Káñina, Rev. Artes y Letras, Univ. Costa Rica XL (Extraordinario): 175-199, 2016 / ISSN:2215-2636

# ALINEACIÓN FORZADA SIN ENTRENAMIENTO PARA LA ANOTACIÓN AUTOMÁTICA DE CORPUS ORALES DE LAS LENGUAS INDÍGENAS DE COSTA RICA

*FORCED ALINGMENT WITHOUT TRAINING FOR AUTOMATIC ANNOTATION OF ORAL CORPORA OF INDIGENOUS LANGUAGES OF COSTA RICA*

*Rolando Coto-Solano*[*]

*Sofía Flores Solórzano*[**]

**RESUMEN**

La alineación forzada provee un ahorro drástico de tiempo al segmentar grabaciones de habla. Esto es particularmente útil para las lenguas indígenas, las cuales carecen de recursos para su estudio desde la lingüística computacional. Este artículo presenta un método para alinear grabaciones en bribri, cabécar y malecu usando modelos acústicos entrenados para inglés y francés. Se usaron los sistemas FAVE-align e EasyAlign para producir TextGrids de Praat, y se obtuvieron errores de 2~3 milisegundos para el centro de las palabras en bribri y malecu (8~13% de la duración de las palabras) y de 7 milisegundos para el cabécar (37% de la duración de las palabras). Los fonemas también tuvieron un desempeño adecuado; para el bribri y el malecu el 40% de los fonemas estaban alineados con un error igual o menor a 1 milisegundo, mientras que esta cifra es de 24% para el cabécar. El desempeño más bajo del cabécar puede deberse a que usó una grabación con más ruido ambiental. Estos sistemas de alineación forzada pueden ayudar al estudio automatizado de las lenguas de Costa Rica mediante la generación de corpus alineados que puedan usarse para estudios fonéticos y para entrenamiento de modelos acústicos y de reconocimiento del habla.
**Palabras clave**: Bribri, cabécar, malecu, alineamiento forzado, fonética.

**ABSTRACT**

Forced alignment provides drastic savings in time when aligning speech recordings. This is particularly useful for Indigenous languages, which lack tagged corpora and resources for their computational study. In this article we present a method for the alignment of Bribri, Cabecar and Malecu recordings using acoustic models trained for English and French. We used the FAVE-align and EasyAlign to produce Praat TextGrids, and obtained error rates of 2~3 milliseconds when marking the center of Bribri and Malecu words (8~13% of average word duration), and of 7 milliseconds when marking Cabécar words (37% of average word duration). Phoneme alignment also showed an adequate performance: An average of 40% of Bribri and Malecu phonemes were aligned with an error of 1 millisecond or less; while 24% of Cabécar phonemes had the same error rate. The lower performance when aligning Cabécar might have been caused by a higher level

* Universidad de Arizona. Phd Student, Linguistics. Estados Unidos.
Correo electrónico: rcoto@email.arizona.edu

** Universidad de Costa Rica, Profesora de la Sede Regional del Atlántico. Costa Rica.
Correo electrónico: sofia.flores@ucr.ac.cr
*Recepción: 15/1/2016.Aceptación: 16/3/2016.*

of environmental noise in the recording used. These forced alignment systems can assist in the study of the Indigenous languages of Costa Rica, in particular in the generation of aligned corpora for phonetic study and for the training of acoustic and speech recognition models.

**Keywords**: Bribri, Cabecar, Malecu, forced alignment, phonetics.

## 1. Introducción

La *alineación forzada* es una familia de algoritmos que toman como entrada un archivo de audio y su transcripción, y calculan los puntos temporales del archivo de audio que corresponden a cada palabra, e incluso a cada fonema, contenido en la transcripción (Wightman y Talkin 1997, Schiel y Draxler 2003). La figura 1 muestra un ejemplo para el inglés, en donde la alineación está representada en un *TextGrid* del programa *Praat* (Boersma 2001). En la primera fila aparecen los intervalos correspondientes a las palabras en la grabación; y en la segunda fila, se observan los intervalos para los fonemas dentro de cada palabra. Estos algoritmos se entrenan usando modelos ocultos de Markov (HMM), los cuales aprenden *ventanas* ("*frames*") que contienen las características auditivas y espectrales de cada fono, por ejemplo la altura de los formantes y la intensidad de la onda. Después del entrenamiento estas ventanas son usadas para recorrer la señal de audio y encontrar los límites potenciales entre dos fonos (Yuan 2013).

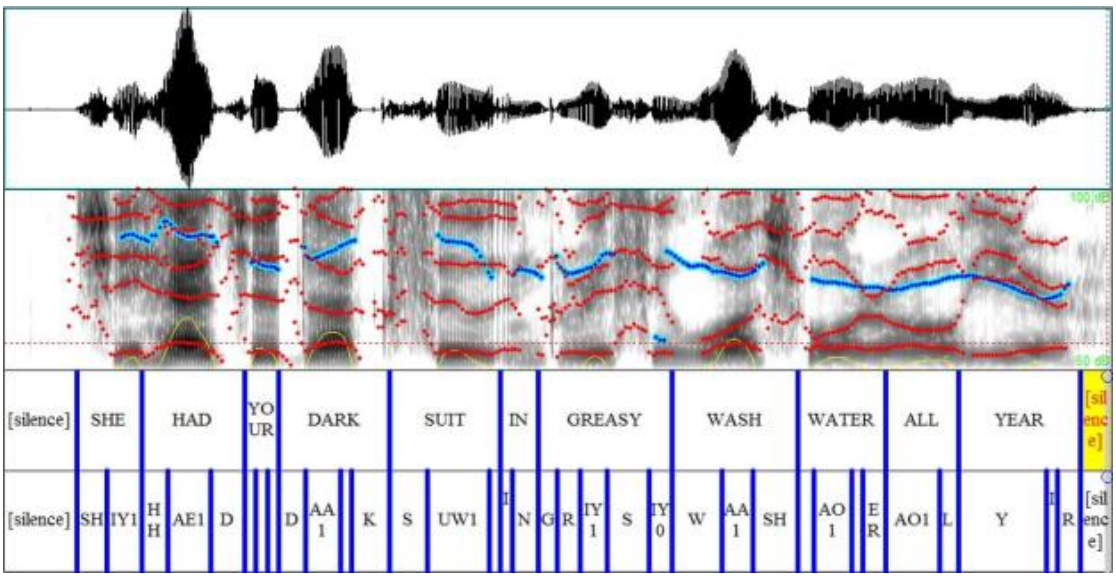


FIGURA 1

Ejemplo de una alineación automática de una grabación de habla en inglés (Yoon 2008).

La segmentación de un corpus oral facilita el uso de los datos en investigaciones lingüísticas, desde la fonética hasta la sociolingüística, y la alineación forzada provee un ahorro drástico de tiempo a la hora de segmentar y anotar grabaciones (Yuan & Liberman 2009, Adda-Decker & Snoeren 2011, Lin et al. 2005). Labov et.al. (2013), por ejemplo, reportan que con 40 horas

de trabajo manual se pueden procesar alrededor de 300 vocales, mientras que usando el sistema de alineación *FAVE-align* se pueden estudiar hasta 9000 vocales durante las mismas 40 horas. Esta familia de algoritmos es particularmente útil para las lenguas minoritarias e indígenas, las cuales carecen de corpus para entrenar modelos acústicos de las mismas, y en general de recursos para su estudio desde la lingüística computacional (DiCanio et al. 2013, Strunk et al. 2014, Brognaux et al. 2012). Esto es posible ya que, a pesar de que se usen modelos acústicos entrenados para lenguas mayoritarias, como el inglés o español, existe un alto nivel de transferencia cuando estos modelos se aplican a otras lenguas (DiCanio et al. 2013). Por ejemplo, se pueden usar modelos acústicos del húngaro para reconocer fonemas en checo debido a las similaridades entre sus sistemas fonológicos (Sim y Li 2008). Muchos de los fonemas de lenguas indígenas como el mixteco de Yoloxóchitl resultan ser suficientemente cercanos a los fonemas del inglés como para obtener resultados satisfactorios (DiCanio et al 2013).

En la sección 2, *Metodología*, exponemos las manipulaciones necesarias para tomar datos del bribri, cabécar y malecu y alinearlos usando modelos acústicos de lenguas mayoritarias como el inglés y el francés. En la sección 3, *Resultados*, comparamos la alineación resultante con la alineación corregida a mano y determinamos que hay un alto grado de transferencia de estos modelos acústicos para las lenguas indígenas de Costa Rica. En las secciones 4 y 5, *Discusión* y *Conclusiones*, comparamos la preparación de los datos requerida por dos sistemas de alineación, FAVE-align e EasyAlign, y sugerimos continuar la alineación forzada para generar corpus que sirvan de insumo para investigaciones fonéticas y para entrenamiento de modelos acústicos para las lenguas indígenas de Costa Rica.

## 2. Metodología

En este trabajo buscamos alinear habla conversacional de lenguas indígenas costarricenses con su respectiva señal sonora y transcripción. Este paso expande la metodología de DiCanio et.al. (2013), quienes usaron sistemas de alineación con modelos acústicos del inglés para procesar listas de palabras separadas en mixteco de Yoloxóchitl (Otomangue, México). Para este trabajo, usamos un modelo acústico del inglés en el sistema de alineación *FAVE-align* (Rosenfelder et al. 2011), un derivado del sistema *Penn Phonetics Lab Forced Aligner Toolkit* o *P2FA* (Yuan & Liberman 2008) entrenado en la Universidad de Pennsylvania. En la sección 3.1.3 explicamos un segundo modelo, un modelo acústico del francés en el sistema de alineación *EasyAlign* (Goldman 2011), entrenado en la Universidad de Ginebra, para comparar el desempeño del modelo acústico del inglés con el de otros modelos. [1,2]

### 2.1. FAVE-align

A continuación detallamos el proceso de preparación de los datos para procesamiento con FAVE-align. Siguiendo la metodología de DiCanio et al. (2013), este es un proceso semi-guiado, en el que el usuario transforma la transcripción ortográfica en una transcripción fonológica, y luego el sistema FAVE-align alinea la transcripción fonológica a la onda de audio.

q. Para comenzar se debe contar con una grabación de calidad en un formato de audio sin comprimir (por ejemplo, en WAV *WaveForm Audio File*), así como la transcripción digitalizada de la grabación.

b. La transcripción digitalizada se guarda en un archivo de texto dividido en secciones, que especifican los límites temporales de cada sección. Este archivo *contiene la lista de frases*. Como se muestra en la tabla 1, la transcripción tiene cinco columnas separadas por tabulaciones: (1) el identificador del sujeto (e.g. *S01*), (2) el nombre del sujeto (e.g. *AG*), (3) el tiempo de inicio de la frase en segundos, con un punto como separador de decimales (e.g. *4.5*), (4) el final de la frase en segundos, con un punto como separador de decimales (e.g. *8.3*), y (5) la frase en cuestión, en la transcripción que el usuario desee usar [3].

TABLA 1

Lista de frases, con los tiempos de inicio y fin de cada frase en la transcripción

| | | | | |
|---|---|---|---|---|
| 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 |

c. Creamos un segundo archivo de texto, donde esté almacenada la lista de palabras únicas que se usan en la transcripción. Este archivo será el *diccionario*. Como se muestra en la tabla 2, el diccionario tiene dos columnas separadas por tabulaciones: (1) la transcripción ortográfica de las palabras, y (2) la transliteración de las palabras en el sistema Arpabet, con los fonos separados por espacios. (La sección 2.2 provee detalles para la conversión entre la lengua indígena y Arpabet).

TABLA 2

Diccionario con las palabras transliteradas en el sistema Arpabet

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

d. Hay que crear un tercer archivo de texto, idéntico al diccionario excepto porque tiene una columna adicional. Este tercer archivo será el *diccionario expandido*. En la tercera columna está la transcripción ortográfica pero cada letra está separada por un espacio. Adicionalmente, cada sílaba está separada por un punto. Esto se usará en el paso *F* para cambiar el Arpabet del TextGrid y devolver los intervalos a la transcripción elegida, y en el paso *G* para generar la división por sílabas.

TABLA 3

Diccionario expandido con transcripción Arpabet y transcripción del usuario

| | | |
|---|---|---|
| be' | B EH1 | b e' |
| bë`rie | B IH1 R Y EH0 | b ë` . r i e |
| bikéitse | B IY0 K EH1 EH0 T S EH0 | b i . k é i . t s e |
| bö`k | B UH1 K | b ö` k |

e. La lista de frases y el diccionario constituyen la entrada del sistema FAVE-align. Este se puede descargar como código Python [enlace] o se puede ejecutar mediante una interfaz en línea [enlace]. La salida es un archivo TextGrid de Praat con dos filas: (1) en la primera, los fonemas en Arpabet aparecen alineados y (2) en la segunda fila, las palabras en la transcripción elegida por el usuario quedan alineadas. Un ejemplo de la salida se muestra en la figura 2.

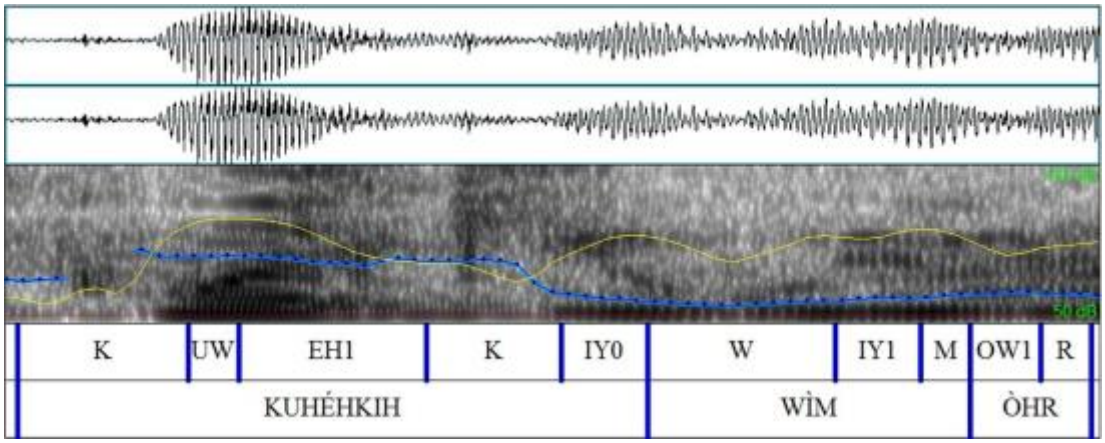


FIGURA 2

TextGrid bribri con fonemas en Arpabet.

f. El TextGrid resultante del paso anterior puede convertirse para que la transcripción esté, por ejemplo, en el sistema ortográfico y no en Arpabet. Para esto, hay que utilizar el código de C# programado por los autores, disponible en GitHub [http://github.com/rolandocoto/alineacion-lenguas-cr]. Este código toma como entrada el TextGrid y el diccionario expandido, y genera un nuevo TextGrid con la transcripción de los fonemas en el sistema ortográfico. La figura 3 es un ejemplo de esta salida.

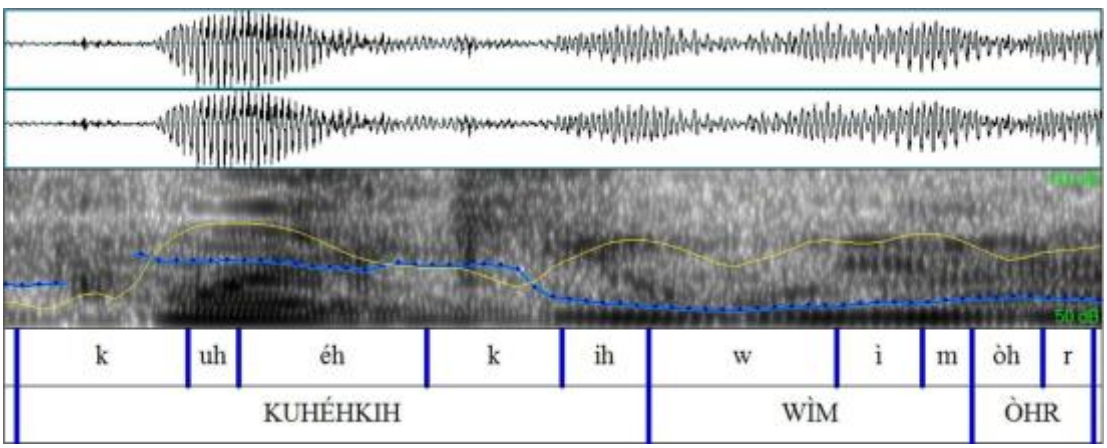


FIGURA 3

TextGrid bribri con fonemas en la transcripción del usuario.

g. Finalmente, el código de los autores permite añadir una fila adicional al TextGrid, en donde aparezcan las palabras separadas por sílabas. La entrada es el TextGrid y el diccionario expandido y la salida es un nuevo TextGrid con tres filas: (1) los fonemas, (2) las sílabas y (3) las palabras en la grabación.

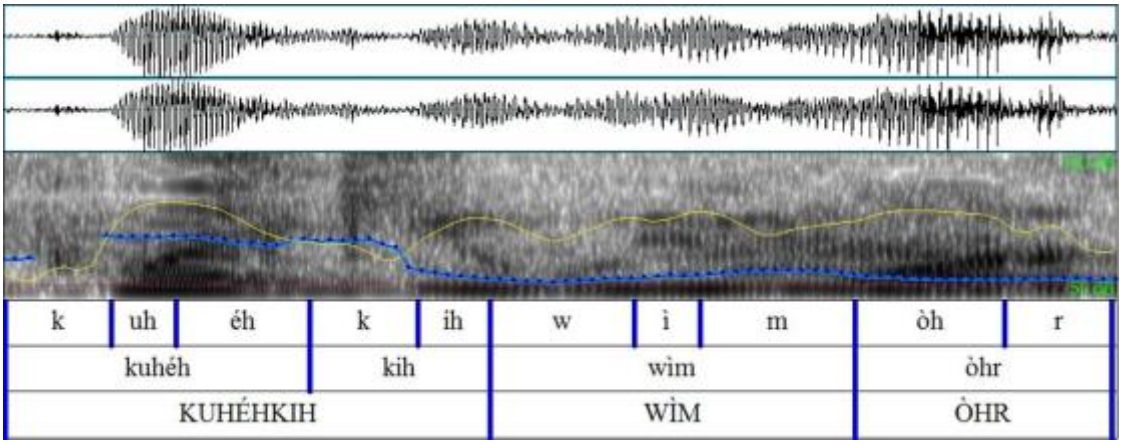


FIGURA 4

TextGrid bribri con fonemas y sílabas.

Al final de este proceso, el usuario dispone de un TextGrid que contiene una alineación a nivel de palabra, sílaba y fonema.

## 2.2. Arpabet

El sistema de alineación FAVE-align necesita que las palabras estén convertidas al sistema de transcripción *Arpabet*, como se muestra en el paso *C* de la sección 2.1. Esta transcripción, creada originalmente para el inglés, reemplaza las formas ortográficas por formas fonológicas. Para este artículo, hay que convertir la ortografía de las lenguas indígenas a este sistema, y tomar decisiones de transcripción para los fonemas de los que carezca el inglés.

La tabla 4 muestra la transcripción para las vocales. (Los valores en alfabeto fonético internacional fueron tomados de Constenla (1998) y Flores Solórzano (2010) para el bribri, González Campos (2011) para el cabécar y Constenla (1999) para el malecu). El sistema propuesto carece de representación de tonos en bribri y cabécar; la marcación del tono para cada sílaba se recupera cuando la transcripción Arpabet se reemplaza usando el diccionario expandido (ver sección 2.1 arriba). El sistema carece además de representaciones para la nasalización en bribri y cabécar y para la longitud vocálica en malecu, pero el colapsar estas categorías no degrada la calidad del alineamiento; esto se discutirá en detalle en la sección 3. Finalmente, en los tres idiomas los grafemas <*i*> y <*u*> pueden representar tanto las vocales /i/ y /u/ como las semiconsonantes /j/ y /w/; esto se trata de forma acorde en la transcripción Arpabet.

Además, de la vocal, cada representación Arpabet tiene un número. Este representa el sistema acentual del inglés, y puede tener tres valores, 1 para sílabas acentuadas, 2 para sílabas para con acento secundario (*secondary stress*), y 0 para sílabas no acentuadas. Para estas pruebas, todas las vocales se señalaron como *1* [4].

TABLA 4

Vocales y su transcripción en Arpabet

| Bribri | Arpabet | | Cabécar | Arpabet | | Malecu | Arpabet |
|---|---|---|---|---|---|---|---|
| a /a/ | AE1 | | a /a/ | AE1 | | a /a/ | AE1 |
| e /ɛ/ | EH1 | | ä /ɤ/ | AH1 | | e /e/ | EH1 |
| ë /ɪ/ | IH1 | | e /ɛ/ | EH1 | | i /i/ | IY1 ~ Y |
| i /i/ | IY1 ~ Y | | ë /ɪ/ | IH1 | | o /o/ | OW1 |
| o /ɔ/ | OW1 | | i /i/ | IY1 | | u /u/ | UW1 ~ W |
| ö /ʊ/ | UH1 | | o /ɔ/ | OW1 | | | |
| u /u/ | UW1 ~ W | | ö /ʊ/ | UH1 | | | |
| | | | u /u/ | UW1 | | | |

La tabla 5 muestra la transcripción propuesta para las consonantes. Para el bribri, se tomaron tres decisiones importantes: (1) usar el glifo Arpabet 'N' para representar dos fonemas diferentes, *n* /n/ y ñ /ɲ/, (2) usar el glifo Arpabet 'R' para representar dos fonemas diferentes, *l* /ɭ/ y *r* /ɾ/, y (3) dividir el fonema *ts* /ʦ/ en dos glifos Arpabet, 'T S'. Además, se tomó la decisión de no representar explícitamente al alto glotal como un fonema separado, y se le trató como un tono parte de la vocal; esto no resultó en una degradación de la exactitud del sistema (véase la sección 3.1.1) [5]. Para el malecu, se decidió usar el glifo 'R' para representar dos fonemas, *r* /ɾ/ y *rr* /r/, usar el glifo 'F' para representar al fonema *f* /ɸ/, y usar el glifo 'SH' para representar al fonema *lh* /ɬ/. Finalmente, para el cabécar se usó el glifo Arpabet 'N' para el fonema /ɲ/ y se dividió el fonema /tk/ en dos dos Arpabet, 'T K'.

TABLA 5

Consonantes y su transcripción en Arpabet

| Bribri | Arpabet | Cabécar | Arpabet | Malecu | Arpabet |
|---|---|---|---|---|---|
| b /b/ | B | b /b/ | B | c /k/ | K |
| ch /ʧ/ | CH | ch /ʧ/ | CH | ch /ʧ/ | CH |
| d /d/ | D | d /d/ | D | f /ɸ/ | F |
| j /x/ | HH | j /h/ | HH | j /x/ | HH |
| k /k/ ~ [kʰ] | K | k /k/ | K | l /l/ | L |
| l /ɭ/ | R | l /ɭ/ | R | lh /ɬ/ | SH |
| m /m/ | M | m /m/ | M | m /m/ | M |
| n /n/ | N | n /n/ | N | n /n/ | N |
| ñ /ɲ/ | N | ñ /ɲ/ | N | nh /ŋ/ | NG |
| p /p/ | P | p /p/ | P | p /p/ | P |
| r /ɾ/ | R | r /ɾ/ | R | qu /k/ | K |
| rr /r/ | R | s /s/ | S | r /ɾ/ | R |
| s /s/ | S | sh /ʃ/ | SH | rr /r/ | R |
| sh /ʃ/ | SH | t /t/ | T | s /s/ | S |
| t /t/ | T | tk /tk/ | T K | t /t/ | T |
| ts /ʦ/ | T S | w /w/ | W | y /ʤ/ | JH |
| w /w/ | W | y /ʤ/ | JH | | |
| y /ʤ/ | JH | | | | |

### 2.3. EasyAlign

Como se mencionó al inicio de la sección 2, EasyAlign (Goldman 2011) es un sistema de alineación entrenado en la Universidad de Ginebra, que cuenta con modelos acústicos del francés, inglés, español y portugués. Este sistema se instala como un complemento de *Praat*, y hace posible segmentar automáticamente un texto para el cual se provea un archivo de audio y una transcripción ortográfica. Los desarrolladores del sistema invitan a los interesados a enviar datos para entrenar nuevos modelos acústicos, pero no brindan la metodología de forma abierta.[6]

A continuación detallamos los pasos para configurar un texto y procesarlo en EasyAlign:

a. Al igual que se hizo para FAVE-align, para comenzar, hay que tener una grabación de calidad en un formato de audio sin comprimir, por ejemplo, en WAV *WaveForm Audio File*, así como una la transcripción ortográfica digitalizada de la grabación.

b. Se alinea manualmente la transcripción a nivel de frases. Esto se hace creando un TextGrid de Praat y creando una fila (*tier*) con nombre *NAT [ortho]*.

c. Se copian los intervalos en la fila *NAT [ortho]* a una fila de nombre *ortho*. En esta segunda fila se eliminan todas las marcas diacríticas del sistema de transcripción ortográfico.

d. La alineación automática forzada se realizará en la fila *ortho* del TextGrid. Se ejecuta el script de *fonetización* del EasyAlign desde la interfaz de Praat, indicando en la configuración que *ortho* es la fila que se desea alinear, e indicando el modelo acústico que se desea usar (para este experimento usaremos el francés). El resultado es un nuevo TextGrid intermedio con una fila que contiene las palabras en el sistema de transcripción fonética SAMPA.

e. Finalmente, se selecciona el TextGrid intermedio y el archivo de audio y se ejecuta el script de *segmentación fonética*. El resultado de esta operación es un TextGrid con cinco filas: *ortho* para la frase ortográfica, *phono* para la frase transcrita en SAMPA, *words* para las palabras en SAMPA, *syll* para las sílabas transcritas en SAMPA y *phones* para los fonemas en SAMPA. La figura 5 muestra un ejemplo del TextGrid resultante.

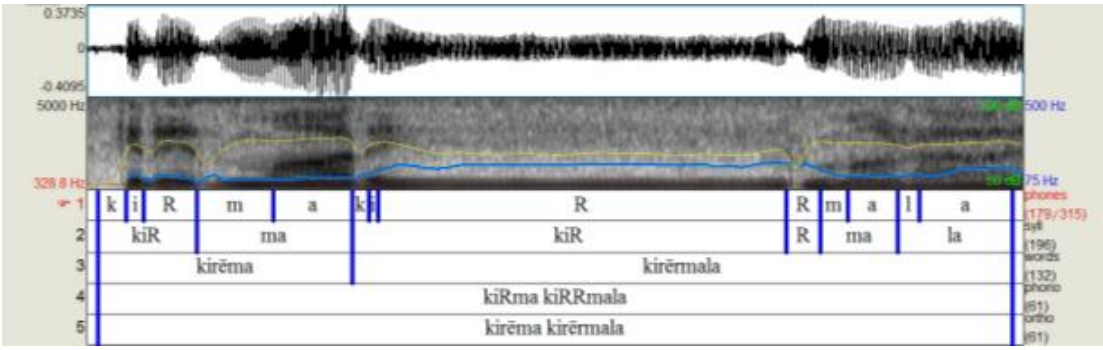


FIGURA 5

TextGrid bribri alineado con EasyAlign.

## 2.4. Cálculo del error

Después de generar los TextGrids automáticos tanto en FAVE-align como en EasyAlign, estos fueron corregidos manualmente para determinar la exactitud de los algoritmos de alineación automática. Se usó el algoritmo de *promedio de raíz de cuadrados* ("*root mean squares*") (Biadsy y Hirschberg 2009) para calcular la diferencia en el centro de las palabras entre el TextGrid automático y el TextGrid corregido a mano. Como ejemplo vamos a tomar la palabra òhr (õ̀r 'grita') en la figura 6. El centro de la palabra marcada automáticamente está en 1,887 segundos, mientras que el centro de la palabra corregida está en 2,039 segundos. Entonces, la diferencia en el centro es de $\sqrt{(2{,}039 - 1{,}887)^2} = 152$ milisegundos. (Nótese que se usa el cuadrado para obtener el valor absoluto de las diferencias negativas, en caso de que un intervalo corregido resulte estar antes del intervalo manual). Para tener mayor claridad sobre la relación entre la diferencia y la palabra en cuestión, se calculó además el porcentaje del tamaño de la palabra corregida que representa la diferencia entre los centros. Por ejemplo, la palabra corregida òhr mide 196 milisegundos, por lo que la diferencia es 152/196 = 78% del tamaño de la palabra òhr.

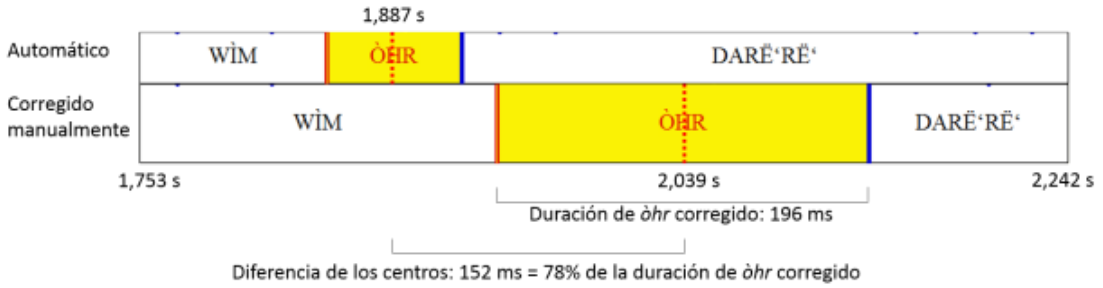


FIGURA 6

Cálculo del error en la alineación del centro de una palabra.

El método de promedio de raíz de cuadrados también se usó para determinar la diferencia entre el inicio del fonema en el TextGrid automático y el inicio del mismo en el TextGrid corregido. Como se muestra en la figura 7, la diferencia para el inicio del fonema *uh* (ũ) es de milisegundos. De la misma forma, este método también se aplicó para calcular la diferencia al final de cada fonema.

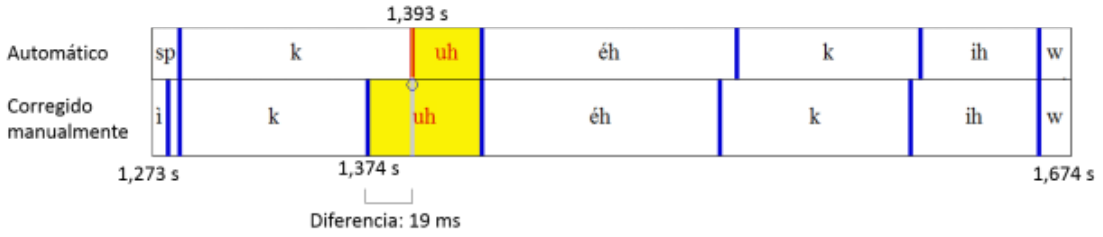


FIGURA 7

Cálculo del error en la alineación del inicio de un fonema.

### 2.5. Métodos estadísticos

Para determinar si hay diferencias significativas entre los datos se usó el modelo de *análisis de la varianza de Fischer* (ANOVA) de efectos fijos. Los modelos se calcularon usando el programa R (R Core Team 2013).

### 2.6. Materiales

Para realizar los experimentos se usaron varias grabaciones en bribri y en malecu. Para el bribri se usaron cuatro minutos de la narración Ì kũẽ́kĩ wìm õ̀r darëẽ̀rẽ̀ 'Porqué el mono congo grita tan fuerte' (Jara & García 2009), y un minuto y siete segundos de la canción inédita de Natalia Gabb Ísela i-yӧ́ sӧla i-yӧ́ "Pobrecitos nosotros, bebamos, bebamos". Para el cabécar se utilizaron dos minutos de la narración inédita *Yẽ́ bulẽ́* 'Cazadorcete' (González Campos, comunicación personal). Para el malecu se utilizaron tres minutos de la narración *Muérra o curíjuríyu óje* 'El ogro que se robó a una mujer' (Constenla 2015). La tabla 6 detalla el número de palabras, vocales y consonantes que se alinearon para cada grabación, así como el promedio de sílabas por segundo para tener una comparación de la velocidad del habla en cada grabación.

TABLA 6

Número de ítemes extraídos de las grabaciones

| Grabación | Palabras | Vocales | Consonantes | Sílabas/segundo |
|---|---|---|---|---|
| Narración bribri ("mono") | 119 | 196 | 153 | 5,8 |
| Canción bribri ("pobrecitos") | 76 | 141 | 114 | 2,7 |
| Narración cabécar ("cazadorcete") | 159 | 287 | 265 | 6,7 |
| Narración malecu ("ogro") | 98 | 231 | 192 | 7,0 |

## 3. Resultados

Como se ve en la tabla 7, el sistema FAVE-align tuvo un error promedio de entre 3,8 milisegundos al marcar el centro de las palabras, lo que representa entre un 19% de la duración promedio de las palabras. Sin embargo, hay una diferencia significativa entre los tres idiomas ($F(2,373)=22.2$, $p<0.000001$): El bribri y el malecu tienen una diferencia promedio de 2,2 milisegundos en el centro, mientras que el cabécar tiene una diferencia promedio de 7,1 ± 0,8 milisegundos. Nótese que la velocidad del habla no es la causante de este efecto, ya que tanto la grabación malecu como la cabécar tienen velocidades promedio de aproximadamente 7 sílabas por segundo. La diferencia puede deberse a la diferente calidad de las grabaciones; tanto la grabación bribri como la malecu están enfocadas en un solo hablante, mientras que la grabación cabécar está hecha en un espacio abierto, con un hablante que interactúa con otras personas y que incluso se ríe al contar la historia.

TABLA 7

Diferencia del centro de las palabras entre los intervalos automáticos y corregidos a mano, en milisegundos y en porcentaje de duración de la palabra

| | Bribri | Cabécar | Malecu | ANOVA |
|---|---|---|---|---|
| Diferencia (ms) [7] | 1,9 ± 0,2 | 7,1 ± 0,8 | 2,6 ± 0,6 | $F(2,373)=22.2$, $p<0.000001$ |
| % Diferencia | 8% ± 2% | 37 ± 6% | 13% ± 5% | |

### 3.1. Bribri

A continuación, se analizarán los resultados para los fonemas del bribri. La sección está dividida en tres partes: los resultados para fonemas del FAVE-align, la alineación del canto comparada con la de la narración, y la narración de FAVE-align comparada con la de EasyAlign.

#### *3.1.1. Bribri hablado alineado con FAVE-align*

La figura 8 muestra los resultados de la alineación para el inicio y el final de las vocales. Para el inicio de los intervalos, el 42% de las vocales tiene una diferencia de 1 milisegundo o menos entre el intervalo automático y el intervalo corregido, y el 80% de las vocales tiene una diferencia de 30 milisegundos o menos entre las versiones corregida y automática (la duración promedio de las vocales bribri es de 11 milisegundos). Para el final de los intervalos, el 43% de las vocales tienen una diferencia de 1 milisegundo o menos entre el intervalo automático y el intervalo corregido y el 80% de las vocales tienen una diferencia de 45 ms o menos entre las versiones automática y corregida.

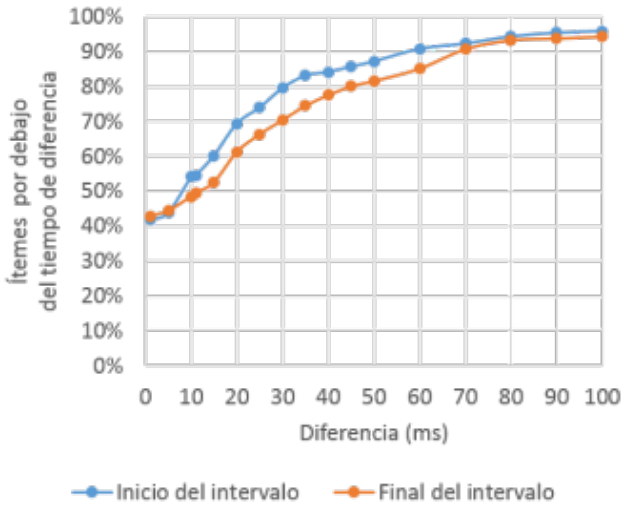


FIGURA 8

Diferencia entre alineación automática y corrección manual para vocales bribri.

La figura 9 muestra los resultados de la alineación para el inicio y el final de las consonantes. Para el punto inicial, el 52% de las consonantes tienen una diferencia de 1 milisegundo o menos, y 80% de las consonantes tiene una diferencia de 32 milisegundos o menos. Para el punto final del intervalo, el 46% de las consonantes tienen una diferencia de 1 milisegundo o menos, mientras que el 80% de las consonantes tienen una diferencia de 25 milisegundos o menos (las consonantes bribri miden en promedio 8 milisegundos).

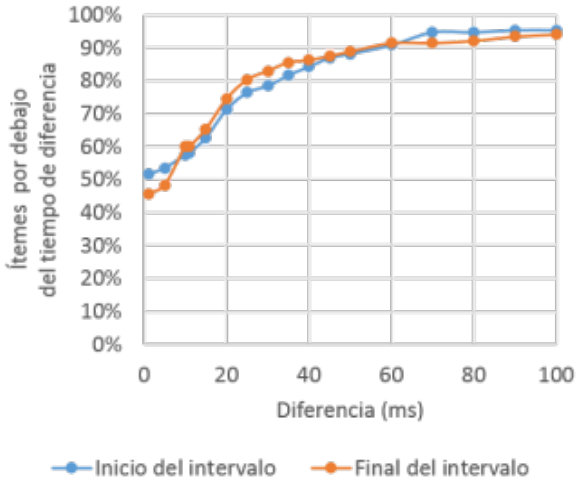


FIGURA 9

Diferencia entre alineación automática y corrección manual para consonantes bribri.

Para verificar la validez de la transcripción Arpabet, se comparó el error entre las vocales orales y nasales. Para el inicio de los intervalos las vocales orales tienen un promedio de error de 2,2 ± 0,4 milisegundos, versus 2,2 ± 0,5 milisegundos para las vocales nasales, lo cual no representa una diferencia significativa ($F(1,194)=0$, $p=0,99$). De igual forma, para el final de los intervalos las vocales orales tienen un promedio de error de 2,7 ± 0,4 milisegundos, versus 2,6 ± 0,5 milisegundos para las vocales nasales, lo que tampoco constituye una diferencia estadísticamente significativa ($F(1,194)=0.02$, $p=0,88$).

Al igual que se hizo con la nasalidad, se comparó a las vocales según su tono para comprobar que esto no causa una diferencia a la hora de transliterar los datos en Arpabet. La tabla 8 detalla los promedios de error para los diferentes tonos. Aunque, los tonos de contorno (el ascendente y el descendente) tienen tasas de error menores, los intervalos vocálicos no presentan diferencias significativas de acuerdo con el tono ni para el inicio del intervalo ($F(3,192)=1.9$, $p=0,13$) ni para el final ($F(3,192)=1.4$, $p=0,25$).

TABLA 8

Promedio de error para tonos bribri

| Tono | Promedio de error Inicio del intervalo (ms) | Final del intervalo (ms) |
|---|---|---|
| Alto | 3,2 ± 1,0 | 3,1 ± 1,0 |
| Bajo/Neutro [8] | 2,2 ± 0,4 | 3,0 ± 0,4 |
| Ascendente (glotal) | 2,1 ± 0,3 | 1,3 ± 0,3 |
| Descendente | 0,7 ± 0,2 | 1,9 ± 0,6 |
| ANOVA | F(3,192)=1.9, p=0.13 | F(3,192)=1.4, p=0.25 |

Para concluir, el estudio de las vocales se analizó el error dependiendo de la vocal específica, pero el sistema no presenta diferencias significativas de alineación ni al inicio ($F(4,191)=1,0$; $p=0,4$) ni al final ($F(4,191)=0,8$; $p=0.48$) del intervalo. Tampoco muestra diferencias significativas dependiendo de la glotalización de la vocal, ni al inicio ($F(1,194)=0.04$; $p=0,84$) ni al final ($F(1,194)=2.9$; $p=0,09$).

Con respecto a las consonantes, FAVE-align no produjo diferencias significativas dependiendo del punto de articulación de la consonante (inicio: $F(4,148)=1,0$; $p=0,43$; final: $F(4,148)=0,43$; $p=0.79$). Sin embargo, sí se detectaron diferencias significativas dependiendo del modo de articulación al inicio del intervalo ($F(5,147)=4,9$, $p<0,0005$). (No hay efectos significativos para el error del final del intervalo dependiendo del modo de articulación: $F(5,147)=1,7$, $p=0.14$). La tabla 9 muestra los promedios para diferentes tipos de consonantes, y se puede observar que las consonantes líquidas {/ɾ/, /r/, /l/} tienen un error de aproximadamente 6,7 milisegundos, mucho mayor que el promedio de 1,3 milisegundos para los otros modos. Esto puede deberse a que ninguno de estos fonemas existe en inglés, por lo que el equivalente en Arpabet no provee una conversión tan buena como para otros fonemas.

TABLA 9

Promedio de error para consonantes bribri por modo de articulación

| Modo de articulación | Promedio de error: inicio (ms) |
|---|---|
| Oclusiva | 1,7 ± 0,4 |
| Fricativa | 1,0 ± 0,3 |
| Africada | 1,0 ± 0,4 |
| Nasal | 1,6 ± 0,5 |
| Líquida | 6,7 ± 2,2 |
| Aproximante | 1,2 ± 0,6 |
| ANOVA | $F(5,147)=4,9$, $p<0,0005$ |

Las consonantes también presentan diferencias significativas al inicio de sus intervalos dependiendo de la sonoridad ($F(1,151)=4,5$; $p<0,04$). (No hay efecto significativo para el final del intervalo: $F(1,151)=1,6$; $p=0,20$). Como se muestra en la tabla 10, las consonantes sonoras presentan más del doble de error que las consonantes sordas (3,0 versus 1,4) al inicio de la marcación. Este problema puede estar relacionado a las diferencias en la sonorización de las consonantes entre el inglés y el bribri. En inglés la diferencia entre una consonante sonora y sorda no se marca con presonorización como en español, sino con aspiración (Lisker y Abramson 1964, Cho y Ladefoged 2000), o con la falta de aspiración en el caso de las consonantes sordas. Así, una /t/ se producirá como [$t^h$] y una /d/ como [t]. El bribri, por el contrario, podría comportarse como el español en cuanto a las pistas acústicas para la sonoridad consonántica, lo que haría que FAVE-align ignore el inicio de las consonantes sonoras, donde la presonorización toma lugar.

TABLA 10

Promedio de error para consonantes bribri por sonoridad

| Sonoridad de consonantes | Promedio de error: inicio (ms) |
|---|---|
| Sorda | 1,4 ± 0,3 |
| Sonora | 3,0 ± 0,7 |
| ANOVA | $F(1,151)=4,5$; $p<0,04$ |

Finalmente, se investigó la exactitud del sistema dependiendo del tipo de contacto entre fonemas. El contacto a la izquierda del fonema no resultó en resultados significativos ($F(5,343)=0.6$, $p=0,67$), pero el contacto a la derecha sí tuvo diferencias significativas ($F(5,343)=3.0$, $p<0,02$). Como se muestra en la tabla 11, la interacción entre un fonema y un silencio (#) resultó en una tasa de error mayor que otros tipos de contactos; 4 milisegundos para vocal y silencio (C#) y 11 milisegundos para consonante y silencio (C#), comparado con un promedio de 2,2 milisegundos para otros tipos de contactos. En el caso de las consonantes esto puede deberse a la aspiración final en consonantes bribris como /k/, que parecen tener un alófono [$k^h$] cuando están al final de la palabra. Esta aspiración con frecuencia era ignorada en la alineación forzada, aumentando el error. En el

caso de las vocales, el sistema parecía alinear de forma consistentemente conservadora, cubriendo el inicio más estable de la sílaba pero ignorando el final, en zonas con menor sonoridad.

TABLA 11

Promedio de error por tipo de contacto a la derecha del fonema

| Tipo de contacto a la derecha del fonema | Promedio de error: final (ms) |
|---|---|
| C# | 10,9 ± 6,6 |
| CC | 2,0 ± 0,9 |
| CV | 2,1 ± 0,4 |
| V# | 4,2 ± 1,0 |
| VC | 2,4 ± 0,5 |
| VV | 2,6 ± 0,5 |
| ANOVA | F(5,343)=3,0; p<0,02 |

En resumen, el sistema FAVE-align muestra un desempeño aceptable al alinear bribri hablado. El 42% de las vocales y el 52% de las consonantes tienen tasas de error de 1 milisegundo o menos, y los tipos de segmentos para los que sí hay diferencias de desempeño (e.g. consonantes líquidas, consonantes sonoras, segmentos seguidos de silencio) tienen tasas de error mayores pero todavía comparables a las de otros segmentos. En la siguiente subsección analizaremos el desempeño del FAVE-align al tratar de alinear una canción en la lengua bribri.

#### *3.1.2. Bribri cantado alineado con FAVE-align*

Comparamos el desempeño entre palabra hablada y canción, comparando el relato bribri con la canción Ísela i-yö söla 'Pobrecitos nosotros, bebamos'. FAVE-align tuvo un desempeño significativamente mejor al encontrar el centro de las palabras habladas, con una diferencia de 1,8 ± 0,2 milisegundos (8%±2% de la duración de las palabras), comparado con 19 ± 3,3 milisegundos (65%±28%) para las palabras cantadas ($F(1,193)=12,8$; $p<0,00001$).

#### *3.1.3. Comparación para el bribri entre FAVE-align e EasyAlign*

Debido a que la canción Ísela i-yö söla fue la grabación con el desempeño relativamente más bajo en FAVE-align, la usamos para comparar el desempeño de este sistema contra el EasyAlign. Sin embargo, no encontramos diferencias significativas en la exactitud de los dos sistemas: El promedio de error es 19 ± 3 milisegundos para FAVE-align y 20 ± 3 milisegundos para EasyAlign ($F(1,154)=0,6$;$p=0,4$).

### 3.2. Cabécar

La figura 10 contiene un ejemplo de cabécar con la alineación forzada generada por FAVE-align.

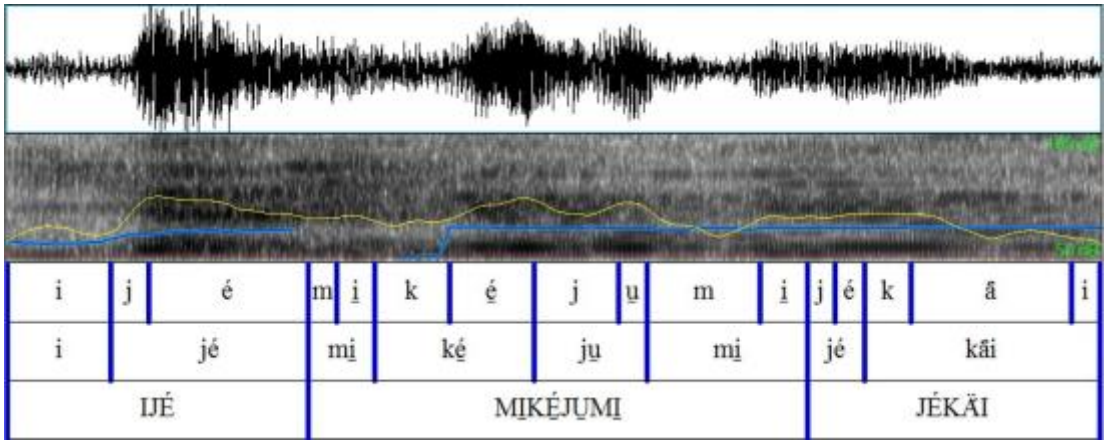


FIGURA 10

TextGrid cabécar alineado con FAVE-align.

La figura 11 muestra los resultados de la alineación para el inicio y el final de las vocales en cabécar. Para el inicio de los intervalos, el 24% de las vocales tiene una diferencia de 1 milisegundo o menos entre el intervalo automático y el intervalo corregido, y el 80% de las vocales tiene una diferencia de 110 milisegundos o menos entre las versiones corregida y automática (la duración promedio de las vocales cabécar es de 10 milisegundos). Para el final de los intervalos, el 20% de las vocales tienen una diferencia de 1 milisegundo o menos entre el intervalo automático y el intervalo corregido, y el 80% de las vocales tienen una diferencia de 140 ms o menos entre las versiones automática y corregida.

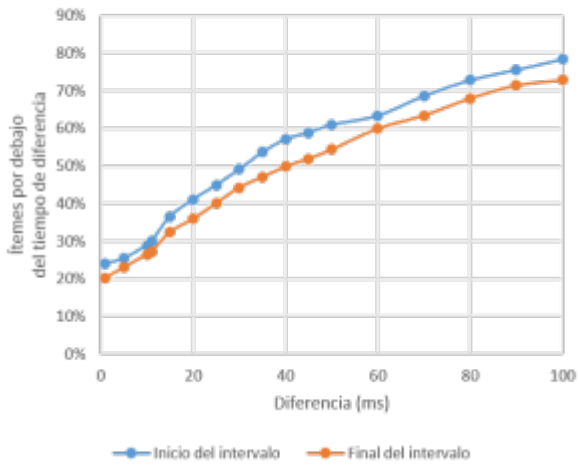


FIGURA 11

Diferencia entre alineación automática y corrección manual para vocales cabécar.

La figura 12 muestra los resultados de la alineación para el inicio y el final de las consonantes cabécar. Para el punto inicial, el 25% de las consonantes tienen una diferencia de 1 milisegundo o menos, y 80% de las consonantes tiene una diferencia de 120 milisegundos o menos. Para el punto final del intervalo, el 25% de las consonantes tienen una diferencia de 1 milisegundo o menos, mientras que el 80% de las consonantes tienen una diferencia de 110 milisegundos o menos. Cabe mencionar que las consonantes cabécar miden en promedio 6 milisegundos.

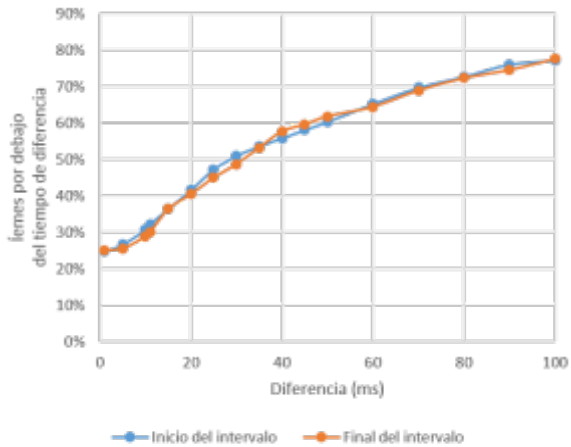


FIGURA 12

Diferencia entre alineación automática y corrección manual para consonantes cabécar.

Las interacciones significativas observadas en bribri entre el error y características fonéticas de las vocales y consonantes no están presentes en cabécar. No hay diferencias significativas de acuerdo con la vocal específica (inicio: F(4,282)=0,01; p=0,99, final: F(4,282)=0,3; p=0,88), al tono de la vocal (inicio: F(1,285)=0,3; p=0,59, final: F(1,285)=0,05, p=0,84) o a su nasalidad (inicio: F(1,285)=0,1; p=0,74; final: F(1,285)=0,05; p=0,83). Tampoco, hay diferencias en cuanto al modo de articulación de la consonante (F(5,259)=0,5; p=0,78, final: F(5,259)=0,4; p=0,81), a su punto de articulación (inicio: F(4,260)=0,5; p=0,73, final: F(4,260)=0,8; p=0,50) o a su sonoridad (inicio: F(1,263)=0,3; p=0,56, final: F(1,263)=0,4; p=0,51). Finalmente, no hay diferencias significativas dependiendo del contacto con otros fonemas (inicio: F(5,343)=0,4; p=0,88, final: F(5,343)=1,4; p=0,21).

### 3.3. Malecu

La figura 13 contiene un ejemplo de malecu con la alineación forzada generada por FAVE-align. Nótese que la grabación carece de frecuencias mayores a los 4000 hertz.

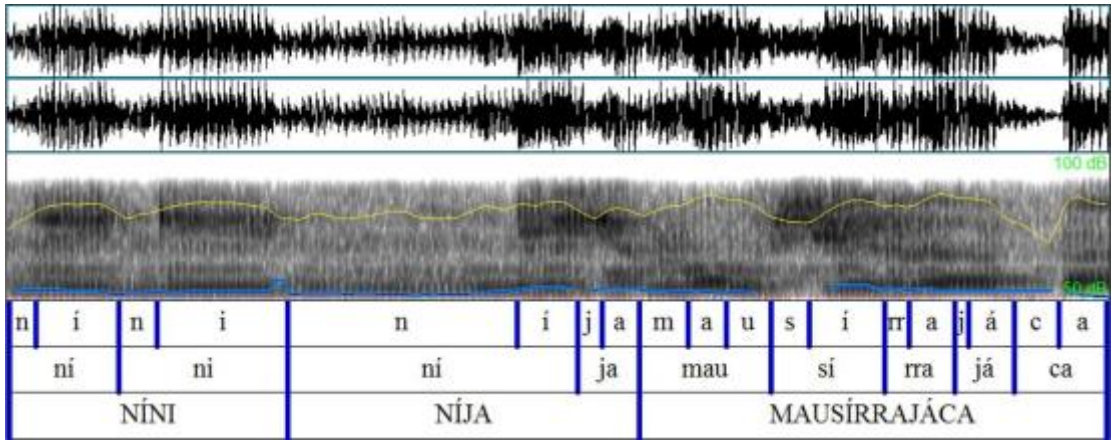


FIGURA 13

TextGrid malecu alineado con FAVE-align.

La figura 14 muestra los resultados de la alineación para el inicio y el final de las vocales en malecu. Para el inicio de los intervalos, el 37% de las vocales tiene una diferencia de 1 milisegundo o menos entre el intervalo automático y el intervalo corregido, y el 80% de las vocales tiene una diferencia de 40 milisegundos o menos entre las versiones corregida y automática. Cabe destacar que la duración promedio de las vocales malecu es de 10 milisegundos. Para el final de los intervalos, el 13% de las vocales tienen una diferencia de 1 milisegundo o menos entre el intervalo automático y el intervalo corregido, y el 80% de las vocales tiene una diferencia de 120 milisegundos o menos entre las versiones corregida y automática.

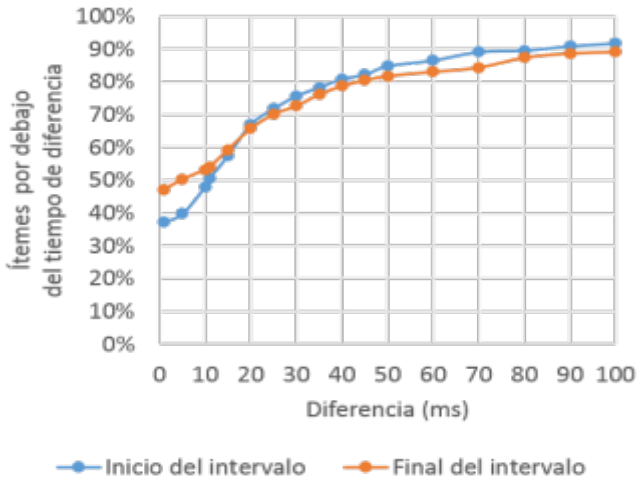


FIGURA 14

Diferencia entre alineación automática y corrección manual para vocales malecu.

La figura 15 muestra los resultados de la alineación para el inicio y el final de las consonantes. Para el punto inicial, el 53% de las consonantes tienen una diferencia de 1 milisegundo o menos, y 80% de las consonantes tiene una diferencia de 45 milisegundos o menos. Para el punto final del intervalo, el 35% de las consonantes tienen una diferencia de 1 milisegundo o menos, mientras que el 80% de las consonantes tienen una diferencia de 45 milisegundos o menos (este desempeño no es tan bueno como el de las vocales, porque las consonantes malecu miden en promedio 6 milisegundos).

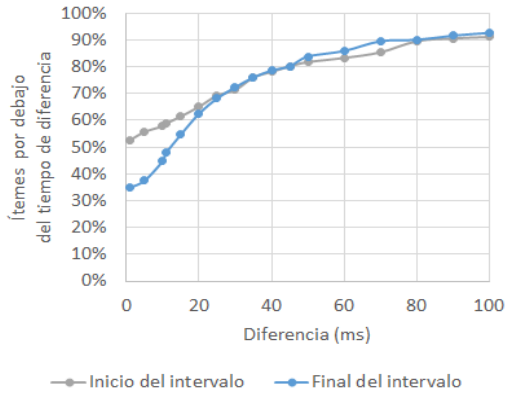


FIGURA 15

Diferencia entre alineación automática y corrección manual para consonantes malecu.

Para verificar la validez de la transcripción Arpabet, se comparó el error entre las vocales largas y cortas. Para el inicio de los intervalos las vocales largas tienen un promedio de error de 2,6 ± 0,4 milisegundos, versus 3,4 ± 0,6 milisegundos para las vocales cortas, lo cual no representa una diferencia significativa ($F(1,229)=1,1$; $p=0,29$). De igual forma, para el final de los intervalos las vocales largas tienen un promedio de error de 3,1 ± 0,6 milisegundos, versus 3,6 ± 0,6 milisegundos para las vocales orales, lo que tampoco constituye una diferencia significativa ($F(1,229)=0.3$, $p=0.58$).

Las interacciones significativas observadas en bribri entre el error y características fonéticas de las vocales y consonantes tampoco están presentes en malecu. No hay diferencias significativas de acuerdo con la vocal (inicio: $F(4,226)=1.2$, $p=0.30$, final: $F(4,226)=1.1$, $p=0.36$), al modo de articulación de la consonante (inicio: $F(4,187)=1,5$, $p=0.21$, final: $F(4,187)=2,0$, $p=0.10$), a su punto de articulación (inicio: $F(3,188)=0,94$, $p=0.42$, final: $F(3,188)=0.93$, $p=0.43$) o a su sonoridad (inicio: $F(1,190)=0.16$, $p=0.69$, final: $F(1,190)=0.02$, $p=0.89$). Tampoco hay diferencias significativas dependiendo del contacto con otros fonemas (inicio: $F(5,417)=2.1$, $p=0,06$, final: $F(6,416)=1.2$, $p=0,31$).

En resumen, la alineación del FAVE-align para el malecu es comparable a la del bribri, y la marcación de tipos de fonemas individuales para influir menos de lo que hace en el caso del bribri, donde se notan diferencias significativas entre tipos de consonantes y tipos de contacto entre fonemas.

### 3.4. Resumen de los resultados de FAVE-align

La tabla 12 resume el desempeño de FAVE-align al alinear los intervalos vocálicos y consonánticos. Como se comentó al principio de la sección 3, el desempeño del alineador para

el bribri y el malecu es significativamente mejor que para el cabécar, probablemente porque la grabación en cabécar incluye una mayor cantidad de ruido ambiental.

TABLA 12

Error promedio para los fonemas

| | | Bribri (ms) | Cabécar (ms) | Malecu (ms) | ANOVA |
|---|---|---|---|---|---|
| Vocales | Inicio del intervalo | 2,2 ± 0,3 | 7,3 ± 0,6 | 3,0 ± 0,4 | $F(2,711)=32.9$, $p < 0,000001$ |
| | Final del intervalo | 2,7 ± 0,3 | 8,3 ± 0,6 | 3,4 ± 0,4 | $F(2,711)=35.5$, $p < 0,000001$ |
| Consonantes | Inicio del intervalo | 2,2 ± 0,4 | 7,2 ± 0,6 | 3,1 ± 0,5 | $F(2,607)=23.8$, $p < 0,000001$ |
| | Final del intervalo | 2,2 ± 0,4 | 7,3 ± 0,6 | 3,1 ± 0,4 | $F(2,607)=26.0$, $p < 0,000001$ |
| Promedio | | 2,3 | 7,5 | 3,2 | |

La tabla 13 resume el desempeño de FAVE-align al alinear los intervalos con una exactitud de un milisegundo o menos. El bribri tiene un número mayor de intervalos con una exactitud igual o menor a un milisegundo, un 45% en promedio comparado a un 34% para el malecu y un 2% para el cabécar. El promedio para las tres lenguas es de 34%.

TABLA 13

Porcentaje de los intervalos que tienen un error de 1 ms o menos

| | | Bribri | Cabécar | Malecu |
|---|---|---|---|---|
| Vocales | Inicio del intervalo | 42% | 24% | 37% |
| | Final del intervalo | 43% | 20% | 13% |
| Consonantes | Inicio del intervalo | 52% | 25% | 53% |
| | Final del intervalo | 46% | 25% | 35% |
| Promedio | | 45% | 24% | 34% |

La tabla 14 resume los máximos de error al alinear con FAVE-align, y en particular detalla el tiempo de error debajo del cual están el 80% de los intervalos alineados. De nuevo el bribri superior al de los otros idiomas: El 80% de sus intervalos tienen un error promedio de 33 milisegundos o menos, mientras que esta cifra es de 62 milisegundos para el malecu y de 120 milisegundos para el cabécar. El promedio para las tres lenguas es de 72 milisegundos.

TABLA 14

Error (en milisegundos) debajo del cual están el 80% de los intervalos

| | | Bribri | Cabécar | Malecu |
|---|---|---|---|---|
| Vocales | Inicio del intervalo | 30 ms | 110 ms | 40 ms |
| | Final del intervalo | 45 ms | 140 ms | 120 ms |
| Consonantes | Inicio del intervalo | 32 ms | 120 ms | 45 ms |
| | Final del intervalo | 25 ms | 110 ms | 45 ms |
| Promedio | | 33 ms | 120 ms | 62 ms |

### 3.5. Resumen

Los resultados alcanzados al usar la alineación forzada son aceptables, especialmente para grabaciones en ambientes controlados y en las que una sola persona está hablando o narrando. Estas tasas de error hacen posible llevar a cabo correcciones manuales con una duración mucho menor a la marcación del texto de forma completamente manual. El desempeño, sin embargo, depende del ruido ambiental así como del tipo de discurso, dado que la grabación del canto produjo un desempeño más bajo que las narraciones.

## 4. Discusión

Dado el desempeño aceptable al alinear fonemas y palabras en lenguas indígenas, inferimos que hay un alto grado de transferencia entre los modelos en inglés y francés y los fonemas de las lenguas indígenas de Costa Rica. Ambos sistemas, FAVE-align e EasyAlign, producen TextGrids que necesitan corrección manual después de su ejecución, pero ambos tuvieron un desempeño similar, a pesar de usar modelos acústicos de idiomas diferentes. Como se ve en la tabla 15, los dos sistemas tienen ventajas y desventajas diferentes, que deben ser evaluadas por el usuario antes de implementar un sistema de alineamiento. En general, EasyAlign hace el procesamiento de los datos más rápido, pero FAVE-align permite más control sobre los TextGrids resultantes.

TABLA 15

Comparación de los sistemas FAVE-align e EasyAlign en el pre y postprocesamiento de los datos

| Sistema | Ventajas | Desventajas |
|---|---|---|
| FAVE-align | - Mayor control de la transcripción; el sistema alinea exactamente el número de unidades y la transcripción fonológica que el usuario ingresa | - Hay mayor carga de trabajo al preparar el documento para su procesamiento. |
| EasyAlign | - Mayor rapidez dado que el sistema alinea la transcripción automáticamente y convierte la transcripción ortográfica a una transcripción fonológica<br>- Los creadores de EasyAlign ofrecen entrenar modelos acústicos específicos para otras lenguas (pero las herramientas no son libres ni disponibles para el usuario) | - El sistema puede sustraer o añadir unidades fonológicas sin que estas correspondan a unidades en la lengua indígena<br>- El sistema no es consistente en el uso de su trascripción (SAMPA) (e.g. la letra SAMPA *O* puede mapearse a las letras bribri {o,ö})<br>- La falta de consistencia hace difícil recuperar los rasgos que se pierden en la conversión de ortografía a SAMPA (e.g. nasalidad, tono) |

El principal obstáculo a la hora de usar EasyAlign es la falta de consistencia entre el número de unidades de entrada (i.e. el número de fonemas representados en la ortografía de la lengua indígena) y las unidades de salida. La figura 16 muestra un TextGrid generado con EasyAlign. En la primera y segunda filas (*tiers*), se puede ver cómo las sílabas bribri í+*se* se han transcrito como *iz*. Con esto se pierde la segunda vocal, haciendo necesario el añadirla manualmente para una subsiguiente conversión de SAMPA a fonología bribri. En la figura 17 se observa el caso contrario. Aquí, el sistema EasyAlign ha añadido un fonema *E* al inicio de la palabra *sö*, haciendo necesaria su eliminación manual en la etapa de corrección antes de poder convertir el SAMPA a una transcripción fonológica propia del bribri.

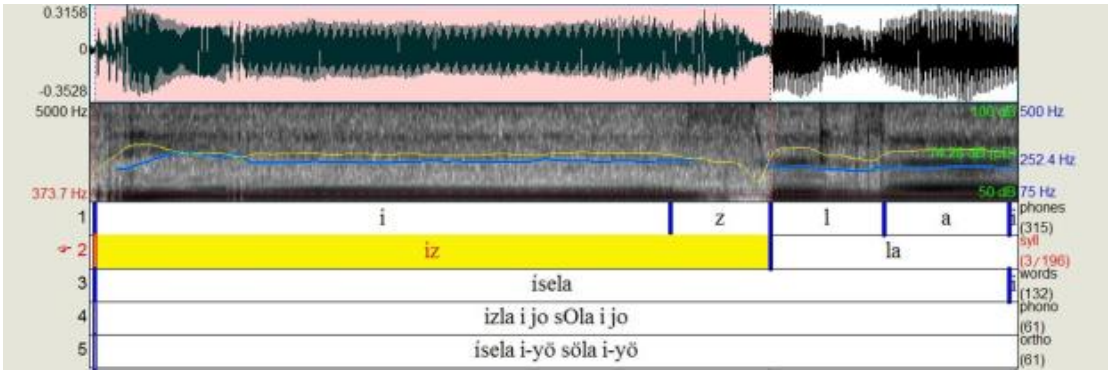


FIGURA 16

TextGrid bribri en EasyAlign donde una vocal no aparece en la alineación.

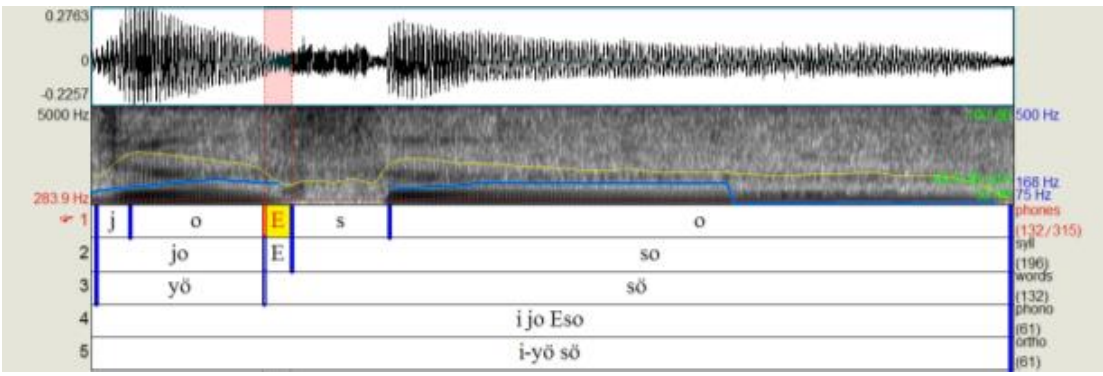


FIGURA 17

TextGrid bribri en EasyAlign donde se insertó una vocal en la alineación.

FAVE-align, por su parte, conserva todos los segmentos que el usuario haya especificado en el archivo *diccionario* (véase la sección 2.1), lo que permite el estudio incluso de segmentos que presenten reducciones severas (Ernestus y Warner 2011). Esto representa una ventaja al hacer estudios fonéticos detallados, pero tiene un precio a la hora el pre-procesamiento: el preparar una grabación para su alineación forzada con FAVE-align conlleva un tiempo mayor que simplemente introducir la grabación al sistema EasyAlign.

Ambos sistemas, con sus ventajas y desventajas, representan un ahorro importante de tiempo y esfuerzo a la hora de generar transcripciones alineadas en lenguas indígenas. Este procedimiento ayudaría al estudio de estas lenguas desde la lingüística computacional, en particular en tres aspectos. Primero, permitiría un mejor estudio de su fonética, y de aspectos como la sonorización, la prosodia, la reducción y otros que son susceptibles al estudio automatizado. Segundo, permitiría la creación de un corpus de transcripciones alineadas para el entrenamiento

de modelos acústicos específicos a estas lenguas. Tercero, estos corpus también se pueden usar el entrenamiento de sistemas de reconocimiento automático del habla (*ASR* o *automatic speech recognition*), que automatizarían la etapa de transcripción y permitirían acceso a un volumen de datos aún mayor.

Finalmente, estos avances en la inclusión de las lenguas indígenas de Costa Rica en plataformas tecnológicas tienen el potencial de generar un impacto positivo en la revitalización de dichas lenguas. El bribri es una lengua vulnerable y el malecu es una lengua en extremo peligro según las mediciones de vitalidad de la UNESCO (Moseley 2010, Sánchez, 2013), por lo que realzar la visibilidad digital de estas lenguas tendría consecuencias positivas para los hablantes y los esfuerzos de los activistas por conservar su lengua. En particular, el permitir interacciones más ricas con las interfaces humano-computadora, en particular el traducir las interfaces a las lenguas indígenas y permitir a las computadoras comprender estas lenguas, crearía una serie de motivaciones adicionales para la conservación de la lengua (Buzsard-Welcher 2001; Gasser 2006; Font Llitjós, Levin y Aranovich 2010).

## 5. Conclusiones

En este artículo usamos modelos acústicos ya existentes para alinear narraciones y canciones en las lenguas indígenas bribri, cabécar y malecu. Los resultados son alentadores, ya que el error de marcación oscila entre 8% y 13% para el centro de las palabras en bribri y en malecu, y para estas dos lenguas aproximadamente el 40% de los fonemas tienen un error de marcación de 1 milisegundo o menos. Los resultados para el cabécar reflejaron desempeños más bajos; el error para marcar el centro es de 37%, y el 24% de los fonemas tienen un error de 1 milisegundo o menos.

Estos sistemas de alineación forzada pueden ayudar al estudio automatizado de las lenguas indígenas de Costa Rica mediante la generación de corpus alineados, que puedan luego usarse para estudios fonéticos y para el entrenamiento de modelos acústicos y de reconocimiento del habla. Como trabajo futuro nos proponemos replicar estos experimentos con más grabaciones en la lengua cabécar así como con un mayor número de cantos y de grabaciones con más ruido ambiental para comprobar la aplicabilidad general del sistema.

## 6. Agradecimientos

Los autores desean agradecer a la Dra. Monica McCauley de la Universidad de Wisconsin-Madison por su presentación en la conferencia *LSA* en enero de 2015, y al Dr. Douglas Whalen de Laboratorios Haskins en la Universidad de Yale por discutir los resultados obtenidos por su equipo. Igualmente agradecemos al M.L. Guillermo González Campos de la Universidad de Costa Rica por facilitarnos grabaciones en cabécar, al M.A. Dane Bell y a la M.Sc. Samantha Wray de la Universidad de Arizona por sus comentarios en los borradores de este proyecto, así como a dos revisores anónimos por sus comentarios. Cualquier error en el artículo es responsabilidad de los autores.

## 7. Notas

1. Los dos sistemas, FAVE-align e EasyAlign, usan el set de herramientas HTK para sus modelos ocultos de Markov (Young et al. 2010).

2. Ambos sistemas presuponen que el habla se puede segmentar en partes discretas, y que es posible marcar el límite preciso entre un fonema y otro. Esto es fácil en casos como el de una consonante sorda al inicio de palabra, y casi imposible al tratar de determinar el límite entre una vocal y otra, o cuando hay casos de reducción segmental (Ernestus y Warner 2011). A pesar de este obstáculo, presuponemos que la segmentación es una herramienta útil para estudiar no solo fonos individuales, sino las transiciones entre diferentes fonos.

3. Para estas pruebas usamos una transcripción que no dependiera de caracteres Unicode. Por lo tanto, las nasales están representadas por una ‘h’ después de la vocal, y los tonos alto y bajo después de una vocal laxa están representados por una tilde después de la vocal.

4. Los autores reconocen que, conforme se entienda mejor el sistema acentual del bribri y el malecu,

podrían usarse los números dos y cero para representar otras intensidades en las sílabas de una palabra.

5. El alto glotal como fonema es particularmente problemático para los modelos acústicos en inglés, dado que este fono no existe como fonema en esa lengua (DiCanio et. al 2013; Whalen, comunicación personal).

6. Al momento de redactar este artículo los desarrolladores de EasyAlign no parecían estar activamente desarrollando otros modelos. Tampoco es posible hacer modificaciones a la herramienta, ya que se proporcionan solo los archivos binarios y no el código abierto.

7. La cifra después del '±' representa el error estándar de la media. Esta convención se usará a través de toda la sección de resultados.

8. El bribri tiene dos fonemas tonales que se han entendido tradicionalmente como el "bajo": el bajo fonológico y el neutro, que se realiza como bajo o medio dependiendo de la altura (registro) del tono siguiente (Coto-Solano 2015).

## 8. Bibliografía


Adda-Decker, Martine y Natalie Snoeren. 2011. "Quantifying temporal speech reduction in French using forced speech alignment". En: *Journal of Phonetics* XXXIX:261–270.

Biadsy, Fadi y Julia Hirschberg. 2009. *Using Prosody and Phonotactics in Arabic Dialect Identification*. Interspeech 2009.

Boersma, Paul. 2001. "Praat, a system for doing phonetics by computer". En: *Glot International* V(9/10):341-345.

Brognaux, Sandrine et al. 2012. *Train&Align: A New Online Tool for Automatic Phonetic Alignment*. Spoken Language Technology Workshop (SLT) 2012, IEEE. 416-421.

Buszard-Welcher, Laura. 2001. "Can the web help save my language?". En: *The Green Book of Language Revitalization in Practice*. Pgs. 331-45.

Constenla, Adolfo. 1998. *Curso básico de bribri*. San José: Editorial de la Universidad de Costa Rica.

Constenla, Adolfo. 1999. *Gramática de la lengua guatusa*. San José: EUNA.

Constenla, Adolfo. 2015. *Muérrajá Mausírrajáca Pláticas sobre ogros*. San José: Editorial de la Universidad de Costa Rica. Pgs. 51-52, 116-118.

Coto-Solano, Rolando. 2015. *The Phonetics, Phonology and Phonotactics of the Bribri Language*. 2nd International Conference on Mesoamerican Linguistics. Los Angeles: California State University. https://www.academia.edu/11365794/The_phonetics_phonology_and_phonotactics_of_the_Bribri_Language. Consulta: 25 de enero, 2016.

Cho, Taehong y Peter Ladefoged. 1999. "Variation and universals in VOT: evidence from 18 languages". En: *Journal of Phonetics* XXVII:207–229.

DiCanio, Christian et al. 2012. "Assessing agreement level between forced alignment models with data from endangered language documentation corpora". En: *Proceedings of InterSpeech 2012*.

DiCanio, Christian et al. 2013. "Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment". En: *Journal of the Acoustic Society of America*, CXXXIV(3):2235-2246.

Ernestus, Mirjam y Natasha Warner. 2011. "An introduction to reduced pronunciation variants". En: *Journal of Phonetics* XXXIX(3):253-260.

Flores Solórzano, Sofía. 2010. "Teclado Chibcha: Un software lingüístico para los sistemas de escritura de las lenguas bribri y cabécar". En: *Revista de Filología y Lingüística* XXXVI(2):155-161.

Font Llitjós, Ariadna, Lori Levin y Roberto Aranovich. 2010. Building Machine translation systems for indigenous languages. CILLA 2. http://www.cs.cmu.edu/~aria/Papers/FontAranovich_CILLA2_mapuche_quechua(2).pdf. Consulta: 20 de enero, 2016.

Gasser, Mike. 2006. *Machine Translation and the Future of Indigenous Languages.* I Congreso Internacional de las Lenguas y Literaturas Indoamericanas. ftp://ftp.cs.indiana.edu/pub/gasser/cilli.pdf. Consulta: 16 de enero, 2016.

Goldman, Jean-Philippe. 2011. "EasyAlign: an automatic phonetic alignment tool under Praat". En: *Proceedings of InterSpeech 2011.*

González Campos, Guillermo. 2011. "Dificultades para normalización ortográfica y problemas de escritura entre los cabécares de Chirripó". En: *Lingüística Chibcha* XXX:7-35.

Jara Murillo, Carla y Alí García Segura. 2009. *Se' ẽ' yawö bribri wa Aprendemos la lengua bribri*. San José: Editorial de la Universidad de Costa Rica. Pgs. 155-159.

Labov, William, Ingrid Rosenfelder y Josef Fruehwald. 2013. "One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis". En: *Language*, LXXXIX(1):30–65.

Lin, Cheng Yuan, Jyh-Shing Roger Jang y Kuan-Ting Chen. 2005. "Automatic segmentation and labeling for Mandarin Chinese speech corpora for concatenation-based TTS". En: *Computational Linguistic Chinese Language Processing* X(2):145–166.

Lisker, Leigh y Arthur Abramson. 1964. "A cross-language study of voicing in initial stops: acoustical measurements". En: *Word* XX:384-422.

R Core Team. 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Viena, Austria. http://www.R-project.org/. Consulta: 20 de enero, 2016.

Rosenfelder, Ingrid et al. *FAVE (Forced Alignment and Vowel Extraction) Program Suite*. http://fave.ling.upenn.edu. Consulta: 20 de enero, 2016.

Sánchez Avendaño, Carlos. 2013. "Lenguas en peligro en Costa Rica: Vitalidad, documentación y descripción". En: *Káñina* XXXVII(1):219-250.

Schiel, Florian y Christoph Draxler. 2003. *The production of speech corpora*. Bavarian Archive for Speech Signals.

Sim, Khe Chai y Haizhou Li. 2008. "Robust phone mapping using decision tree clustering for cross-lingual phone recognition". En: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*: 4309–4312.

Strunk, Jan, Florian Schiel y Frank Seifart. 2014. "Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS". En: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*:3940-3947.

Wightman, Colin y David Talkin. 1997. "The Aligner: Text to speech alignment using

Markov Models". En: *Progress in Speech Synthesis*. Nueva York: Springer Verlag. Pgs. 313-323.

Yoon, Tae-Jin. 2008. *HTK-TIMIT Forced Alignment Toolkit*. http://web.uvic.ca/~tyoon/resource/htk_utt.m. Consulta: 15 de enero, 206.

Young, Steve et al. 2010. *The HTK Book*. Cambridge University Engineering Department. http://htk.eng.cam.ac.uk/. Consulta: 15 de enero, 2016.

Yuan, Jiahong et al. 2013. "Automatic Phonetic Segmentation using Boundary Models". En: *Proceedings of Interspeech 2013*:2306-2310.

Yuan, Jiahong y Mark Liberman. 2008. "Speaker identification on the SCOTUS corpus". En: *Proceedings of Acoustics '08*.

Yuan, Jiahong y Mark Liberman. 2009. "Investigating /l/ variation in English through forced alignment". En: *Proceedings of InterSpeech* 2009:2215–2218.