Características cuantitativas de la flexión verbal del Chuj

Alfonso Medina
Grupo de Ingeniería Lingüística
Instituto de Ingeniería, UNAM
Elsa Cristina Buenrostro
Instituto de Investigaciones Antropológicas, UNAM

This paper presents the application of two word-segmentation methods — measurement of information content or entropy (Shannon and Weaver 1949) and that of the economy of signs (de Kock and Bossaert 1974, 1978)— to a small corpus of Chuj, a Mayan language spoken in Chiapas and Guatemala. The motive for this is to determine whether or not it is possible to discover automatically some set of inflexional affixes, in spite of the small size of the corpus. Results show that information content is adequate for the discovering of about 83% (precision measure) of verbal inflexional affixes, whereas application of the economy principle requires a greater sized corpus.

En este trabajo se presenta la aplicación de dos métodos de segmentación automática de palabras —contenido de información o entropía (Shannon y Weaver 1949) y el principio de economía de los signos (de Kock y Bossaert 1974; 1978)— en un corpus pequeño del chuj, una lengua maya hablada en Chiapas y Guatemala. El objetivo es determinar si es posible descubrir automáticamente, a partir de un corpus reducido, por lo menos algunos sufijos de flexión verbal. Los resultados indican que el contenido de información es una medida suficientemente fina para llevar a cabo este objetivo con un 83% de éxito (porcentaje de precisión) para los afijos de flexión verbal, mientras que la medida de economía requiere de un corpus de mayor tamaño.

Palabras clave: *morfología, chuj, afijos, corpus, entropía, economía.*
Fecha de recepción del manuscrito: marzo del 2003

Alfonso Medina.
Grupo de Ingeniería Lingüística, Instituto de Ingeniería, UNAM
Torre de Ingeniería, Cubículo 3, Basamento, Circuito Interior, 04510 México D. F.
correo electrónico: amedinau@iingen.unam.mx.

Elsa Cristina Buenrostro.
Instituto de Investigaciones Antropológicas
UNAM, 04510 México D. F.
correo electrónico: cristina_buenrostro@hotmail.com

Introducción

Dentro de los trabajos de lingüística automática existen en realidad pocos métodos de descubrimiento de morfemas mediante computadoras, aquellos que casi siempre se han aplicado a lenguas muy conocidas de origen indoeuropeo, especialmente el inglés. La riqueza de lenguas de América no ha llamado la atención de los lingüistas que utilizan máquinas para investigar al lenguaje. Sin duda uno de los mayores obstáculos para aplicar estos métodos en estas lenguas es la recolección de corpus apropiados. La recopilación de tales herramientas de investigación científica es muy costosa (en tiempo y recursos de todos tipos) y raras veces recibe el reconocimiento que merece. Por eso los lingüistas a menudo atesoran los textos que ellos mismos obtienen de su interacción con los hablantes. Por si fuera poco, los corpus que logran reunir con tanto trabajo son de tamaño tan reducido que se cuestiona, con razón, su carácter de muestras estadísticamente representativas de las lenguas en cuestión, en especial al compararlos con aquellos corpus gigantescos que se utilizan en los grandes proyectos de lingüística automática del mundo.

En este trabajo se presentan los resultados de la aplicación de algunos métodos de segmentación automática de palabras a un corpus pequeño de una lengua maya. La idea es *descubrir* automáticamente los afijos de flexión de esa lengua, es decir, identificarlos mediante sus características cuantitativas sin la intervención del ojo humano. Luego, para evaluar este procedimiento, se comparan los resultados con la información proporcionada por el especialista.

Antecedentes

Fue en el marco del distribucionalismo que Zellig Harris examinó en corpus de diversas lenguas las evidencias formales de las fronteras entre morfemas (Harris 1955). Su método consiste en determinar el número de fonemas que preceden o siguen una frontera morfológica hipotética. A mayores cuentas de fonemas, más morfológica es dicha frontera. Este método inspiró muchos procedimientos para segmentar el discurso. Aunque Harris lo aplicó a varias lenguas, los procedimientos automáticos basados en éste y en otros métodos se ocuparon predominantemente de la lengua inglesa (véase, por ejemplo, Hafer y Weiss (1974)).

La primera aplicación propiamente computacional para descubrir fronteras entre morfemas de lenguas no sólo indoeuropeas fue quizá la dirigida por el ruso N. D. Andreev en los años setenta (Cromm 1996). Además del descubrimiento de fronteras morfológicas, este método se diseñó para determinar automáticamente los paradigmas de flexión en el que intervienen los afijos descubiertos. El criterio para identificarlos es básicamente la frecuencia de las secuencias de caracteres en los corpus analizados (los afijos de flexión son los más frecuentes). El procedimiento se aplicó al vietnamita y al húngaro, entre otras lenguas.

Hoy en día, la necesidad de segmentar palabras se hace patente especialmente en aplicaciones relacionadas con recuperación de información (*information retrieval*). En este

marco, cada vez más lenguas son objeto de diversos procedimientos de segmentación morfológica, especialmente las indoeuropeas y las de los pueblos industriales de la cuenca del Pacífico [1].

Típicamente, cuando todos estos métodos se hacen operativos mediante computadoras (en lugar de contar manualmente, por ejemplo, fonemas anteriores y posteriores, como en algún momento tuvo que hacerlo Harris), se requieren grandes cantidades de texto. Si bien en los inicios de los corpus electrónicos, se consideraba que aquellos de uno o dos millones de palabras (como el Corpus Brown o el Corpus del Español Mexicano Contemporáneo) eran de buen tamaño, los corpus de hoy en día suelen ser mucho más extensos, esto es, de varios millones de palabras gráficas (Manning y Schütze 1999:118-120).

En este marco, la riqueza de lenguas de México no ha sido motivación suficiente para que estas lenguas se investiguen de manera automática, ni para construir modelos computacionales de análisis o síntesis, ni para estudiar sus características cuantitativamente a partir de corpus. Sin duda, como se apuntó arriba, uno de los mayores obstáculos en la aplicación automática de métodos cuantitativos en estas lenguas es la recolección de corpus de tamaño apropiado. Sin embargo, todavía no está claro lo que significa "tamaño apropiado".

Corpus

El chuj es una lengua maya que se habla en ambos lados de la frontera entre México y Guatemala. El corpus utilizado en este trabajo (Buenrostro 2000) fue compilado en diversas estancias de trabajo de campo en el estado de Chiapas. Se trata de una colección de cinco narraciones [2], todas con intercambios conversacionales, por lo que de entrada no podemos hablar ni de un corpus balanceado ni representativo, véanse discusiones sobre esto en Lara (1990:85-106) y Manning y Schütze (1999:119-120). Además, dadas las características de los corpus electrónicos de hoy en día, podemos calificar a este corpus del chuj de minúsculo: en un archivo plano de sólo 86Kb caben alrededor de 15,485 palabras gráficas, las que corresponden a poco más de 2,300 vocablos (o tipos de palabras).

De todas maneras, mediante este corpus se pueden examinar las características cuantitativas de los morfemas de flexión, que esperaríamos formaran parte de un conjunto pequeño y muy regular de segmentos relativamente fáciles de descubrir automáticamente (incluso a partir de un corpus reducido). En cuanto a los afijos de derivación, que deben formar parte de un conjunto menos organizado y más irregular, se puede esperar que sea más difícil descubrirlos automáticamente en un corpus tan pequeño.

Lo importante es que al determinar automáticamente algunos sufijos de flexión y corroborar su condición de morfemas del chuj, podremos evaluar los resultados y deter-

[1] Estas últimas con escritura no alfabética; véase por ejemplo Kageura (1999) para la aplicación de estadísticas de diagramas en la segmentación de cadenas de caracteres Kanji.

[2] Los títulos de las narraciones son "María", "El alcalde", "A nok' chich yet' nok' okes", "Te' chum" y El éxodo".

minar si un corpus así de reducido es suficiente por lo menos para descubrir automáticamente la morfología flexiva de una lengua.

Metodología

En este trabajo se aplicaron dos métodos para determinar cuantitativamente la mejor frontera entre bases y afijos de cada una de las palabras del corpus. El primero es el cálculo de entropía, uno de los temas de la teoría de la información (Shannon y Weaver, 1949). El segundo está inspirado en el principio de economía de signos (de Kock y Bossaert 1974, 1978).

Pertinencia de las medidas de entropía y economía

Probablemente fue Joseph Greenberg el primero en reflexionar sobre la cantidad de información (en el sentido técnico de la teoría de la información) como característica distintiva de las raíces de palabras:

> both in the technical sense of information theory and in the nontechnical meaning of information, the utterance of a member of a root class of morphemes gives more information (Greenberg 1957: 91).

Desde entonces se ha reportado repetidamente que medir la cantidad de información o entropía asociada a los radicales de palabras es un método más o menos exitoso a la hora de determinar fronteras entre bases y afijos (véanse, por ejemplo, Hafer y Weiss (1974), Frakes y Baeza (1992), Medina Urrea (2000) y el reporte del trabajo de Joula, Hall y Boggs en Oakes (1998: 86-87)).

Conceptualmente, las cantidades de entropía corresponden a los altibajos de información que formalmente puede esperar un lector u oyente al leer un texto o escuchar la cadena hablada. La conocida fórmula de Shannon, presentada abajo, es un método muy popular para medir el contenido de información o entropía. Sin embargo, esos altibajos de información no siempre corresponden con las fronteras morfológicas en la cadena escrita o hablada. La situación pragmática, la posición del cuerpo, las manos, las interrupciones, muecas, gestos, sonrisas, etc. también proporcionan información importante. Es decir, hay comunicación (y por lo tanto entropía) sin estructuras propiamente lingüísticas. Por otra parte, sabemos que en un corpus de lengua natural existe implícita una estructura de signos que también debe servir como evidencia de las fronteras entre morfemas.

Por eso es pertinente el principio de economía de los signos. Si el sistema lingüístico es económico, podemos esperar que las relaciones de economía entre los signos nos proporcionen indicios sobre la estructura que sirve de vehículo para la transmisión de información. Una manera de concebir el concepto de economía es considerar la propiedad de ciertos signos (afijos) de combinarse con otros (bases) para producir un número

virtualmente infinito de signos del nivel siguiente (palabras). Así, el número de signos debe ser menor que el número de cosas nombradas sin que se produzca ambigüedad alguna (de Kock y Bossaert 1978: 15).

Los afijos permiten precisamente eso: al combinarse con las bases forman palabras nuevas (tanto los lemas de un diccionario, como las palabras flexionadas del discurso); además, las bases se pueden combinar con otros afijos, sin que el nuevo contexto resulte en ambigüedad. Es claro que los afijos no se combinan con cualquier base, unos con más, otros con menos, pero tiene sentido esperar que a mayores posibilidades combinatorias, mayor economía de signos y mayor su cualidad de ser afijos. Además, si los afijos forman conjuntos que al combinarse con las bases alternan con otros afijos (paradigmáticamente), sus relaciones tendrán que considerarse todavía más económicas. Esto es pertinente tanto para afijos derivativos como flexivos; es decir, tanto para los afijos típicos de una lista de lemas, como para aquellos de las palabras flexionadas del discurso. Si bien en un corpus como el utilizado en este experimento los primeros serán más escasos que los segundos, ambos tipos guardan la misma relación con las bases a las que se unen, aunque quepa esperar que la de los segundos sea más económica (porque son menos, más frecuentes y se adhieren a muchas más cosas).

Simplificando, los afijos pueden concebirse como un conjunto pequeño de morfos (formas que los morfemas exhiben en el habla/escritura) muy frecuentes que se combinan con otros tipos de morfos; específicamente, raíces y bases. En contraste, estas últimas constituyen un grupo enorme (potencialmente infinito) de morfos de baja frecuencia en un corpus. Así, el número de afijos multiplica el número potencialmente infinito de bases para crear nuevas maneras de referirse a las cosas del mundo. Esto es, mientras menos signos sirvan para designar más cosas, más economía habrá en el sistema.

Si bien el contenido de información o entropía ya ha sido reconocido como indicio de frontera morfológica apto de aplicarse a la segmentación automática de morfemas, la razón de incorporar otro método como el de economía es muy sencilla. Por un lado, la entropía no distingue entre las ocurrencias de un afijo como morfema y las ocurrencias de la mera secuencia de caracteres que constituyen su forma. Así, en la palabra española 'aumente', el método de entropía *descubre* erróneamente el sufijo adverbial ***-mente***. Por otro lado, el método basado en la economía de signos propone acertadamente un sufijo ***-e***.

Cálculo de índices de entropía y economía

El contenido de información de un grupo de fragmentos de palabras se mide típicamente mediante la fórmula siguiente[3]:

$$H(p_1, p_2, ..., p_n) = -\sum_{i=1}^{n} p_i \log_2 (p_i) \qquad (1)$$

el corpus. La tendencia de ese fragmento particular a ser afijo residiría en que sus *acompañantes* (en relación sintagmática) fueran considerablemente muchos más que sus *alternantes* (en relación paradigmática). Esto se puede medir, por ejemplo, mediante la siguiente fórmula que es una simplificación de aquella propuesta por de Kock (Medina Urrea 2003: 280):

$$k = \frac{acompañantes}{alternantes}$$

De esta manera, dado el fragmento de un extremo de alguna palabra gráfica, un número muy grande de *acompañantes* y un número reducido de *alternantes* resultaría en una medida alta de economía, mientras que un número reducido de *acompañantes* y uno alto de *alternantes* de dicho fragmento indicaría una medida de economía baja y, por lo tanto, una reducida probabilidad de que represente un morfema. Naturalmente, el número de acompañantes a la derecha es diferente al de la izquierda. Además, el número de alternantes varía según el número de acompañantes, por lo que, al igual que con la entropía, hay dos valores distintos para un mismo fragmento según se tome en cuenta lo que le sigue o le antecede.

Un ejemplo simplificado del español sería el sufijo derivativo ***-idad*** (en, por ejemplo, 'nacional**idad**') que, por un lado, alterna con el morfema nulo, **-Ø**, (en 'nacional'), con el sufijo del plural ***-es*** ('nacional**es**') y con la secuencia de dos sufijos ***-idades*** ('nacional**idades**'); y, por el otro, se sufija a una gama enorme de adjetivos, que en sí constituyen una clase abierta (potencialmente infinita). En este contexto, la medida de economía correspondería al número de palabras de un corpus con el sufijo ***-idad*** (el número de signos que lo acompañan a la izquierda) divido entre el número de signos con que alterna en el corpus (en este caso cuatro). Eso con respecto a la izquierda del sufijo. Por otra parte, con respecto a la derecha de otro segmento, considérese el vocablo 'nacionalidad'. Diríamos que el fragmento ***nacional-*** alterna con un número enorme de formas (en '**comun**idad', '**oportun**idad', '**vorac**idad', '**final**idad', etc.) y que, como base morfológica, se le afijan pocas cosas: además de ***–idad*** y la secuencia ***–idades***, están muchas veces los sufijos ***-mente*** (adverbial), ***-es*** (plural) e, incluso, el morfema nulo, **-Ø**. Esto corresponde a una medida ***k*** muy reducida que resulta de dividir un número relativamente pequeño (5) entre uno considerablemente mayor (el número de formas en ***–idad*** del corpus examinado). De todo esto tendríamos que concluir que 'nacional' (en 'nacional**idad**') no puede ser un afijo.

Lo importante no es cuál de estos dos métodos es el mejor, sino que ambos se pueden combinar para obtener una estimación de la *afijalidad* de un afijo con respecto a una base determinada. Por ejemplo, las cantidades que resulten de estos procedimientos se pueden multiplicar o se pueden promediar. En este experimento, como veremos adelante, los valores resultantes se normalizaron y luego se promediaron.

Conviene enfatizar que los valores se calculan automáticamente dos veces para cada segmentación de cada palabra, la primera considerando al inicio de la palabra como base y el final como sufijo y la segunda considerando al inicio de la palabra como prefijo y el

final como base. Como vimos arriba y veremos de nuevo adelante, es de notarse que en este procedimiento el fragmento afijal puede contener más de un afijo (por ejemplo, ***-idades***). Lo importante es que los valores más altos, aquellos de segmentaciones entre prefijo y base y entre base y sufijo, sirven como criterios para la inclusión de los fragmentos de palabras examinados en alguno de dos catálogos, uno de prefijos y otro de sufijos de la lengua examinada.

Construcción de catálogos de afijos del chuj

Los métodos presentados arriba permiten examinar automáticamente cada una de las palabras gráficas del corpus. Si la mejor segmentación de cada palabra permite identificar un fragmento de ésta con carácter morfológico (afijo o grupo de afijos), la afijalidad de cada fragmento se puede capturar automáticamente en una estructura de información que aquí llamaremos catálogo, donde se registra la frecuencia de dicho fragmento como afijo o grupo afijal (secuencia de afijos). Los pormenores de una estructura de esta naturaleza se presentan en Medina Urrea 2000 y 2003. Los aspectos importantes son:

- Se construyen dos catálogos diferentes, uno para prefijos y otro para sufijos;
- Cada entrada del catálogo corresponde a un afijo o grupo afijal, con los promedios de los valores de entropía y economía calculados para cada vocablo en el que resultó ser el mejor afijo.
- Estos valores se promedian para obtener un índice de afijalidad que permite ordenarlos de más a menos afijal.
- Todos los valores están normalizados, es decir, son valores entre 0 y 1. Esto se lleva a cabo al dividir el valor obtenido para cada afijo (o grupo afijal) entre el máximo valor obtenido para algún afijo (o grupo de afijal) en la construcción del catálogo.
- Hay varias posibilidades para seleccionar qué afijo o grupo afijal entra a formar parte del catálogo. En este experimento sólo se tomó el fragmento más afijal de cada vocablo del corpus; esto es, sólo aquel con el valor más alto de afijalidad.

Los elementos de estos catálogos se pueden ordenar de diferentes maneras. Para este trabajo se ordenaron por valores de afijalidad, aquí estimada (como se dijo arriba) mediante el promedio de los valores normalizados de entropía y economía inherentes a cada fragmento de palabra. Con este orden, se concentran entre los primeros los elementos más afijales de los catálogos. Además, como es de esperarse, el umbral entre las formas más afijales y las menos afijales no es ni evidente ni claro. De todas maneras, por inspección observamos que dentro de los primeros treinta de cada catálogo se encuentra la mayor concentración de verdaderos afijos y grupos de afijos. De hecho, como veremos en las tablas 3 y 4, que consignan los 30 afijos de flexión verbal del chuj (18 prefijales y 12 sufijales [4]), treinta deben ser más que suficientes. Por eso la siguiente discusión está basada en las primeras treinta formas de cada catálogo.

[4] Sin considerar sus posibilidades combinatorias, es decir los posibles grupos afijales.

Así, para evaluar los resultados se calcularon las medidas de *recall* y de precisión [5]. Para este trabajo la primera indica el porcentaje de aciertos dentro de las treinta formas más afijales de cada catálogo. Aunque no todas son flexivas, fueron consideradas como aciertos cuando se observó que corresponden a alguna forma afijal. Esto es, se obtuvieron formas afijales de todos tipos (afijos y secuencias de afijos; derivativos y flexivos) que se contrastaron con formas residuales (errores o ruido). Aunque en este experimento el objetivo haya sido descubrir los afijos de flexión verbal, la validez de los derivativos dentro del sistema lingüístico impide que los segundos deban considerarse residuales a la hora de evaluar el procedimiento, en parte porque no hay una frontera nítida entre un tipo y el otro, pero sobre todo porque no dejan de constituir un componente definitivamente afijal de la lengua chuj.

Por otra parte, la medida de precisión representa aquí la proporción de formas detectadas automáticamente mediante este procedimiento con respecto al conjunto de afijos de flexión verbal del chuj descubiertos después de años de investigar dicha lengua; esto es, la proporción de afijos consignados en las tablas 5 y 6 que fueron aislados al construir los catálogos del chuj descritos arriba. Evidentemente, la noción de residuo no es pertinente aquí porque se está midiendo lo obtenido frente a lo que debió ser obtenido; es decir, no se trata de un contraste entre aciertos y errores, sino entre aciertos y omisiones.

Catálogo de prefijos

En la tabla 3 se consignan los fragmentos de palabra gráficas más afijales según los métodos descritos arriba. La primera columna muestra el rango de afijalidad (a menor rango, mayor afijalidad). La segunda columna consigna los supuestos prefijos; la tercera el número de vocablos en los que obtuvieron los valores más altos de afijalidad; la cuarta es la proporción de entropía con respecto al máximo obtenido por algún prefijo. En la última se exhibe la afijalidad estimada.

La tabla muestra los 30 fragmentos de palabra más prefijales del corpus. Como ya se dijo, están ordenados por cantidad de afijalidad. Sin embargo, los valores de economía no aparecen porque las bases y afijos no exhibieron relaciones particularmente económicas. En un corpus tan pequeño, las ocurrencias de los afijos son comparables a las de las bases. Al no haber una diferencia significativa, sus relaciones económicas son mínimas. De allí que la afijalidad no sea mayor de 0.5 en ningún caso (la mayor cantidad de entropía equivale solamente a la mitad de la afijalidad correspondiente a ese prefijo).

En cuanto a la lengua chuj, es de notarse que entre las primeras 22 entradas ocurren todos los prefijos temporales del paradigma verbal ***ix-, tz-, ol-*** (núms. 6, 7 y 22) y ***x-*** (núm. 30, alomorfo de ***ix-***). También hay una muestra significativa de los pronombres personales absolutivos y ergativos que se suelen prefijar al verbo: ***a-, s-, in-, e-, ach-, ko-*** y ***ku-*** (núms. 1, 2, 3, 5, 17, 9 y 13, los dos últimos alomorfos).

[5] Para los detalles de estas medidas véase, por ejemplo, Manning y Schütze (1999).

Los mismos se prefijan a bases nominales como marcas de posesión. Los prefijos temporales se adhieren a los personales. Por eso ocurren grupos prefijales temporales y personales: ***tz.in-, ol.in-, ix.in-, ix.s-, tz.s-, tz.onh-, ol.e-, ol.ach-*** y ***tz.a-*** (núms. 4, 8, 14, 15, 16, 20, 23, 27 y 28). De hecho, si los personales aparecen como prefijos aislados, es porque los temporales alternan con **Ø-**. Un prefijo interesante es el que sirve para negar oraciones ***ma-*** (núm. 11), cuya forma se observa también en las cinco maneras de negar que hay en chuj: en los grupos prefijales ***ma.j-, ma.x-*** y ***ma.n-***, así como en las formas ***malaj*** y ***ma'ay***. Por otra parte, la forma ***to-*** (núm. 18) que también ocurre libre, sirve entre otras cosas para introducir oraciones subordinadas.

Otro grupo de formas no menos importante es el de aquellas que no representan ningún prefijo conocido (núms. 10, 12, 19, 21, 24, 25, 26 y 29 de la tabla 3). Este grupo constituye un reto interesante para el lingüista, sobre todo si los mismos ocurren en muestras de mayor tamaño. Por un lado, la naturaleza del error debe examinarse seriamente, principalmente porque los errores son inevitables. Por el otro, convendría estudiar lo que aparece como tal para determinar si podría considerarse, aunque fuera incipientemente, como un tipo de morfema; es decir, examinar si son errores verdaderamente. En el caso de este experimento, el corpus es demasiado pequeño como para preocuparse mucho por los residuos. De todas maneras, al tomarlos como errores, se puede calcular la proporción de aciertos de la tabla (medida de *recall*), 22 ÷ 30 = 0.73, mientras que 0.27 es la proporción de formas residuales. En un corpus tan pequeño como el utilizado en este experimento los residuos representan quizá la interferencia ocasionada por la escasez de datos.

Catálogo de sufijos

Con respecto a los sufijos y grupos sufijales reunidos en este experimento, en la tabla 4 se muestran las 30 formas más importantes según los criterios descritos arriba. El primer grupo es el de las vocales temáticas ***-a*** e ***-i*** (núms. 5 y 10) que permiten distinguir los verbos transitivos de los intransitivos. Además, indican el final de la frase. Otro grupo interesante es el de los sufijos ***-ok*** y ***-nak*** (núms. 3 y 26) que son respectivamente marcas de modo y tiempo. Están en distribución complementaria y ocurren entre las vocales temáticas y la base verbal. Entre estos sufijos y la base ocurren las marcas de voz que pueden ser de dos tipos, pasiva y antipasiva. Veamos primero los de la voz pasiva. En la tabla 4 sólo aparecen dos, ***-chaj*** y ***-aj*** (núms. 22 y 25). Los otros miembros de este paradigma, ***-ji, -nax*** y ***-b'il*** no ocurren solos entre los primeros 290 grupos sufijales (sólo los primeros 30 se muestran en la tabla 4). Sin embargo, sí ocurren en grupos sufijales tales como ***-a.ji*** y ***-ak'.nax*** (con rangos 66 y 290). El carácter de estos dos últimos es dudoso dada su baja frecuencia (13 y 2 respectivamente) como afijos (de hecho, como veremos, ***-ak'*** es una raíz). Luego están los sufijos de voz antipasiva. De los tres que forman el paradigma, ***-an, -wi*** y ***-waj***, sólo el primero aparece en la tabla 4 (núm. 16). Los otros dos ocurren después (***-wi*** con rango 50 y frecuencia de 14 y ***-waj*** con rango 150 y frecuencia

Tabla 3. Catálogo de prefijos más afijales

rango	prefijo	frecuencia	entropía	afijalidad
1	A~	160	1.0000	0.5000
2	S~	177	0.9874	0.4937
3	IN~	80	0.9828	0.4914
4	TZIN~	42	0.9338	0.4669
5	E~	63	0.9173	0.4587
6	IX~	166	0.9088	0.4544
7	TZ~	338	0.8861	0.4430
8	OLIN~	25	0.8818	0.4409
9	KO~	64	0.8783	0.4392
10	AL~	16	0.8740	0.4370
11	MA~	30	0.8722	0.4361
12	KA~	31	0.8496	0.4248
13	KU~	11	0.8303	0.4152
14	IXIN~	27	0.8183	0.4092
15	IXS~	23	0.8154	0.4077
16	TZS~	45	0.8101	0.4051
17	ACH~	10	0.8034	0.4017
18	TO~	13	0.8010	0.4005
19	AK'~	10	0.7881	0.3940
20	TZONH~	15	0.7866	0.3933
21	JA~	12	0.7808	0.3904
22	OL~	176	0.7807	0.3903
23	OLE~	12	0.7761	0.3880
24	NA~	8	0.7761	0.3880
25	U~	20	0.7682	0.3841
26	TA~	13	0.7670	0.3835
27	OLACH~	26	0.7663	0.3831
28	TZA~	41	0.7612	0.3806
29	YO~	16	0.7600	0.3800
30	X~	41	0.7568	0.3784

de 3). El sufijo ***-in*** (núm. 9) es, al igual que la forma prefijada, un pronombre absolutivo de primera persona. Por otra parte, el carácter afijal del sufijo ***-an*** se debe seguramente a que, además de ser muy productivo, es una forma en extremo polisémica: aparte de ser marca de antipasiva, es marca de subordinación, marca de agente en foco y de continuidad de tópico. Además, al igual que en otras lenguas mayas, también es sufijo de posicionales. No es de sorprenderse que morfemas tan polisémicos obtengan valores altos de afijalidad [6].

Con respecto a los sufijos ***-al*** e ***-il*** (núms. 1 y 4), se trata de alomorfos que sirven en sustantivos de marcas de genitivo o absolutivo, según el contexto. Otro grupo de sufijos digno de comentarse es el de los direccionales ***-kan, -ek', -k'och, -b'at, -el, -em***, y ***-pax*** (núms. 8, 11, 12, 13, 14, 15 y 20). Se trata de verbos de movimiento que se sufijan y sirven como clasificadores verbales. No están todos pero sí los principales. Lo interesante de estos sufijos es que deben considerarse más sufijos derivativos que de flexión, cosa significativa porque con un corpus tan pequeño era de esperarse que cuando mucho sólo los paradigmas de flexión se identificaran. Al final de la tabla están los sufijos ***-nak***, (núm. 26), participio de verbos intransitivos; ***-e*** (núm. 27), clasificador numeral de inanimados; ***-oj*** (núm. 29), marca de infinitivo en oraciones de complemento. Los sufijos con carácter adverbial son ***-ta'*** (núm. 18) que significa "inmediata o recientemente", ***-alan*** (núm. 17) "debajo" y ***-nej*** (núm. 28) "solamente".

Algunas entradas de la tabla son irreconocibles como morfemas del chuj, pero no podemos considerarlas propiamente errores porque son formas que ocurren al final de los préstamos españoles muy abundantes y profusos en el corpus: ***-o*** (núm. 2) en 'remedio', 'konejo', 'ciento', 'puro', 'ejersito', 'exodo', 'templo', 'cuatro', 'bueno', 'mismo', 'pero', 'San Pransisko', etc.; y ***-es*** (núm. 19) en 'tres', 'tonces', 'entonces', 'despues', 'jues'. Lo interesante es que en varias palabras estas formas tienen carácter morfológico (sobre todo ***-o***) [7]. La discusión sobre si deben considerarse o no sufijos del chuj está fuera del alcance de este trabajo. Por lo pronto, no podemos considerarlos afijos del chuj, pero tampoco pueden considerarse errores: si tienen relaciones afijales con objetos de un corpus, no pueden descartarse como parte del sistema implícito en ese corpus.

Otra cosa que se observa de las formas de la tabla 4 es que varias contienen verbos. Así, el sufijo adverbial ***-alan*** (núm. 17), arriba citado, tiene la misma forma que la secuencia ***-al.an***, donde ***-al*** significa "decir". Por otra parte, ***-ak'*** (núm. 24), que también ocurre en ***-ak'.an*** (núm. 23), significa "dar". La forma ***-cham*** (núm. 21) es la raíz de "matar" y forma verbos compuestos con significados como "golpear" y "acabar". Con rango menor están ***-tak*** (núm. 6) que es la raíz del verbo "aceptar" y ***-ab'*** (núm. 7) que es la raíz del verbo "oír" y suele utilizarse como sufijo citativo. Por último, está la forma residual ***-ek*** (núm. 30) sin un valor morfológico obvio.

[6] De hecho, es de esperarse que las formas más frecuentes sean las más polisémicas (a mayor número de contextos, más sentidos). Sin embargo, ese tema, por cierto bastante complejo, está fuera del alcance de este trabajo.

[7] De manera similar, la ocurrencia de los préstamos españoles terminados en 'a' (como 'semana', 'pena', etc.) debe haber contribuido al rango de la vocal temática ***-a*** en la Tabla 4.

Tabla 4. Catálogo de sufijos más afijales

rango	sufijo	frecuencia	entropía	afijalidad
1	~AL	82	1.0000	0.5000
2	~O	123	0.9634	0.4817
3	~OK	68	0.9374	0.4687
4	~IL	62	0.9347	0.4673
5	~A	142	0.9306	0.4653
6	~TAK	19	0.9062	0.4531
7	~AB'	49	0.9059	0.4530
8	~KAN	68	0.9029	0.4515
9	~IN	46	0.8917	0.4458
10	~I	205	0.8769	0.4384
11	~EK'	23	0.8740	0.4370
12	~K'OCH	28	0.8670	0.4335
13	~B'AT	63	0.8659	0.4329
14	~EL	68	0.8643	0.4321
15	~EM	15	0.8282	0.4141
16	~AN	233	0.8271	0.4135
17	~ALAN	13	0.8225	0.4112
18	~TA'	70	0.8203	0.4102
19	~ES	8	0.8140	0.4070
20	~PAX	15	0.8093	0.4046
21	~CHAM	16	0.8039	0.4020
22	~CHAJ	14	0.8037	0.4018
23	~AK'AN	11	0.7946	0.3973
24	~AK'	43	0.7922	0.3961
25	~AJ	51	0.7867	0.3934
26	~NAK	18	0.7812	0.3906
27	~E	60	0.7803	0.3901
28	~NEJ	24	0.7698	0.3849
29	~OJ	11	0.7673	0.3837
30	~EK	11	0.7673	0.3837

Lo interesante es que al identificar todos estos sufijos podemos calcular la proporción de aciertos (*recall*) de la tabla 4: *29* ÷ 30 = 0.97 (con un porcentaje de ruido residual del 0.03)[8]. Además, al tomar en cuenta las dos tablas (3 y 4), obtenemos un índice de aciertos de 0.85; esto es, 51 aciertos dentro de las 60 formas más afijales. Así, la proporción total de residuos es de 0.15, cuestión nada desalentadora al considerar el tamaño del corpus.

El paradigma de flexión verbal del chuj

Otra manera de evaluar la pertinencia de los fragmentos de palabras seleccionados —y, por ende, la del procedimiento descrito arriba— es identificar aquellos que pertenecen a los paradigmas de flexión verbal de la lengua en cuestión, aunque no estén dentro de los 30 más afijales. Lo importante es verificar que lo seleccionado forme parte de los morfemas más afijales del chuj, así como determinar lo que no se seleccionó pero debió haber ocurrido entre los resultados por su ya conocido carácter morfológico.

Como quedó establecido arriba, en el chuj hay prefijos y sufijos de flexión verbal. En las tablas 5 y 6 se exhiben estos afijos. La primera columna de la primera tabla muestra las marcas de tiempo, que son los prefijos más alejados de la base. Entre éstos y la base ocurren morfemas con carácter pronominal que pueden ser absolutivos o ergativos.

Tabla 5. Paradigma de prefijos de flexión verbal

tiempo	**persona**			
		absolutivos	**ergativos**	
	1a	in	in-	***w-***
tz-	**2a**	ach	a-	Ø-
ix-, x-	**3a**	Ø	s-	y-
ol-	**1a**	onh	ko-, ku-	k-
Ø-	**2a**	ex	e-	***ey-***
	3a	Ø … eb'	s- … eb'	y- … eb'

Como los pronombres absolutivos ocurren prefijados, sufijados o como morfemas libres, en la tabla 5 no se muestran con el guión que usamos para representar a los prefijos. En cuanto a los prefijos ergativos, los de la primera columna son preconsonánticos y los de la segunda son prevocálicos. Las formas que no se identificaron automáticamente están en bastardillas y negritas. Como puede verse, solamente faltan dos pronombres ergativos ***w-*** y ***ey-***. Sin contar el morfema nulo, **Ø-**, se aprecia que se aislaron automáticamente 16 de las 18 formas prefijales posibles (88.89%).

[8] Si los sufijos no flexivos y las formas españolas se consideraran errores, cosa a nuestro juicio inapropiada, tendríamos 16 de 30 "aciertos" o una medida de *recall* de 0.53 (0.47 de "residuos"). De todas maneras, estas cifras no parecen nada despreciables dadas las limitaciones del experimento.

Por otra parte, los sufijos verbales del chuj marcan especialmente voz, modo y final del enunciado. En la tabla 6 se muestran dichos sufijos:

Tabla 6. Paradigma de sufijos de flexión verbal

	voz	modal/ temporal	vocal temática
pasiva	-chaj		
	-b'il		
	-nax		
	-aj		
	-ji	-ok	-a
		-nak	-i
antipasiva	-waj		
	-an		
	-wi		

También aquí los sufijos que no fueron identificados se muestran en negritas y bastardillas. Como se dijo arriba y se aprecia en la tabla 6, faltaron entre los resultados del procedimiento automático sólo tres marcas de voz pasiva (***-b'il, -nax*** y ***-ji***). Esto significa que se aislaron automáticamente 9 de las 12 formas sufijales posibles (75.00%). Si tomamos a los dos grupos de afijos de flexión como uno solo, vemos que el procedimiento automático permitió aislar 25 de 30. Esto indica que 83.33% de los afijos pertinentes (medida de precisión) ocurrió entre los fragmentos más afijales de la palabra chuj. Mientras que la cantidad de residuos es pertinente en la medida de *recall*, en la de precisión lo importante es por definición que el procedimiento sencillamente no incluyó dentro del catálogo el 16.66% de los afijos de flexión del chuj. Lo que, al considerar el reducido tamaño del corpus, no habla nada mal del procedimiento.

Conclusiones

En este trabajo se presentaron los resultados de la aplicación de algunos métodos de segmentación automática en un corpus pequeño de una lengua indígena, en concreto el chuj, que se habla en la frontera entre Chiapas y Guatemala. La idea era determinar si por lo menos la morfología verbal flexiva podía descubrirse automáticamente, a pesar del tamaño del corpus. Los resultados indican que el 83% (medida de precisión) de los morfemas de flexión verbal se identificaron automáticamente, evidencia de que el tamaño reducido de los costosos corpus de lengua indígena no es un impedimento para investigar las propiedades cuantitativas de los componentes de esas lenguas (siempre y cuando se utilice algo más que las meras frecuencias). Esto fue posible porque los morfemas de flexión forman parte de un conjunto pequeño y muy regular de fragmentos de palabra relativamente

fáciles de descubrir automáticamente. Cabe señalar que además de los de flexión, algunos afijos de tipo derivativo también fueron identificados automáticamente (los direccionales), lo que indica su importancia relativa en el chuj.

Finalmente, aunque se aplicaron dos métodos de segmentación (las medidas de contenido de información y de economía de los signos), solamente el cálculo de entropía sirvió para los propósitos de este experimento. Esto significa que calcular la cantidad de información es un método más apropiado para corpus relativamente pequeños, por lo menos en el caso del chuj. Por otra parte, si lo que se requiere es medir la naturaleza económica de las relaciones entre los signos para, por ejemplo, determinar cuando la forma de un afijo reconocido ocurre sólo como una secuencia de fonemas (o caracteres) y cuando ocurre como verdadero morfema afijal, es indispensable un corpus de mayor tamaño.

Referencias

BUENROSTRO, Elsa Cristina (2000) *Corpus de la lengua chuj*, 1997-2000.

_____. (2003) *La voz en el chuj de San Mateo Ixtatán*, borrador de tesis doctoral, México, El Colegio de México.

_____. (1992) *Morfología verbal del chuj*, tesis licenciatura, México, ENAH.

CROMM, Oliver (1996) *Affixerkennung in deutschen Wortformen. Eine Untersuchung zum nicht-lexikalischen Segmentierungsverfahren von N. D. Andreev*, Abschluss des Ergänzungsstudiums Linguistische Datenverarbeitung, Francfort del Meno.

DE KOCK, Josse y Walter Bossaert (1974) *Introducción a la lingüística automática en las lenguas románicas*, Gredos, Madrid, (*Estudios y Ensayos* 202).

_____. (1978) *The Morpheme: an experiment in quantitative and computational linguistics*, Amsterdam/Madrid, Van Gorcum.

FRAKES, William y Ricardo Baeza (1992) "Stemming algorithms" en Frakes, William, ed., *Information retrieval, data structures and algorithms*, Prentice Hall, New Jersey, 1992, pp. 131-160.

GREENBERG, Joseph H. (1967) *Essays in linguistics*, The University of Chicago Press, Chicago, 1967 [1957].

HAFER, Margaret y Stephen Weiss (1974) "Word segmentation by letter successor varieties", *Information Storage and Retrieval*, 10 (1974), pp. 371-385.

HARRIS, Zellig (1955) "From phoneme to morpheme", *Language* 31:2, pp. 190-222.

KAGEURA, Kyo (1999) "Bigram statistics revisited: a comparative examination of some statistical measures in morphological analysis of japanese kanji sequences", *Journal of Quantitative Linguistics*, 6(1999), pp. 149-166.

LARA, Luis Fernando (1990) "Caracterización metódica del *corpus* del DEM" en *Dimensiones de la lexicografía. A propósito del Diccionario del español de México*, El Colegio de México, México, (Jornadas 116) 1990, pp. 85-106.

MANNING, Christopher y Hinrich Schütze (1999) *Foundations of statistical natural language processing*, Cambridge (Mass.), The MIT Press.

MEDINA URREA, Alfonso (2000) "Automatic discovery of affixes by means of a corpus: a catalog of Spanish affixes", *Journal of quantitative linguistics* 7:2, pp. 97-114.

_____. (2003) *Investigación cuantitativa de afijos y clíticos del español de México: glutinometría en el Corpus del Español Mexicano Contemporáneo*, tesis doctoral, México, El Colegio de México.

OAKES, Michael P. (1998) *Statistics for corpus linguistics*, Edinburgh, Edinburgh UP.

SHANNON, Claude y Warren Weaver. (1949) *The mathematical theory of communication*, University of Illinois Press, Urbana, 1964 [1949].