Segmentación de morfemas con cálculo de entropía en pa ipai

Morpheme segmentation with entropy measure in pa ipai

Manuel Alejandro Sánchez Fernández
Facultad de Pedagogía e Innovación Educativa
Universidad Autónoma de Baja California
msanchez12@uabc.edu.mx

RESUMEN: Este trabajo propone implementar el cálculo de entropía de la información en la lengua pa ipai para determinar si esta medida puede ser útil para sugerir lindes morfológicos. Se recopilaron y homogenizaron textos en lengua pa ipai para realizar dos experimentos: el primero para detectar posibles morfemas y el segundo con un texto glosado para calificar la validez de la medida de entropía para predecir los límites morfológicos. Los resultados muestran una relación significativa entre el cálculo de entropía de la información y las segmentaciones sugeridas por una lingüista. La propuesta resulta novedosa tanto para el estudio de lenguas yumanas como para la implementación de herramientas computacionales con lenguas que cuenten con muy pocos recursos.

Palabras clave:
entropía de la información, pa ipai, morfología computacional, lingüística de corpus, lenguas yumanas

ABSTRACT: This work proposes implementing the calculation of information entropy in the pa ipai language to determine if this measure can be useful in suggesting morphological boundaries. Texts in the Pa ipai language were collected and homogenized to conduct two experiments: the first to detect suspicious morphemes and the second with a glossed text to assess the validity of the entropy measure in predicting morphological boundaries. The results show a significant relationship between the calculation of information entropy and the segmentations suggested by a linguist. The proposal is innovative both for the study of Yuman languages and for the implementation of computational tools with languages that have very few resources.

KEYWORDS:
Information entropy, Paipai language, computational morphology, corpus linguistics, Yuman languages.

Fecha de recepción: 25 de marzo de 2023
Fecha de aceptación: 8 de mayo de 2023

Introducción

El presente trabajo tiene como objetivo implementar el cálculo de entropía de la información (*H(X)*) (Ackerman y Malouf 2013: 437) a palabras en lengua pa ipai, con la finalidad de determinar si tal medida puede ser útil para sugerir lindes morfológicos. Los trabajos computacionales que han buscado implementar estrategias de delimitación automática de morfología en lenguas indígenas son escasos y han utilizado diversas herramientas (Mager, Carrillo y Meza 2018; Osvaldo Porta 2010). Sin embargo, no existe ningún antecedente que haya implementado alguna estrategia computacional para determinar de manera semisupervisada o supervisada los límites morfológicos dentro de la palabra en pa ipai o de alguna lengua yumana. Esto, entre otras razones, se debe a la naturaleza de los corpus que se requieren para realizar la experimentación computacional (Mager *et al.* 2018).

Con miras a lograr el objetivo planteado, se recopilaron y homogenizaron distintos textos en lengua pa ipai y se realizaron dos experimentos. El primer experimento consistió en determinar, con textos sin glosa (Peralta 1994), la posibilidad de detectar posibles morfemas; el segundo experimento incluyó un texto glosado por Judith D. Joël (Joël 1976), una de las primeras investigadoras en describir la lengua pa ipai. La glosa ayudó a calificar la validez de la medida de entropía para predecir los límites morfológicos. La propuesta resulta novedosa tanto para el estudio de lenguas yumanas como para la implementación de herramientas computacionales con lenguas que cuenten con muy pocos recursos.

El trabajo está dividido de la siguiente manera. En primer lugar, se presenta una caracterización de la lengua pa ipai hablada en Santa Catarina, así como un breve repaso de lo que se conoce sobre la estructura de su palabra. Posterior a esto, se expone el concepto de entropía de la información y el cálculo que se utiliza en trabajos de lingüística computacional. A esta sección le continúa el procedimiento que dio lugar a los distintos experimentos, desde la limpieza del corpus hasta la implementación en los algoritmos

informáticos. De tal manera, en la penúltima sección, se muestran los resultados. En primer lugar, se explicitan aquellos obtenidos de la experimentación de textos sin glosa (Peralta 1994), y luego se comparan con los obtenidos del análisis del texto glosado de Joël (1976) a partir de un análisis de varianza y una prueba de correlación biserial puntual. Este análisis estadístico permite sostener que sí existe una relación significativa entre el cálculo de entropía de la información y las segmentaciones sugeridas por Joël, lo que proporciona una pista a los datos encontrados sin glosa para los textos de Benito Peralta. El capítulo finaliza con unas conclusiones, en donde se trazan los siguientes pasos de experimentación.

PA IPAI DE SANTA CATARINA

El pa ipai (ISO 639-3 ppi) es una lengua de la familia lingüística yumana hablada en Baja California, México. Comparte filiación con las lenguas de la rama pai de esta familia, en donde se encuentran el havasupai, el hualapai y el yavapai. La familia completa consiste en alrededor de 12 lenguas, con distintas variantes dialectales identificadas. Aunque la distribución de la familia ha podido ser descrita con amplia precisión desde la década de los 60 del siglo pasado (Joël 1964; Langdon 1970), aún hay discusiones sobre la identificación de otras variantes. En la figura 1, se observa la distribución de la familia lingüística, en donde se señala la ubicación del pa ipai en negritas en el árbol.

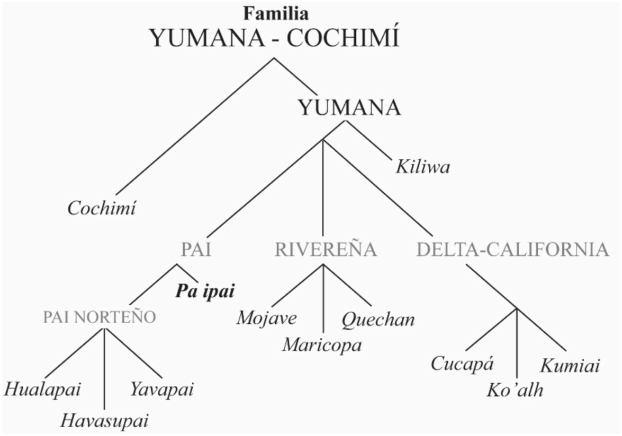


Figura 1. Distribución de la familia lingüística yumana-cochimí. Creación propia.

En la actualidad, la comunidad de habla pa ipai más vital se encuentra en la localidad de Santa Catarina, ubicada alrededor del kilómetro 95 de la carretera Ensenada-San Felipe, en el municipio de Ensenada, Baja California, México (Carbajal 2002). Distintos registros apuntan a que quedan alrededor de 30 a 60 hablantes de pa ipai, además de señalar que ya no hay niños que hablen la lengua (González Castillo 2020; Sánchez-Fernández y Rojas-Berscia 2016; Gómez Serna 2010). Debido a estas circunstancias, la UNESCO y el Instituto Nacional de Lenguas Indígenas (Moseley 2010; Embriz Osorio y

Zamora Alarcón 2012) han catalogado a esta lengua, así como a todas las otras lenguas de la familia yumana, como en *grave peligro de desaparecer*. A este escenario se suma la escasa cantidad de investigaciones de documentación y descripción lingüística. En trabajos posteriores (Sánchez-Fernández 2022) se ha hecho énfasis en que existen muy pocos recursos accesibles en esta lengua, especialmente para el trabajo computacional, el cual demanda cierto rigor en la forma de manejar las bases informáticas. El trabajo que se presenta a continuación busca contribuir al manejo y uso de los recursos a los cuales se tiene acceso, y poner el acento en generar espacios digitales para volverlos alcanzables, tanto a investigadores como a miembros de la misma comunidad, lo que puede ayudar en los esfuerzos de documentación de esta lengua amenazada.

Palabra en pa ipai

No existen antecedentes para definir el concepto de *palabra* en la lengua pa ipai. La lengua no cuenta con trabajos lexicográficos, pero sí con ciertos antecedentes que inician en la discusión sobre la noción de palabra como estrategia para identificar mecanismos para nombrar la naturaleza, los astros y el tiempo (González Castro *et al.* 2016; Martínez Arellano, Bertha Uribe y Sánchez-Fernández 2013; Owen 1963; Cortés Rodríguez 2013).

Siguiendo lo definido por Lara (2015), podemos entender *palabra* desde tres perspectivas: como palabra gráfica, morfológica y fonológica. La primera perspectiva tiene que ver con grupos culturales que han desarrollado una tradición escrita. Las comunidades yumanas, y, en particular, los pa ipai, se han distinguido, en contraste, por una fuerte tradición oral. Muchos de los ritos y relatos se mantienen gracias a esta oralidad, y siguen pautas y reglas que difieren de lo concebido en la tradición escrita (Bright 1992).

La palabra gráfica es la que se suele usar como fundamento en los análisis computacionales. La demarcación de los espacios en blanco son el recurso por excelencia para la alimentación del algoritmo informático. Los lingüistas o los hablantes pueden establecer espacios en blanco por distintas razones, y es precisamente esta situación la que vuelve compleja la experimentación en algoritmos informáticos. Aunque en esta etapa de la presente investigación no hay preferencia por algún tipo de separación, será suficiente con delinear el criterio que se pudo haber utilizado para la transcripción de cada texto. Es decir, parte del método para este caso consiste en examinar los textos transcritos y asociar grafema con fonema, lo cual permite bosquejar las intenciones de escritura del autor. Sobre esto, también será suficiente partir, no de fonemas, sino de grafemas; de unidades de representación gráfica dentro de la palabra.

Una ventaja que tiene esta investigación es que los grafemas usados en pa ipai siguen pautas similares al uso del alfabeto latino para el español, lo que ha permitido salvar parcialmente el problema que implica la demarcación del grafema. Sin embargo, el grafema también tiene su propia problemática en cuanto a su delimitación semiótico-visual (Morin, Kelly y Winters 2020), además de los problemas metodológicos que implica la ausencia en pa ipai de una norma de escritura. Sin lugar a duda, en un futuro se podrá

realizar una investigación que describa y explique el proceso de normalización, adaptación y uso del sistema de escritura en pa ipai (Barriga Villanueva 2001).

Avanzando en la descripción de la palabra en pa ipai, y en cuanto a su morfología y fonología, esta lengua se categoriza como de tipo sintética, con bajo índice de fusión (Sánchez-Fernández 2016). No es común encontrar varias bases verbales en una palabra, pero sí está documentada la estrategia de incorporación de nominal a las bases (Mithun 1986). En (1a) se observa esta estructura sintética en la base verbal *ʔu:* 'ver', la cual carga con varios morfemas que expresan relaciones sintácticas de persona y número. Nótese el bajo índice de fusión: los morfemas apenas y adquieren un solo valor.

(1)

a. *machy sach' paa mch'uub*
má-ʃ-j sa-ʃ-j pa:-m-ʃ-ʔú:-β
2-PL-NOM[1] 3-PL-ACC PL.OBJ-2 > 3-PL.SBJ-VER-PL.SBJ
'Ustedes los ven a ellos'.

b. *ŕituyum chkñaay jwak muñaay nyam*
ʂituyúm ʃ-k-ɲa:j χwak mu-ɲá:j nyam
una_vez ʃ-REL-cazar.PL dos borrego-cazar.PL ir
'En alguna ocasión, dos cazadores salieron a borrego-cazar'.

En (1b) se muestra una frase que inicia el cuento de "Los Cazadores" en pa ipai (Sánchez-Fernández 2016). Se nota que la base verbal *ɲa:j* 'cazar.PL' tiene incorporado el nominal *mu* 'borrego cimarrón' que cumpliría con la función de objeto directo en la oración. En pa ipai, se espera la presencia de un determinante en una frase nominal completa. La ausencia de este elemento, así como de acentos y modificadores, indica que la forma no se refiere a una entidad en específico, sino que proporciona información de una categoría para el verbo. Esto complejiza la palabra a nivel morfológico.

Sobre esta incorporación, ahora para los nominales, podemos encontrar un ejemplo de estructuras morfológicamente complejas en palabras que se utilizan como unidades de denominación, como la que se presenta en (2).

(2)

chipay yelpatchkyob ñwa
ʃipaj-jəlpátʃ-k-joβ-ɲ-wa
animal-miel-REL-hacer.PL-POSS-casa
'Panal' *Lit.* 'La casa de las que hacen la miel'.

[1] 1 = primera persona; 2 = segunda persona; 3 = tercera persona; ACC = acusativo; NOM = nominativo; PL = plural; POSS = posesivo; REL = relativa; SBJ = sujeto; SG = singular.

Nótese que la hablante, en la primera línea, preferirá dividir esta misma unidad de denominación en tres piezas gráficas distintas. Aunque en pa ipai se advierte esta complejidad morfológica, con anterioridad se ha sugerido que la palabra fonológica tiende a tener la estructura mínima CVC, tanto en esta como en todas las otras lenguas yumanas. Margaret Langdon, una experta yumanista, reconstructora del protoyumano, hace el llamado a sospechar que si encontramos en una palabra yumana una estructura del tipo CCVC, tal estructura probablemente contenga afijos (Langdon 1975).

Años después, la investigadora Ibáñez Bravo (2015), en un estudio fino sobre la fonología de esta lengua, evidencia que sí podemos encontrar bases un poco más complejas fonológicamente que la estructura CVC sugerida por Langdon (1975). Ibáñez Bravo identifica hasta cinco consonantes prevocálicas y dos posvocálicas para palabras que forman verbos finitos. Es el caso de (3a), aunque nótese (3b) y (3c) para casos parecidos (Ibáñez Bravo 2015: 81).

(3)

a. CCCCCVC
/ʂxmkwiɾ/ ‘celosa’
b. CCCCVC
/xɾkyet/ ‘alisar’
c. CCCCV
/tʂkwa/ ‘lleno’

Es fundamental resaltar que el pa ipai cuenta con un inventario fonológico de 28 fonemas, 18 consonánticos y 10 vocálicos, para poder contrastar los experimentos de manera adecuada (Ibáñez Bravo 2015). Se notará en los textos que conforman el corpus que, de acuerdo con cada autor, se cuenta con una variedad distinta de grafemas. Esto se cuida en la interpretación de los resultados de cada experimento y en las comparaciones. Lo antes mencionado permite trazar la complejidad de alimentar a un algoritmo informático con una lista de palabras, ya sean delimitadas por un lingüista —con intenciones de especificar fonología o morfología— o por los mismos hablantes, a través de sus esfuerzos por escribir la lengua.

CÁLCULO DE ENTROPÍA

De acuerdo con Jurafsky y Martin (2009), se entiende ENTROPÍA DE LA INFORMACIÓN como “una unidad de medida de información […] que nos dice el límite inferior del número de bits que se necesitarían para codificar una determinada decisión o pieza de información en el esquema de codificación óptima” (p. 222). Esta propuesta, a su vez, encuentra su base en lo expuesto en su momento por Shannon (1948). El concepto que maneja este autor tiene distintas variaciones, con un amplio espectro de usos tecnológicos e interpretaciones matemáticas. De acuerdo con Medina Urrea (2021: sec. 2.3.2), la

entropía se ha entendido como el caos que puede existir dentro un sistema y la capacidad de *sorprendernos* por la aparición de una letra en particular dada una secuencia. Por ejemplo, supongamos que tenemos un inventario de tres unidades léxicas:

(4)

a. GATO
b. TATO
c. SALTO

Un primer supuesto que se maneja es que estas unidades forman parte de un sistema de comunicación. Esto nos adelanta a que estas secuencias no pueden tender a la aleatoriedad máxima. De tal manera que si se presenta un grafema como A podemos decir que lo que le seguirá sólo puede ser T o L, pero nunca O —lo anterior, a partir de este microsistema de tres palabras. De hecho, es posible calcular esta probabilidad: una vez presente A, L aparece con un 33.33% de probabilidad y T con un 66.66% probabilidad. Si ordenamos las probabilidades de todo el inventario, tomando como pivote el grafema A, e incluimos los casos con probabilidad cero, tendríamos una representación como se muestra en la tabla 1.

Tabla 1. Distribución de probabilidades dada la letra A

A	A	0
	O	0
	S	0
	G	0
	T	0.66
	L	0.33

Obsérvese que el grafema A resulta muy informativo: su aparición nos permite descartar más de la mitad del inventario. Condiciona en gran medida la aparición de otros elementos. Esta capacidad informativa es la que podemos calcular a partir de *H(X)*, la cual se desarrolla a partir de la siguiente fórmula (Shannon 1948).

(5)

$$H(X) = -\sum_{i=1}^{n} p_i p_i$$

Una MÁXIMA ENTROPÍA implicaría que, por ejemplo, dado el grafema A, cada una de las letras del inventario tuviera la misma probabilidad de aparecer. La presencia de A no sería nada informativa, no nos permite saber qué seguirá. Supongamos el caso anterior, y resolvamos la fórmula. Esto supone que cada letra tiene un 0.16 de probabilidad de aparecer.

(6)

$$H(X) = -\sum_{i=1}^{6} p_i p_i = -\sum_{i=1}^{6} (0.16)(0.16) \approx 2.585$$

Este número se lee como que la máxima entropía del sistema representado por las tres palabras en (4) sería de 2.585 bits. Si resolvemos el caso para el grafema A, pero ahora sólo con las probabilidades de la tabla 1, tendríamos lo siguiente:

(7)

$$H(X) = -\sum_{i=1}^{2} p_i p_i = -(0.330.33 + 0.660.66) \approx 0.918$$

La entropía absoluta de A sería 0.918 bits. Para poder tener una lectura general, se normaliza el resultado anterior con el 2.585 que obtuvimos de la máxima entropía. Es decir, usamos la máxima entropía como suelo común, en donde 1 representaría la máxima incertidumbre posible, y 0 la mínima. Para obtener la entropía normalizada (también llamada relativa), realizaríamos lo siguiente:

(8)

$$\frac{H(X)}{2.585} \approx \frac{0.918}{2.585} \approx 0.355$$

El anterior paso es el que permite realizar la comparación con otras piezas del sistema, ya sea con cada grafema o, como se llevará a cabo en los experimentos a continuación, de cada posible combinación de grafemas dados los arreglos dispuestos en el inventario léxico de (4). Esto quiere decir que, para los experimentos que se desarrollan en este trabajo, sólo se calcula la entropía de los *posibles afijos* producto de todos los posibles cortes lógicos a partir de tomar como pivote algún lado de la palabra.

Evaluación de los cortes

El resultado del cálculo de entropía relativa no dice mucho por sí solo. Es una huella, en contraste con los resultados de los otros afijos posibles. En todo caso, el dato sólo nos ayudaría a confirmar que el sistema no tiende al caos, aunque se espera que un sistema de comunicación no lo haga (Shannon 1948). La primera etapa de la experimentación propuesta es exploratoria, en donde una vez realizado el cálculo, se contrasta una lista de posibles afijos encontrados en el primer conjunto de textos. Sin embargo, pronto se descubren las limitaciones de un experimento así. Por ello, se propuso un segundo experimento en este trabajo, lo que permite una evaluación más precisa.

En otros experimentos se usa una versión del cálculo de entropía para determinar si hay correspondencia entre la entropía del sistema estudiado y la entropía propia de una lengua, en contraste con la entropía de, por ejemplo, el ADN o de un lenguaje de

programación (*vid*. Rao *et al.* 2009). El segundo experimento del presente trabajo busca realizar una correlación entre el cálculo de entropía y la delimitación que efectivamente hace un lingüista. Por poner un ejemplo, podríamos predisponer la búsqueda de correlación de *H(X)* de -O a partir de los datos de (4), debido a que suele ser analizada como un sufijo en los nominales de español. No esperamos que salga una entropía 0 o 1, sino un rango de entropía relativa que ayude a sugerir que antes de O hay un corte morfológico. Con lo anterior, en este punto del trabajo se plantea la siguiente pregunta de investigación: ¿hay alguna correlación entre alguna cantidad de bits y los cortes morfológicos? Lo que nos arroja la hipótesis a evaluar:

H1 Existe una correlación significativa entre los cortes propuestos de los afijos y su cálculo de entropía.

Se utilizará un análisis de varianza (ANOVA) para determinar si la diferencia en la entropía entre los diferentes cortes propuestos de los afijos es estadísticamente significativa, así como una prueba de correlación biserial puntual como segunda prueba de correlación. La hipótesis nula (H0) a descartar es que no hay diferencia significativa entre la entropía de los diferentes cortes propuestos de los afijos y los cortes propuestos por el lingüista. Lo anterior, también permitirá generar una base para contrastar los resultados del primer experimento, lo cual se explica a continuación.

Metodología

El método propuesto sigue los siguientes pasos:

I.-Ubicación y limpieza de corpus. Es importante tener un inventario léxico en formato CSV[2] con un sistema de escritura homogéneo para facilitar la lectura por una computadora. A veces, se usan símbolos difíciles de leer o que representan dos grafemas, como dobles vocales en pa ipai o dos puntos en la transcripción fonológica. Para simplificar, estos casos de doble símbolo se sustituyen por un solo símbolo que representa un fonema de la lengua.

I.a.- Los textos de Benito Peralta. El primer conjunto de textos está conformado por 16 relatos del libro *Relatos Pai pai. Kuriut' trab pai pai* (1994). Estos textos fueron transcritos

[2] CSV (*Comma Separeted Values*) es un tipo de archivo utilizado para guardar información en forma de tablas. En un archivo CSV, cada línea representa una fila de datos y las columnas se separan por comas. Es ampliamente utilizado para compartir datos entre diferentes programas, debido a su facilidad de lectura y escritura tanto para personas como para máquinas. Los archivos CSV pueden abrirse y modificarse con software de hojas de cálculo y bases de datos.

en archivos de texto, y luego homogenizados para el tratamiento informático. El corpus final cuenta con 4 379 palabras gráficas, un inventario de 32 grafemas, los cuales, una vez homogenizados, llegaron a 27. En la tabla 2 se muestra la equivalencia analizada entre grafemas (G), fonemas (F) descritos de manera articulatoria de acuerdo con Ibáñez Bravo (2015), y los grafemas usados en el algoritmo informático (G.AL). Nótese que los fonemas /ʐ/ y /ʂ/ tienen una misma representación en los textos, por lo que también sólo tiene una representación en el algoritmo.

Tabla 2. Equivalencias entre fonemas, grafemas y grafemas de uso informático de Peralta

F	G	G.AL	F	G	G.AL	F	G	G.AL	F	G	G.AL
/p/	p	p	/x/	h, x	x	/a/	a	a	/a:/	aa	á
/t/	t	t	/β/	v, b	b	/i/	i	i	/i:/	ii	í
/k/	c, k	k	/s/	s	s	/u/	u	u	/u:/	uu	ú
/q/	q	q	/ʃ/	ch	$	/e/	e	e	/e:/	ee	é
/l/	l	l	/ʐ/	z	z	/o/	o	o	/o:/	oo	ó
/ɾ/	r, rr	r	/ʂ/	z	z	/w/	g, w	w	/j/	y	y
/m/	m	m	/n/	n	n	/ɲ/	ñ	ñ	/ʔ/	’	’

I.b.- Los textos de Judith Joël. El segundo experimento se realizó a partir de un solo texto de Judith Joël, The Earthquake of ’57: a paipai text (Joël 1976). Este relato es el único glosado por la investigadora, quien es considerada la primera lingüista en reportar de manera detallada aspectos morfológicos y fonológicos de la lengua. El procedimiento para este texto fue un poco distinto al conjunto anterior. Primero, se transcribieron todas las palabras registradas en la línea de la lengua del texto en una tabla de Excel. Se tomó como criterio de división de palabra los espacios y la asociación con un conjunto de morfemas en la línea de glosa del texto.

2. mat?in-ha-Y ñə-va:-m ?ñe-Y ?-ma:n kaqwar ?-čpa: ?i-?i-k ?a-?am
e.quake-the-sb wh-come-ds I-sb 1-get=up outside 1-go=out 1-say-ss 1-go=round

Figura 2. Ejemplo de fragmento glosado de Joël (1976: 85)

La figura 2 muestra esta doble línea de análisis presente en el texto. Tomado como ejemplo ese fragmento, se tendría un total de siete palabras. Esta transcripción se realizó respetando los guiones de división de morfemas. Una vez realizado lo anterior, se homogenizó el inventario de grafemas y luego se integró en el algoritmo un paso que genera dos conjuntos: uno con los guiones y otro sin ellos. Se obtuvieron 490 palabras y 34 grafemas, de los cuales, una vez homogenizados, se llegaron a 27 grafemas[3]. El

[3] En la transcripción del texto de Joël (1976), hay un caso sumamente extraño que parece ser un error tipográfico. En la página 87 aparece la palabra *ñə-nka:v-kz* ‘WH-arrive = PL-k**z**’. Incluso, en el documento original, la letra *z* aparece en negritas y ligeramente más arriba de la línea normal de

inventario de grafemas, junto con los fonemas asociados y los grafemas usados en el algoritmo informático se despliega en la tabla 3.

Tabla 3. Equivalencias entre fonemas, grafemas y grafemas de uso informático de Joël

F	G	G.AL	F	G	G.AL	F	G	G.AL	F	G	G.AL
/p/	p	p	/x/	h, x	x	/a/	a	a	/a:/	a:	á
/t/	t, d	t	/β/	v, b	b	/i/	i	i	/i:/	i:	í
/k/	c, k	k	/s/	s, c	s	/u/	u	u	/u:/	u:	ú
/q/	q	q	/ʃ/	č	$	/e/	e, ə	e	/e:/	e:, ə:	é
/l/	l	l	/ʐ/	š	z	/o/	o	o	/o:/	o:	ó
/ɾ/	r, rr	r	/ʂ/	š	z	/w/	g, w	w	/j/	y	y
/m/	m	m	/n/	n	n	/ɲ/	ñ	ñ	/ʔ/	ʔ	ʔ

Tanto los textos de Peralta como los de Joël contenían algunas palabras prestadas de español. En especial, hay dos escenarios que implicaron soluciones distintas: por un lado, en Peralta se encuentra la palabra *Mexicali* en donde la forma *c* en realidad hace referencia a la /k/, escrita como *k* de manera extensiva en su propio texto. Para este caso fue sencillo hacer la sustitución. En el caso de Joël (1976: 85), se encontró la palabra *fevre:ro*. Ni la forma *f* ni el fonema /f/ están presentes en la lengua, por lo que tampoco es posible trazar una relación alofónica. Debido al cálculo, esto podría afectar el sistema completo, por lo que se eliminó la palabra *fevre:ro* para este experimento. Los demás casos de préstamo no presentaron problema para mantener el sistema homogenizado.

II.- Procedimiento para el cálculo de entropía. Una vez realizado el ajuste y homogenización de los textos, se procedió a construir el algoritmo que realizó el corte de afijos y el cálculo de entropía. El proceso completo está codificado en Python, con la diferencia de que, para los textos de Benito Peralta, se lee un documento terminación .txt, mientras que para los de Joël, se lee un .xlsx. A grandes rasgos, el proceso del algoritmo es el siguiente:

a.-Limpieza interna de tokens. Se limpia puntuación, se pasa todo a minúsculas y se eliminan posibles saltos de línea y tabuladores. En este paso es en donde se realizan las sustituciones a G.AL.

b.-Segmentado de posibles afijos. Se generan todos los posibles afijos. El programa pasa dos veces por esta tarea: una para analizar de izquierda a derecha, to-

texto, lo que parece sugerir que se agregó al final, despues de terminado el manuscrito. Debido a que es la única instancia de la forma *z*, no tiene glosa y sería muy extraño en pa ipai encontrar una fricativa retrofleja sonora en esa posición, se eliminó el grafema para estos experimentos. En su lugar, se retoma en el algoritmo el grafema *z* para hacer referencia al grafema *š* del texto.

mando como pivote el primer grafema del lado izquierdo (prefijos), y otra de derecha a izquierda (sufijos), tomando como pivote el último grafema del lado derecho.

c.- *Cálculo de entropía.* Se busca, en el inventario léxico inicial, todas las apariciones de cada afijo posible segmentado. Para cada caso, se contabiliza la cantidad de letras posteriores o anteriores, sea que se estén analizando prefijos o sufijos. Se realiza el procedimiento de cálculo de entropía explicado en §3 (fórmula en (6), (Jurafsky y Martin 2009; Ackerman y Malouf 2013)).

III.-Evaluación y correlación. Una vez realizado el procedimiento anterior con el texto de Joël, se procede a un tercer paso en el algoritmo informático. Se toma la lista de palabras y se extraen todos los afijos a partir del corte sugerido por la investigadora, tomando como indicio el guion corto (-). Esta nueva lista de afijos es entonces comparada con el inventario de potenciales afijos. A cada una de las piezas determinadas por Joël y encontradas por el programa se le asigna el valor 1, y a todas aquellas que el programa encontró, pero Joël no las codifica como afijos, se le asigna 0. Este valor binario ayudará para proponer un análisis de varianza (ANOVA) (Urdan 2005: 118; Hernández Campoy y Almeida 2005: 221) y un análisis de correlación biserial puntual (Urdan 2005: 85) entre todo el conjunto de afijos que sí están presentes en Joël con los que no lo están. Se reporta el estadístico de correlación, el valor *p* y se obtiene el gráfico de la distribución de los dos conjuntos, lo que ayuda a la examinación de la normalidad de los conjuntos. Al final, se comparan los gráficos de distribución del cálculo de entropía normalizado de los textos de Benito Peralta y el obtenido de Joël.

RESULTADOS

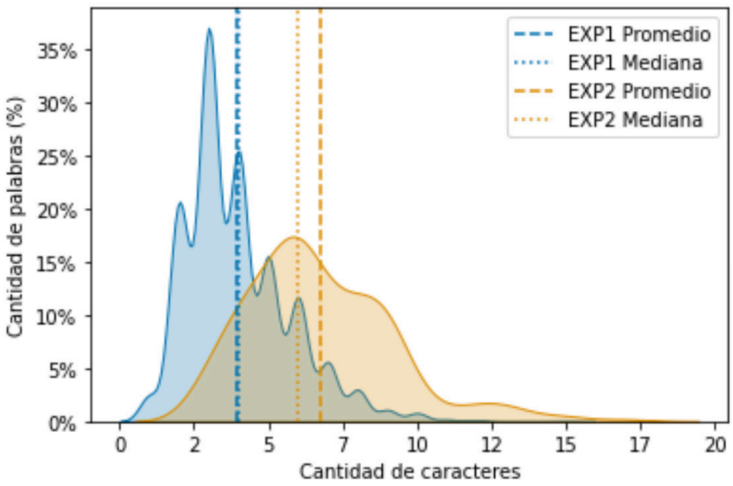
Para la exposición de los resultados, primero se describirán datos descriptivos generales de cada uno de los experimentos, incluyendo las longitudes de palabras posterior a la homogenización. A partir de este momento se referirá a EXP1 a los resultados de los experimentos que utilizaron los textos de Benito Peralta, mientras que EXP2 se referirá a aquellos obtenidos a partir del ejercicio realizado con los datos de Judith Joël.

Datos descriptivos generales

El análisis de segmentos en EXP1 arrojó lo siguiente. De las 4 379 palabras del corpus se obtuvieron 12 958 segmentos, tanto para prefijos como para sufijos. Una vez creado el conjunto cardinal, resultaron 2 480 para prefijos y 2 692 para sufijos. Finalmente, se descartaron todos aquellos afijos que tuvieran entropía 0, lo que resultó en 659 prefijos y 680 sufijos. Para el análisis de segmentos en EXP2, se obtuvo lo siguiente. De las 490

palabras extraídas se obtuvieron 2 366 cortes posibles para los prefijos y los sufijos. De ellos, una vez realizado el conjunto cardinal, quedaron 929 segmentos para prefijos y 941. Al eliminar aquellos prefijos con entropía 0, la cuenta final quedó en 174 potenciales prefijos y 171 potenciales sufijos.

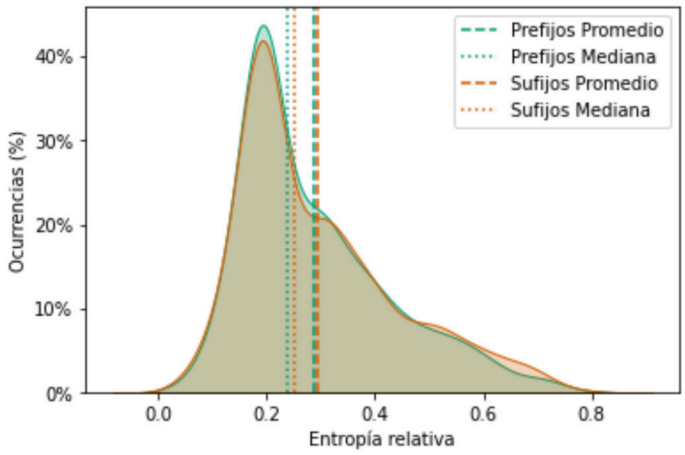
Es interesante notar que las palabras en EXP1 tienen un rango de extensión que oscila entre 1 a 15 grafemas ($\overline{x} = 3.96$; $Me = 4$; $Mo = 3$), mientras que en EXP2 tienen un rango de extensión que oscila entre 3 a 17 grafemas ($\overline{x} = 6.73$; $Me = 6$; $Mo = 6$). Esto apunta a que la forma en que la lingüista analiza la palabra en pa ipai es distinta a la forma en que lo llegó a hacer el hablante, y quien lo haya asistido, al momento de plasmar la palabra gráfica en su lengua. También es interesante notar que en EXP1 contamos con palabras de hasta un grafema. Esto se refiere a la forma *ee,* que una vez homogenizado pasa a *é.* Esta forma en pa ipai se refiere a la conjunción coordinante. Una síntesis de esta distribución interna de grafemas en la palabra se observa en la gráfica 1.



Gráfica 1. Distribución de palabras a partir de la cantidad de grafemas (longitud)

Exploración del primer experimento

Para el primer experimento, la distribución de los afijos segmentados luce de la manera en que se muestra en la gráfica 2. Se hace notar una preferencia en ambas segmentaciones (de prefijos y afijos) por medidas de entropía bajas y media bajas (que oscilan entre 0.1 y 0.4). Son pocos afijos los que obtienen medidas que sobrepasan el 0.5 de entropía relativa. También es interesante notar que, en ambos casos, la distribución es prácticamente la misma. Esto se debe a que los cortes en ambas direcciones tienen la misma variabilidad de inventario de grafemas posible. No obstante, nada garantiza que el corte *az-* obtenga la misma entropía relativa que el corte *-az*. Siguiendo con este ejemplo, *az-* obtuvo 0.16 de entropía mientras que *-az* llegó a 0.28.



Gráfica 2. Distribución de afijos en EXP1 de acuerdo con su entropía relativa

Exploración de prefijos de EXP*1*. Los niveles más altos de entropía los alcanzan los afijos presentes en la tabla 4.

Tabla 4. Entropías relativas más altas en EXP1

Prefijo	*Entropía relativa*	*Entropía*
i-	0.79	3.97
u-	0.73	3.67
a-	0.72	3.62
um-	0.72	3.61
$-	0.72	3.61
ki-	**0.71**	**3.5**
b-	0.70	3.5
i$-	0.70	3.5
o-	0.69	3.49
mi-	**0.68**	**3.43**

El prefijo *i-* es interesante destacarlo ya que es el que se utiliza en partes del cuerpo como *iway* 'corazón' pero que no tiene un valor posesivo como el prefijo *ɲ-* (Joël 1966; Sánchez-Fernández 2016). Esta forma también aparece en nombres para plantas, e incluso se ha sugerido que es un clasificador que tiene su origen en la forma *ʔiʔi* 'rama' (Silver 1974). En cuanto a los otros casos, es interesante notar las formas *mi-* y *ki-*. En pa ipai, la *m-* expresa la segunda persona, ya sea como marca en el verbo de objeto directo o como sujeto, independientemente de si en el evento que formaliza la oración está involucrada una tercera o primera persona; esto contrasta con la primera persona, cuyas formas sí varían de acuerdo con esta relación entre *ɲ-* y *ʔ-*. De las segmentaciones, 62 (9.4%) comienzan con *m*.

Aunque no aparece en los primeros diez prefijos con entropía más alta, también se buscó la entropía de la forma *ñ-*. En pa ipai, esta forma es la que indica la primera persona y es marca de posesivo en nominales. De los 659 segmentos, 93 (14%) inician con *ñ*. Sólo la forma *ñ-* alcanzó un 0.59 de entropía relativa, junto con formas como *ñi-* que tiene un 0.64 de entropía relativa. Para el caso de la *k-,* 92 (13.9%) de los prefijos segmentados comienzan con esta forma. Este prefijo es muy productivo en la lengua para indicar oraciones relativas o encontrarse en linde entre el verbo principal y el verbo auxiliar. Su ubicuidad podría reafirmar su carácter como afijo. Los anteriores datos cubren alrededor del 50% de los resultados. Como exploración, se aconseja que, a menos que se tengan preguntas particulares sobre afijos, hasta este momento podría ser suficiente la búsqueda. Por otro lado, debido a que no se tiene aún un parámetro para sospechar que, por ejemplo, la forma *o-* en pa ipai es un prefijo (dado que tiene una de las entropías relativas más altas), no podemos continuar con el análisis. Es decir, por un lado, tenemos el ejercicio de comparar lo que ya se sabe de la prefijación en esta lengua, y por otro, el que las entropías nos sugieran afijos que tal vez en la investigación de esta lengua no estaban contemplados. Esto se retomará para EXP2, por lo pronto, se continuará con la exploración de los sufijos.

Exploración de sufijos de EXP*1*. Los diez sufijos con entropía más alta en el análisis se muestran en la tabla 5.

Tabla 5. Entropías relativas más altas en EXP1

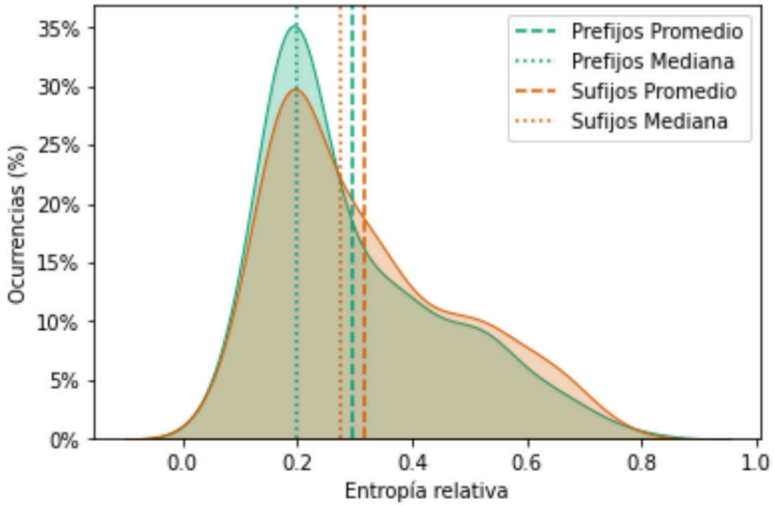
Sufijo	Entropía relativa	Entropía
-i	0.79	3.96
-a	0.74	3.70
-ik	0.71	3.58
-á	0.71	3.56
-u	0.71	3.55
-’	0.70	3.53
-il	0.69	3.49
-um	**0.69**	**3.49**
-p	0.69	3.46
-ma	**0.69**	**3.45**

De la misma manera que en la exploración con los prefijos, para los sufijos se encuentran ciertas agrupaciones que, al comparar a lo que se sabe de la lengua pa ipai, parecen tener coherencia. Por ejemplo, la forma *m* aparece en dos casos en la tabla 5, además de encontrarse en 100 de los 680 posibles sufijos (14.70%). La *-m* como sufijo, junto con la *-k*, son parte de un conjunto de sufijos polisémicos. Expresan el cambio de referencia (*switch-reference*), así como el caso locativo en frases nominales (*-k*), y el mediativo y comitativo (*-m*). Si revisamos la distribución de la *-k*, parece extenderse también en los

sufijos analizados (90 casos, 13.23%). Es interesante destacar dos sufijos de la tabla 3. Primero, está la forma *-ik*, que es muy productiva en estilos indirectos del tipo "alguien dijo que…" o "se dice que". El otro es el cierre glotal *-'* (/ʔ/) que puede vincularse con el caso de acusativo. De la misma manera que con los prefijos, esta exploración puede ayudar a reconocer formas que tal vez no se tenían contempladas como afijos. Es el caso de la *-p*; o confirmar sospechas como son el caso del prefijo *-a* (de acuerdo con Joël, la marca de *irrealis*) o la *-u* (sufijo deverbal para indicar que algo sirve para realizar la acción que describe el verbo) (Joël 1966).

Contraste y análisis con Joël Judith

El procedimiento para el análisis de EXP2 fue un poco distinto a la exploración de EXP1. A diferencia de los textos de Benito Peralta, con Joël sí tenemos una forma de evaluar las segmentaciones sugeridas y correlacionar si la entropía relativa nos sirve para sospechar de lindes morfológicos. Para ello, se procedió en un principio de la misma manera que el análisis anterior. Se muestra una síntesis de lo encontrado en la gráfica 3.



Gráfica 3. Distribución de afijos en EXP2 de acuerdo con su entropía relativa

En la gráfica se observa, por un lado, que en este caso las entropías para los prefijos sugeridos no coinciden tanto como lo hicieron para los textos de Benito Peralta. También nos permite notar que hay mayor dispersión, aunque sigue habiendo una preponderancia en las entropías relativas de medias a bajas. De nuevo, parece que la mayoría de los datos se concentran entre 0.2 y 0.4 de entropía relativa. La exploración visual entre las gráficas de distribución de entropía relativa (gráfica 2 y 3) nos permitiría arrojar la hipótesis de que entropías media bajas son indicadores de posibles lindes para prefijos y sufijos. No obstante, sólo con esto, es muy arriesgada tal aseveración. Para ello se

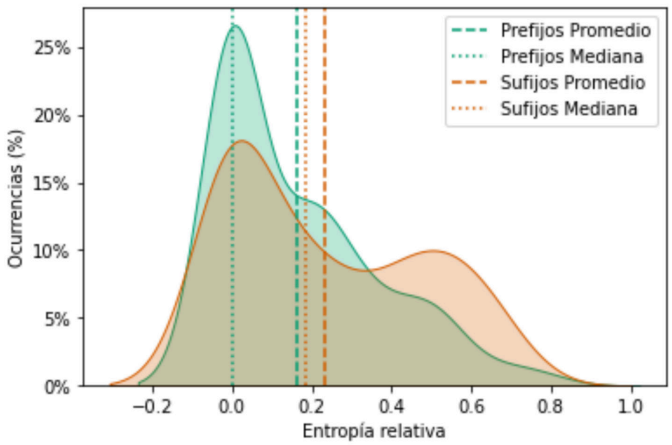
propone un análisis de varianza y uno de correlación para corroborar si tales cortes se relacionan significativamente con esta medida.

La segmentación de Joël. Para este experimento, se retomó el texto de Joël (1976) en su texto y, a partir de los afijos que ella glosa, se creó un inventario, distinto al propuesto por el algoritmo. En contraste con los 174 potenciales prefijos y 171 sufijos del algoritmo, Joël propone 131 prefijos y 99 sufijos, para un total de 230 afijos analizados. De estos, sólo 93 (39.7%) tienen un valor de entropía mayor a 0. Es relevante notar lo anterior, ya que el alcance del experimento no incluye la identificación de morfemas que funcionen como bases. Casos de entropía 0 podrían revelar este tipo de morfemas, pensando en que, dentro de la palabra, la cantidad de morfemas con estructuras inagrupables correspondería a más del 60%, mientras que aquellos que pueden agruparse (como, por ejemplo, las variaciones *-um*, *-em*, *-m*) corresponderían al 40% restante. La exploración de este hecho sobrepasa el objetivo de la presente investigación, pero se dejará para futuros trabajos.

Posterior a esta segmentación y al cálculo de entropía relativa, se emparejaron los 230 afijos analizados por Joël con la entropía calculada por el algoritmo, incluso si fue cero. Esto debido a que el peso en este momento de la experimentación es el hecho de que son afijos que, en efecto, fueron analizados por Joël.

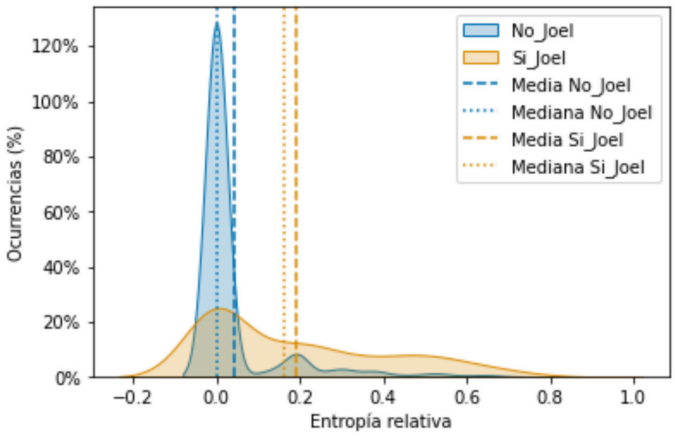
En este punto surgió otro problema que ya había sido mencionado líneas arriba. Existen en la lengua ciertos morfemas que no toman como pivote el extremo derecho o izquierdo de la palabra. En el corpus, son morfemas que sólo aparecen entre otros morfemas. Estos afijos no son segmentados por el programa, pero sí por Joël. De los 230 afijos glosados, el programa no calculó la entropía de 50. Ahora, nótese que no son 50 casos únicos. Por ejemplo, aparece el morfema *oli*, el cual es propuesto como prefijo y sufijo (*oli-*, *-oli*), debido a que aparece sólo entre morfemas en la segmentación de Joël. Por lo que esos 50 casos, podrían ser aproximadamente 25 formas por explorar. De la misma manera que el primer problema enunciado, este escenario se salva retirando de la exploración los casos que no fueron analizados por el programa, pero se dejarán para otras exploraciones los casos de posibles intrafijos —afijos que van entre la base y algún otro sufijo o prefijo— tomando como criterio el análisis de Joël, o incluso, el de otras glosas hechas por otros investigadores.

Correlación. Al final, el conjunto para realizar la correlación corresponde a 180 afijos que sí fueron segmentados por Joël, y otros 1690 afijos que el programa segmentó; para ambos grupos tenemos su cálculo de entropía que incluye también los casos de entropía 0. A manera de contraste, en la gráfica 4 se muestra la distribución de los 180 afijos segmentados tomando como parámetro su entropía relativa calculada. Nótese la diferencia con respecto a las distribuciones anteriores (gráfica 2 y 3). Se observa que las curvas no están tan homogenizadas como en los otros dos casos anteriores. Además, también sobresale que la tendencia de los datos de los sufijos es a ser bimodal (dos crestas en el gráfico). Esto podría sugerir una división entre los sufijos. Esto se deja para evaluación en futuros trabajos.



Gráfica 4. Distribución de afijos en EXP2, filtrado por afijos segmentados por Joël, de acuerdo con su entropía relativa

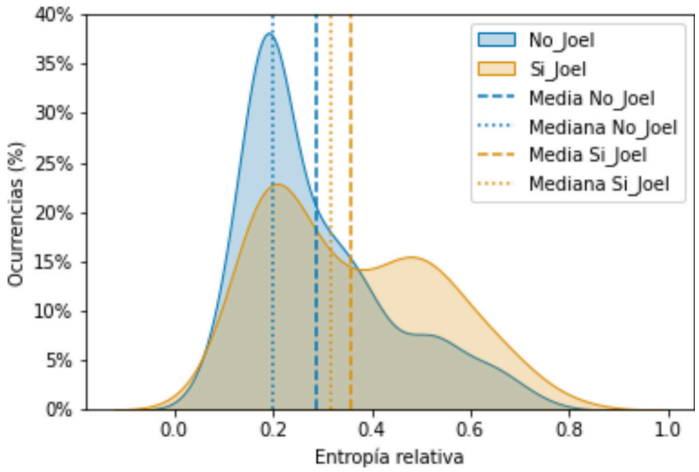
Una vez emparejados los datos, se creó una nueva tabla en donde se estableció como variable dependiente si Joël segmentó el afijo (Si_Joel) o no (No_Joel), es decir, asociamos una variable binaria. Como variable independiente se asoció el cálculo de entropía. La distribución de ambos grupos de afijos se muestra en la gráfica 5.



Gráfica 5. Distribución de afijos en EXP2 de acuerdo con su entropía relativa

Con esta base de datos fue posible realizar el análisis de varianza (ANOVA), el cual arrojó un estadístico F de 210.67, con un valor *p* de 2.59^{-45}. Debido a que $p < 0.05$, se descarta la hipótesis nula, es decir, sí existe una diferencia entre los dos conjuntos, implica que sí es relevante agrupar la entropía con respecto a afijos determinados por la lingüista. Sin embargo, resulta problemático realizar esta prueba con grupos que tienen varianzas tan poco homogéneas. Este mismo ejercicio se repitió, pero quitando los ceros. Este nuevo ANOVA arrojó un F de 13.78, con un valor *p* de 0.0002 ($p < 0.05$). A juzgar

por las distribuciones en el gráfico 6, se observa que, en este caso, las varianzas tienden a ser más homogéneas, lo que permite seguir sosteniendo que la división de los grupos fue pertinente.



Gráfica 6. Distribución de afijos en EXP2 de acuerdo con su entropía relativa sin incluir entropía 0

Para evaluar la fuerza de la correlación entre el cálculo de entropía y la segmentación de Joël, se integró en Python la paquetería Scipy (*scipy.pointbiserialr*), con la que se realizó una prueba de correlación biserial puntual. Se encontró una correlación significativa entre la variable Segmentado de Joël y el cálculo de entropía relativa ($r\_pb = 0.31$, $p < 0.05$, $n = 1870$). Según el coeficiente de correlación de biserial puntual, se observó una relación moderadamente positiva entre las variables, lo que indica que, a medida que la entropía relativa aumenta, también lo hace la probabilidad de que una unidad sea un afijo, según el criterio de Joël.

CONCLUSIONES

El objetivo del presente trabajo fue implementar el cálculo de entropía de la información ($H(X)$) a palabras en lengua pa ipai y evaluar su capacidad para sugerir límites morfológicos. La exploración inicial de los textos (EXP1) permitió identificar segmentos relevantes y sugerir la agrupación de aquellos con entropía relativa mayor a 0 basada en los grafemas de inicio o fin. En la exploración posterior con el texto glosado (EXP2) se propuso una evaluación mediante análisis de varianza (ANOVA) que resultó significativa, así como una correlación biserial puntual que demostró una relación moderada entre el cálculo de entropía y los cortes propuestos por la lingüista Judith Joël. En consecuencia, se cumple con el objetivo planteado y se confirma la hipótesis inicial (H1) de que existe una correlación significativa entre los cortes propuestos de los afijos y su cálculo de entropía.

A medida que se realizaban estas experimentaciones, surgieron tres desafíos que deberán abordarse en futuras investigaciones. El primero implica un trabajo enfocado al concepto de palabra en pa ipai, lo que permitiría una mayor comprensión de su significado. Esto contribuiría en la aplicación de técnicas de lingüística cuantitativa para analizar cómo se ha expresado la palabra a lo largo de diferentes textos y autores. El segundo desafío se relaciona con la falta normalización de un sistema de escritura en pa ipai. Si bien hay esfuerzos en curso para establecer distintos sistemas de escritura, será necesario describir este proceso desde una perspectiva lingüística, antropológica y sociológica en el futuro. El tercer desafío se refiere a la necesidad de incorporar la segmentación de intrafijos en el corpus, un aspecto que inicialmente se pasó por alto debido a los objetivos del proyecto. Sin embargo, la naturaleza de la lengua y las segmentaciones sugeridas por Joël hicieron evidente el abordar este aspecto en una investigación futura.

A manera de cierre, se requiere más investigación en lingüística cuantitativa de todas las lenguas yumanas en términos más amplios. A diferencia de otros idiomas originarios en el centro de México, las lenguas yumanas han recibido muy poca atención en cuanto a la aplicación de tecnologías de lenguaje. Como resultado, un futuro proyecto derivado de esta investigación podría implicar el uso del cálculo de entropía en otras lenguas de la familia y, en última instancia, de la región. Para lograr esto, será necesario construir corpus digitales accesibles que puedan ser utilizados en algoritmos informáticos. Se tiene planeado que, para facilitar futuras investigaciones, se construya un archivo digital que contenga los programas utilizados en este experimento y los textos en sus distintas versiones de tratamiento. Ello permitiría a otros investigadores construir sobre este trabajo y seguir avanzando en el campo de la lingüística computacional en las lenguas yumanas.

BIBLIOGRAFÍA

ACKERMAN, Farrell y Robert MALOUF. 2013. "Morphological organization: The low conditional entropy conjecture", *Language* 89, núm. 3: 249-64. <DOI: 10.1353/lan.2013.0054>

BARRIGA VILLANUEVA, Rebeca. 2001. "Oralidad y escritura: una encrucijada para las lenguas indígenas", *Caravelle*, núm. 76/77: 611-21.

BRIGHT, William. 1992. *A Coyote Reader*. Los Angeles: University of California Press.

CARBAJAL, Norma. 2002. *Misión de Santa Catarina*. Ensenada: Instituto Nacional Indigenista, Delegación Baja California.

CORTÉS RODRÍGUEZ, Edna Alicia. 2013. *Conocimiento tradicional herbolario pa ipai y perspectiva de desarrollo local en Santa Catarina, B.C.,* tesis de doctorado. Mexicali: Universidad Autónoma de Baja California.

EMBRIZ OSORIO, Arnulfo y Óscar ZAMORA ALARCÓ (eds.). 2012. *México. Lenguas indígenas nacionales en riesgo de desaparición: variantes lingüísticas por grado de riesgo. 2000*. México: Instituto Nacional de Lenguas Indígenas.

GÓMEZ SERNA, Ana María. 2010. *Proyecto de documentación inicial de la lengua paipái*. Ciudad de México: Instituto Nacional de Lenguas Indígenas.

GONZÁLEZ CASTILLO, Ivette Selene. 2020. *Propuesta metodológica para la documentación lingüística de la lengua paipai*, tesis de maestría. México: Universidad de Sonora.

GONZÁLEZ CASTRO, Armandina, Noboru TAKEUCHI, Manuel Alejandro SÁNCHEZ-FERNÁNDEZ y Nina Alejandra MARTÍNEZ ARELLANO. 2016. *Números y cielo paipai, Chribchu ee myaa paipai*. Ensenada, B.C.: Universidad Nacional Autónoma de México.

HERNÁNDEZ CAMPOY, Juan Manuel y Manuel ALMEIDA. 2005. *Metodología de la investigación sociolingüística*. Granada, España: Editorial Comares.

IBÁÑEZ BRAVO, María Elena. 2015. *Descripción fonológica de la lengua pa'ipá:y*, tesis de licenciatura. México: Escuela Nacional de Antropología e Historia.

JOËL, Dina Judith. 1964. "Classification of the Yuman languages", en William Bright (ed.), *Studies in California linguistics*. Berkeley: University of California Press, pp. 99-105.

JOËL, Dina Judith. 1966. *Paipai phonology and morphology*, tesis de doctorado. Los Ángeles: University of California, Los Ángeles.

JOËL, Dina Judith. 1976. "The earthquake of '57: a Paipai text", *International Journal of American Linguistics. Native American Texts Series* 1, núm. 3: 84-91.

JURAFSKY, Daniel y James H. MARTIN. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. Upper Saddle River, N.J.: Pearson Prentice Hall.

LANGDON, Margaret. 1970. "Review: A comparative study of Yuman consonantism. By Alan Campbell Wares", *Language* 46, núm. 2: 533-44. <DOI: 10.2307/412302.>

LANGDON, Margaret. 1975. "Boundaries and lenition in Yuman languages", *International Journal of American Linguistics* 41, núm. 3: 218-33.

LARA, Luis Fernando. 2015. *Curso de lexicología*. Ciudad de México: El Colegio de México.

MARTÍNEZ ARELLANO, Nina Alejandra, Ana Bertha URIBE y Manuel Alejandro SÁNCHEZ-FERNÁNDEZ. 2013. "Decir el tiempo. Orientaciones temporales a través de la lengua en la cultura Pa ipai", en Heriberto Cairo Carou y Lucila Finkel Morgenstern (eds.), *Memorias del XI Congreso Español de Sociología. Crisis y cambio: propuestas desde la sociología*. Madrid: Universidad Complutense de Madrid.

MAGER, Manuel, Diónico CARRILLO e Ivan MEZA. 2018. "Probabilistic finite-state morphological segmenter for the Wixarika (Huichol) language", *Journal of Intelligent & Fuzzy Systems* 34, núm. 5: 30813087.

MAGER, Manuel, Ximena GUTIERREZ-VASQUES, Gerardo SIERRA e Ivan MEZA. 2018. "Challenges of language technologies for the indigenous languages of the Americas", *ArXiv Preprint ArXiv:1806.04291*, junio. <DOI: 10.48550/arXiv.1806.04291>

MEDINA URREA, Alfonso. 2021. *El signo afijal en la muestra textual*. Ciudad de México: El Colegio de México.

MITHUN, Marianne. 1986. "On the nature of noun incorporation", *Language* 62, núm. 1: 32-37.

MOSELEY, Christopher (ed.). 2010. *Atlas of the World's Languages in Danger*. 3ra ed. Paris: United Nations Educational, Scientific and Cultural Organization, en <http://www.unesco.org/culture/en/endangeredlanguages/atlas> [consultado el 23 de marzo del 2023].

OSVALDO PORTA, Andrés. 2010. "The use of formal language models in the typology of the morphology of Amerindian languages", *Proceedings of the ACL 2010 Student Research Workshop*: 109-114, en <https://aclanthology.org/P10-3019> [consultado el 23 de marzo del 2023].

OWEN, Roger C. 1963. "The use of plants and non-magical techniques in curing illness among the Paipai, Santa Catarina, Baja California, Mexico", *América Indígena* 23: 319-44.

PERALTA, Benito. 1994. *Relatos Pai pai. Kurit' trab pai pai.* México: Consejo Nacional para la Cultura y las Artes. Lenguas de México 1.

RAO, Rajesh P. N., Nisha YADAV, Mayank N. VAHIA, Hrishikesh JOGLEKAR, R. ADHIKARI y Iravatham MAHADEVAN. 2009. "Entropic Evidence for Linguistic Structure in the Indus Script", *Science* 324 (5931): 1165-1165. <DOI: 10.1126/science.1170391.>

SÁNCHEZ-FERNÁNDEZ, Manuel Alejandro. 2016. *Deixis espacial y demostrativas de la lengua paipai*, tesis de maestría. México: Universidad de Sonora.

SÁNCHEZ-FERNÁNDEZ, Manuel Alejandro. 2022. "La investigación lingüística de las lenguas yumanas en México (LYUM)", *Expedicionario. Revista de estudios en Antropología* 2, núm. 4: 31-43.

SÁNCHEZ-FERNÁNDEZ, Manuel Alejandro y Luis Miguel ROJAS-BERSCIA. 2016. "Vitalidad lingüística de la lengua paipai de Santa Catarina, Baja California", *LIAMES - Línguas Indígenas Americanas* 16, núm. 1: 157-183. <DOI: 10.20396/liames.v16i1.8646171.>

SHANNON, C. E. 1948. "A mathematical theory of communication", *Bell System Technical Journal* 27, 3: 379-423. <DOI: 10.1002/j.1538-7305.1948.tb01338.x.>

SILVER, Shirley. 1974. "Some Northern Hokan plant-tree-bush forms", *The Journal of California Anthropology* 1, núm. 1: 102-9.

URDAN, Timothy C. 2005. *Statistic in Plain English*. New Jersey: Lawrecen Erlbaum Associates, Publishers Inc.