Comparison of Two Forced Alignment Systems for Aligning Bribri Speech

Rolando Coto-Solano
School of Linguistics and Applied Language Studies, Te Kura Tātari Reo
Victoria University of Wellington, Te Whare Wānanga o te Ūpoko o te Ika a Māui
PO Box 600, Wellington 6140, New Zealand
rolando.coto@vuw.ac.nz (corresponding)

and

Sofía Flores Solórzano
Universidad de Costa Rica
Sede del Atlántico, Turrialba, Cartago
sofia.floressolorzano@ucr.ac.cr

Received: June 30, 2016; Revised July 22, 2016; Accepted March 13, 2017

Abstract

Forced alignment provides drastic savings in time when aligning speech recordings and is particularly useful for the study of Indigenous languages, which are severely under-resourced in corpora and models. Here we compare two forced alignment systems, FAVE-align and EasyAlign, to determine which one provides more precision when processing running speech in the Chibchan language Bribri. We aligned a segment of a story narrated in Bribri and compared the errors in finding the center of the words and the edges of phonemes when compared with the manual correction. FAVE-align showed better performance: It has an error of 7% compared to 24% with EasyAlign when finding the center of words, and errors of 22~24 ms when finding the edges of phonemes, compared to errors of 86~130 ms with EasyAlign. In addition to this, EasyAlign failed to detect 7% of phonemes, while also inserting 58 spurious phones into the transcription. Future research includes verifying these results for other genres and other Chibchan languages. Finally, these results provide additional evidence for the applicability of natural language processing methods to Chibchan languages and point to future work such as the construction of corpora and the training of automated speech recognition systems.

Keywords: Bribri, natural language processing, forced alignment, phonetics, Chibcha, linguistics

Resumen

El alineamiento forzado provee un ahorro drástico de tiempo al alinear grabaciones del habla, y es útil para el estudio de las lenguas indígenas, las cuales cuentan con pocos recursos para generar corpus y modelos computacionales. Aquí comparamos dos sistemas de alineamiento, FAVE-align e EasyAlign, para determinar cuál provee mayor precisión al alinear habla en la lengua chibcha bribri. Alineamos una narración y comparamos el error al tratar de encontrar el centro de las palabras y los bordes de los fonemas con sus equivalentes en una corrección manual. FAVE-align tuvo mejor rendimiento, con un error de 7% comparado con 24% de EasyAlign para el centro de las palabras, y con errores de 22~24 ms para el borde de los fonemas, comparado con 86~130 ms con EasyAlign. Además, EasyAlign no pudo detectar el 7% de los fonemas, y al mismo tiempo añadió 58 sonidos espurios a la transcripción. Como trabajo futuro verificaremos estos resultados con otros géneros hablados y con otras lenguas chibchas. Finalmente, estos resultados comprueban la aplicabilidad de los métodos de procesamiento de lengua natural a las lenguas chibchas, y apuntan a trabajo futuro en la construcción de corpus y el entrenamiento de sistemas de reconocimiento automático del habla.

Palabras clave: Bribri, procesamiento de lenguaje natural, alineamiento forzado, fonética, Chibcha, lingüística

1 Introduction

Forced alignment is a family of algorithms that take as their input an audio file and its transcription, and calculate the time points in the audio file that correspond to each word, and even to each phoneme in the transcription [1,2]. Figure 1 shows an example in English, where the alignment is displayed in the visual representation used by the *Praat* program [3]. The first row shows the intervals corresponding to words, while the second row shows intervals corresponding the phonemes in each word. These algorithms are trained using Hidden Markov Models (HMM) which learn by imposing temporal *frames* on the recording, extracting the spectral information for each phone, such as formants and wave intensity. After training, these frames are used to traverse the signal and determine the potential points of transition between two phones [4].

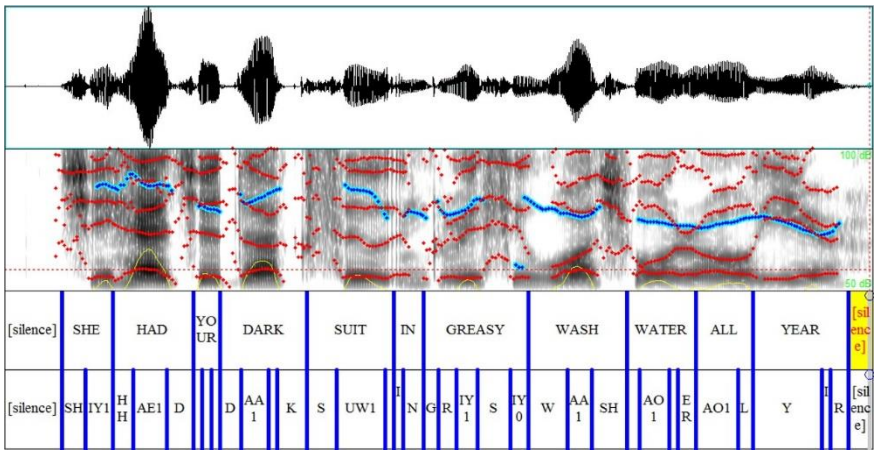

Figure 1. Example of forced alignment for an English recording [5]

Oral corpora segmentation facilitates the use of data in linguistic research ranging from phonetics to sociolinguistics, and forced alignment provides drastic saving in time when segmenting and annotation recordings [6,7,8]. For example, Labov et al. [9], report that up to 300 vowels can be processed with 40 hours of manual work, whereas up to 9000 vowels can be studied in the same time using the *FAVE-align* alignment system. This family of algorithms is particularly useful when studying minority and Indigenous languages, which lack corpora and acoustic models, and in general resources for their study from computational linguistics [10,11,12]. This is possible because, even though existing systems using acoustic models are trained for languages such as English or Spanish, there is a high level of transfer when these models are applied to other languages [10]. For example, acoustic models of Hungarian can be used to recognize Czech phonemes because of the similarities in their phonological systems [13]. Likewise, many of the phonemes in the Yoloxóchitl Mixtec language are close enough to English that forced alignment produces satisfactory results [10].

The authors of this paper are developing the first oral corpus of spontaneous speech in the Bribri language. This corpus, whose compilation is in progress and which will contain approximately 4 hours of recordings, will document speech from a wide range of social situations encountered in Bribri daily life, such as greetings, dialogues, forms of politeness, prayers, insults and praise, and storytelling among others. The outcome of this project will provide an accurate reflection of Bribri speech, and will contribute to creating resources for acoustic modeling and other natural language processing applications. We believe that the method of untrained forced alignment is a rapid way to construct this type of materials in under-resourced languages such as Bribri.

The structure of the paper is as follows: Section 2 of this paper will describe the methodology used for alignment, including the preparation of the Bribri input, the transformations needed to normalize the output, the procedures to measure the error rate for the aligners, and the materials used for the comparison. Section 3 describes the results in terms of error rates, and provides evidence that FAVE-align has better performance in a number of conditions. Sections 4 and 5 discuss the pros and cons of each alignment system and propose new experiments to study forced-alignment in Chibchan languages.

2 Methodology

In this work we aligned conversational speech in the Bribri language (Chibchan, Costa Rica) to match their transcriptions and their sound waves. This step expands on the methodology of DiCanio et al. [10], who used forced transcription with acoustic models of English to process spoken word lists in Yoloxóchitl Mixtec (Otomanguean, Mexico). It also expands on the methodology of Coto-Solano & Flores-Solórzano [14], who first applied untrained

forced-alignment to running speech in Bribri. In this paper we will compare the performance of two acoustic models: (i) The English model in the *FAVE-align* system [15], a derivate of the *Penn Phonetics Lab Forced Aligner Toolkit* or *P2FA* [16], trained at the University of Pennsylvania, and (ii) a French acoustic model in *EasyAlign* [17], trained at the University of Geneva.[1] These two systems were chosen because they both represent the state of the art of the numerous systems trained using the Hidden Markov Models in the HTK toolkit [18], which is the most tested alignment toolkit available. Another factor was that these systems have two of the easiest interfaces when interacting with users.[2]

2.1 Data preparation

Following the method in Coto-Solano & Flores-Solórzano [14], we prepared the data for alignment by FAVE-align and EasyAlign. For FAVE-align, this involved manually marking the limits of phrases, as shown in figure 2, converting standardized Bribri spelling[3] to the Arpabet transliteration code, as illustrated in table 1, and preparing a file with the list of unique words in the recording and their Arpabet equivalents, as shown in figure 3. These files are then processed using Python code [link], or through an online interface [link].

```
S01     AG     1      3      Ì kuhéhkih wìm òhr darë`rë`
S01     AG     4.5    6      Wìm che kë´këpa tö
S01     AG     6.6    8.3    tö wìm dör pë' wë´m táhih
```

Figure 2. List of phrases with start and end times

Graphemes	IPA	Arpabet	Graphemes	IPA	Arpabet
a	/a/	AE1	b	/b/	B
e	/ɛ/	EH1	ch	/tʃ/	CH
ë	/ɪ/	IH1	d	/d/	D
i	/i/	IY1	j	/x/	HH
i	/j/	Y	k	/k/, [kʰ]	K
o	/ɔ/	OW1	l	/ɭ/	R
ö	/ʊ/	UH1	m	/m/	M
u	/u/	UW1	n	/n/	N
u	/w/	W	ñ	/ɲ/	N
			p	/p/	P
			r	/ɾ/	R
			rr	/r/	R
			s	/s/	S
			Sh	/ʃ/	SH
			t	/t/	T
			ts	/ʦ/	T S
			w	/w/	W
			y	/ʤ/	JH

Table 1. Bribri phonemes and their Arpabet equivalents

```
be'       B EH1
bë`rie    B IH1 R Y EH0
bikéitse  B IY0 K EH1 EH0 T S EH0
bö`k      B UH1 K
```

Figure 3. Examples from a dictionary with words transliterated into Arpabet

Table 1 shows the transliteration decisions for all the Bribri segmental phonemes; the IPA values are taken from Constenla [20] and Flores Solórzano [21]. The proposed system lacks a way to represent the phonemic

[1] Both systems presuppose that speech can be segmented into discrete elements, and that it is possible to mark the precise boundary between two phonemes. This presupposition is reasonable when there is, for example, a voiceless consonant in word-initial position, but almost impossible to defend when marking the boundary between two vowels, or when there are cases of segmental reduction with phonemes sharing features [19]. Despite this obstacle, we will use these tools keeping in mind that we are not just studying individual phones, but also the transitions between different types of phones.

[2] Systems like those based on Kaldi [33], for example, have considerable more involved configuration and data-preparation processes.

[3] Figure 2, and indeed all the transcriptions, use a Bribri orthography based on Constenla's [20] standard. The system presented here is modified so that nasals are represented by 'h' and not by a line underneath the vowel (e.g. *a̱* → *ah*), and that the complex diacritics are represented next to the vowel (e.g. ë́ → ë´). These modifications were made because the FAVEalign system does not use Unicode characters.

suprasegmentals of Bribri, including tones, nasalization and glottalization, so an additional routine was programmed to convert the Arpabet symbols back to Bribri orthography (available at: http://github.com/rolandocoto/alineacion-lenguas-cr) and recover these suprasegmentals. Collapsing these categories in the Arpabet does not degrade the quality of the alignment; this will be discussed in detail in section 3.3. The graphemes '*i*' and '*u*' are used for both the vowels /i/ and /u/ and the semiconsonants /j/ and /w/; this difference is captured in the Arpabet transliteration. In addition to the vowel per se, each representation of an Arpabet vowel carries a number. This is meant to represent the accentual system of English and it can be set to three values: 1 for accented syllables, 2 for syllables with secondary stress, and 0 for non-accented syllables. In these tests all of the syllables were set to 1 to avoid keep the system from associated any Indigenous vowels with English schwas and reduced vowels.[4]

Regarding the consonants, three decisions were taken when representing them in Arpabet: (i) to use the Arpabet glyph 'N' to represent the two Bribri phonemes '*n*' /n/ and '*ñ*' /ɲ/, (ii) to use the glyph 'R' to represent the phonemes '*l*' /l/ and '*r*' /ɾ/, and (iii) to split the phoneme '*ts*' /ts/ into two glyphs, 'T' and 'S' because English lacks /ts/. In addition to this, the decision was made to not explicitly represent the glottal stop as a separate phoneme, and it was instead treated as a part of the vowel (which it is in most dialects and phonological environments); this did not result in any significant degradation of the system's accuracy (see section 3.4) [5].

In the case of EasyAlign [17], we used the French model as it is the primary model in the system and the one that provides better performance [17, pp. 4]. It is installed as an add-on module of Praat, and the system's developers invite users to send data to train new acoustic models, but do not provide this methodology as an open-source mechanism.[6] In order to prepare the data, the user manually provides the phrase boundaries and the transcription in a standardized orthography. This step, however, requires the elimination of all diacritics, which have to be reestablished manually after the alignment. After the phrase boundaries are set, a Praat script returns a TextGrid with the text aligned for words, syllables and phonemes. This TextGrid has phonemes transcribed in the SAMPA format, and has to manually be converted to a user-selected transcription system.

2.3 Alignment Results and Error Estimation

Figure 4 shows the results of alignment using FAVE-align, while figure 5 shows the alignment obtained with EasyAlign.

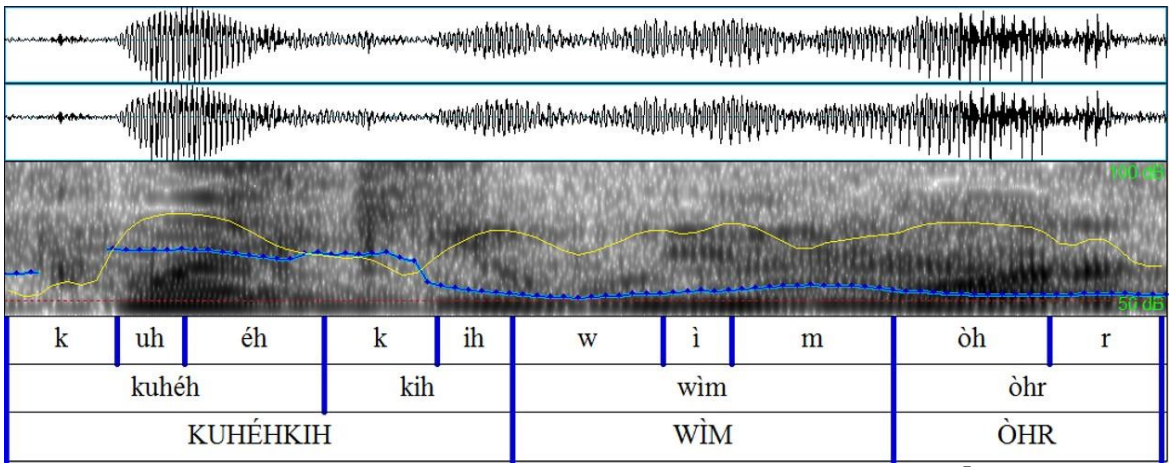

Figure 4. Praat TextGrid with Bribri text force-aligned by FAVE-align [7]

[4] The authors acknowledge that, as understanding of the prosody of Bribri improves, researchers might be able to use the other digits (two and zero) to represent different degrees of accentual marking.
[5] The glottal stop is a problematic phoneme for English acoustic models because this sound does not exist as a phoneme in this language [10].
[6] At the time of writing the developers of EasyAlign are not actively developing other acoustic models. Neither is it possible to modify the tool because it is provided as binary files and not as open source code.
[7] In these tests we used transcriptions that didn't depend on Unicode characters. Therefore, nasals are represented with an 'h' after the vowel, and the high and low tones on a lax vowel are represented using an accent after the vowel.

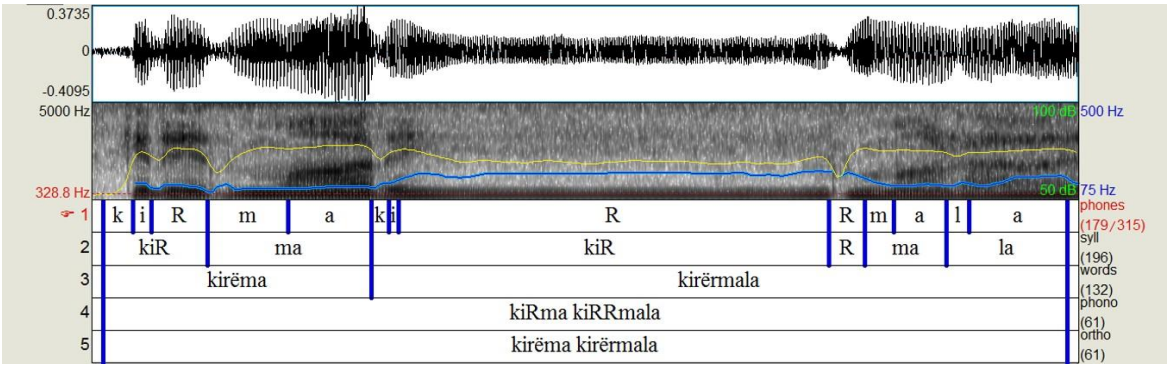

Figure 5. Praat TextGrid with Bribri text force-aligned by EasyAlign

After generating the automatic TextGrids these were manually corrected to determine the accuracy of the forced alignment algorithms. We used the *root mean squares* method [22] to calculate the difference between the center of the words in the automatic and the hand-corrected TextGrids. The word *òhr* (*ồr* 'it shouts') in figure 6 provides an example. The center of the automatically aligned word is at 1.887 seconds, while the center of the corrected word is at 2.039 seconds. Then, the difference of the center is $\sqrt{(2.039 - 1.887)^2}$ = 152 milliseconds. Note that we use the square root to calculate the absolute difference, in case one of the corrected intervals is located before the automatic one, which would result in a negative difference. In order to provide more clarity about the relationship between a time difference and the specific word, we calculated the percentage of the duration of each word that corresponds to the difference in the center between the automatic and hand-corrected versions. For example, the corrected word *òhr* is 196 milliseconds long, so the percent difference of the center is 152/196 = 78% of the duration of the word *òhr*.

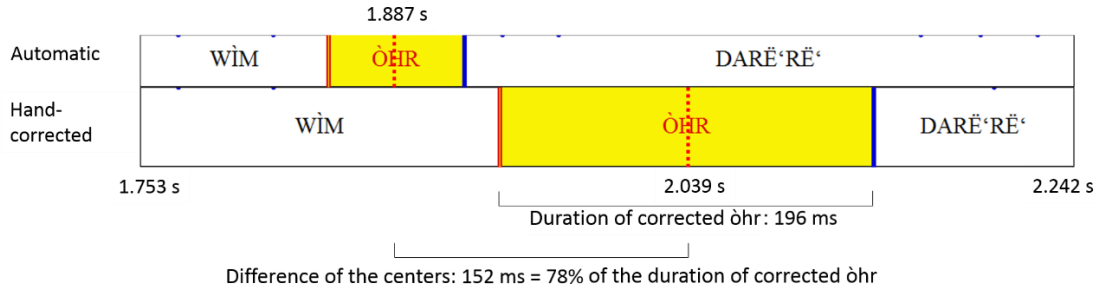

Figure 6. Error estimation for the alignment of the center of a word

The root mean squares method was also used to calculate the difference in the beginning and the end of the phonemes. As shown in Figure 7, the difference at the beginning of the phoneme '*uh*' /ũ/ is 19 milliseconds.

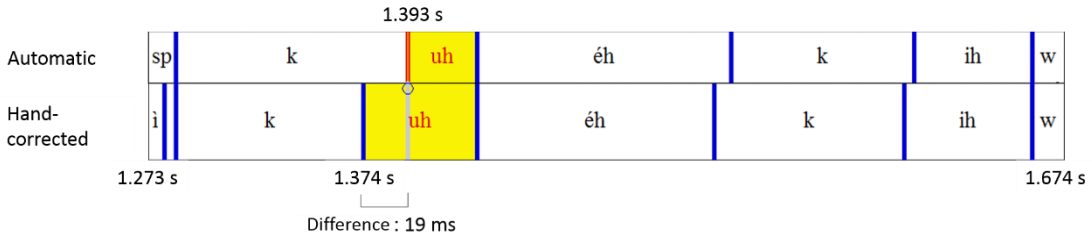

Figure 7. Error estimation for the alignment of the beginning of a phoneme

2.4 Statistical methods

To determine if there were significant differences in error rates, we used Linear Mixed Effects models (LMER) [23] with word as a random variable. The models were calculated using the R program [24].

2.5 Materials

In this study, we analyzed the first fifty seconds of narration from the story *Ì kṹkĩ̀ wìm ồr darèrè* 'Why does the monkey howl so loudly' [25]. Table 2 provides details of the number of words and segments studied, as well as the rate of speech for the recording.

Recording	Words	Vowels	Consonants	Syllables/Second
Bribri: Monkey	94	154	120	3.1

Table 2. Items extracted from the recording (same for both aligners)

3 Results

This section compares the error rates of words, vowels and consonants aligned with FAVE-align versus those aligned with EasyAlign. It will also report on incongruences between the input and the output text (section 3.2) and on the effects of boundaries between segments on system performance.

3.1 Words

Words in this story have an average length of 287±149 milliseconds long[8]. The FAVE-align system had an average error of 19 ms when aligning the center of the words, which is equivalent to 7% of the average duration of words. This is significantly more precise than EasyAlign, which has an error of 50 ms (24% of duration) when aligning the center of words ($\chi^2(109)=-3.3$, $p<0.005$). The model with aligner as an independent variable is significantly different from the null model ($\chi^2(1)=8.2$, $p<0.005$).

	FAVE-align	EasyAlign	LMER
Error (ms)	19 ± 27	50 ± 102	$\chi^2(109)=-3.3$, $p<0.005$
% Error	7% ± 12%	24 ± 49%	

Table 3. Difference in the center of words between automatic and hand-corrected Intervals (error), in milliseconds and in percentage of the duration of the word

3.2 Missing phonemes

One of the main obstacles in using EasyAlign is the lack of consistency between the number of input units (the number of phonemes represented in the user's transcription system) and the number of intervals generated. Figure 8 shows such a problem in an EasyAlign generated TextGrid. In the first and second tiers the Bribri syllables *í+se* are transcribed as *iz*. Both a syllable and a phoneme are lost, making it necessary to add these again the manual correction. Figure 9 shows the opposite case, where EasyAlign has added a phoneme '*E*' to the word *sö* 'they (are)', which would need to be deleted in a manual correction stage prior to converting the SAMPA back to the user's chosen phonological representation.

In the sample studied, the EasyAlign output was missing 21 out of 274 segments: 7 consonants out of 120 (6%) and 14 vowels out of 154 (9%). In addition to this, there were 58 spuriously added segments, 33 vowels and 25 consonants. Only segments that were detected by both aligners were included in the error comparisons.

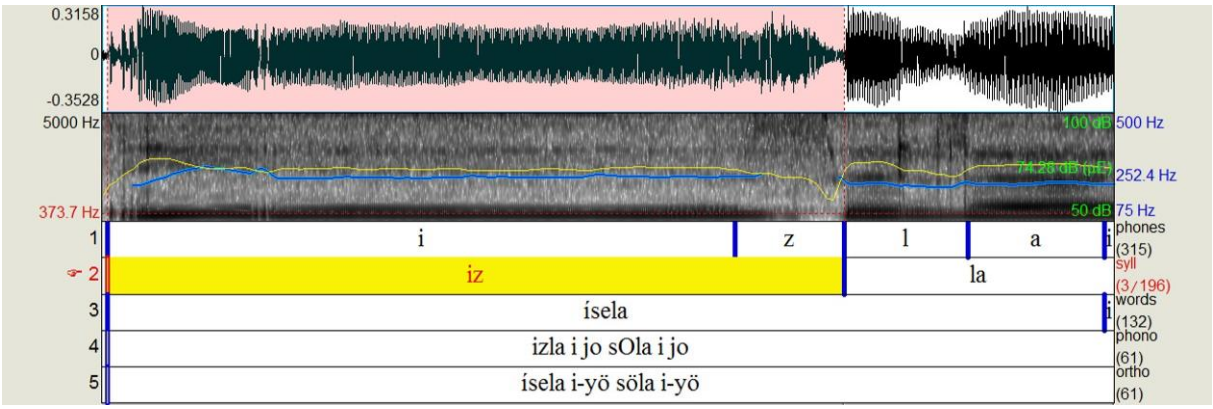

Figure 8. Bribri TextGrid where EasyAlign has deleted a vowel (*íse -> iz*)

[8] The numbers reported next to the results (after the plus-minus sign) are the standard deviations.

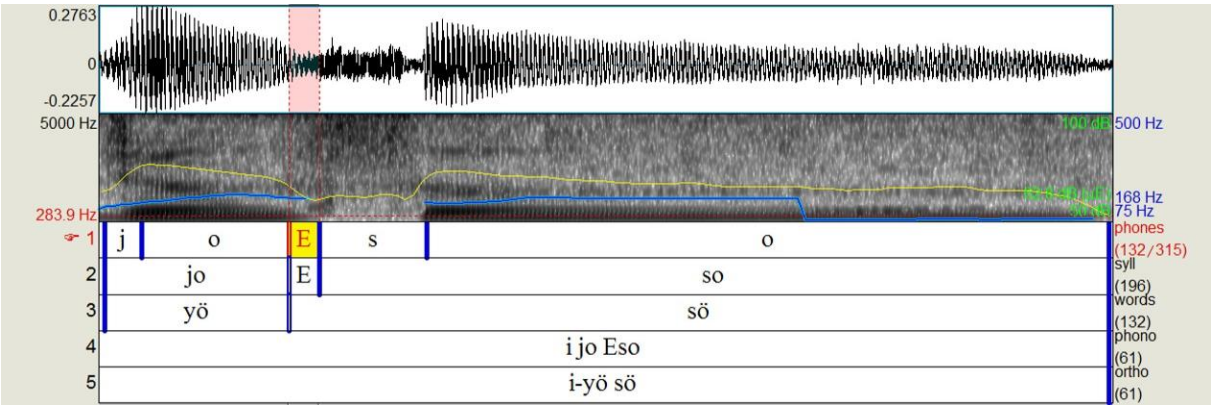

Figure 9. Bribri TextGrid where EasyAlign has inserted a vowel (*sö -> Eso*)

FAVE-align preserves all of the segments that the user specified in the initial transcription (see section 2.1), which allows for the study of segments even when they present severe reductions [19]. This is an advantage when studying phonetic details, but it comes at a price at the preprocessing stage: Preparing a file for alignment with FAVE-align implies preparing phrase boundaries and making the dictionary, which takes longer than merely marking the phrases as is done in EasyAlign.

3.3 Vowels

Vowels are, on average, 105±46 milliseconds long. As shown in table 4, FAVE-align has significantly smaller error rates, between 22 and 26 ms at the edges, compared to the 86~102 ms of EasyAlign. The model with aligner as an independent variable is significantly different from the null model for both the beginning ($\chi^2(1)=22.3$, $p<0.0001$) and the end of the vowels ($\chi^2(1)=31.6$, $p<0.0001$).

	FAVE-align (ms)	EasyAlign (ms)	LMER
Beginning	22 ± 44	86 ± 159	$\chi^2(258)=-4.8$, $p<0.000001$
End	26 ± 48	102 ± 156	$\chi^2(258)=-5.7$, $p<0.000001$

Table 4. Average error for vowels

The averages, however, hide the fact that the convergence rate onto the correct alignment is very different between the two systems. Figure 10 shows the cumulative percentage of items below a difference time. For the beginning of vowels, 80% of the items of FAVE-align have an alignment error of 31 ms or less, whereas 80% of the EasyAlign items have an alignment error of 150 ms or less. As for the end of the vowels, 80% of FAVE-align items have an error of 44 ms or less, whereas 80% of EasyAlign items have an error of 190 ms or less. The convergence rates also show differences in the starting accuracy for both systems: 57% of the FAVE-align items had an error of 1 ms or less at the beginning of the vowels, compared to 42% of the EasyAlign items. When we look at the end of the vowels, we see that 49% of the FAVE-align items had an error of 1 ms or less, compared to 43% for the EasyAlign items. In summary, FAVE-align vowels have more items with negligible error rates (1 ms or less), and more of the items have smaller error rates compared to the EasyAlign vowels. Finally, figure 11 illustrates the complete spread of responses for the vowel alignment. The distribution of EasyAlign errors shows more variability and outliers than that for FAVE-align.

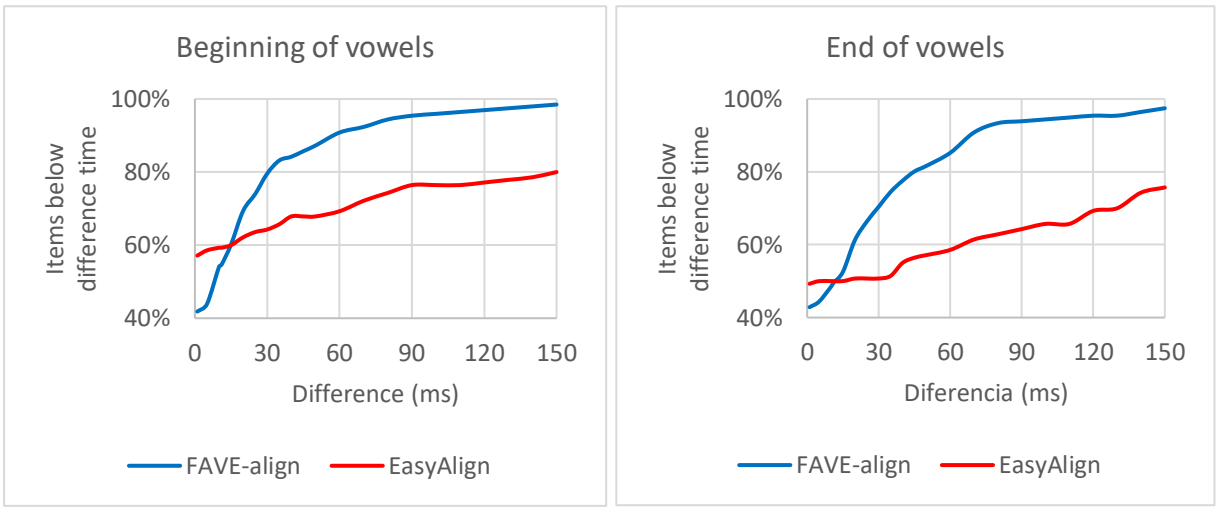

Figure 10. Difference between FAVE-align and EasyAlign in the alignment of vowels

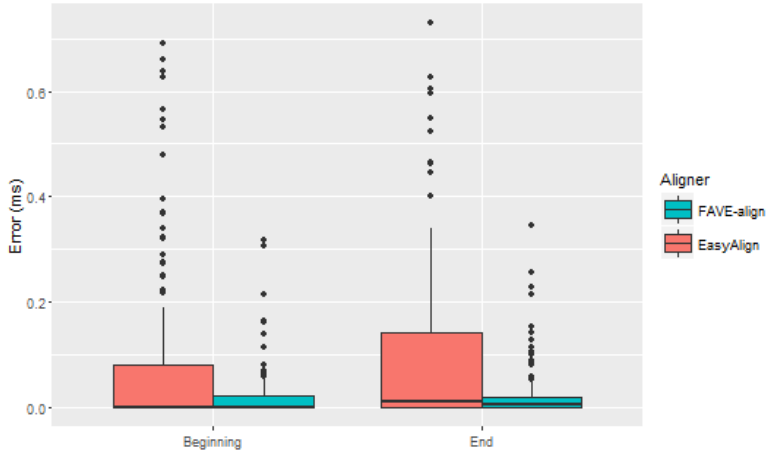

Figure 11. Difference in the distribution of error rates for vowels

We will now turn to specific phonological features of Bribri vowels, including glottalization, nasality, tone and vocalic quality. We found no interaction between aligner and the marking of glottalized vowels. We did find a significant interaction between aligner and nasality ($\chi^2(257)=-2.5$, $p<0.02$): In FAVE-align nasal vowels perform better, whereas in EasyAlign oral (non-nasal) vowels perform better. This is an outcome completely opposite to what we expected, because the EasyAlign model would have nasal vowels as part of its French input[9]. The difference between null and non-null models in this comparison is $\chi^2(3)=39.8$, $p<0.0001$.

	FAVE-align (ms)	EasyAlign (ms)
End of nasal vowel	19 ± 34	155 ± 196
End of non-nasal vowel	29 ± 51	87 ± 141

Table 5. Differences in the alignment error depending on nasality

We also found an interaction between aligner and the error at the end of the vowels depending on the tone. The difference in precision between the two aligners is much less with the low and neutral tones, where FAVE-align is approximately twice as precise as EasyAlign, than with the other tones, where FAVE-align is between 5 and 10 times more precise. The difference between null and non-null models in this comparison is $\chi^2(9)=43.1$, $p<0.0001$.

[9] One possibility is that there is a mismatch in the repertoire of nasal vowels between the two languages. French has four nasal vowels (/ã/, /ɛ̃/, /õ/ and /œ̃/), while Bribri a somwhat different set of five nasals (/ã/, /ɛ̃/, /ĩ/, /ɔ̃/ and /ũ/).

	FAVE-align (ms)	EasyAlign (ms)	LMER Interaction Aligner*Tone
High	29 ± 64	147 ± 194	- [10]
Falling	20 ± 35	138 ± 151	$\chi^2(252)=0.0$, $p=1.0$
Rising / Glottal	14 ± 17	120 ± 173	$\chi^2(252)=0.3$, $p=0.8$
Low	40 ± 56	75 ± 79	$\chi^2(252)=1.5$, $p=0.13$
Neutral [11]	27 ± 47	72 ± 138	* $\chi^2(252)=2.1$, $p<0.04$

Table 6. Differences in the alignment error at the end of the vowel depending on its tone

Finally, there is a main effect for vocalic quality. Both aligners perform better with *i* /i/ and *ö* /ʊ/. This warrants further investigation of the exact phonetic distribution of Bribri vowels, and why would these two be a good match to categories in English and French.

	FAVE-align (ms)	EasyAlign (ms)	LMER
[a]	35 ± 63	154 ± 174	-
[i]	19 ± 24	67 ± 112	$\chi^2(28)=-2.3$, $p<0.03$
[ö]	20 ± 27	60 ± 140	$\chi^2(250)=-2.1$, $p<0.04$

Table 7. Differences in the alignment error at the end of the vowel depending on vocalic quality

In this section we have compared the performance of the two aligners when aligning vowels, and FAVE-align produces results with smaller error rates. In the next section we will compare the performance of both aligners when processing consonants.

3.4 Consonants

Consonants are, on average, 82±38 milliseconds long. As shown in table 8, FAVE-align has smaller error rates, 23~24 ms at the edges, compared to the 112~130 ms of EasyAlign. The model with aligner as an independent variable is significantly different from the null model for both the beginning ($\chi^2(1)=29.5$, $p<0.0001$) and the end of the vowels ($\chi^2(1)=25.1$, $p<0.0001$).

	FAVE-align (ms)	EasyAlign (ms)	LMER
Beginning	23 ± 52	130 ± 204	$\chi^2(214)=-5.6$, $p<0.000001$
End	24 ± 54	112 ± 180	$\chi^2(214)=-5.1$, $p<0.000001$

Table 8. Average error for consonants

As was the case with the vowels, the convergence rate for the consonants is shown in figure 12. In general, the error rates for consonants are again smaller in the FAVE-align TextGrids. For the beginning of consonants, 80% of the FAVE-align items have an alignment error of 29 ms or less, whereas 80% of the EasyAlign items have an alignment error of 251 ms or less. As for the end of the consonants, 80% of FAVE-align items have an error of 31 ms or less, whereas 80% of EasyAlign items have an error of 225 ms or less. Regarding the starting accuracy of the aligners, 52% of the FAVE-align items had an error of 1 ms or less at the beginning of the consonants, compared to 46% of the EasyAlign items. The end of the consonants is the only case where EasyAlign shows a better performance than FAVE-align: 52% of EasyAlign items had an error of 1 ms or less, compared to 50% of FAVE-align items. Finally, figure 13 illustrated the spread of responses for consonant alignment. The error distribution of the EasyAlign output shows more variability and outliers than the FAVE-align one.

[10] The Linear Mixed-Effects Model in R compares the levels of each variable with the first level in alphabetical order. A dash in the LMER column indicates that this was the level the others were compared against.

[11] Bribri has two tonal phonemes that have been traditionally described as a single "low" tone: A phonetically low tone, and a tone with an underspecified register which can surface as either low or mid depending on the register of the following tone [26].

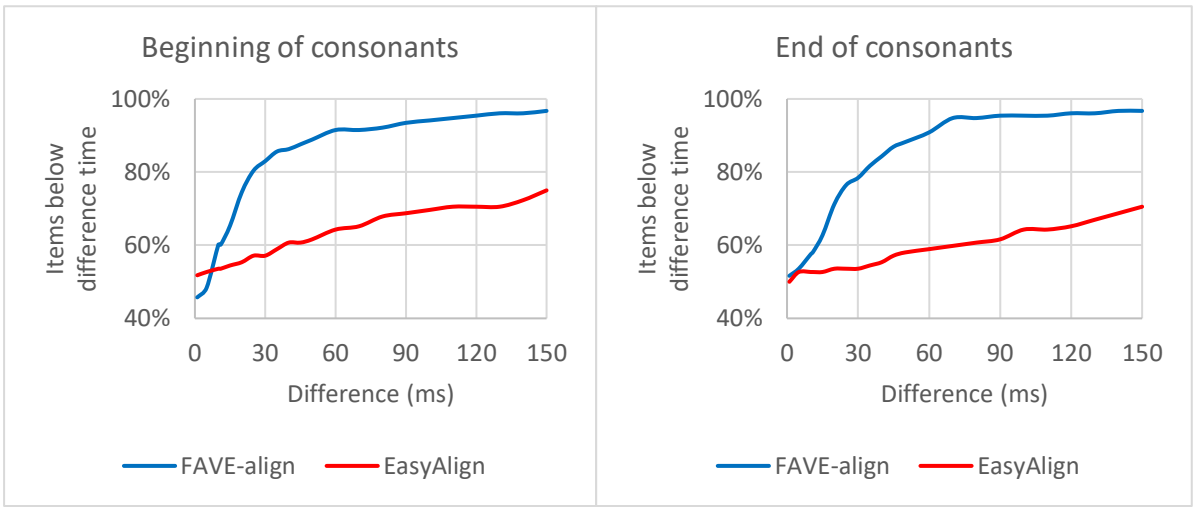

Figure 12. Difference between FAVE-align and EasyAlign in the alignment of consonants

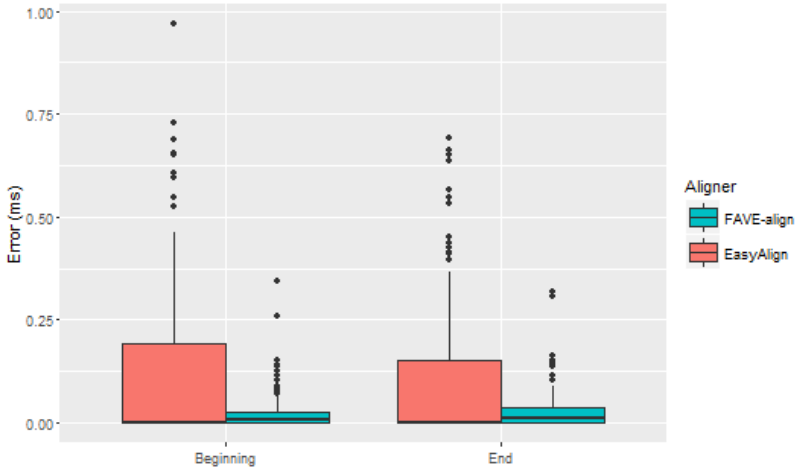

Figure 13. Difference in the distribution of error rates for consonants

We examined several phonological features of consonants, namely voicing, manner and place of articulation. There were no effects of voicing dependent on the aligner. There was an interaction between aligner and manner of articulation: FAVE-align has better performance overall, but the fricatives perform much better in that aligner, with almost no error (average 2 ms). The difference between the null and full models was significant ($\chi^2(5)=34.2$, $p<0.0001$).

	FAVE-align (ms)	EasyAlign (ms)	LMER aligner*manner
Stops	24 ± 48	80 ± 154	* $\chi^2(212)=2.1$, $p<0.04$
Fricatives	2 ± 7	181 ± 256	-
Sonorants	31 ± 67	140 ± 187	$\chi^2(212)=1.2$, $p<0.2$

Table 9. Error at the end of consonants depending on manner of articulation

There was also a significant interaction between aligner and place of articulation. Performance is better for FAVE-align in almost all categories except for velar sounds. For velars, the error rates are practically equal for the two aligners (28 vs. 31 ms). The full model was significantly different from the null model ($\chi^2(9)=36.7$, $p<0.0001$).

	FAVE-align (ms)	EasyAlign (ms)	LMER aligner*manner
Labial	18 ± 26	129 ± 179	$\chi^2(210)=-0.1$, p=0.9
Alveolar	27 ± 66	134 ± 196	-
Post-alveolar	11 ± 14	138 ± 208	$\chi^2(210)=-0.3$, p=0.8
Velar	28 ± 55	31 ± 96	$\chi^2(210)=2.2$, p<0.03

Table 10. Error at the end of consonants depending on place of articulation

Finally, we found no effects on the boundaries of vowels or consonants that correlate to the difference in aligners.

4 Discussion

Previous experiments [14] compared the two systems when aligning songs, but found that both had equally unreliable performances, with errors of up to 190 ms when finding the center of words. The results in this paper provide evidence that FAVE-align achieves better performance than EasyAlign when aligning Bribri running speech. Both systems provide output where 43% to 52% of segments have an error of 1 ms or less, but FAVE-align items have smaller average errors (22~26 ms) than those of EasyAlign (86~130 ms). The relatively small error rates of FAVE-align mean that it can provide a usable tool for aligning Bribri text, demonstrating that these systems can be used for the initial alignment of Indigenous speech. These error rates permit much faster alignment than would be possible with completely manual alignment.

Given the better performance of FAVE-align when aligning phonemes and words, we infer that there was a higher transfer from the English acoustic model to Bribri that there was from the French model. The experiment also continues to provide evidence in favor of using automatic alignment to work with the Indigenous languages of Costa Rica. Table 11 provides some of the advantages and disadvantages of each of the systems used, which should be evaluated by the user before implementing an alignment system. In general, EasyAlign makes it quicker to pre-process the data, but FAVE-align provides greater control over the resulting TextGrids.

System	Advantages	Disadvantages
FAVE-align	Greater control of the transcription: The system aligns exactly the number of units that the user provides as input, and it uses the transliteration units that the user provides.	There is a greater workload when preparing the document for processing.
EasyAlign	Initial alignment is quicker because the system attempts to automatically generate the phonological transcription.	The system can erase or add phonological units when generating the transcription. These units might be unrelated to the actual units in the Indigenous language. The system is not consistent in its use of SAMPA, and this has to be corrected manually (e.g. the SAMPA letter 'O' is mapped to the Bribri phonemes /ɔ/ and /ʊ/) The lack of SAMPA consistency makes it difficult to retrieve phonetic features (e.g. nasality, tone, etc).

Table 11. Comparison of the FAVE-align and EasyAlign systems in the pre and post-processing of data

Even with their disadvantages, both systems contribute important savings in time and effort when generating aligned transcriptions of Indigenous languages. These procedures will help bring computational linguistics techniques to the study of these languages, particularly in three areas. First, it would allow for better studies of their phonetics and of features such as voicing, prosodic realization and speech reduction. Second, it would help in the creation of speech corpora to train acoustic models specific to these languages. Third, these corpora can also be used for the training of automatic speech recognition systems, which would automate transcription and allow access to a larger volume of data, including data in legacy media such as audio cassettes and other older recordings.

Finally, these advances in the computational linguistics of Indigenous languages have the potential of generating a positive impact in the revitalization of these languages. According to UNESCO's vitality measures [27,28], Bribri is a vulnerable language. Enhancing the digital visibility of Bribri, in particular through means where speakers

could interact with interfaces in their own languages, would greatly contribute to the efforts done by activists and linguists to improve the status of these languages and create additional environments for their learning and use [29,30,31,32].

5 Conclusions

We used English and French language acoustic models to align running speech in the Indigenous language Bribri. The results show that FAVE-align has better performance than EasyAlign when aligning Bribri, both in finding the center of words (7% versus 24% of error) and when finding the edges of phonemes (22~26 ms versus 86~130 ms). At this time we recommend the use of the FAVE-align system when marking Bribri text, but we also recommend that further testing is necessary, for example comparing the English language model in both systems, and determining if the FAVE-align advantage remains in other Costa Rica Chibchan languages such as Cabécar and Malecu. Regarding the use of forced aligners, we still strongly believe that they can aid in the study of the Indigenous languages of Costa Rica by accelerating the alignment of speech corpora, which can be later used in phonetic studies, and in the training of language-specific acoustic models and automated speech recognition systems.

6 Acknowledgments

The authors wish to thank Dr. Monica McCauley of University of Wisconsin-Madison for her LSA presentation in January 2015, as well as Dr. Douglas Whalen from Haskins Lab at Yale University for discussing the results obtained in this study. We also wish to thank M.A. Dane Bell and M.Sc. Samantha Wray from the University of Arizona for their comments on previous drafts of this work, and two anonymous reviewers for their comments on previous versions of this work. Any errors in the paper remain our own.

7 References

[1] C. Wightman and D. Talkin. 1997. “The Aligner: Text to speech alignment using Markov Models”. *Progress in Speech Synthesis*, pp. 313-323. 1997. https://doi.org/10.1007/978-1-4612-1894-4_25.

[2] F. Schiel and C. Draxler. 2003. *The production of speech corpora*. Bavarian Archive for Speech Signals. 2003.

[3] P. Boersma. “Praat, a system for doing phonetics by computer”. *Glot International*, volume V, number 9/10, pp. 341-345, 2001.

[4] J. Yuan *et al*. “Automatic Phonetic Segmentation using Boundary Models”. *Proc. Interspeech* 2013, pp. 2306-2310.

[5] T.J. Yoon. “HTK-TIMIT Forced Alignment Toolkit”. 2008. [Online]. Available: http://web.uvic.ca/~tyoon/resource/htk_utt.m. Accessed on: January 15, 2016.

[6] J. Yuan and M. Liberman. “Investigating /l/ variation in English through forced alignment”. *Proc. InterSpeech* 2009, pp. 2215–2218.

[7] M. Adda-Decker and N. Snoeren. “Quantifying temporal speech reduction in French using forced speech alignment”. *J. Phonetics*, vol. XXXIX, pp. 261–270, 2011. https://doi.org/10.1016/j.wocn.2010.11.011.

[8] C.Y. Lin, R.J. Cheng Yuan and K.T. Chen. “Automatic segmentation and labeling for Mandarin Chinese speech corpora for concatenation-based TTS”. *Computational Linguistics and Chinese Language Processing* vol. X, num 2, pp. 145–166, 2005.

[9] W. Labov, I. Rosenfelder and J. Fruehwald. 2013. “One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis”. *Language*, vol. LXXXIX, num. 1, pp. 30–65, 2013. https://doi.org/10.1353/lan.2013.0015.

[10] C. DiCanio *et al*. “Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment”. *J. Acoustic Society of America*, vol. CXXXIV, num. 3, pp. 2235-2246, 2013. https://doi.org/10.1121/1.4816491.

[11] J. Strunk, F. Schiel and F. Seifart. “Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS”. *Proc. 9th Int. Conf. on Language Resources and Evaluation (LREC'14)*, pp. 3940-3947, 2014.

[12] S. Brognaux *et al*. “Train&Align: A New Online Tool for Automatic Phonetic Alignment”. *IEEE Spoken Language Technology Workshop (SLT)*, pp. 416-421. 2012. https://doi.org/10.1109/SLT.2012.6424260.

[13] K.C. Sim and H. Li. “Robust phone mapping using decision tree clustering for cross-lingual phone recognition”. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4309–4312, 2008.

[14] R. Coto-Solano and S. Flores Solórzano. "Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de Costa Rica". Káñina. In press.
[15] I. Rosenfelder *et al*. "FAVE (Forced Alignment and Vowel Extraction) Program Suite". [Online]. Available: http://fave.ling.upenn.edu. Accessed on: January 20, 2016.
[16] J. Yuan and M. Liberman. "Speaker identification on the SCOTUS corpus". *Proc. Acoustics '08*. 2008.
[17] J.P. Goldman. , Jean-Philippe. "EasyAlign: an automatic phonetic alignment tool under Praat". *Proc. InterSpeech*, 2011.
[18] S. Young *et al*. *The HTK Book*. Cambridge University Engineering Department. [Online] Available: http://htk.eng.cam.ac.uk/. Accessed on: January 15, 2016.
[19] M. Ernestus and N. Warner. "An introduction to reduced pronunciation variants". *J. of Phonetics*, vol. XXXIX, num. 3, pp. 253-260, 2011. http://dx.doi.org/10.1016/S0095-4470(11)00055-6.
[20] A. Constenla. *Curso básico de bribri*. San José: Editorial de la Universidad de Costa Rica. 1998.
[21] S. Flores Solórzano. "Teclado Chibcha: Un software lingüístico para los sistemas de escritura de las lenguas bribri y cabécar". *Revista de Filología y Lingüística*, vol. XXXVI, num. 2, pp. 155-161, 2010.
[22] F. Biadsy and J. Hirschberg. "Using Prosody and Phonotactics in Arabic Dialect Identification". *Proc. Interspeech*, 2009.
[23] D. Bates, M. Maechler, B. Bolker and S. Walker. "Fitting Linear Mixed-Effects Models Using lme4". *J. Statistical Software*, vol. 67, num. 1, pp. 1-48, 2015. https://doi.org/10.18637/jss.v067.i01.
[24] R Core Team. 2013. "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Viena, Austria. [Online]. Available: http://www.R-project.org/. Accessed on: January 20, 2016.
[25] C. Jara Murillo and A. García Segura. *Se' ẽ' yawö bribri wa Aprendemos la lengua bribri*. San José: Editorial de la Universidad de Costa Rica. Pp. 155-159, 2009.
[26] R. Coto-Solano. "The Phonetics, Phonology and Phonotactics of the Bribri Language". Presented at 2nd Int. Conf. on Mesoamerican Linguistics. Los Angeles: California State University. [Online]. Available: https://www.academia.edu/11365794/The_phonetics_phonology_and_phonotactics_of_the_Bribri_Language. Accessed on: January 25, 2016.
[27] C. Moseley. "Atlas of the World's Languages in Danger". UNESCO, 2010.
[28] C. Sánchez Avendaño, Carlos. "Lenguas en peligro en Costa Rica: Vitalidad, documentación y descripción". *Káñina*, vol. XXXVII, num. 1, pp. 219-250. 2013.
[29] L. Buszard-Welcher. "Can the web help save my language?". *The Green Book of Language Revitalization in Practice*, pp. 331-45, 2001. https://doi.org/10.1163/9789004261723_027.
[30] A. Kornai. "Digital language death". *PLoS ONE*, vol. 8, num 10:e77056, 2013. https://doi.org/10.1371/journal.pone.0077056.
[31] V. Golla. "What does it mean for a language to survive? Some thoughts on the (not-so-simple) future of small languages". Lectures on Endangered Languages 2, Kyoto Conference 2000. Tokyo: ELPR Publication Series C002, pp. 171-177, 2001.
[32] M. Gasser. "Machine Translation and the Future of Indigenous Languages". I Congreso Internacional de las Lenguas y Literaturas Indoamericanas. 2006. [Online] Available: ftp://ftp.cs.indiana.edu/pub/gasser/cilli.pdf. Accessed on: January 16, 2016.
[33] E. Chodroff. "Corpus Phonetics Tutorial: Kaldi Forced Alignment". [Online] Available: http://pages.jh.edu/~echodro1/tutorial/kaldi/kaldi-forcedalignment.html. Accessed on: August 10th, 2016.