Alfonso Medina Urrea. *El signo afijal en la muestra textual. Claves para entender el descubrimiento automático de morfemas*. El Colegio de México, México, 2021; 316 pp.

Carlos Francisco Méndez Cruz
Universidad Nacional Autónoma de México
cmendezc@ccg.unam.mx

Seguramente, no hay mejor frase para describir el espíritu de esta obra que aquélla citada por el autor y atribuida a Galileo Galilei: "Medir todo lo medible e intentar hacer medible todo lo que aún no lo es" (p. 282). Así, ésta es una obra que nos invita a conciliar dos posturas aparentemente opuestas: lo descriptivo y lo cuantificable.

El objetivo del libro es mostrar, mediante aproximaciones cuantitativas basadas en corpus, la factibilidad de medir propiedades lingüísticas de unidades morfológicas (específicamente de afijos) antes descritas sólo desde una perspectiva cualitativa en la tradición lingüística. Este hecho de cuantificar conlleva, como lo demuestra el autor, un acto de *descubrimiento* o *aprendizaje* de esas unidades. Las unidades están allí, lo sabemos, nos lo han dicho las gramáticas; pero conviene dejar que emerjan, esto es, observar que están allí con la ayuda de un procedimiento automático desprovisto de una concepción *a priori* de sus propiedades.

Este libro no pone de manifiesto sólo una propuesta metodológica, sino que también es una muestra clara de una línea de investigación que se ha trabajo por años, durante los cuales se han realizado experimentos no sólo para demostrar su pertinencia, sino también su utilidad en aplicaciones tecnológicas, como lexematizadores y sintetizadores de voz; y que se ha extendido, además, al estudio de cuestionamientos lingüísticos, como los cambios diacrónicos en el español y la madurez en la escolaridad. El libro está dirigido a lingüistas, especialmente morfólogos, quienes encontrarán en esta obra una verdadera muestra de la intersección de dos mundos: la lingüística y la computación. Consideramos que su lectura no requiere de conocimientos técnicos o computacionales, ya que el autor logra explicar con la mayor claridad posible todos los términos usados en su propuesta.

Lingüística Mexicana. Nueva Época, v (2023), núm. 1: 135-139
ISSN: 2448-8194

La obra está conformada por cuatro capítulos, la introducción, un apartado con observaciones finales y un apéndice con ejemplos de código de programación de computadoras para implementar algunos procedimientos presentados en el libro. El primer capítulo, "Algunos temas de la morfología computacional", ofrece un panorama general del descubrimiento automático de morfemas en el marco de la morfología computacional. El segundo capítulo, "El signo afijal", tiene como objetivo presentar el aparato conceptual y metodológico de la propuesta del autor, así como su aplicación para descubrir automáticamente los afijos de las palabras de un corpus textual del español mexicano.

En el tercer capítulo, "Aplicaciones del descubrimiento de afijos", el autor discute el problema de cómo evaluar su propuesta y muestra ejemplos de evaluaciones basadas en su capacidad para aplicarse en otras lenguas, en su utilidad para desarrollar tecnología y en el uso de medidas estándar de evaluación. Finalmente, el cuarto y último capítulo, "Hacia el cálculo de la variación morfológica", tiene por objeto demostrar que el conjunto de afijos descubiertos puede verse como un perfil morfológico que permitiría una comparación entre estados o variantes de una misma lengua.

El subtítulo de esta obra promete darnos las claves para entender el descubrimiento automático de los morfemas. Estas claves, que a su vez sustentan la propuesta del autor, no son sino conceptos bien conocidos en la lingüística, y se exponen en la introducción de su obra: *secuencia*, *combinación*, *sorpresa*, *descubrimiento* y *afijalidad*. Dice el autor: "Sabemos que la estructura de lo que oímos o leemos está dada por la secuencia de los signos y por cómo se combinan unos con otros" (p. 16). Luego, esa secuencia y combinación producen un sistema de posibles signos anteriores y posteriores a cualquier signo, donde la aparición de alguno de ellos producirá mayor o menor *sorpresa* a nuestro interlocutor o lector. Esta cantidad de sorpresa es un reflejo de que los signos pueden llevar información gramatical o semántica, en mayor o menor grado. Además, para que podamos nombrar un sinfín de cosas, no necesitamos recordar una enorme cantidad de signos: aprovechamos que la lengua es un sistema económico, y combinamos formas de diferente nivel para crear nuevos signos.

Uno de los principios de la propuesta del autor es dejar que el corpus *hable* por sí mismo, poniendo de lado, por un momento, las unidades que *a priori* consideramos que pertenecen a la lengua. La idea es concentrarnos en crear un procedimiento automático que, motivado y lingüísticamente, descubra (aprenda) las unidades morfológicas presentes en el corpus. El descubrimiento, advierte el autor, no debe estar guiado por ninguna información explícita puesta deliberadamente en el corpus, por lo que se tratará de un procedimiento *no supervisado*. Por otro lado, se resalta que el análisis gramatical automático conlleva cuatro aspectos a considerar: uno empírico para determinar unidades de interés, uno conceptual de definición de esas unidades, uno metodológico para seleccionar el mejor procedimiento para hacer lo anterior y uno evaluativo para definir y usar criterios que califiquen los resultados del procedimiento.

Medina Urrea propone el término *afijalidad* como "la cualidad que una porción de la palabra, ya sea inicial o final, pueda tener de ser un afijo de la lengua a la que pertenece" (p. 24). Este concepto es clave para entender su propuesta y es, de hecho, la cualidad que el autor mide a través de la cuantificación de tres propiedades de un afijo. Antes

de formalizar su propuesta de medición de la afijalidad, presenta en el primer capítulo, "Algunos temas de la morfología computacional", antecedentes del descubrimiento de morfemas. Lo hace en el contexto de la lingüística computacional, subdisciplina dedicada al estudio de los fenómenos morfológicos de las lenguas naturales mediante procedimientos automáticos. Coincidimos en lo que Medina Urrea atribuye como el objetivo principal de la morfología computacional: "Adquirir un mejor y más explícito entendimiento de la morfología en general" (p. 37).

Una cuestión interesante de este capítulo es la discusión a propósito de las ventajas y desventajas de los métodos de análisis morfológico computacional basados en reglas (supervisados) en comparación con los cuantitativo-estadísticos (no supervisados). Los primeros parecen pecar de cierto *dogmatismo*, pues generalmente evitan que se revisen los postulados específicos que sustentan las reglas elaboradas o usan la información morfológica de un especialista para guiar el análisis automático. Por otra parte, los no supervisados, al tomar en cuenta sólo los patrones estadísticos presentes en el corpus, no son *dogmáticos*, sino *escépticos* con respecto a esos postulados e información *a priori*.

El capítulo resulta muy ejemplificador de los distintos caminos que se han tomado para estudiar, con ayuda de la computadora, los fenómenos morfológicos de distintas lenguas, como el alemán, español, finés, francés, inglés, japonés y rumano. El inventario de métodos de segmentación morfológica no supervisada no sólo da cuenta de las aproximaciones interesantes elaboradas a lo largo de décadas de investigación, sino que además consigna aquéllas que inspiraron el método desarrollado por el autor.

Podemos mencionar, por ejemplo, la propuesta de Zellig Harris (1955) basada en el conteo de fonemas anteriores y posteriores a una posible segmentación para determinar si ésta es morfológica. La propuesta conlleva la idea de que la mayor o menor sorpresa medida en el número de fonemas diferentes es un indicador de frontera morfológica. Por otro lado, se resalta en el libro el papel que juega la *entropía*, concepto tomado de la teoría de la información, para medir esa sorpresa o incertidumbre, de manera que las segmentaciones que exhiban valores altos de entropía pueden ser indicios de cortes morfológicos. Conviene resaltar aquí la clara explicación que ofrece el autor del término técnico de *entropía*. El último método antecedente de la propuesta de Medina Urrea es el llamado *principio de economía*, desarrollado por Josse de Kock y Walter Bossaert (1978), para descubrir fronteras morfológicas entre bases y afijos.

En el segundo capítulo, "El signo afijal", el autor define y formaliza los elementos teóricos y procedurales que sustentan su propuesta metodológica. Nos dice que la afijación se refiere "a los procesos de formación de palabras mediante la concatenación de ciertos tipos de signos mínimos llamados afijos a otros llamados raíces o bases" (p. 75), y que estos afijos (i) no ocurren aislados, sino como parte de las palabras, (ii) ocurren en muchas palabras de baja frecuencia, y (iii) su significado es más gramatical que el de la base o lexema al que se pegan.

Estas propiedades abstractas de un afijo se formalizan y cuantifican con tres medidas respectivamente: (i) el número de cuadros en el que participan, donde un cuadro es un conjunto de cuatro segmentos de palabras que, combinados, forman cuatro palabras presentes en el corpus; (ii) una medida basada en el principio de economía; y (iii) la

cantidad de entropía. Después de ofrecernos la definición detallada de estas medidas y su formalización matemática, el autor propone la combinación de estas medidas en un índice de *afijalidad* que permite describir de forma cuantitativa tal propiedad cualitativa. Es claro cómo Medina Urrea continúa la tradición de lingüistas cuantitativos inspirándose en sus propuestas y actualizándolas. Una innovación presente en esta obra es la construcción de un *catálogo de afijos* cuya función es consignar los prefijos y sufijos extraídos (aprendidos) por el método.

Hasta aquí, de los cuatro aspectos propuestos por el autor para llevar a cabo un análisis gramatical automático (empírico, conceptual, metodológico y evaluativo), se han cumplido los tres primeros; falta el último: la evaluación de los resultados del método. Para tal fin, Medina Urrea calcula el índice de afijalidad de una muestra aleatoria de vocablos del *Corpus del español mexicano contemporáneo* (CEMC). Después de una revisión manual de las segmentaciones propuestas, se observó que el método logró segmentar correctamente 764 de los 836 vocablos analizados, lo que representa un 90.41% de aciertos. Así, el autor logra una caracterización automática y cuantitativa del sistema sufijal del español de México.

Entre los hallazgos interesantes que resultan de observar los afijos extraídos, tenemos que los afijos de flexión verbal obtuvieron los valores más altos de afijalidad, lo que, en nuestra opinión, es un reflejo de que la cuantificación de propiedades fue pertinente. Otros descubrimientos interesantes expuestos en este libro son las diferencias cuantitativas entre prefijos y sufijos del español, y entre segmentos flexivos y derivativos; y la observación de que las diferentes combinaciones de medidas (cuadros, entropía y economía) podrían dar indicios de estos fenómenos, lo que representaría un avance en esta área, ya que se estaría cuantificando no sólo la afijalidad, sino también otros fenómenos morfológicos, como la flexión, la derivación, la composición, etcétera.

Una de las ventajas del desarrollo de este tipo de propuestas lingüísticas computacionales es que pueden tener, además, una aplicación práctica (tecnológica). Así, el tercer capítulo, “Aplicaciones del descubrimiento de afijos”, presenta cómo la propuesta ha sido empleada en la tecnología para el análisis de textos, como lexematizadores (que separan el lexema o la raíz de los afijos), analizadores gramaticales y sintetizadores de voz (recepción de palabras de entrada y emisión de su sonido correspondiente). Si bien todos estos proyectos mostraron un impacto benéfico en cada tecnología, llama la atención la última. Se trató del experimento de incluir un catálogo de sufijos descubiertos en un sintetizador de voz para la lengua ralámuli. Vemos este ejemplo como una muestra clara de la aplicación de la lingüística en beneficio de la sociedad y de una lengua con pocos recursos electrónicos.

Las aproximaciones no supervisadas, al carecer de información explícita e integrada premeditadamente en las muestras textuales de las que aprenden, permiten su aplicación a muestras de otras lenguas de morfología relativamente similar. Medina Urrea demuestra este hecho por medio de la aplicación exitosa de su método a muestras textuales de distintas lenguas no emparentadas, como el checo, para extraer sus prefijos; el ralámuli, para extraer sus sufijos, y el chuj, para descubrir sus prefijos y sufijos. En todos los casos se obtuvieron porcentajes de aciertos relevantes: 100% para el checo, 71% para el ralámuli y 86% para el chuj.

Esta obra termina con una propuesta interesante descrita en su cuarto capítulo, "Hacia un cálculo de la variación morfológica". Se postula considerar el conjunto de afijos extraídos de una muestra textual como un *perfil morfológico* de la lengua representada por tal muestra, y usarlo para comparar con otros perfiles. Como anota el autor, las diferencias entre estos perfiles se pueden ver como diferencias morfológicas, y una medida de esa diferencia (distancia) puede representar una medida de variación a nivel morfológico. Con esta idea, el autor lleva a cabo experimentos de medición de distancias morfológicas en diversas muestras textuales, con lo cual confirma cuantitativamente las diferencias esperadas entre las muestras. Por ejemplo, mide la distancia entre dos muestras textuales de niños de cuarto y sexto grado de Santiago de Cuba, y encuentra que, de acuerdo con la distancia entre perfiles morfológicos, hay "un incremento de la competencia morfológica en los estudiantes más avanzados" (p. 220).

Dos interesantes experimentos más se nos presentan: uno sobre las distancias entre perfiles morfológicos de cuatro lenguas mayas (chuj, tojolabal, yucateco y huasteco) y otro sobre la distancia entre el español, de lo que hoy es México, de los siglos XVI, XVIII y XIX y el español de España de los siglos XVIII y XX. A propósito del primero, el resultado es muy interesante porque las distancias afijales calculadas entre las lenguas coinciden con la separación geográfica de los lugares donde se hablan, por ejemplo, el huasteco es el más alejado del grupo, y la distancia entre el huasteco y el yucateco es mayor que entre el huasteco y las demás. Lo anterior deja claro que la propuesta del autor puede ayudar a estudios de relaciones genéticas entre lenguas.

Es claro que el método presentado aún tiene limitaciones, siempre señaladas por el autor a lo largo de su obra, y también que queda camino por recorrer para llevarlo a sus últimas consecuencias. Pero, siendo justos, estas limitaciones también son un reflejo de la complejidad de la morfología. La lectura de este libro nos deja entusiasmados por sumarnos al esfuerzo de "hacer medible lo que aún no lo es" y por saber cuál será la próxima unidad lingüística o fenómeno morfológico que el autor intentará cuantificar.

Bibliografía

De Kock, Josse y Walter Bossaert. 1970. *The Morpheme. An Experiment in Quantitative and Computational Linguistics*. Amsterdam: Van Gorcum.

Diccionario del español de México. *Corpus del español mexicano contemporáneo* (CEMC), en <http://www.corpus.unam.mx/cemc>.

Zellig, Harris. 1955. "From phoneme to morpheme", *Language* 18: 190-222.