computed by dividing the number of hits by the sum of the hits and misses (*i.e.* the number of correctly predicted boundaries over the total actual boundaries):

(15) $\text{Recall} = \frac{\text{\# of hits}}{\text{\# of hits + \# of misses}}$

The best models will balance these two factors, identifying a large number of correct boundaries but not over predicting sequences that are within a domain. This is measured by the F-score, the harmonic mean[6] of precision and recall:

(16) $\text{F-score} = 2\frac{\text{Precision}\times\text{Recall}}{\text{Precision + Recall}}$

The F-Score ranges between 0, the worst score, and 100, the best score (because precision and recall are being reported as percentages); a higher value represents a better result.

In order to to test the null hypothesis, a class of models that randomly predict if a bigram has a boundary between its members or not was created. These models know how many boundaries there are, but randomly assign which bigram has a boundary.

Since the distinction between forward and backward TP is not used as a predictor by these models, we only need to consider the unit of analysis and the boundary type about which a prediction is being made. This means there are four comparison categorizers to consider: a categorizer that randomly assigns morpheme boundaries to syllabic bigrams, a categorizer that randomly assigns word boundaries to syllabic bigrams, a categorizer that randomly assigns morpheme boundaries to segmental bigrams, and a categorizer that randomly assigns word boundaries to segmental bigrams.

These categorizers are built by generating a random prediction, sampled from a normal distribution between 0 and 1, for each bigram. These numerical values are then turned into categorical decision in the same way as the categorizers that use TP as a predictor. However, instead of turning the probability of a bigram containing a boundary into a categorical response, like the categorizers that use TP as a predictor, the random categorizers select values less than or equal to the probability of a boundary and convert these into a categorical response indicating a boundary. This means that these categorizers will posit a boundary for any bigram which was (randomly) assigned a value less than the probability of a boundary existing in that dataset.

The precision, recall and F-score of these categorizers converges on the probability of a boundary; the probability of a morpheme boundary between syllables is %25. The probability of a word boundary between syllables is %29. The probability of a morpheme boundary between segments is %22. The probability of a word boundary between segments is %16.

The chart below summarizes the precision, recall, and F-score of the four comparison categorizers and eight fitted categorizers in predicting whether a bigram occurs over a boundary using the transitional probability of the two members of that bigram. The categorizers are separated by unit of analysis.

[6] Harmonic means are appropriate for taking the mean of two rates. The harmonic mean of AB $= 2\frac{A\times B}{A+B}$.

(17) Summary of categorizer performance

a. Syllabic bigrams

Categorizer		Precision	Recall	F-Score
Morph	∼ Random	%25.78	%25.70	25.74
Word	∼ Random	%29.01	%29.13	29.07
Morph	∼ Fwd TP	%2.57	%42.61	4.85
Word	∼ Fwd TP	%7.19	%40.21	12.20
Morph	∼ Bkwd TP	%0	%0	0
Word	∼ Bkwd TP	%9.07	%50.78	15.39

b. Segmental bigrams

Categorizer		Precision	Recall	F-Score
Morph	∼ Random	%21.82	%21.86	21.84
Word	∼ Random	%16.15	%16.26	16.20
Morph	∼ Fwd TP	%0.007	%6.0	0.0014
Word	∼ Fwd TP	%3.56	%24.82	6.22
Morph	∼ Bkwd TP	%0	%0	0
Word	∼ Bkwd TP	%0	%0	0

3.3 Overview

F-scores for the categorizers that use TP to predict boundaries range between 0 and 15. In comparison to the random models, the categorizers all performed worse with respect to F-score. While performing poorly, the categorizers that predict boundaries in syllabic bigrams have better F-scores than those that predict boundaries in segmental bigrams.

It should be noted that three categorizers failed to predict any boundaries. This is reflected in the F-score of zero of the categorizers that used backward TP of segmental bigrams and the categorizer that used backward TP of syllabic bigrams to predict morpheme boundaries. Since these models had zero hits, they had precision and recall of zero, resulting in an F-score of zero.

3.4 Syllabic bigram categorizers

The syllabic bigram categorizers had F-scores ranging between 0 and 15.39; the categorizers that predicted word boundaries had higher F-scores than the categorizers that predicted morpheme boundaries. For the categorizers that predicted any boundaries, recall was much higher than precision.

3.5 Segmental bigram categorizer

The segmental bigram categorizers had F-scores ranging between 0 and 6.22; the categorizers that predicted word boundaries had higher F-scores than the categorizers that predicted morpheme boundaries. For the categorizers that predicted any boundaries, recall was much higher than precision.

4 Discussion

One of the major contributions of this study is investigating the identification of word-internal morpheme boundaries. Because Sesotho has so many polymorphemic words, it provides a good test corpus for comparing word boundary identification and morpheme boundary identification. Transitional probability was not found to be a better than random predictor of boundaries by any measure. The results indicate that transitional probability between bigrams is a better indicator of word boundaries than of morpheme boundaries; this is found even when there is a higher number of morpheme boundaries than word boundaries (segmental bigrams). The TP of syllabic bigrams is a slightly better predictors of boundaries than the TP of segmental bigrams.

One factor which was thought to create a difference between backward and forward transitional probability was the order in which affixes are attached to roots[7]; this follows from the observation that roots are a much larger class of morphemes and therefore less predictable in contrast to the smaller set of possible affixal morphemes. In Sesotho, most words contain a string of inflectional prefixes. All things being equal, the forward and backward TPs between the prefixes will be equivalent. Therefore, we predicted that backwards transitional probability, from the root to the prefixes, would be higher than forward TP, from the prefixes to the root. This would make morpheme boundaries have generally higher backward TP than forward TP, meaning there is a greater distinction between in-morpheme bigrams and cross morpheme bigrams when considering backward TP as compared to forward TP. This prediction was not borne out. Backward TP was not a clearer indicator of what type of bigram exists in a bigram. In fact, backward TP was such an ambiguous indicator that three of the four categorizers which used it as a predictor indicated that all bigrams lacked any sort of boundary. This would be equivalent to noticing that there are proportionally more bigrams without boundaries and therefore predicting all bigrams to lack a boundary. In otherwords, backward TP was a very coarse way to predict the type and existance of boundaries in bigrams.

The categorizers did not perform very well at predicting word boundaries in comparison to other studies. In a comparable study, Daland (2009) reported models with an F-score of 62.7 for finding English word boundaries from forward transitional probability between segments. For Russian he reports an F-score of 58. The chart below summarizes the precision, recall and F-score of these categorizers.

Note that these categorizers all have higher precision than they do recall; this is the opposite of our categorizers, which all have higher recall than precision. This fact is related to the ratio of bigrams with boundaries to those without. When the ratio is low, as in Sesotho, there are fewer total hits needed since there is a low number of total boundaries to predict.

(18) Comparison of segmental bigram categorizers for word boundaries

Categorizer	Precision	Recall	F-score
English - Fwd TP (Daland)	%87.4	%48.9	62.7
Russian - Fwd TP (Daland)	%95	%40	58
Korean - Fwd TP (Daland&Zuraw)	%28	%11	15
Sesotho - Fwd TP	%3.56	%24.82	6.22

[7] The order of words in a phrase also creates a distinction between backward and forward TP.

Daland and Zuraw (2013) report much less success predicting word boundaries from transitional probability in Korean (F-score = 15). They attribute the shortcomings of their model to the fact that Korean generally tolerates the same sort of sequences word-internally and across word boundaries. This is related to the restrictive syllable structure of Korean. The same reasoning holds for Sesotho; because segmental sequences are largely constrained by the valid syllable types of Sesotho, segmental bigrams are not uniquely found within a word or across word boundaries, but rather they are equally likely within words as across word-boundaries.

However, there are other cues which may be informative in languages where segment sequencing is restricted by syllable shape, such as non-adjacent relationships between segments. Since the corpus did not encode a contrast between advanced and retracted mid vowels, none of the observed effects can be based on mid vowel quality. Tongue root harmony in Sesotho acts within words to make all mid vowels within parts of the word have the same quality, either advanced or retracted. Therefore, in a more robust dataset, we expect the syllabic bigram models to be improved when predicting word boundaries; syllables with mid vowels that are not both advanced or both retracted are less likely to be within a word than those that do agree. We do not expect that segmental bigrams would improve. This is due to the fact that when considering two syllables, the difference in mid vowel quality is inaccessible. However, when considering adjacent segments, this is not the case. A more nuanced model could include bigrams of adjacent vowels, in essence ignoring intervening consonants.

5 Conclusion

This paper demonstrates that transitional probability between adjacent segments and between adjacent syllables is not useful as a predictor of boundaries in Sesotho. Boundary identification based on TP is worse in Sesotho than in the worst previously reported study on Korean. Interestingly, these two languages both have restrictive syllable shapes.

The measures used in this investigation fail to tell us why transitional probability is a poor predictor in Sesotho. The overlapping distribution of morpheme internal, cross-morpheme and cross-word bigrams for both segments and syllables contributed to the poor results of using TP as a predictor.

In order to better investigate the role of phonological units in boundary prediction, a larger sample of languages must be considered. Important factors identified by this study and previous work include: morpheme shape, syllable shape, syllable inventory, segment inventory, and boundary frequency.

In order to investigate the role of directionality in TPs when predicting boundaries, it would be necessary to know with what frequency each type of bigram is being misclassified and what it is being classified as. For instance, in order to confirm that higher backward TP between roots an prefixes results in better distinction between within morpheme and cross morpheme bigrams, it would be necessary to check the number of cross morpheme bigrams classified as within morpheme by a categorizer using forward TP versus a categorizer using backward TP.

References

Bertoncini, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development*, 4:24–260.

Bijeljac-Babic, R., Bertoncini, J., and Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances. *Developmental Psychology*, 29:711–721.

Daland, R. (2009). *Word Segmentation, word recognition, and word learning: a computational model of first language acquisition.* PhD thesis.

Daland, R. and Pierrehumbert, J. (2011). Learning diphone-based segmentation. *Cognitive Science*, 35:119–155.

Daland, R. and Zuraw, K. (2013). Does korean defeat phonotactic word segmentation? In *ACL-2013*.

Demuth, K. (1983). *Aspects of Sesotho language aquisition*. PhD thesis, Indiana University.

Demuth, K. (1992). Acquisition of Sesotho. In Slobin, D., editor, *The Cross-Linguistic Study of Language Acquisition*, volume 3, pages 557–638. Lawrence Erlbaum Associates, Hillsdale, N.J.

Demuth, K. (2007). Sesotho speech acquisition. In *The international guide to speech acquisition*. Thomson Delmar Learning, Clifton Park, NY.

Doke, C. M. and Mofokeng, S. M. (1957). *Textbook of Southern Sotho Grammar*. Longmans, Cape Town.

Harris, Z. (1955). From phoneme to morpheme. *Language*, 31:190–222.

Harris, Z. (1967). Morpheme boundaries within words: report on a computer test. In *Transformational and Discourse Analysis Papers*.

Lewis, P. M., Simons, G. F., and Fennig, C. D., editors (2013). *Ethnologue: Languages of the World, Seventeenth edition*. SIL International, Dallas, Texas, online version: http://www.ethnologue.com edition.

Manning, C. D., Raghavan, P., and Schutze, H. (2009). Introduction to information retrieval. Online.

Pelucchi, B., Hay, J. F., and Saffran, J. R. (2009). Learning in reverse: 8-month-old infants track backward transitional probabilities. *Cognition*, 2.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.

Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Languages*, 35:606–621.

Zerbian, S. (2007). Phonological phrasing in northern sotho (bantu). *The Linguistic Review*, 24:233–262.