El signo afijal en la muestra textual

Claves para entender el descubrimiento automático de morfemas

Alfonso Medina Urrea

EL COLEGIO DE MÉXICO

EL SIGNO AFIJAL EN LA MUESTRA TEXTUAL
CLAVES PARA ENTENDER EL DESCUBRIMIENTO AUTOMÁTICO DE MORFEMAS

SERIE
ESTUDIOS DE LINGÜÍSTICA
XXXVII

CENTRO DE ESTUDIOS LINGÜÍSTICOS Y LITERARIOS

EL SIGNO AFIJAL EN LA MUESTRA TEXTUAL
CLAVES PARA ENTENDER EL DESCUBRIMIENTO AUTOMÁTICO DE MORFEMAS

Alfonso Medina Urrea

EL COLEGIO DE MÉXICO

415.92
M4914s

Medina Urrea, Alfonso

El signo afijal en la muestra textual : claves para entender el descubrimiento automático de morfemas / Alfonso Medina Urrea. – Primera edición. -- Ciudad de México : El Colegio de México, Centros de Estudios Lingüísticos y Literarios, 2021.

316 p. : il., mapa, gráfs. ; 22 cm. – (Serie Estudios de lingüística ; 37)

ISBN 978-607-564-307-6

1. Gramática comparada y general – Afijos. 2. Gramática comparada y general – Morfología. 3. Lingüística computacional. I. t. II. Ser.

Primera edición, 2021

Carretera Picacho Ajusco núm. 20
Ampliación Fuentes del Pedregal
Alcaldía Tlalpan
14110, Ciudad de México
www.colmex.mx

ISBN 978-607-564-307-6

Impreso en México

ÍNDICE

Índice de tablas

Índice de figuras

INTRODUCCIÓN:
LA CADENA HABLADA Y LA CADENA ESCRITA

Cuando escuchamos con atención a alguien que queremos o admiramos o cuyo mensaje nos interesa o nos seduce, solemos identificar qué cosas, personas, ideas o sensaciones nombra, qué dice que hacen o sucede con ellas, dónde ocurre lo que sucede y cómo y cuándo ocurre. Gracias a que compartimos un idioma, esto es, a que conocemos los signos que nuestro interlocutor produce y la manera en que los va estructurando, logramos comprender lo que va diciendo. Si no estamos familiarizados con alguna de sus palabras, intentamos inferir su significado a partir del mensaje, a partir de lo dicho y de lo que suponemos que se va a decir. Por otra parte, si algo ocurre en un orden que no esperábamos, podemos pensar que nuestro interlocutor se equivocó o está tratando de decir algo distinto a lo que anticipábamos y nos esforzamos en darle un nuevo sentido.

Desconocer alguno de los signos en la cadena hablada es una experiencia diferente a la de encontrarnos con un ordenamiento inusual o equivocado de los signos. Una palabra desconocida nos recuerda que no sabemos muchas cosas del mundo y nos invita a inferir su significado o a investigarlo en otro lado, como en un diccionario, en una enciclopedia, en Internet, o preguntándole a alguien. Una estructura sintáctica poco típica o mal formada nos fuerza a reanalizar el ordenamiento de los signos para reinterpretar el mensaje y buscar que tenga sentido. Casi siempre lo logramos, puesto que somos máquinas de descartar ambigüedades y encontrar significados, incluso de inventarlos donde no los hay. Por otra parte, si percibimos que está mal formada, puede ser que en un instante deduzcamos las causas o conjeturemos sobre las consecuencias de esa malformación.

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. Hay signos que nos revelan más sobre el mensaje del interlocutor y menos sobre su estructura. También hay signos que dan pistas sobre la estructura y nos hablan menos del mensaje mismo. Conocemos a los primeros como signos de contenido y solemos concentrar nuestra atención en ellos. En un momento dado del discurso, cuando tratamos de adivinar qué signo es el siguiente, los de contenido son los que menos podemos predecir y, una vez que han sido pronunciados, son los que más pueden habernos sorprendido. Es interesante que, como veremos adelante, el tamaño de esa sorpresa crece y decrece con la cantidad de información que intenta transmitir nuestro interlocutor.

En cuanto a los signos con función gramatical, aquellos que nos dan pistas sobre la estructura del mensaje, solemos estar menos pendientes de ellos. Sabemos que ocurren mucho, sus significantes suelen ser breves y comunican relativamente menos información que los de contenido. La que comunican tiende a ser de tipo gramatical. Por eso, los conocemos como partículas o signos gramaticales. Ocurren tanto que ya ni siquiera los vemos conscientemente, aunque sin duda los percibimos en tanto organizan el mensaje. De repente, cuando los encontramos en la posición equivocada o en lugar de otros que inconscientemente esperábamos o, simplemente, cuando no están donde consideramos que deberían estar, se vuelven otra vez visibles.

De esta manera, cada signo gramatical suele ocurrir en ciertos lugares y no en otros. Se aglutina con los demás de maneras determinadas para estructurar los enunciados del discurso: una preposición precede una expresión nominal; un adverbio orbita alrededor de una expresión verbal; un enclítico sigue al verbo y un proclítico lo precede; un afijo se pega a un signo de contenido o a otros signos afijales, etc. Esta morfotáctica es un síntoma de que la lengua es un sistema económico, lo cual se manifiesta en el hecho de que los hablantes, como nuestro interlocutor, puedan nombrar un número potencialmente infinito de cosas,

acciones y situaciones, sin tener que conocer o recordar una cantidad descomunal de signos.

La lengua hablada es diferente a la escrita en muchas maneras, pero la experiencia de escuchar a un interlocutor guarda ciertas similitudes con la de leer un texto. Las palabras gráficas se suceden una tras otra, unas son de contenido y otras gramaticales. De nuevo, las primeras transmiten el mensaje y las segundas lo estructuran. Mucho se pierde en lengua escrita: dejamos de ver los gestos del interlocutor y no hay pistas de la entonación. En cambio, contamos con signos de puntuación que orientan al lector en cuanto a la estructura sintáctica del mensaje y dan ritmo a su lectura. Los espacios separan las palabras gráficas, aunque no separan las bases o lexemas de los afijos. Las comas marcan fronteras entre frases o sintagmas y los puntos separan enunciados simples o complejos. La lengua hablada es fugaz. Si acaso, puede grabarse con un dispositivo de audio y volverse a escuchar o transcribirse. La lengua escrita y transcrita permanece plasmada en algún medio físico. De allí que, como veremos adelante, podamos estudiar mejor con ella las relaciones económicas entre los signos. En particular, veremos que estas relaciones permiten descubrir las fronteras entre bases y afijos y que, si no existieran los espacios entre las letras, tal vez nos permitirían descubrir las fronteras entre las palabras mismas.

Las relaciones económicas y la fuerza de atracción entre los signos

La alusión más antigua al carácter económico de la lengua es probablemente la de Marco Terencio Varrón, quien, algunas décadas antes de nuestra era, se percató de la carga que supone, para la memoria, la carencia de afijos derivativos y flexivos:

> La "declinación" [tanto de nombres y adjetivos como de conjugación verbal] se ha aplicado no sólo a la lengua latina, sino a la de todos los

> hombres, por una razón útil y necesaria. De no haberlo hecho así no podríamos aprender un número tan grande de palabras (ya que las formas naturales en que los vocablos se declinan son infinitas), y aunque las hubiéramos aprendido no podríamos descubrir a partir de ellas qué sistema [sic] las relaciona entre sí[1]. En cambio, ahora sí podemos percibirlo porque se trata de algo semejante, de algo que ha derivado; […] Dos son, en general, los orígenes de las palabras: la imposición y la flexión. La primera viene a ser la fuente; la segunda, el río. Los hombres quisieron que las formas "impuestas" fueran las menos posibles, con el fin de aprenderlas cuanto antes; y que las "flexionadas" fueran el mayor número posible, para que todos pudiesen emplear aquellas que fuera necesario utilizar[2].

Muchos siglos después, entre los lingüistas mecanicistas estadounidenses, para quienes era natural suponer que los fenómenos del lenguaje, sus causas y sus efectos son observables, el trabajo con corpus textuales o transcripciones de lengua hablada se volvió esencial. Sin embargo, como se sugirió arriba, en los corpus no se puede observar todos los fenómenos lingüísticos. Además de que faltan los gestos del interlocutor y las pistas de la entonación, se escapan otras pistas lingüísticas y extralingüísticas, pasando por los procesos mentales más diversos que, si bien no se manifiestan abiertamente en la escritura, pueden dejar en ella huellas de diversas índoles, poco evidentes, incluso para la mirada del experto.

Por otra parte, entre aquellos lingüistas para los que el lenguaje no se puede caracterizar exclusivamente en términos de causas y efectos visibles al investigador, se ha observado que las palabras se relacionan entre sí para combinarse en estructuras más complejas y que la fuerza

[1] Del latín: "nisi enim ita esset factum, neque discere tantum numerum verborum possemus (infinitae enim sunt naturae in quas ea declinantur) neque quae didicissemus, ex his, quae inter se rerum cognatio esset, appareret" [edición de Roland G. Kent (Varro 1938 [47-45 a.C.], 373)].

[2] Varrón, *De lingua Latina*, tr. Manuel-Antonio Marcos Casquero, VIII, 3-5 (1990 [ca. 40 a.C.], 292-295).

o energía de esas relaciones puede cambiar con el tiempo. Por ejemplo, Edward Sapir imagina una atracción entre palabras y elementos que se han expresado en cierto orden y que pueden solidificarse en palabras complejas y formar una sola masa y que aún después de cristalizada puede volverse a aflojar:

> Así, pues, las palabras y los elementos, una vez que se han expresado en cierto orden, no sólo tienden a desarrollar algún tipo de relación entre sí, sino que son atraídos más o menos el uno al otro. Se puede presumir que precisamente este "más o menos" es lo que, en resumidas cuentas, da origen a aquellos grupos de elementos, firmemente solidificados (elemento o elementos radicales más uno o más elementos gramaticales), que hemos estudiado como palabras complejas. Con toda verosimilitud, no son sino series de elementos que se han contraído, formando una sola masa, a partir de otras series, o de elementos aislados en la corriente del habla. Mientras están plenamente vivos, o, dicho en otras palabras, mientras son funcionales en cada punto de su estructura, pueden mantenerse a una distancia psicológica de sus vecinos. A medida que van perdiendo su vida individual, caen en brazos de la frase en cuanto conjunto, y la serie de las palabras independientes vuelve a adquirir la importancia que había transferido, en parte, a los grupos cristalizados de elementos. De esta manera, el lenguaje está apretando y aflojando sin cesar sus concatenaciones de palabras. En sus formas más sintéticas (como en latín o en esquimal), la "energía" de la secuencia queda encerrada, en gran parte, en complejas formaciones de palabras, viene a transformarse en una especie de energía potencial que quizá no se libere durante milenios. En sus formas más analíticas (como en chino o en inglés), esta energía es móvil, pronta para ser empleada en el servicio que se exija de ella[3].

Como se ve, Sapir concibe esto —aunque metafóricamente— como una energía, una fuerza que crece y decrece a través de los mile-

[3] Sapir, *El lenguaje*, tr. Frenk y Alatorre, FCE, México (1992 [1921], 131).

nios. En este trabajo, se busca mostrar, mediante corpus textuales, cómo las relaciones entre los signos, dentro y fuera de las palabras, crecen y decrecen con el tiempo y cómo podemos medir algunas propiedades de estas relaciones para descubrir o aprender los signos afijales de una lengua representada en un corpus.

El análisis automático de la cadena hablada

Aunque no sabemos todavía cómo logramos analizar —y menos cómo logramos entender— lo que escuchamos y leemos, en las últimas décadas se han dado avances considerables en el análisis sintáctico-gramatical automático mediante computadoras. De hecho, hace mucho que construir un analizador automático para una lengua natural, ya sea morfológico o sintáctico, dejó de ser un problema indescifrable. Existen numerosos métodos para llevar a cabo, con mayor o menor éxito, diferentes tipos de análisis automático[4].

Un reto no menos interesante es determinar cómo y de dónde viene el inventario de componentes del fenómeno en que se basa ese análisis, lo cual implica el descubrimiento automático de estos componentes mismos, los morfemas, por ejemplo, sin depender de la información previa que el especialista pueda tener a su disposición.

[4] Por ejemplo, véase la diversidad de métodos en manuales como los de Allen, *Natural Language Understanding*, Benjamin/Cummings, Redwood (1995) y Jurafsky y Martin, *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Pearson Prentice Hall, Upper Saddle River, NJ (2009). Otros manuales son, de métodos simbólico-cualitativos: Naumann y Langer, *Parsing: Eine Einführung in die maschinelle Analyse natürlicher Sprache*, Teubner, Stuttgart (1994), Gazdar y Mellish, *Natural Language Processing in Prolog*, Addison-Wesley, Wokingham, Reino Unido (1989); y de métodos probabilísticos: Charniak, *Statistical Language Learning*, The MIT Press, Cambridge, Mass. (1993) y Manning y Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Mass. (1999).

El análisis gramatical automático es una tarea de varios aspectos: primero, uno empírico[5] de determinar, a partir de un corpus, las unidades lingüísticas pertinentes; segundo, el aspecto conceptual de definir esas unidades para poder extraerlas y utilizarlas; tercero, el problema metodológico de escoger las técnicas convenientes para llevar a cabo lo anterior; y, finalmente, el problema de seleccionar y evaluar los criterios que nos permitan calificar los resultados.

Muchos métodos de análisis automático gramatical y sintáctico se basan en reglas formuladas por especialistas y resuelven los dos primeros aspectos del problema presuponiendo que lo que se sabe de la gramática es bien sabido; es decir, que la gramática es algo dado, que las unidades lingüísticas son conocidas y transparentes y que la introspección es un método empírico[6], suficiente para la investigación del lenguaje. Así, estos métodos dependen de reglas que representan la concepción que el analista tiene *a priori* de la lengua en cuestión, esto es, antes del análisis automático. Para formular estas reglas, normalmente se consultan las gramáticas disponibles o se recurre a la introspección del algún hablante[7]. No es que el método introspectivo no sea valioso para la ciencia, pero al depender sólo de éste se desatiende necesariamente una porción importante de los hechos reales. Por otra parte, los métodos de análisis automático construidos a partir de hechos lingüísticos (datos documentados) requieren casi siempre de corpus etiquetados, esto es, marcados previamente con anotaciones gramaticales aplicadas a cada una de sus palabras, muchas veces de manera manual.

[5] En el sentido de *factual*: "que se refiere a estados de hecho"; véase Abbagnano, *Diccionario de filosofía*, tr. Alfredo N. Galletti, Fondo de Cultura Económica, México (1991 [1961]), *s.v.*, Empírico.

[6] Pero en el sentido de experiencia *intuitiva*; véase *ibid.*, *s.v.*, Empírico.

[7] En sus inicios, el campo del análisis automático heredó de la lingüística generativa la idea de que los juicios e intuiciones de un hablante nativo de una lengua (derivados de un proceso de introspección, es decir, basados en la gramática mental internalizada de ese hablante) eran suficientes para el quehacer del lingüista. Así, aunque a veces se recurría a la consulta de las gramáticas tradicionales de la lengua estudiada, muchos trabajos de corte computacional se basaban en la formulación de reglas gramaticales o simbólicas a partir de los juicios e intuiciones de quienes las formulan.

Así, los sistemas de análisis lingüístico reciben de entrada dos tipos de información: texto o audio. En esencia, la información textual es una serie de caracteres alfanuméricos y puntuación, que representa la cadena escrita o transcrita, mientras que el audio es la grabación digital de esa cadena hablada o leída. Cuando a esos datos se les agrega información lingüística en forma de etiquetas, como las categorías gramaticales o semánticas de la palabra gráfica, o, en el audio, las marcas de frontera entre fonos, fonemas o patrones de entonación, se dice del sistema que los procesa que es *supervisado*. En cambio, cuando no se proporciona ninguna información adicional, esto es, cuando la clasificación de dichos datos no está disponible, se habla de sistemas de aprendizaje o entrenamiento *no supervisado* (*unsupervised training or learning*)[8].

Otra cuestión relativa al trabajo empírico inherente al análisis automático es la selección del nivel lingüístico que se quiere investigar. Por ejemplo, un nivel descuidado hasta hace relativamente poco es el morfológico. Cuando siquiera se consideraba, normalmente se presuponían los morfemas antes de empezar la investigación. Una razón determinante era que la lengua inglesa, la más trabajada en el campo de la automatización, tiene una morfología muy sencilla y muy conocida[9], cosa que también explicaba el relativo poco interés en los métodos automáticos de descubrimiento de unidades morfológicas en general[10]. Dada la di-

[8] Para mayor profundidad en los conceptos de supervisión y no supervisión del análisis lingüístico automático, véase por ejemplo Manning y Schütze, *op. cit.* (1999, 232).

[9] De allí la popularidad de esquemas como el de Porter que se basan en un número relativamente pequeño de reglas, y que se examinará en el capítulo siguiente; véase Porter, "An Algorithm for Suffix Stripping", *Program*, 14:3 (1980, 130-137); Frakes, "Stemming Algorithms" en Frakes y Baeza, *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, New Jersey (1992, 131-169). A pesar de sus deficiencias, el algoritmo de Porter funciona relativamente bien para lenguas indoeuropeas, especialmente la inglesa.

[10] Como se verá adelante, existen desde los años cincuenta métodos de segmentación automática de palabras —véanse por ejemplo Hafer y Weiss, "Word Segmentation by Letter Successor Varieties", *Information Storage and Retrieval*, 10 (1974, 371-385), basado en Harris, "From Phoneme to Morpheme", *Language* 31:2 (1955, 190-220); véanse también Cromm, *Affixerkennung in deutschen Wortformen. Eine Untersuchung zum nicht-lexikalischen Segmentierungsverfahren von N. D. Andreev* (sobre el método del ruso Andreev), Abschluß des Ergänzungsstudiums Linguistische

versidad tipológica de las lenguas, esto es lamentable, porque cualquier tecnología del lenguaje depende en alguna medida de la morfología de la lengua enfocada; en unas más que en otras.

Esto muestra la necesidad de estudiar los métodos para determinar morfemas automáticamente a partir de corpus electrónicos. Afortunadamente, ya se han visto investigaciones importantes para descubrir de manera no supervisada los morfemas de lenguas concatenativas y de bajos recursos, y su morfotáctica. Por ejemplo, Hammarström elaboró, a principios de este siglo, visiones generales de los estudios computacionales de la morfología de lenguas de bajos recursos (2009) y del aprendizaje no supervisado de la morfología en general (2010) que permiten contemplar la variedad de investigaciones no supervisadas que se llevaron a cabo en la primera década de este siglo[11].

Aparte del aspecto empírico y muy en relación con la selección del nivel lingüístico que se busca investigar, está el problema conceptual de definir lo que se puede y quiere extraer de un corpus. Para un trabajo de procesamiento cuantitativo del nivel morfológico que suponga la intervención mínima del analista, tiene sentido optar por los afijos, que son los fragmentos de palabra[12], morfológicamente pertinentes, que ocurren adheridos a la base, ya sea al principio (prefijos) o al final (sufijos) de los vocablos que forman, esto es, que tienden a acompañar, subordina-

---

Datenverarbeitung, Fráncfort del Meno (Cromm 1996); de Kock y Bossaert, *Introducción a la lingüística automática en las lenguas románicas*, Gredos, Madrid (1974) y *The Morpheme. An Experiment in Quantitative and Computational Linguistics*, Van Gorcum, Amsterdam/Madrid (1978).

[11] Véanse Hammarström, "A Survey of Computational Morphological Resources for Low-Density Languages", *Journal of the Northern European Association for Language Technology* (2009, 105-130) y Hammarström y Borin "Unsupervised Learning of Morphology", *Computational Linguistics* (2010, 309-350).

[12] El término *palabra* o *palabra gráfica* se referirá la serie de caracteres (letras o representaciones de fonemas) que ocurren en la cadena textual, esto es, en un corpus. El término *vocablo* se referirá al conjunto de ocurrencias de las palabras en esa cadena. Vale la pena recordar que la palabra hablada no corresponde exactamente a la palabra escrita y que las palabras gráficas no pueden considerarse todas del mismo tipo; de nuevo, unas son palabras función o gramaticales y las otras son palabras de contenido.

damente, a las bases y raíces de las palabras. Más adelante se retomará este concepto.

Sobre la *afijalidad* de los signos

En la cadena hablada y en la escrita se encuentran rastros de una multitud de fenómenos. El problema de extraer de allí información lingüística, además de empírico, es conceptual, metodológico y evaluativo. Concebir unidades, por ejemplo, morfológicas, implica determinar las características formales que las definen. Cuando esos rasgos son en efecto de tipo formal o estructural, estas unidades se pueden contar y sus cualidades se pueden medir para corroborar, o refutar, la presunción inicial de que esos rasgos en efecto caracterizan las unidades concebidas.

Con el término *afijalidad* nos referiremos a 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. Este término parece especialmente adecuado porque el sufijo *~idad*, además de formar vocablos con significados que hacen referencia a la *cualidad* de aquello calificado con el adjetivo al que se adhiere, forma también vocablos como *pluviosidad*, *mortalidad*, *natalidad*, etc., que se refieren a la *cantidad* de lluvia, muertes o nacimientos[13]. De esta manera, el sentido del término *afijalidad*, aunque cualitativo, será aquí, por la naturaleza de los métodos que se utilizan para estimar esta cualidad, en esencia cuantitativo[14]. Además, como se verá, este término puede caracterizar afijos en vocablos con diferentes grados de lexicalización, como *~illo* en *armadillo*, *membrillo*, *blanquillo*,

[13] Véase Rainer, *Spanische Wortbildungslehre*, Niemeyer, Tübingen (1993, 530).

[14] En lo que respecta a las alternativas para nombrar a este concepto, *afijidad* y *afijabilidad*, el primero va contra la tendencia general del sufijo *~idad* de adherirse a adjetivos, a pesar de las poquísimas excepciones de sustantivos que también son adjetivos; por ejemplo, *hermandad* y *complicidad*. El segundo, *afijabilidad*, implica un proceso posible —del verbo *afijar*— o la posibilidad de su resultado, lo cual es más complejo que la cualidad de ser un afijo. De hecho, este término podría ser útil para referirse a las diferentes medidas de probabilidad que se pueden calcular de un afijo (véase la sección de *Probabilidades de los afijos*, en el capítulo sobre el signo afijal).

estanquillo, *bocadillo*, tan lexicalizados que cuentan con entradas en los diccionarios, y *poquillo* y *lapicillo*, que son formaciones circunstanciales, resultado de la productividad de este sufijo y, principalmente, de la creatividad del hablante.

Intuiciones sobre cómo medir la *afijalidad*

Como se dijo antes, cuando oímos hablar a alguien, puede ser que escuchemos cosas que nos sorprendan, que nos intriguen, que nos informen o, en su defecto, que notemos que ya sabemos. Similarmente, cuando leemos un texto nos encontramos con cadenas de caracteres que nos desconciertan por inesperadas, que nos intrigan porque no las entendemos, porque no las conocemos o porque ocurren en lugares inesperados. Pero muchas de las cosas que oímos o leemos no nos causan asombro ni nos intrigan porque ya intuíamos que iban a ser dichas, es decir, las estábamos esperando; sabíamos que tarde o temprano tenían que ocurrir. Algunas constituyen una especie de vehículo gramatical que facilita la transmisión de los mensajes. Son, como se mencionó, los segmentos gramaticales (artículos, preposiciones, marcas de flexión y derivación, etc.) los que constituyen este vehículo que facilita la comunicación de prácticamente cualquier mensaje que pueda contener un texto hablado o escrito. De nuevo, los segmentos gramaticales proporcionan la estructura que matiza a los de contenido, que típicamente son sustantivos y verbos. De esta manera, al escuchar a alguien hablar, experimentamos una oscilación entre lo común y lo inesperado, que corresponde al tránsito entre signos gramaticales y de contenido.

En el nivel pragmático, la estructura informativa de la oración se estudia con categorías como tópico y foco, que a grandes rasgos corresponden, respectivamente, a la información previamente dada en el discurso (esto es, aquello de lo que ya se habló) y a la información nueva (aquello que se menciona por primera vez). Sin embargo, la oscilación entre lo común y lo incierto a la que se refiere este trabajo no está rela-

cionada con este tipo de estructura informativa. Más bien se percibe en el nivel morfológico, incluso morfosintáctico, de la cadena de signos; esto es, al notar qué tan predecible es la ocurrencia de una base o un afijo, o un clítico y una palabra, en un sintagma determinado.

Como veremos, esta oscilación puede medirse. Tradicionalmente, se ha utilizado la mera frecuencia como método para distinguir entre lo esperado y lo inesperado, así como para medir lo gramatical y lo agramatical o lo productivo y lo no productivo. De hecho, desde por lo menos los años cincuenta del siglo pasado se sabe que medir la impredecibilidad de ocurrencia de ciertos segmentos en ciertos contextos ayuda a descubrir fronteras morfológicas[15]. Sin embargo, esta oscilación entre lo esperado y lo incierto no es exclusiva de la cadena de signos: así como hay comunicación sin palabras, también hay palabras que comunican poco o adquieren su significado en sus contextos situacionales.

Como sea, el concepto de *entropía* de Claude Shannon[16] se usa para medir la impredecibilidad en general, por lo que también sirve para medir esta oscilación, especialmente en el interior de las palabras. Como veremos, la entropía corresponde a la cantidad de información que contiene un conjunto de signos. De hecho, si medimos la cantidad de información de cada partícula gramatical, que evidentemente son muy frecuentes, podemos esperar que esa cantidad de información sea menor que la de cualquier palabra de contenido[17]. En otras palabras, cabe esperar que

[15] Véanse Harris, art. cit. (1955, 190-20), Hafer y Weiss, art. cit. (1974, 371-385), Frakes, art. cit. (1992, 131-160), el reporte que hace Michael Oakes del trabajo de Joula, Hall y Boggs (1994) en *Statistics for Corpus Linguistics*, Edinburgh University Press (1998, 86-87), etcétera.

[16] Shannon y Weaver, *The Mathematical Theory of Communication*, University of Illinois, Urbana (1964 [1949], 14, 50). Este concepto se usa para medir caos, desorganización, incertidumbre, etc. y, en aparente contradicción, también información, organización, sorpresa, etc., porque la medida de los primeros corresponde a la de los segundos: la información necesaria para organizar el caos es del mismo tamaño que la incertidumbre que causa ese caos o la sorpresa que experimentamos al contemplarlo.

[17] Por ejemplo, considérese el sufijo *-mente*, que funciona básicamente como derivativo pero que originalmente era una base para formar compuestos. Su significado, hoy en día de naturaleza sobre todo gramatical, se captura en un párrafo breve que describe para qué sirve: "Sufijo que sirve para formar adverbios de modo, añadiéndolo a los adjetivos femeninos: 'prontamente', 'sabiamen-

los signos que le dan estructura al léxico y al sintagma —como los afijos y los clíticos— contengan menos información que los signos plenos o de contenido —los lexemas libres o ligados.

Además, la capacidad combinatoria de los signos también puede tomarse en cuenta: si los segmentos gramaticales dicen por sí mismos relativamente poco sobre el contenido del mensaje y del discurso, los segmentos de contenido dicen mucho más al combinarse con los primeros. En otras palabras, muchos lexemas podrán aglutinarse con algunos pocos afijos derivativos para formar un número mayor de signos de contenido del nivel léxico.

Evidentemente, no cualquier raíz se combina con cualquier afijo: los patrones combinatorios pueden dar lugar a estructuras relativamente rígidas, particulares a cada lengua en complejidad y flexibilidad. De todos modos, la aglutinación de signos de contenido con signos afijales resulta en estructuras que hacen, de los sistemas lingüísticos, sistemas económicos. Esto es muestra de que en un corpus hay más que cadenas de letras. De hecho, si hay ahí algo más que una secuencia intermitente de palabras gráficas, vale la pena examinar cómo los datos cuantitativos que se pueden obtener de allí sirven para descubrir los afijos de la lengua de ese corpus.

De esta manera, para contar la presencia de estructuras combinatorias en un corpus, éstas se pueden definir de diversas maneras; por ejemplo, contando las combinaciones de signos llamados *cuadrados* o

---

te'. Puede añadirse acomodaticiamente a todos los adjetivos que lo admiten por su significado", *Diccionario de uso del español*, Gredos, Madrid, *s.v.* (Moliner 1992). Como se ve, se trata sobre todo de información de los contextos en que aparece; poca información de carácter combinatorio. Por otra parte, el sustantivo 'mente', de naturaleza léxica plena, es más informativo y suele ser mucho menos frecuente que el sufijo, en prácticamente cualquier muestra textual. Se define en el mismo diccionario en términos de vocablos de contenido con significados como 'inteligencia', 'facultad', 'pensamiento', 'intimidad', etc. En el *Diccionario del español de México* (2010) todavía no hay una entrada para el sufijo, pero la definición del sustantivo utiliza también términos con gran cantidad de información: 'pensamiento', 'inteligencia', 'memoria', 'conciencia', 'juicio', etc. Como veremos adelante, la definición técnica de entropía o sorpresa permite estimar la diferencia de contenidos informativos: lo más raro contiene más información y lo más común contiene menos.

cuadros[18]. Además, está la caracterización de esas combinaciones según su capacidad de generar nuevos signos de los niveles siguientes; por ejemplo, el mayor o menor número de objetos a los que los afijos se adhieren es una aproximación a su cualidad de ser económicos. Lo importante es que esto también se puede contar: mientras se adhieran a más signos, más económicas serán sus relaciones[19].

Además, podemos concebir los afijos como palabras que se han desgastado fonológica y semánticamente a tal grado que, después de ser piezas léxicas plenas e independientes, ahora sólo aparecen adheridos a otras formas. Por consiguiente, podemos hipotetizar que —además de ser muy frecuentes y ocurrir, por lo tanto, en un gran número de estructuras combinatorias— contienen poca información en el sentido técnico del término (porque son muy probables en la cadena hablada) y se adhieren a muchas bases léxicas para darles estructura a las palabras y al discurso en que aparecen. La ventaja de estas propiedades es que pueden medirse y eso es el tema de este trabajo.

En otras palabras, los afijos —en su calidad de signos que le dan estructura a las palabras[20]— contienen menos información (esto es, al ocurrir en la cadena hablada, sorprenden menos) que las bases con que se asocian, ya que estas últimas cargan el grueso del contenido transmitido por el texto. Además, los afijos, que suelen ser pocos, son considerablemente más frecuentes que las bases. De allí que el número de combinaciones en las que participan sea enorme. Finalmente, el hecho de ser pocos implica una relación de economía entre los signos del nivel

[18] Joseph Greenberg, *Essays in Linguistics*, The University of Chicago Press, Chicago (1967 [1957], 20). La definición formal de estas estructuras se examinan en el capítulo sobre el signo afijal.

[19] Como veremos adelante, un método interesante para medir estas relaciones económicas es el cociente propuesto por Kock y Bossaert, *op. cit.* (1978, 21-26, 30-32).

[20] Los afijos, tanto derivativos como de flexión, codifican información pertinente de la estructura interna de la palabra (por ejemplo, informan sobre la categoría gramatical—adj., adv., etc.) y de la externa, pertinente al mensaje y al discurso (por ejemplo, fenómenos anafóricos y de concordancia).

morfológico y los signos del nivel léxico: unos pocos del primero contribuyen a la formación de muchas palabras.

En resumen, podemos adelantar que los afijos se caracterizan cuantitativamente por su número limitado, su alta frecuencia, sus muchos contextos, su baja entropía (menor contenido de información) y su alta aparición en muchos vocablos; y que se adhieren a numerosos segmentos de contenido, bases o lexemas, que son poco frecuentes, lo que resulta en una mayor economía de signos.

Por otra parte, los vocablos formados mediante afijos pueden tener, como se mencionó, varios grados de lexicalización, ya sea porque los hablantes los perciben como parte del vocabulario de su lengua típicamente documentado en un diccionario (por ejemplo, *armadillo, mosquito, payasito, pañuelo*, etc.) o como construcciones cuyo significado se infiere de la concatenación de sus partes (*arbolito*, *poquito*, *popeyezco*, etc.). Los primeros son más que la suma de sus partes, porque han adquirido un nuevo sentido (en México, un *payasito* no es un payaso pequeño). En cambio, el significado de los segundos sí se infiere de sus partes (un *arbolito* sí es un árbol pequeño), por lo que no necesitan documentarse en un diccionario. Evidentemente, un método de segmentación de palabras basado solamente en relaciones económicas y entrópicas, como las esbozadas aquí, no será suficiente para determinar el nivel de lexicalización de las palabras, pero, mientras la base y los afijos conserven su forma, independientemente de su grado de lexicalización, este tipo de método de segmentación encontrará, como veremos, la frontera morfológica entre ellos.

Las muestras textuales

Otra cuestión interesante es la de los corpus lingüísticos. Un corpus textual es una cadena de signos, una porción de habla escrita y finita, una muestra de lo que los hablantes suelen escuchar o leer. Como se dijo, esa simple secuencia de signos tiene una estructura gramatical que también

comunica significados concretos y abstractos, particulares y generales, comunes y extraordinarios, y significados estructuradores del discurso, o gramaticales, aquellos que contribuyen a revelar el contenido de los mensajes que estructuran.

Por eso, mediante el análisis de corpus textuales u orales, se pueden aprender muchas cosas sobre las lenguas del mundo. Los corpus electrónicos de hoy en día suelen ser colecciones de documentos que representan el habla, oral o escrita, de una o varias lenguas. Típicamente, se recolectan como muestras que buscan garantizar la representatividad de algún aspecto o fenómeno lingüístico o extralingüístico. Normalmente, esto implica que se construyen siguiendo criterios de selección de géneros textuales y de equilibrio entre los mismos para lograr la representatividad del fenómeno que se desee estudiar[21]. Así, los corpus textuales son grandes colecciones de documentos escritos o transcritos, disponibles en algún medio electrónico, seleccionados y ordenados según criterios lingüísticos, determinados por los usuarios, para representar una variedad particular del uso de una lengua[22]. Hoy en día, un corpus de este tipo es pequeño cuando cuenta con dos millones de palabras. Sin embargo, circulan entre especialistas muestras textuales pequeñas, sobre todo de lenguas de bajos recursos, que, aunque muy valiosas y reunidas con mucho trabajo, están confeccionadas con criterios relativamente laxos por lo que su representatividad resulta limitada. Son particularmente valiosas aquellas muestras de habla natural, textos orales editados, o textos escritos por hablantes, etc. Vale la pena investigar la calidad de los resultados que se pueden obtener a partir de esas muestras, a pesar de ser pequeñas, al aplicarles métodos cuantitativos para descubrir afijos.

Muchas veces se llama corpus a cualquier muestra textual, especialmente si se encuentra en un medio electrónico. Sin embargo, conviene distinguir entre las muestras textuales construidas con criterios muy

[21] Véase Sierra Martínez, *Introducción a los corpus lingüísticos*, Instituto de Ingeniería, UNAM, México (2017, 3-5).

[22] McEnery y Wilson, *Corpus Linguistics*. Edinburgh UP, Edinburgh (2001).

precisos para representar aspectos puntuales de una lengua, como su uso o algún otro fenómeno lingüístico, y aquellas que se confeccionaron sin un método definido y que son generalmente pequeñas en comparación con las primeras o son recursos ya existentes elaborados con otros objetivos, como narrar historias o desarrollar un tema social o cultural.

En las primeras, podemos observar patrones lingüísticos que luego es posible, con cierto nivel de confianza, generalizar como fenómenos de las lenguas o de las áreas del conocimiento representadas en ellas. En cambio, en las segundas encontramos cosas que no podemos generalizar como distintivas de las lenguas muestreadas, porque no fueron compiladas con criterios de representatividad o fueron compiladas con objetivos distintos. Los primeros suelen ser proyectos ambiciosos de largo aliento y buenos apoyos institucionales, por lo que en realidad son pocos. Los segundos son esfuerzos aislados de estudiantes e investigadores, muchas veces con poco apoyo institucional, y suele haber muchos. Aunque ambos tipos son muestras textuales, en este trabajo nos referiremos a los primeros como corpus y a los segundos como muestras textuales. Naturalmente, la frontera entre unos y otros no es clara. Lo importante es que podemos aprender de ambos tipos, aunque no podamos hacer inmediatamente generalizaciones sobre las lenguas que representan. En este trabajo, utilizaremos ambos tipos de materiales de lenguas de México y Europa.

Mapa del libro

El primer capítulo contiene el panorama general en el que se enmarca este trabajo. Esto es, se describen algunos enfoques de carácter computacional que se ocupan de algunas técnicas de reconocimiento y segmentación morfológica de palabras. En el segundo capítulo, se presenta una investigación del nivel morfológico, destinada a determinar o descubrir automáticamente el conjunto de signos afijales de las palabras gráficas de un corpus. Específicamente, se utiliza el *Corpus del Español Mexicano*

Contemporáneo (CEMC)[23] que, como es bien sabido, se compiló el siglo pasado para determinar la nomenclatura base del Diccionario del Español de México. Luego, en el tercer capítulo se muestran los resultados de varios experimentos y se examinan algunas maneras de evaluar el método utilizado; por ejemplo, observando su utilidad en aplicaciones específicas. De hecho, se examinan resultados de experimentos con diversas muestras textuales, como una lista de vocablos nominales de la lengua checa, obtenida del *Corpus Nacional Checo* (Český národní korpus 2005), una colección de textos de la lengua ralámuli (Parra 2003) y otra de narraciones de la lengua chuj (Buenrostro Díaz, Corpus de la lengua chuj 2002), entre otras.

En el cuarto y último capítulo, se busca comparar diversas extracciones de afijos de textos de una misma lengua o familia de lenguas. Por ejemplo, se comparan los afijos de estudiantes cubanos de primaria de cuarto año con los de sexto (Pérez Marqués 2003). Luego se examinan ciertos cambios en el interior de la frase nominal definida posesiva (*las nuestras provincias*, *el vuestro pecado*, etc.), que desapareció del español alrededor del siglo XVI. También, se miden distancias entre extracciones de afijos de lenguas emparentadas. Específicamente, se comparan los afijos extraídos de algunos cuentos del tojolabal (Gómez Hernández, Palazón y Ruz 1999), con los de algunos cuentos del yucateco de la Universidad Autónoma de Yucatán (Centro de Investigaciones Dr. Hideyo Noguchi s.f.), y éstos con los de ciertas narraciones del huasteco (Meléndez Guadarrama 2010) y, finalmente, con los de algunos cuentos en lengua chuj (Buenrostro Díaz 2002).

Por último, en la dimensión diacrónica, se presentan los resultados de comparar diversas extracciones de sufijos de muestras del español del siglo XVI al XX. Específicamente, se comparan los resultados extraídos del CEMC con aquellos de los siglos XVI y XVIII del Corpus Histori-

[23] Véanse Lara, Ham Chande y García Hidalgo, *Investigaciones lingüísticas en lexicografía*, El Colegio de México, México (1979) y Lara, *Dimensiones de la lexicografía. A propósito del Diccionario del Español de México*, El Colegio de México, México (1990).

co del español de México (CHEM)[24]. Los documentos de este último corpus provienen de las versiones electrónicas de los *Documentos Lingüísticos de la Nueva España, Altiplano Central* (Company Company 1994), *Los procesos inquisitoriales contra indígenas que realizó Fray Juan de Zumárraga en Nueva España* (Buelna Serrano 2009) y *El habla de Diego de Ordaz: contribución a la historia del español americano* (Lope Blanch 1985). También se utilizan datos de consultas hechas a los corpus electrónicos de la Real Academia Española (http://corpus.rae.es), CORDE y CREA, y al Corpus del español de Mark Davies (http://www.corpusdelespanol.org).

[24] El CHEM se desarrolló en el Grupo de Ingeniería Lingüística del Instituto de Ingeniería de la UNAM (DGAPA PAPIIT IX 402204, IN400905 y 402008, 2005-2007). El objetivo de ese proyecto fue desarrollar un corpus textual representativo de los siglos XVI al XX y, sobre todo, las herramientas necesarias para analizarlo estadísticamente. El corpus está descrito en Medina Urrea y Méndez Cruz, "Arquitectura del Corpus Histórico del Español de México", en Hernández Aguirre y Zechinelli Martini, eds., *Avances en la Ciencia de la Computación*, Sociedad Mexicana de Ciencia de la Computación, México (2006, 248-253) y "El Corpus Histórico del Español en México", *Revista Digital Universitaria*, 12, nº 7 (2011) y se puede consultar mediante previo registro en http://www.iling.unam.mx/chem o en http://www.corpus.unam.mx:8080/unificado/index.jsp?c=chem#.

CAPÍTULO 1
ALGUNOS TEMAS DE LA MORFOLOGÍA COMPUTACIONAL

En este capítulo se presentan algunos antecedentes del campo del reconocimiento y descubrimiento de morfemas. Estas áreas de investigación son parte de la morfología computacional, que se ocupa del tratamiento de los fenómenos morfológicos de las lenguas naturales mediante procedimientos automáticos, que pueden ser: 1) simbólicos (de reglas), 2) cuantitativo-estadísticos o 3) combinaciones de ambos acercamientos.

Dada la complejidad de la morfología de las lenguas, estos estudios son de muy diversas naturalezas. Por ejemplo, algunos tienen como fin la determinación de conjuntos de reglas de reconocimiento y generación de palabras. Otros buscan encontrar reglas para describir los cambios en la morfofonología y morfofonémica de los vocablos. Otros más se ocupan del reconocimiento y descubrimiento de morfemas e investigan las técnicas de segmentación morfológica. Estos últimos recibieron mucha atención en la primera década de este siglo[1].

Podemos afirmar, al igual que Hammarström y Borin (2010, 310), que un procedimiento de aprendizaje automático de la morfología que toma de entrada texto *natural* y espontáneo, sin etiquetado ni marcas agregadas al original, se define como un ejercicio de aprendizaje *no supervisado* de la morfología[2]. Este tipo de aprendizaje produce una des-

[1] Véase, por ejemplo, el capítulo de segmentación morfológica "Segmentation and Morphology" en el *Computational Linguistics and Natural Language Processing Handbook*, Blackwell (Goldsmith 2009).

[2] Manning y Schütze distinguen los trabajos supervisados de los no supervisados, según se agreguen o no a los corpus etiquetas que representen la información estructural de la lengua: "The distinction is that with supervised learning we know the actual status (here, sense label) for each

cripción de la estructura morfológica de la lengua, con la menor *supervisión* posible, esto es, con la menor intervención del especialista, que típicamente los *supervisa* clasificando la información, seleccionando los valores de parámetros y umbrales, aplicando etiquetas, etcétera.

De esta manera, los procedimientos de aprendizaje supervisado de la morfología reciben de entrada documentos etiquetados con diferentes tipos de información. Incluso, la información de entrada puede consistir en textos elegidos *artificialmente* —esto es, seleccionados con un objetivo específico que trastoca el orden natural del habla o discurso para representar algún fenómeno de interés para el analista, como las formas de un paradigma verbal o listas de sustantivos en singular junto con sus formas plurales (Hammarström y Borin 2010, 311).

Este capítulo consta de tres apartados. En el primero, se esbozan los inicios del campo de la morfología computacional. En el segundo, se presentan algunos métodos de reconocimiento de morfemas; se trata de técnicas tempranas, típicamente supervisadas, para dividir palabras en morfemas. En el tercero, se describe el campo de la segmentación morfológica no supervisada y se presentan algunos trabajos sobresalientes de los inicios de esta disciplina.

1.1 Morfología computacional

Antes de examinar los métodos de reconocimiento y segmentación morfológica, conviene hacer un breve resumen de lo que se puede llamar morfología computacional, es decir, del estudio mediante computadoras de los fenómenos lingüísticos en el interior de la palabra. En este apartado se presentan algunas de las técnicas más conocidas de este campo. Esto es importante porque nos brinda un panorama global de la investigación automática de los fenómenos morfológicos en general.

---

piece of data on which we train, whereas with unsupervised learning we do not know the classification of the data in the training sample", *op. cit.* (1999, 232).

Más concretamente, la morfología computacional es aquella rama de la lingüística computacional dedicada al estudio del análisis y síntesis (generación o producción) de palabras[3].

El análisis de las palabras implica el reconocimiento de formas implícitas (o subyacentes) a partir de las palabras flexionadas o derivadas; es decir, la identificación de los lexemas y segmentos afijales. La síntesis, por otra parte, implica la construcción o producción automática de formas derivadas a partir de lexemas o bases y, especialmente, de formas flexionadas a partir de los lexemas. De allí que el objetivo principal de la morfología computacional sea adquirir un mejor y más explícito entendimiento de la morfología en general. Esto está relacionado con, por un lado, el análisis y la producción de oraciones y, por el otro, con el reconocimiento y la síntesis automáticos del habla (*speech recognition and synthesis*).

Su articulación con el análisis sintáctico está íntimamente ligada a la aplicación de etiquetas de categorías gramaticales a las palabras en sus contextos de ocurrencia. Por eso, muchos sistemas de análisis gramatical tienen como componentes centrales las descripciones morfológicas de las lenguas particulares que analizan.

El inglés es una lengua de flexión sencilla, por lo que a menudo se soslaya este aspecto en la mayoría de los sistemas de procesamiento de esta lengua[4] y simplemente se incluyen todas las formas flexionadas y derivadas en los componentes léxicos de esos sistemas. Por eso, los trabajos que se ocupan de lenguas más complejas morfológicamente, como el ruso, el finlandés, el árabe, las lenguas romances, etc. suelen ser más interesantes en el campo de la morfología computacional.

[3] Véase Koskenniemi, "Computational Morphology" (1992) en Bright, ed., *The International Encyclopedia of Linguistics*. Oxford, Oxford UP (1992, 291-293),

[4] De hecho, aunque la derivación en inglés es más compleja que su flexión, también tiende a soslayarse. Más adelante revisaremos el algoritmo de Porter, muy utilizado para el inglés, que con unas pocas reglas es suficiente para desnudar las palabras inglesas de sus afijos flexivos y derivativos, cosa que no se puede decir de lenguas con derivación más compleja.

Se ha propuesto que el análisis y generación de palabras dentro del marco de la morfología computacional incluya las siguientes tareas (Koskenniemi 1992, 291):

1. Estudiar los procesos fonológicos y morfofonológicos que causan variación en fonemas (o letras en la lengua escrita). Esto corresponde al dominio de la fonología.
2. Identificar morfemas de bases y afijos (prefijos, infijos, sufijos, etc.).
3. Describir las estructuras morfotácticas posibles, es decir, las secuencias y combinaciones posibles de morfemas (la manera en que se combinan prefijos, raíces y sufijos para formar palabras completas).
4. Identificar los rasgos morfosintácticos y las descripciones semánticas de las palabras a partir de las descripciones de los morfemas que las forman.
5. Describir el proceso de lexicalización, donde ciertas configuraciones de morfemas tienen propiedades que no pueden deducirse de sus componentes individuales.

A pesar de que en tiempos preinformáticos el punto 2 fue objeto de gran atención —la identificación o descubrimiento de morfemas—, fue hasta finales del siglo pasado el más descuidado en la morfología automática, en parte por la resistencia a trabajar con estructuras más complejas que las del inglés y en parte por la escasez de corpus electrónicos para la mayoría de las lenguas, especialmente las no europeas. Sin embargo, en los últimos años se han llevado a cabo numerosas investigaciones que han buscado compensar esta deficiencia[5]. El presente trabajo se circunscribe a este punto.

[5] Véanse Goldsmith, art. cit. (2009, 364-394) y Hammarström y Borin, art. cit. (2010, 309-350).

El asunto medular de los enfoques basados en reglas es la representación del conocimiento morfológico y su estructura. Típicamente, el conocimiento que el analista tiene de la morfología se representa mediante la formulación de reglas: cuando un sistema de reglas funciona, hay cierta presunción de que el conocimiento codificado en ellas es descriptivo del fenómeno representado por esas reglas. Anderson[6] apunta (1994, 375):

> the virtue of a computer is that it knows nothing other than what it has been told, and so when one reaches a point at which a computational procedure operates correctly, it is reasonably certain that all of the underlying assumptions and subprocedures involved in the description have indeed been made fully explicit.

Para Anderson, esto es aceptable en sistemas cuyo objetivo no es investigar la morfología, pero no es una justificación de su estudio científico mediante computadoras. El problema está en que rara vez se revisan los presupuestos y, aunque éstos no sean centrales al trabajo, los teóricos no se pueden dar el lujo de dejar de investigarlos. La condición que Anderson propone para aceptar esta dinámica como científicamente válida es que la naturaleza de las reglas que se formulen refleje los principios de la disciplina lingüística. Y considera que la morfología cuenta con una gran ventaja en esto, ya que confía en que una vasta bibliografía sobre los conocimientos de la psicología y de la ciencia cognoscitiva se ha constituido en parte de esta disciplina.

Así que muchos trabajos de morfología computacional se basan en la formulación de reglas como metodología privilegiada para desarrollar formalismos lingüísticos. Presumen que mediante estos formalismos se pueden construir "precise and elegant descriptions of morphological phenomena"[7]. Quizá el método de reglas morfológicas más

[6] Anderson, *A-Morphous Morphology*, Cambridge UP, Cambridge (1994).

[7] Ritchie *et al.*, *Computational Morphology*, MIT Press, Cambridge, Mass. (1992, 11).

conocido sea el de los trabajos de Martin Kay y Kimmo Koskenniemi. Aunque hoy en día no es un campo muy popular de la fonología, propiamente dicha, estos trabajos se conocen bajo el término de *fonología de estados finitos*, porque se basan en la aplicación de un tipo de máquina abstracta llamada *transductor* (*transducer*)[8]. En esencia, un transductor es un tipo de autómata que translitera los símbolos de un sistema de escritura en los de otro. Los autómatas son máquinas de estados finitos que equivalen a sistemas de reglas y son temas centrales de la computación en general y del procesamiento del lenguaje natural en particular.

A principios de los años ochenta, en el marco de la fonología de reglas generativas, se concibieron los primeros transductores de estados finitos para el reconocimiento y generación de la estructura superficial de las palabras, en oposición a la forma léxica o subyacente. Como se dijo, un transductor es una máquina de estados finitos que —como los aparatos que se usan en física para transformar algún tipo de señal (movimiento, onda, excitación, etc.) en otro tipo de señal— sirve para convertir las cadenas de caracteres de un alfabeto en cadenas de otro[9]. Para esto, un transductor consiste en, además de un número finito de estados, un conjunto finito de símbolos de entrada, otro de símbolos de salida y una especificación de correspondencias (*mapping*) entre los primeros y los segundos. En cierta manera, los transductores definen dos lenguajes que son *traducciones* el uno del otro.

En la fonología de estados finitos estos lenguajes corresponden a los niveles léxico y de superficie de una misma lengua. Así, dada una forma del primer nivel se llega a la forma superficial, y dada una forma

[8] Véase la definición formal de transductor en Aho y Ullmann, *The Theory of Parsing, Translation, and Compiling*, Prentice-Hall, Nueva York (1972, 224).

[9] Esto es, son autómatas cuyas transiciones tienen asociadas operaciones de "salida" (Ritchie, y otros 1992, 19). Véanse también Jurafsky y Martin, *op. cit.* (2009), Manning y Schütze, *op. cit.* (1999, 367). Para una introducción a los autómatas en general, véanse Robert Wall, *Introduction to Mathematical Linguistics*, Prentice Hall, Englewood Cliffs, Nueva Jersey (1972, 254-287) y Glück, *Metzler Lexikon Sprache*, Verlag J.B. Metzler, Stuttgart (2000), *s.v.* Automat.

superficial se puede llegar al nivel léxico o subyacente. En la Figura 1, se ilustra un transductor de estados finitos para una regla de este esquema[10].

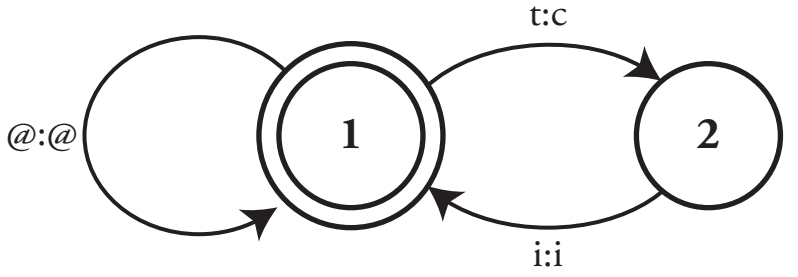


Figura 1. Transductor de estados finitos para la regla t:c ⇒_i

Los círculos son estados posibles. El número 1 es simultáneamente el estado inicial y el final, lo que se simboliza mediante el doble círculo. Los arcos o transiciones muestran las correspondencias de los símbolos entre niveles. Los símbolos del nivel léxico van antes de los dos puntos [:] y los del nivel superficial después. El transductor de la Figura 1 representa la regla de reescritura t:c ⇒ _i. Esta regla especifica que, cuando ocurre /t/ en el nivel léxico, se manifiesta como 'c' en el nivel superficial sólo si /t/ ocurre antes de /i/ (en el nivel léxico); por lo que /i/ se manifestará como 'i' en el superficial. El arco @:@ se refiere a todos los demás símbolos de la cadena a transducir y se especifica para permitir el paso de otros caracteres por esta regla. Con este tipo de notación se pueden especificar contextos más complejos. Permite, por ejemplo, que se definan tipos de símbolos, como vocales o fricativas alveolares. También, las fronteras morfológicas que se manifiestan en el interior de las formas léxicas se pueden marcar mediante el signo [+]. Considérense las siguientes representaciones que Antworth utiliza para ilustrar la aplicación de las reglas (1990, 31):

[10] Ilustración tomada de Antworth, *PC-KIMMO: A Two-level Processor for Morphological Analysis*, Summer Institute of Linguistics, Dallas (1990, 44).

LR (*representación léxica*):	0	t	a	t	+	i
	↕	↕	↕	↕	↕	↕
SR (*representación superficial*):	'	t	a	c	0	i

En esta representación, el [0] es el símbolo nulo que cancela símbolos de un nivel que no ocurran en el otro; ['] es el acento silábico del nivel superficial; y [+] es la marca, como se dijo, de la frontera morfológica, que corresponde al símbolo nulo en el nivel superficial. El procesador analiza cada par de caracteres de un nivel y otro para determinar si se corresponden según el conjunto de reglas que conforma la descripción de la lengua. En este ejemplo, el programa avanza de izquierda a derecha, mediante el arco @:@, hasta llegar a la cuarta columna. Al llegar a /t/ del nivel léxico, avanza al estado 2 mediante el arco t:c copiando una 'c' en el superficial. Luego, como aparece /i/ después de /t/, regresa al estado 1 copiando una 'i' al nivel superficial.

Además, un transductor también puede funcionar como *reconocedor* o como *generador* de cadenas. Como reconocedor, acepta representaciones superficiales para obtener las formas léxicas, por ejemplo[11], *bigger* → 'big+er, *spies* → 'spy+s, *foxes* → 'fox+s. Como generador, acepta las representaciones subyacentes y genera las formas superficiales: 'cat+s → *cats*, 'try+ed → *tried* y 'fox+s+'s → *foxes*', que es el genitivo del plural de *fox*.

La investigación más temprana sobre la relación entre reglas fonológicas y transductores de estados finitos se debe a Martin Kay y Ronald Kaplan y data de principio de los ochenta[12]. Estos autores sugirieron que las reglas ordenadas de la fonología generativa se pueden hacer operativas computacionalmente mediante una secuencia *en cascada* de transductores de estados finitos, *cascading sequence of finite state trans-*

[11] Nótese que se trata de ejemplos que consideran que la forma superficial es la escrita.

[12] Kaplan y Kay, "Phonological rules and finite-state transducers", Annual Meeting of the Linguistics Society of America, Nueva York (1981). Para discusiones sobre este trabajo, véanse Ritchie *et al.*, *op. cit.* (1992, 20); Koskenniemi, art. cit. (1992, 292); y Sproat, *Morphology and Computation*, MIT Press, Cambridge, Mass. (1992).

ducers (Antworth 1990, 6). Una cascada así se puede combinar para formar un solo autómata.

Esto inspiró el modelo de dos niveles de Koskenniemi que se distingue del modelo generativo en que, como se puede apreciar en el ejemplo de arriba, no consta de secuencias de reglas fonológicas que conduzcan de un nivel a otro, sino de reglas que en un solo paso transforman cadenas de un nivel directamente en cadenas del otro. De allí el nombre de *two-level rules*, reglas de dos niveles.

Así, el esquema de estados finitos se distingue del enfoque generativista por las restricciones que se imponen a las reglas para evitar que alguna opere en los resultados de otra (Ritchie, y otros 1992, 21). Como se vio arriba, todas las reglas involucran los dos niveles y son declaraciones lógicas que definen las correspondencias aceptables entre las dos representaciones; por ejemplo, *taci* ⇔ 'tat+i se corresponden —una se reconoce en la otra y la otra genera a la primera— de una sola vez mediante la regla t:c ⇒ _i:i.

Antworth apunta que, además de ser un esquema más económico que el de la fonología generativa (una sola regla de éste corresponde a varias de esta última), es también uno compatible con las ideas de la fonología natural, que rechaza el poder arbitrario e ilimitado de las reglas ordenadas y los resultantes niveles intermedios. El método de Koskenniemi fue divulgado por Karttunen[13] con el nombre de KIMMO, gracias a la publicación de una implementación en LISP acompañada de descripciones de dos niveles del inglés, rumano, francés y japonés[14]

Evidentemente, un enfoque de estados finitos no es capaz de rendir cuenta de todos los fenómenos morfológicos de todas las lenguas. Sin embargo, hubo quienes consideraban que era suficiente por lo menos

[13] Karttunen, "KIMMO: a General Morphological Processor", *Texas Linguistic Forum* 22 (1983, 163-186).

[14] La implementación de Antworth, *op. cit.* (1990) para computadoras personales fue auspiciada por el Instituto Lingüístico de Verano, data de la segunda mitad de los años ochenta y estuvo acompañada de descripciones de muchas lenguas.

para caracterizar lenguas complejas y aglutinantes como el finlandés. Así, por ejemplo, Jäppinen construyó un analizador morfológico de estados finitos para esa lengua, basado en un esquema distinto al de dos niveles de KIMMO, que se difundió comercialmente en los años ochenta[15]. Este analizador generaba descripciones gramaticales (marcas de caso, número, posesión, etc.) de las palabras, a partir de reglas formuladas por el investigador utilizando su conocimiento del finés.

Como se puede ver, los métodos basados en reglas como éstos no resuelven el problema de identificar los morfemas de bases y afijos de las lenguas. En realidad, presuponen que los especialistas los conocen y por eso los utilizan en el desarrollo de máquinas abstractas que permiten navegar de la representación estructural de un nivel a la del otro.

1.2 Reconocimiento supervisado de morfemas

En esta sección, se examinan trabajos que utilizan diccionarios para encontrar morfemas en un texto mediante la comparación de patrones para detectar coincidencias (*pattern-matching*). En esencia, son sistemas donde el especialista codifica el conocimiento de la morfología en alguna estructura de datos que permite reconocer los morfemas en un texto mediante la comparación de cadenas de caracteres. En otras palabras, estos procedimientos reconocen morfemas comparándolos con la estructura de información donde se encuentra codificado *a priori* el conocimiento de la morfología. Normalmente, el especialista mismo es quien codifica manualmente lo que sabe de la lengua o lo que encuentra acerca de ella en las descripciones gramaticales a su disposición.

[15] Véase Jäppinen, "Finite State Computational Morphology" (1992), en Klenk, ed., *Computation Linguae I*, Steiner, Suttgart (1992, 96-109).

1.2.1 Primeros acercamientos al aprendizaje morfológico

Gregor Thurmair propuso un esquema que aplicó al alemán y el inglés a mediados de los ochenta. El objetivo de su trabajo fue la segmentación morfológica automática basada en el *aprendizaje* previo de la estructura morfológica. El objetivo principal[16] era convertir automáticamente grandes cantidades de palabras gráficas, sacadas secuencialmente de textos, a sus formas canónicas mediante la ayuda de reglas generadas a partir de una lista de ejemplos, compilada por el investigador. En este estudio, el término *aprendizaje* se refiere a la codificación de reglas a partir de la información que el lingüista proporciona sobre la lengua (es decir, el lingüista tiene que especificar dónde se segmentan las palabras). Por ejemplo, el analista construye un archivo con una lista de adjetivos alemanes flexionados con marca de dativo como la siguiente:

SCHOENEM - Cut2
REICHEM - Cut2
WILDEM - Cut2
WIRREM - Cut2
⋮
BEQUEM - Cut0

Cada adjetivo de la lista tiene una notación agregada por el lingüista que indica cuántos caracteres de la derecha deben eliminarse para quedarse con la forma básica. En el caso de esta lista el autor del sistema tuvo que agregar la marca 'Cut2' que significa 'quitar los dos últimos caracteres'. Así, para obtener la forma básica del adjetivo con marca de dativo 'SCHOENEM' basta con cortar los dos últimos para quedarse con 'SCHOEN'. Por otra parte, aquellas palabras como 'BEQUEM' que no son adjetivos con marca de dativo se asocian (manualmente) a la notación 'Cut0' que le dirá al codificador que no quite ninguna letra. A partir de listas como ésta, el sistema *aprende* a segmentar las

[16] Thurmair, "Ein Morphologisches Prozessorsegment zur Erzeugung von Grundformen mithilfe von Lernverfahren" (1986) en Schwarz y Thurmair, eds., *Informationslinguistische Texterschließung*, Georg Olms Verlag, Zürich (1986, 8-31).

palabras y debe aprender que, en todos los casos, excepto cuando haya una 'U' inmediatamente antes de los dos últimos caracteres (como en 'BEQUEM'), se eliminan las dos últimas letras, incluso cuando se trate de palabras que no ocurrieron en esta lista.

Este tipo de procedimiento, apunta Thurmair, es apropiado sobre todo para lenguas con morfología y sistema de flexión ricos y permite el manejo de las excepciones con la simple inclusión de reglas que rindan cuenta de ellas. Nótese, sin embargo, que no permitir la segmentación de 'BEQUEM' se debe a que se trata de una raíz completa, no de que el dativo de esa forma se marque de otra manera. Evidentemente, se trata de instrucciones sencillas para eliminar caracteres, sin mucha motivación lingüística.

Esto es pertinente porque, para muchos lingüistas computacionales, las reglas se conciben explícitamente como hipótesis sobre los contextos de las letras al final de la palabra y de cómo se procesan —"Standardhypothesen über den Zusammenhang von Endgraphemen und ihrer Verarbeitung" (Schwarz y Thurmair 1986, 9)—, procesamiento al que han calificado de cognoscitivo, es decir, del hablante.

Estas hipótesis, los renglones de listas como la de arriba, se codifican en árboles que luego se utilizan para producir una etiqueta de la categoría lematizada a la que pertenece cada palabra representada por la cadena de caracteres que se analice. En resumen, el sistema tiene dos fases, una para el aprendizaje y otra para el análisis basado en los datos producidos por la primera. Los resultados erróneos que se den en el análisis se pueden corregir agregándole a la lista de aprendizaje las reglas apropiadas para su manejo correcto.

1.2.2 Codificación de gramáticas

Como se dijo arriba, Thurmair aplicó su procedimiento al alemán y al inglés. Montserrat Meya, por otra parte, ilustra la parte de reconocimiento de patrones del trabajo de Thurmair en su propuesta para el

español[17]. La analista utilizó un corpus pequeño de español de España para comprobar las hipótesis que ella misma formuló y que constituyen una gramática de descomposición (*Zerlegungsgrammatik*).

Una diferencia importante entre los procedimientos de Thurmair y de Meya es que, en el segundo, no se construyen árboles con las reglas hipotéticas de la morfología española, es decir, no recurre a una fase de aprendizaje, que es quizá lo más interesante. En cambio, la gramática de la investigadora describe los tipos de morfemas según el vocablo en que aparecen, si se trata de un prefijo adverbial o denominal, si es o no la raíz de un verbo, etc. (Meya 1986, 138-142).

El análisis se lleva a cabo mediante una lista de 7 000 morfemas seleccionados por la analista[18], tanto libres como ligados (raíces y afijos), los cuales permiten que el análisis se lleve a cabo mediante un sencillo proceso de comparación para detectar patrones (*pattern-matching*). Esto requiere de listas de transformaciones de grafías, alomorfos y formas supletivas que rindan cuenta, entre otras cosas, de las modificaciones vocálicas de la flexión, fonemas epentéticos, etc. (por ejemplo, *feliz* → *felicidad*, *cont~* → *cuent~*, *perd~* → *pierd~*, *produc~* → *produzc~*, etc.). Por último, se toman en cuenta varias propiedades morfológicas específicas a la lengua española, como los modelos de los tres paradigmas verbales (*~ar*, *~er* e *~ir*) y sus marcas de número, modo, tiempo, aspecto, etc. y los cambios de acentuación en palabras derivadas.

Como se ve, este esquema se basa, como la mayoría de los estudios automáticos de la morfología, en varias de las características conocidas de la lengua en cuestión, ejercicio sin duda interesante, pero no descubre los morfemas a partir de los datos contextuales; esto es, de los hechos reales. Lo mejor del trabajo de Thurmair reside en el proceso

[17] Meya, "Morphologische Analyse des Spanischen" (1986) en Schwarz y Thurmair, eds., *op. cit.* (1986, 134-156).

[18] Donde se incluyen todos los morfemas que aparecen en Juilland y Chang Rodríguez, *A Frequency Dictionary of Spanish Words*, Mouton, La Haya (1965), y en la mitad del diccionario de Slabý, Grossmann e Illig, *Wörterbuch der spanischen und deutschen Sprache*, Brandstetter, Wiesbaden (1975).

llamado de aprendizaje. La eliminación de esta fase hace que el trabajo de Meya no difiera mucho de otros trabajos cualitativos, como los que se mencionan a continuación.

1.2.3 Reconocimiento de morfología discontinua

Ursula Klenk presentó, en varios artículos y dos volúmenes editados por ella[19], trabajos de diversos investigadores, dedicados a distintas lenguas, que ilustraban la pertinencia de diferentes métodos, tanto cualitativos como cuantitativos, en la determinación automática de fronteras morfológicas que no dependiera de listados completos del léxico de esas lenguas.

Klenk misma propuso métodos para el español y el árabe[20]. Para este último, por ejemplo, su método se limita al análisis de formas verbales regulares, algunas raíces verbales de tres consonantes. Lo interesante del procedimiento es el carácter discontinuo de la morfología árabe, que se opone al método de estados finitos desarrollado —obviamente— para lenguas de morfología en serie o concatenativa (*anreihender Morphologie*), es decir, que consisten en cadenas de bases y afijos. Los esquemas de estados finitos no son capaces de representar fenómenos de morfología discontinua.

En el método propuesto se especifican las secuencias posibles de vocales y patrones (Schemata: "iX*YaZiZ" —forma en imperativo— donde X, Y, Z representan las consonantes y * indica ausencia de vocal). En estos patrones, se buscan las secuencias particulares de fonemas de las palabras de entrada y sirven para asignar los rasgos morfológicos a

[19] Klenk, ed., *Computation Linguae I* (1992) y *Computation Linguae II*, Franz Steiner, Stuttgart (1994).

[20] Klenk y Langer, "Morphological Segmentation Without a Lexicon", *Literary and Linguistic Computing*, 4:4 (1989); Klenk, "Verfahren morphologischer Segmentierung und die Wortstruktur des Spanischen", en Klenk, ed., *op. cit.* (1992, 110-124); y Klenk, "Automatische morphologische Analyse arabischer Verbformen", en Klenk, ed., *op. cit.* (1994, 84-101).

estas palabras (83 rasgos complejos, por ejemplo, si se trata de formas pasivas, infinitivas, imperativas, etc.). Así, la palabra de entrada se compara con un diccionario y con estos esquemas para determinar si se trata de una secuencia permitida y, en caso de que así sea, señalar sus rasgos morfológicos. Como puede verse, se trata en esencia de reglas para asignar estructuras morfológicas.

1.2.4 Combinaciones de letras y sus frecuencias

Varios métodos de reconocimiento de morfemas se basaron en la observación del lugar donde ocurren los caracteres y sus frecuencias. Por ejemplo, Klenk desarrolló dos programas[21]: MORSPAN y MORGRA. El primero fue un trabajo hecho especialmente para el español y el segundo es una extensión que, basada en los mismos principios, se puede aplicar a otras lenguas. Klenk se basa en el hecho de que la palabra española se puede describir como la siguiente secuencia: afijo de derivación (opcional) + base + afijo de derivación (opcional) + afijo de flexión + clíticos (Klenk y Langer 1989, 248). Su objetivo principal es determinar las fronteras entre el final de la palabra (los sufijos de flexión + enclíticos) y todo lo que los preceda (base compleja). Su método se basa en el hecho de que el segmento final está compuesto generalmente por uno o varios caracteres pertenecientes a un grupo bien definido: r, l, n, s, t, d, b, m, a, á, e, é, i, í, o, ó (se trata de lengua escrita); mientras que la base puede estar formada prácticamente con todas las letras. Esto sirve para determinar un sistema de reglas de segmentación, las cuales se formulan mediante un procedimiento de descubrimiento. Pero el descubrimiento no lo hace la máquina, sino el analista, que laboriosamente tiene que examinar cada secuencia de grafemas del corpus para determinar si forman parte de uno de los segmentos de flexión posibles en español.

[21] Véanse Klenk y Langer, art. cit. (1989) y Klenk, art. cit. (1992).

Otro método interesante fue la determinación de frecuencias de pares de caracteres para utilizarlas como indicadores de fronteras morfológicas. Esta idea en particular se aplicó a varias lenguas, como alemán[22], francés[23] y español[24], entre otras. La idea central fue descrita en Klenk Langer (1989). Su procedimiento se basa en que hay ciertos pares de letras que ocurren exclusiva o predominantemente en ciertas posiciones definidas morfológicamente. Por ejemplo, pares de caracteres como 'cl', 'cr' y 'qu' tienden a ocurrir (o siempre aparecen) al principio de morfemas, lo que significa que puede haber una frontera morfológica inmediatamente antes de ellos; mientras que 'nm' y 'ks' contienen muy a menudo una frontera morfológica en su interior (1989, 250).

Para explotar esta observación, es necesario registrar el porcentaje de ocurrencias de cada combinación posible de caracteres g_1g_2 que ocurre inmediatamente después de una frontera, $A(g_1g_2)$; el de pares que contienen una frontera entre cada componente, $M(g_1g_2)$; aquel de los pares que ocurren seguidos inmediatamente por límites morfológicos, $E(g_1g_2)$; y la proporción de ocurrencias de cada par en contextos que no incluyen ninguno de los tres casos anteriores, $N(g_1g_2)$. Así, la secuencia 'st' en inglés tendrá un porcentaje $A(g_1g_2)$ por su ocurrencia en palabras como *stand*, otro $M(g_1g_2)$ por aparecer en compuestos como *messtin* (recipiente metálico utilizado por militares), un porcentaje $E(g_1g_2)$ al ocurrir en vocablos como *must* y otro $N(g_1g_2)$ por palabras como *custom*. Con esta información se construye una tabla que especifica cada uno de estos valores para cada posible combinación de caracteres. Luego, estos porcentajes se aplican en la segmentación automática de otras palabras. Véase la Tabla 1:

[22] Programa MOSES en Klenk y Langer, art. cit. (1989, 250-251).

[23] Programa MOSEF en Janßen, "Segmentierung französischer Wortformen in Morphe ohne Verwendung eines Lexikons" (1992) en Klenk, ed., *op. cit.* (1992, 74-95).

[24] Programa MOSS en Flenner, "Ein quantitatives Morphsegmentierungssystem für spanische Wortformen" (1994) en Klenk, ed., *op. cit.* (1994, 31-62).

Tabla 1. Cortes posibles del verbo *zerlegen*

g_1g_2	Z	E	R	L	E	G	E	N	
en							89%	1%	85%
ge						63%	34%	61%	
eg					1%	25%	58%		
le				33%	46%	12%			
rl			0%	100%	0%				
er		51%	3%	86%					
ze	51%	49%	11%						
	51%	50%	5%	**73%**	16%	33%	**60%**	31%	85%

Tabla basada en aquella de Langer (Klenk y Langer 1989, 251). El primer valor de cada renglón se refiere a $A(g_1g_2)$, el siguiente a $M(g_1,g_2)$ y el último a $E(g_1g_2)$.

Los valores de las columnas se combinan para determinar donde se puede segmentar la palabra. Por simplicidad, Langer saca un promedio, que se muestra en el último renglón. Nótese que los dos valores más altos (en negritas) coinciden justo con las fronteras morfológicas en el interior de *zerlegen*: entre el prefijo *zer~*, la raíz *~leg~* y el sufijo *~en*. Los resultados muestran para el alemán (Klenk y Langer 1989, 251) alrededor del 90% de palabras segmentadas correctamente (se eliminaron palabras extranjeras, abreviaturas, etc.), para el francés alrededor de 70% (Janßen 1992, 74, 89-93) y entre 68% y 94% para el español (Flenner 1994, 57)[25]. Para la época, estos resultados son muy buenos. El único inconveniente es la laboriosísima intervención del analista para determinar los porcentajes de las posiciones de las fronteras en cada par de caracteres, cosa nada trivial. Finalmente, en este grupo de

[25] Los porcentajes de aciertos para las versiones del francés y del español varían según la aplicación de componentes de reglas para mejorar los resultados logrados con este método.

investigaciones, es también el lingüista quien debe establecer dónde están las fronteras.

1.2.5 Reglas para eliminar afijos: el algoritmo de Porter

Entre los métodos más conocidos de segmentación de palabras están los que se han desarrollado en el marco de recuperación de documentos y que se caracterizan por ser programas pequeños y muy ágiles (ocupan poco espacio y son muy rápidos).

El algoritmo de Porter[26] es probablemente el más conocido y, aunque se han hecho versiones para varias lenguas, se diseñó especialmente para el inglés y refleja la sencillez de su morfología. Se trata de un *stemmer* (programa que desnuda las palabras de sus sufijos) y consiste en un conjunto de reglas que, de cumplirse ciertas condiciones, eliminan de la palabra la cadena de caracteres más larga que pueda recortarse en ese momento, lo que se repite hasta que ya no se pueden eliminar más caracteres (*iterative longest match stemming*). No hay una estructura verdaderamente representativa de la morfología aparte de las instrucciones del algoritmo; esto es, es una representación procesal. En otras palabras, la información morfológica está representada en las reglas del programa de la manera siguiente:

```
if a word ends in "ies" but not "eies" or "aies"
   then "ies" → "y"
if a word ends in "es" but not "aes", "ees" or
   "oes" then "es" → "e"
   ...
```

[26] Véanse Porter, art. cit. (1980, 130-137), Frakes, *op. cit.*(1992) y el apéndice B de Jurafsky y Martin, *op. cit.* (1999, 833-836).

Así que, si una palabra termina en *~ies*, pero no en *~eies* ni en *~aies*, entonces hay que sustituir *~ies* por *~y*, de tal manera que de *parties* se obtiene *party*. Además, se especifican varias condiciones que la base hipotética debe cumplir para que se elimine de la palabra el sufijo que se cree que se ha encontrado. Las condiciones son las siguientes:

- que contenga un número determinado (m) de secuencias VC (donde V puede incluir una o varias vocales y C una o varias consonantes):
  TR, EE, TREE, Y, BY (m = 0)
  TROUBLE, OATS, TREES, IVY (m = 1)
  TROUBLES, PRIVATE, OATEN (m = 2),
- que contenga una vocal: *v*,
- que termine con doble consonante: *d,
- que termine con cierta letra: *<X>, donde X = cierta letra,
- que termine en secuencia CVC y que la última C no sea 'w', 'x' o 'y': *o.

Como se dijo arriba, las reglas especifican que, si se cumple cierta condición en la base y cierto sufijo está involucrado, éste será sustituido por un nuevo segmento (que también puede ser el segmento Ø, "NULL").

Naturalmente la clave del procedimiento es que las reglas deben aplicarse en cierto orden. De hecho, se agrupan en pasos y subpasos. El primer paso (constituido por tres subpasos de no más de cinco reglas el más largo) recorta los poquísimos sufijos de flexión del inglés. Los cuatro pasos restantes se ocupan de los sufijos derivativos y contienen un promedio de diez reglas cada uno (el más largo tiene solamente veinte). En esencia, el corazón de este método es la formulación de reglas, a las que Porter mismo, al formular su algoritmo, llegó revisando sus diccionarios y manuales de gramática y aplicando su sentido común.

Tabla 2. Algunas reglas del algoritmo de Porter

condiciones	sufijo	reemplazo	ejemplos
NULL	-sses	-ss	caresses → caress
(m > 0)	-ed	NULL	plastered → plaster
			bled → bled
(*v*)	-ing	NULL	motoring → motor
			sing → sing
NULL	-s	NULL	cats → cat
(m > 0)	-iveness	-ive	decisiveness → decisive
(m > 0)	-alize	-al	formalize → formal

1.3 Segmentación morfológica no supervisada

Como muchas áreas del procesamiento del lenguaje natural y de la lingüística computacional, el campo del aprendizaje no supervisado de la morfología ha progresado mucho, después de más de medio siglo de investigación. Sin embargo, todavía tiene un camino largo que recorrer. En el estado del arte de este campo ya hay un sinnúmero de proyectos que investigan los procesos de segmentación morfológica[27], pero sigue habiendo poco progreso en otras áreas relacionadas, como en el descubrimiento de los significados mismos de los morfemas.

La expectativa original era determinar un procedimiento formal de descubrimiento de morfemas. Los éxitos recientes en el área han estado

[27] Véase Hammarström y Borin, art. cit. (2010). Se está llevando a cabo mucha investigación a partir de las lenguas poco estudiadas. Hammarström, art. cit. (2009) documenta ampliamente los recursos computacionales de descubrimiento morfológico no supervisado orientados a las lenguas cuyos hablantes tienen pocos recursos económicos. Observa que "computational morphological resources emerge for languages with higher affluence, and we can now also account for the manner in which this happens and for the exceptions to the rule" (2009, 105).

condicionados en gran medida por el desarrollo de metodologías estadísticas con poco conocimiento lingüístico, sobre todo basadas en la teoría de la información. En esta sección, se comentan los trabajos no supervisados de segmentación y descubrimiento de morfemas de más repercusión en las investigaciones morfológicas; específicamente, las investigaciones de N. D. Andreev, Zellig Harris, Claude Shannon, Josse de Kock, John Goldsmith y Mathias Creutz, entre otros.

1.3.1 Frecuencias de caracteres

En esta subsección se comenta el trabajo del equipo ruso dirigido por N. D. Andreev que trabajó en el reconocimiento de afijos de diversas lenguas. Este es el primer procedimiento verdaderamente automatizado de descubrimiento morfológico. Data de la primera mitad de los años sesenta y se llevó a cabo en la entonces Unión Soviética. Andreev y su equipo desarrollaron y aplicaron al ruso, al húngaro, al vietnamita y a otras lenguas un programa capaz de determinar empíricamente afijos de flexión. Cromm (1996) construyó un programa para aplicar el método de Andreev al alemán[28].

El procedimiento de Andreev está basado en el hecho de que los afijos en lengua escrita son cadenas de caracteres que se caracterizan por ser mucho más frecuentes que otras cadenas de caracteres (otros tipos de morfemas) y en que tienden a aparecer sistemáticamente con las mismas bases con que aparecen otros afijos. La idea es encontrar los paradigmas en que ocurren los afijos investigados mediante la comparación de muestras de segmentos de palabras y sus frecuencias, así como mediante la búsqueda de combinaciones de segmentos[29]. Pero en lugar

[28] *Affixerkennung in deutschen Wortformen. Eine Untersuchung zum nicht-lexikalischen Segmentierungsverfahren von N. D. Andreev*, Abschluß des Ergänzungsstudiums Linguistische Datenverarbeitung (Cromm 1996).

[29] Estas estructuras son las que Greenberg llama "cuadros" (*squares*); véase definición de cuadro más adelante.

de contar fonemas anteriores y posteriores como lo hace Harris (véase más adelante), Andreev utiliza frecuencias de caracteres según sus posiciones dentro de la palabra.

El método consiste en buscar en los extremos de las palabras (entre las primeras y las últimas letras) las cadenas de caracteres más frecuentes en esa posición en la totalidad de palabras del texto examinado. La idea es que los afijos son más frecuentes que las bases y eso se debe reflejar en la observación de estas frecuencias. Las secuencias más frecuentes se examinan luego para determinar si son o no afijos. Las secuencias restantes (aquellas que quedan después de eliminar los presuntos afijos de las palabras) se consideran entonces posibles bases. Estas últimas se examinan para determinar si aparecen combinadas con otros segmentos que se intercambian, según la evidencia en el texto examinado, con el primer presunto afijo. Si hay varios candidatos a bases que se combinan con varios presuntos afijos que se intercambian entre sí, se ha encontrado un paradigma.

Con el objeto de llevar todo esto a cabo, se calculan varios parámetros para determinar qué combinaciones se deben menos al azar y para distinguir los paradigmas de flexión de los de derivación, siendo los primeros el objetivo principal del procedimiento. Los índices principales (Cromm 1996, 23-26) son los siguientes: una medición de desnivel (en ruso *mera perepada* o en alemán *Gefällemaß*), una función correlativa (ruso *korrelativnaja funkcija*, alemán *korrelative Funktion*) y una medición de reducción (*mera redukcii* o *Reduktionsmaß*). Las supuestas bases que resultan de este procedimiento también se examinan cuantitativamente (de hecho, hay otras mediciones, como la de descentralización, *Dezentrationsmaß*) con el objeto de detectar los paradigmas a los que los afijos pertenecen.

1.3.2 Cuentas de fonemas anteriores y posteriores

Aquí examinamos el procedimiento propuesto en los años cincuenta por Zellig Harris para segmentar palabras en morfemas. En un artículo

muy leído, "From Phoneme to Morpheme" (1955), Harris exploró la correspondencia de fronteras morfológicas con el número de signos que potencialmente siguen o preceden a algún fonema en una expresión dada. Su método consiste en que, dado un segmento de palabra, se analiza la variedad de fonemas sucesores y predecesores potenciales. Mientras más grande fuera la variedad de fonemas que potencialmente aparecen antes o después de una segmentación, mayor es la incertidumbre de lo que sigue o precede; lo cual significa que hay una buena posibilidad de tener allí una frontera morfológica.

Lo interesante es que Harris comprobó en los años cincuenta que la estructura morfológica de las lenguas se manifiesta en cada punto del discurso, dada la historia de lo dicho hasta ese momento, en el mayor o menor número de posibles fonemas subsecuentes, que se pueden contar en un corpus. Por ejemplo, en un corpus típico del español, después de la cadena de letras *niñ* pueden ocurrir cuatro de las cinco vocales del español (en *niñ~**a***, *niñ~**e**ra*, *niñ~**i**to* y *niñ~**o***), lo que marca con gran probabilidad una frontera morfológica. En cambio, después de *cantab* sólo ocurre una *a*, lo que indica que allí seguramente no hay frontera morfológica. En otras palabras, a mayor variedad de fonemas que pueden ocurrir después de una cadena de caracteres, mayor probabilidad de que termine un morfema y empiece otro. En la Figura 2, podemos apreciar el número de fonemas posibles (eje vertical), según un corpus, en cada punto de la expresión *Dogs were indisputably quicker*.

El procedimiento requiere un corpus de buen tamaño, de preferencia un conjunto de enunciados obtenidos (*elicitados*) de algún informante. Dado que se trata de una investigación morfológica (y no de la distribución de las letras utilizadas en la escritura de la lengua estudiada), todos los enunciados se representan mediante la misma codificación fonética *que no haga referencia*, por supuesto, *a la representación de los morfemas*. La mayoría de los cortes propuestos mediante este método coincide con las fronteras entre palabras y morfemas. Por ejemplo, mediante este procedimiento, /hiyskwikɘr/ (*he's quicker*) se segmenta

/hiy.s.kwik.ɘr/. Nótese que no hay nada que indique el estatus morfológico de los segmentos, es decir, no hay nada que indique si /ɘr/ es afijo o palabra. Por eso, las decisiones en cuanto al estatus morfológico se dejan a los otros métodos distribucionales de determinación de morfemas. Es decir, este procedimiento es solamente un intento de poner orden a aquellos anteriormente propuestos en el marco del distribucionalismo (Z. S. Harris 1955, 191).

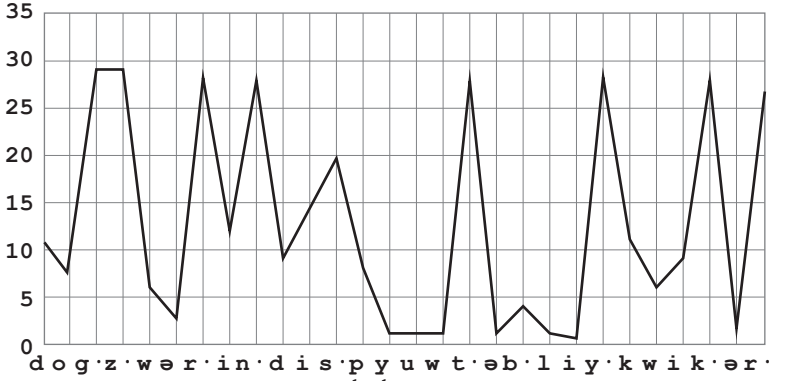


Figura 2. Fronteras entre morfemas en la oración *Dogs were indisputably quicker*.

Gráfica basada en la que aparece en Harris, *A Theory of Language and Information* (1991, 172). Los límites entre segmentos fueron marcados por Harris mediante puntos. Lo interesante es notar que las cuentas altas de fonemas corresponden en gran medida a estos puntos. El punto menos claro está entre 'ab' y 'ly' donde se alcanza apenas un número de cuatro fonemas.

En este método, el primer paso es determinar el número de fonemas que sigue a cada segmentación o corte de palabra examinado. Por ejemplo, al examinar la segmentación entre el primer fonema, común a varios enunciados, y los fonemas que le siguen en esos enunciados, simplemente se cuenta el número de fonemas diferentes que ocurren en la segunda posición de esos enunciados. Luego se toman aquellos enunciados que empiezan con una secuencia común de dos fonemas y se cuenta el número de fonemas diferentes que ocurren en la tercera posición en

esos enunciados. Así, este procedimiento se repite hasta alcanzar el final de un enunciado particular.

En la secuencia de cuentas de fonemas obtenidas para un enunciado de esta manera, se puede ver cómo los números más altos constituyen picos rodeados de hendiduras (la secuencia sube y baja). En cada pico se presume la existencia de un corte morfológico. A este procedimiento base se le pueden hacer ciertas modificaciones para afinar los resultados. La más importante es contar no sólo los fonemas que siguen, sino también los que preceden la segmentación examinada. Lo sorprendente es que los picos de estas cuentas, al revés, también corresponden a cortes morfológicos en el enunciado. Nótese la coincidencia entre las fronteras morfológicas y las cuentas altas en ambas direcciones del enunciado inglés que aparece en la Tabla 3, basada en Harris (1955, 218):

Tabla 3. Cuentas de fonemas anteriores y posteriores en cada corte del enunciado *What did he think of?*

	h	**w**	**ɘ**	**t**	**d**	**ɪ**	**d**	**h**	**i**	**y**	**θ**	**ɪ**	**n**	**k**	**ɘ**	**v**
anteriores	9	5	1	**29**	10	19	**28**	8	12	**28**	5	4	1	**29**	11	**28**
posterio-res	**22**	1	7	18	**23**	1	3	9	19	4	**22**	15	3	12	**23**	6

Las cuentas anteriores muestran picos que coinciden con la segmentación morfológica inmediatamente antes del corte morfológico. Lo mismo las cuentas posteriores que muestran los picos después del corte (con excepción de /hiy/).

Otra modificación es una operación de inserción que consiste simplemente en meter entre el fonema *n* y el fonema *n* + 1 de un enunciado alguna secuencia fonémica, de tal manera que el resultado sea uno de los enunciados que se atestiguan en el corpus (Z. S. Harris 1955, 199). Así, el número de enunciados atestiguados en el corpus que se puedan

obtener mediante esta operación se cuenta como otra medida de la validez de la segmentación. Otra modificación (1955, 199-202) consiste en tomar en cuenta no nada más la variedad de fonemas adyacentes a la segmentación examinada, (n + 1), sino también aquella de los fonemas que se encuentran a dos posiciones de distancia (n + 2). La cuenta de la variedad de estos fonemas es también una medida de incertidumbre (lo poco predecible) de lo que hay al otro lado de la frontera morfológica: de nuevo, a mayor variedad de lo esperado, hay mayor certeza de que el corte examinado sea morfológico.

La última modificación consiste en tomar en cuenta los tipos de fonemas anteriores y posteriores. La idea es que hay tipos más probables después (y antes) de ciertas secuencias. Por ejemplo, después de una secuencia que termina en consonante, hay una gran probabilidad de que la variedad posible de fonemas sea vocálica. Para corregir fluctuaciones debidas a las diferencias en tamaños de los conjuntos de vocales y consonantes, Harris propone fórmulas que reflejan la fonotáctica particular de la lengua examinada.

1.3.3 Métodos de estadística de digramas

A continuación, se comentan las estadísticas de coocurrencia de digramas o bigramas como métodos de descubrimiento de fronteras morfológicas. Estas estadísticas se han aplicado ampliamente tanto en trabajos de extracción de unidades fraseológicas (*collocations*)[30], como en estudios lingüísticos y literarios basados en corpus.

Un digrama o bigrama es sencillamente un par de segmentos que ocurren en un corpus, uno después del otro[31]. Hay varias estadísticas para medir la asociación entre los dos elementos de un digrama. Cada

[30] Las unidades fraseológicas o *colocaciones* son expresiones formadas por dos o más palabras y que corresponden a una manera convencional de decir algo.

[31] Un unigrama es un segmento, un digrama comprende dos, un trigrama tres, etcétera.

una define el concepto de asociación de manera diferente[32], pero comparten el concepto de no asociación que definen en términos de independencia. Es decir, de cada estadística se obtiene un valor que mide la independencia entre segmentos: a menor valor mayor asociación y a mayor valor mayor independencia. En otras palabras, cada una de estas medidas tiene un significado específico, sobre todo si se trata de determinar la asociación de dos elementos puesto que miden diferentes tipos de asociación. Sin embargo, cuando se trata de determinar la no asociación, resultan relativamente comparables, porque, al no haber asociación, no importa qué tipo de asociación se esté buscando.

Las estadísticas más populares son: la prueba de independencia de χ^2 (ji cuadrada), la razón de semejanza (*log likelihood*), el coeficiente de coligación de Yule y la información mutua. Kageura lleva a cabo un experimento para comparar estas medidas y determinar cuál es la más apropiada en la determinación de fronteras morfológicas en secuencias de caracteres Kanji del japonés. En su experimento la razón de semejanza resultó ser la mejor medida, seguida por la prueba de χ^2 (Kageura 1999).

1.3.4 Teoría de la información

Relacionado con el método de Harris, está la aplicación del concepto de entropía que, como ya se dijo, a menudo se menciona como un método para determinar cortes morfológicos. Aquí se presentan los fundamentos del procedimiento.

[32] Véanse Manning y Shütze, *op. cit.* (1999, 169-182); y Kageura, "Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences", *Journal of Quantitative Linguistics* 6 (1999, 149-166) para una explicación detallada.

La teoría de la información, también conocida como teoría de la comunicación[33], es una teoría matemática que permite analizar cuantitativamente la incertidumbre. En esta teoría, *incertidumbre* equivale a *información*, una noción técnica que no es idéntica a la noción intuitiva de información relacionada con el significado[34]. En resumidas cuentas, la teoría de la información se ocupa de la *cantidad* de incertidumbre que lleva una señal y no del contenido de dicha señal.

La idea detrás de asociar incertidumbre con información es que cierta cantidad de información es necesaria para identificar correctamente un mensaje a partir de varios mensajes posibles (para resolver la incertidumbre). Es decir, cuando a partir de una señal recibida, uno tiene que decidir cuál es el correcto de entre un conjunto de mensajes posibles. Así, medir la incertidumbre equivale a determinar cuánta información es necesaria para resolverla, para hacer cierto lo incierto. También se utilizan términos como *sorpresa* (porque cualquier medida de información causa una cantidad proporcional de sorpresa) y, como se dijo antes, *entropía* (caos, desorden o energía desorganizada), debido a la relación evidente de esta idea con el concepto original de la termodinámica.

Así, la entropía o incertidumbre[35] de una variable discreta X —que toma un número finito de valores $x_1, x_2, x_3, \ldots x_n$, cada uno con una probabilidad $p_1, p_2, p_3, \ldots p_n$, donde $0 \leq p_i \leq 1$ (i = 1, 2, 3, …, *n*) y la suma de sus probabilidades es igual a 1 ($\sum_{i=1}^{n} p_i = 1$)— se calcula mediante la siguiente fórmula:

[33] Fue desarrollada a finales de los años cuarenta por Claude Shannon; véase Shannon y Weaver, *op. cit.* (1964 [1949]). Véanse también Meyer-Eppler, *Grundlagen und Anwendungen der Informationstheorie*, Springer, Heidelberg (1969), Manning y Schütze, *op. cit.* (1999, 60-63) y Jurafsky y Martin, *op. cit.* (2009, 114-116).

[34] En este contexto, una cadena aleatoria de signos contiene más información que una de signos encontrada en un texto, donde tenga algún sentido, puesto que la segunda cadena es más predecible o menos incierta.

[35] Véanse Shannon y Weaver, *op. cit.* (1964 [1949], 50) y Weaver, "Recent Contributions to the Mathematical Theory of Communication" (1964), *op. cit.* (1964 [1949], 14).

Ecuación 1. Entropía de la variable X

$$H(X) = H(p_1, p_2, p_3, \ldots p_n) = -\sum_{i=1}^{n} p_i \times \log_2(p_i)$$

donde $p_i \times \log_2(p_i) = 0$, si $p_i = 0$.

Supongamos que tenemos 2 mensajes o símbolos posibles ($X = \{x_1, x_2\}$). La Figura 3 muestra la relación entre la entropía y la probabilidad de uno de esos eventos (p_1); la probabilidad del otro sería, claro está, $p_2 = 1 - p_1$. Se puede apreciar que la entropía es mayor cuando la probabilidad está repartida equitativamente entre los dos valores (es decir, cuando son equiprobables: cuando $p_1 = p_2 = 0.5$), mientras que disminuirá cuando alguna de las dos probabilidades se acerque a cero (por lo que lo otra se acercará a uno).

En otras palabras, la incertidumbre disminuye cuando la probabilidad de alguno de los dos eventos se acerca a cero. En cambio, cuando las probabilidades son equivalentes, la incertidumbre de obtener tal o cual mensaje será la mayor posible (Weaver 1964, 15). De esta manera, la sorpresa será mayor si predecimos correctamente el próximo símbolo.

La unidad de la entropía calculada mediante el logaritmo de base dos se llama *bit* (contracción de *BInary digiT*). Puesto que la función en la Figura 3 fue calculada con el logaritmo de base dos y se trata de un sistema de dos probabilidades, la entropía representada allí no puede tener un valor mayor que uno (1 bit), lo que no necesariamente verdadero en sistemas de más de dos probabilidades. Se puede calcular la entropía con el logaritmo de otra base; por ejemplo, si se utiliza el logaritmo de base 10, se habla de unidades decimales de entropía.

El método para descubrir morfemas consiste en medir la entropía en cada corte posible de cada vocablo. La idea es determinar la frecuencia de todo lo que en el corpus está atestiguado como acompañante de uno de los segmentos y calcular las probabilidades de cada objeto posible después del segmento dado. En la Figura 4 se ilustra esto:

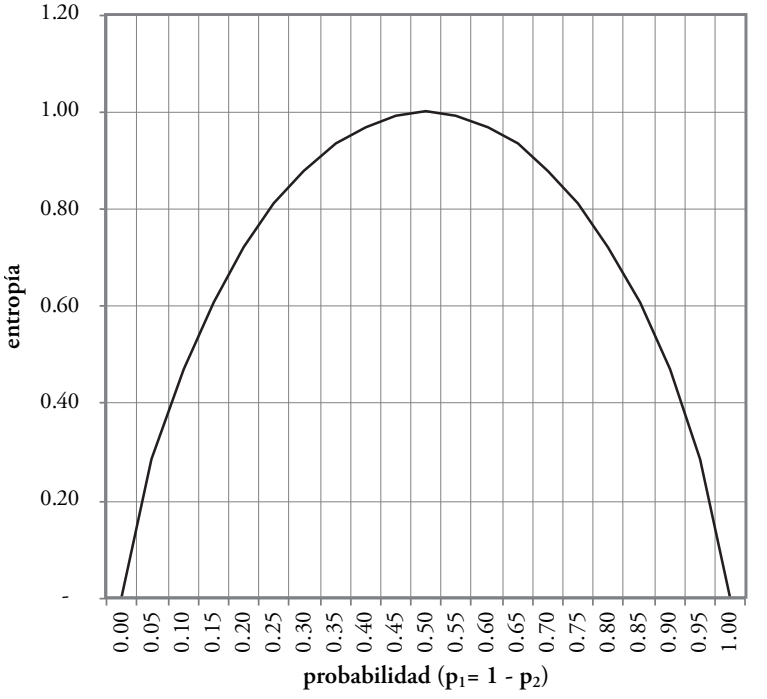


Figura 3. Gráfica de la entropía de dos mensajes posibles

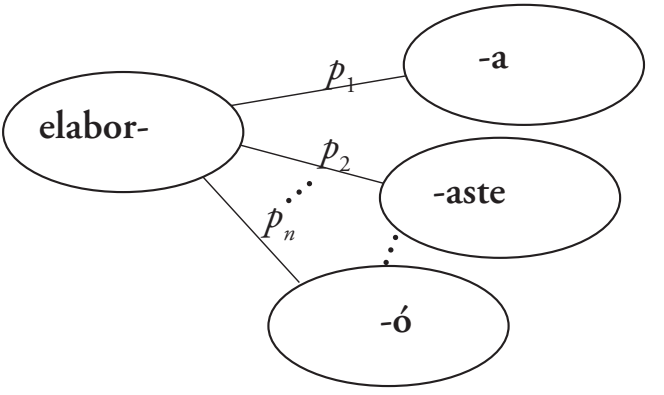


Figura 4. Esquema para ilustrar las probabilidades de los segmentos que según un corpus ocurren después de *elabor~*

La entropía se obtiene al aplicar la Ecuación 1 a estas probabilidades. Como veremos más adelante, las segmentaciones que exhiban las cantidades mayores de entropía seguramente serán el inicio de una raíz, porque son los segmentos más informativos del discurso. Por otra parte, los afijos, si bien también informativos, tienden a llevar apenas la información estructural de la palabra y del discurso. En el capítulo siguiente, se muestran ejemplos de la aplicación de la medición de información en el español de México.

1.3.5 Principio de economía

En esta sección, se presenta el trabajo que Josse de Kock y Walter Bossaert desarrollaron en los años setenta para segmentar palabras del francés y del español. Se trata de otra técnica para determinar empíricamente las fronteras morfológicas entre bases y afijos, tanto de flexión como de derivación. El método coincide en varios aspectos con el de Andreev y, aunque no llega a ocuparse en la determinación de paradigmas, sienta las bases que permiten hacerlo.

La base de todo el procedimiento es el principio de economía de signos o rentabilidad del sistema. Como es bien sabido, este principio tiene diferentes aspectos, pero el aquí pertinente puede parafrasearse de la siguiente manera: el número de signos en todos los niveles del lenguaje debe ser menor al número de cosas nombradas; así, "the code is organized in such a way that a sign can serve in more than one instance without creating any ambiguity" (de Kock y Bossaert 1978, 15). De esta manera, los signos del nivel sintáctico de una lengua como el español son producto de la combinación de los signos del nivel morfológico (se derivan o flexionan), siendo estos últimos pocos en número, pero más frecuentes que los primeros. El que la concatenación de dos signos de un nivel sea económica viene a ser entonces una función de esta diferencia, es decir, mientras menos signos de más frecuencia existan en el nivel morfológico que den lugar a más signos (de baja frecuencia)

del nivel sintáctico, la lengua será más económica. Y si suponemos que las lenguas tienden a la economía de signos, esta sencilla diferencia de carácter formal nos proporciona un mecanismo para determinar los signos morfológicos. Dicho de otra manera y en términos de los procesos de diversificación y unificación[36], si el sistema necesita de un número reducido de signos para ser económico (porque un número pequeño beneficia al hablante de una lengua, al requerirle un menor esfuerzo en el ejercicio de la memoria), estos signos se pueden combinar en otro nivel para dar lugar a un inventario mucho más grande, en beneficio del oyente, al existir mayor diversificación de estructuras que le aclaren el mensaje.

De Kock y Bossaert se basaron en los diccionarios de frecuencias de Alphonse Juilland[37]. En su procedimiento, los investigadores examinan automáticamente cada corte posible de cada vocablo, según se encuentren segmentos que aparezcan en otros vocablos. Así, el número de vocablos con un segmento común a la izquierda, es decir, el número de segmentos diferentes que aparecen a la derecha es m_d (*droit*). El número de vocablos con un segmento común a la derecha, o, lo que es lo mismo, el número de segmentos distintos a la izquierda es m_g (*gauche*)[38]. También, como en el método de Andreev, calculan para cada segmentación un número de combinaciones de cuatro segmentos[39]. El número de combinaciones que se puedan determinar para cada segmentación se llama n_c (*carré*) y es una medida de la validez de la segmentación.

[36] Para una descripción general de relación de estos procesos con el tamaño del inventario del léxico de las lenguas, véase Köhler, "Diversification of Coding Methods in Grammar" (1991), en Ursula Rothe, ed., *Diversification Processes in Language*, Rottmann, Hagen (1991, 47-55).

[37] Para el francés, Juilland, *Frequency Dictionary of French Words*, Gembloux (1965); para el de español, Juilland y Chang Rodríguez, *op. cit.* (1965).

[38] De Kock y Bossaert, *op. cit.* (1978, 18). Estos números se modifican al eliminar ciertos segmentos y tomar en cuenta ciertos criterios como que los segmentos contengan vocales o no, o que haya fonemas en un segmento que pertenezcan con más probabilidad al otro. Los números modificados se representan, con mayúsculas, mediante M_d y M_g (de Kock y Bossaert 1978, 22-23).

[39] Los mismos que Greenberg llama "cuadros" (*squares*); que veremos más adelante.

Tabla 4. Hipótesis para cada corte de los vocablos *capacidad y olvidad*

vocablo	v b,g	÷	v a,d	=	v b,a	v b,d	÷	v a,g	=	v a,b
kapaθida :: d	0		0			1		17		0.059
kapaθid :: ad	0		0			9		6		0.167
kapaθi :: dad	0		0			1		16		0.063
kapaθ :: idad	36		2		18.000	2		51		0.039
kapa :: θidad	1		2		0.500	2		3		0.667
kap :: aθidad	0		0			2		1		2.000
ka :: paθidad	0		0			0		0		
k :: apaθidad	0		0			0		0		
olbida :: d	48		11		4.364	9		16		0.563
olbid :: ad	29		7		4.143	22		25		0.880
olbi :: dad	0		0			21		6		3.500
olb :: idad	0		0			4		1		4.000
ol :: bidad	0		0			0		0		
o :: lbidad	0		0			0		0		

Tabla basada en la de Kock y Bossaert (1978, 31); **v b,g** = número de segmentos distintos (que aparecen en cuadros) de la izquierda bajo la hipótesis de que la base está a la izquierda; **v a,d** = número de segmentos distintos de la derecha bajo la hipótesis de que el segmento de la derecha es un afijo; **v b,a** = valor de la hipótesis de que el segmento de la izquierda sea la base y el de la derecha un afijo, etcétera.

El mecanismo más importante es el examen de dos hipótesis para cada corte: ya sea que el segmento a la izquierda sea un prefijo y el de la derecha la raíz, o que el de la derecha sea un sufijo y el de la izquierda una raíz. Así, se calculan dos valores de segmentación: uno dividiendo el número de segmentos a la izquierda propuestos como raíz entre el número de segmentos a la derecha propuestos como afijo, el otro dividiendo los segmentos a la derecha propuestos como raíz entre los de la izquierda propuestos como afijo. El segmento más probable como raíz será aquél cuyo valor calculado bajo la hipótesis de que es la raíz sea mayor a uno. Así, de Kock y Bossaert muestran los ejemplos que aparecen en la Tabla 4. En el capítulo siguiente, se parafrasea este procedimiento

con detalle, mediante una formalización utilizando conjuntos, y se describe su aplicación en los experimentos de los siguientes capítulos.

1.3.6 Investigaciones recientes

En este siglo, han surgido numerosos métodos de segmentación morfológica no supervisada, entre los que destacan el algoritmo Linguistica, para el aprendizaje no supervisado de morfología de John Goldsmith[40], la serie de técnicas conocidas con el nombre de Morfessor de Mathias Creutz y Krista Lagus[41], para segmentar palabras del finlandés y otras lenguas, y el método ParaMor de Christian Monson[42], aplicable a varias lenguas. En general, lo interesante de estos nuevos acercamientos es el énfasis que ponen en la evaluación de sus resultados.

También algunos investigadores mexicanos o que trabajan en México han llevado a cabo investigaciones sobre segmentación morfológica no supervisada. Por ejemplo, en el Instituto Politécnico Nacional[43],

[40] Goldsmith, "Unsupervised Learning of the Morphology of a Natural Language", *Computational Linguistics* (2001, 153–198) y "An Algorithm for the Unsupervised Learning of Morphology", *Natural Language Engineering* (2006, 353–371)

[41] Creutz y Lagus, "Unsupervised discovery of morphemes", *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, Filadelfia (2002, 21-30), "Induction of a simple Morphology for highly Inflecting Languages", *Proceedings of 7th Meeting of the ACL Special Interest Group in Computational Phonology* (2004, 43-51), "Inducing the Morphological Lexicon of a Natural Language from Unannotated Text", *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning* (2005, 106-113) y *Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0*, Helsinki University of Technology (2005).

[42] Monson *et al.*, "ParaMor: Minimally Supervised Induction of Paradigm Structure and Morphological Analysis", *Computing and Historical Phonology: The Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, Praga (2007) y "ParaMor: Finding Paradigms across Morphology", *Lecture Notes in Computer Science* (2008, 900-907).

[43] Gelbukh y Sidorov, del Centro de Investigación en Cómputo (IPN), también han desarrollado métodos para el análisis morfológico supervisado del español y del ruso. Su implementación determina, además de la base y el afijo de las palabras, el tipo de fenómeno morfológico de las palabras examinadas, a partir un diccionario previamente construido por el especialista. Además,

Lara Reyes (2008) aplicó un algoritmo genético para segmentar palabras del español y descubrir sufijos y prefijos sin información morfológica *a priori*[44]. Unos años antes, Gelbukh *et al.*[45] publicaron un método de segmentación no supervisada de afijos basado en la economía de signos y descubrieron que "it shows surprisingly promising results on different European languages" (2004). Por otra parte, Torres Moreno[46] ha demostrado que, para español, francés e inglés, se pueden normalizar las palabras simplemente cortando todo el final de cada una y dejando sólo las *n* primeras letras. Si bien no logra cortes morfológicos y él mismo menciona que "this technique could be considered a brutal destruction of the lexicon" (2012, 6), resulta interesante y provocador que consiga mejorar los resultados de sus métodos de resumen automático para esas lenguas.

Respecto a Linguistica (Goldsmith 2006), se trata esencialmente de un proyecto con dos objetivos: uno práctico de desarrollar un analizar morfológico para varias lenguas útil en tareas de recuperación de documentos y traducción automática, entre otras; y un objetivo teórico de averiguar cuánto conocimiento lingüístico tiene que codificarse en un programa de análisis morfológico para poder analizar la estructura del lenguaje. Para evaluar los cortes morfológicos hipotetizados automáticamente en un corpus de 300 000 palabras de lengua inglesa, Goldsmith y su equipo los compararon con los de un corpus segmentado manualmente. Poco sorprende que la segmentación manual del corpus de evaluación haya causado dudas sobre el análisis morfológico de las

---

detectan los casos de flexión irregular mediante reglas, que pueden ser recursivas; véase "Approach to construction of automatic morphological analysis systems for inflective languages with little effort" en *Computational Linguistics and Intelligent Text Processing* (CICLing-2003), *Lecture Notes in Computer Science* 2588 (2003, 215-220).

[44] Lara Reyes, *Sistema de segmentación automática de palabras en morfemas para el español.* CIC IPN, México. Tesis de maestría (2008),

[45] "Detecting Inflection Patterns in Natural Language by Minimization of Morphological Model", CIARP 2004, *Lecture Notes in Computer Science* 3287 (Gelbukh, Alexandrov y Han 2004, 432–438).

[46] "Beyond Stemming and Lemmatization: Ultra-stemming to Improve Automatic Text Summarization", arXiv:1209.3126v1 [cs.IR] (Torres Moreno 2012).

palabras inglesas. Como sea, una vez que cotejaron el corpus segmentado automáticamente con el de evaluación, el método obtuvo 72% de éxito en encontrar los cortes morfológicos de las palabras.

Otro acercamiento importante es el conocido como Morfessor. Se trata de una familia de técnicas desarrolladas al menos desde 2002 para segmentar palabras en finlandés y otras lenguas. Inicialmente, desarrollaron dos métodos que llamaron Morfessor baseline (Creutz y Lagus 2002). El primero es un método que genera una lista de segmentos de palabras a partir de cortes aleatorios evaluados por funciones de costo, que es una medida de economía. Un menor costo significa una mejor descripción del corpus analizado. El segundo método consiste en generar una lista de palabras a partir de un corpus y generar un corpus artificial a partir de esta lista mediante procedimientos probabilísticos. El tamaño de la lista de segmentos de palabra, la longitud de cada uno de ellos, los caracteres que los forman, su orden y su frecuencia se calculan con funciones de probabilidad. Sus resultados para la lengua inglesa fueron mejores o semejantes a los de Linguistica. En cambio, en finés, una lengua mucho más compleja morfológicamente, estos métodos superaron al de Goldsmith. Un tercer método (Creutz y Lagus 2004) fue bautizado como Morfessor categories-ML, que está basado en el cálculo de probabilidad máxima (*maximum likelihood*). Éste busca asociarle a cada segmento de palabra alguna de tres categorías: prefijo, sufijo o base. De esta manera, este método descubre una morfotáctica simple, al detectar cambios de una categoría a otra. Para el finés, este método resulta mejor que los métodos de Morfessor anteriores a 2003 y que el algoritmo Linguistica. De hecho, alcanzó el 79% de éxito en 16 millones de palabras de esa lengua.

Por otra parte, el último método mencionado es el de ParaMor. Se trata de un algoritmo que busca determinar paradigmas morfológicos. Luego, a partir de ellos, determina los morfemas. En otras palabras, ParaMor es un algoritmo de segmentación morfológica basado en los paradigmas, lo que funciona muy bien para descubrir la morfología flexiva. Como Morfessor identifica mejor la morfología derivativa, am-

bos métodos se pueden combinar para lograr un análisis más completo (Monson 2008).

Por último, se pueden mencionar los concursos del Morpho Challenge. Estos concursos tuvieron el objetivo general de construir algoritmos estadísticos de aprendizaje no supervisado de la morfología de ciertas lenguas, como el finlandés, el alemán, el turco, el inglés y el árabe que se materializaron en competencias y talleres conocidos con este nombre. Los retos Morpho Challenge[47] se llevaron a cabo entre 2005 y 2010.

1.4 Hacia el descubrimiento de afijos

En este primer capítulo se esbozaron los objetivos del campo de la morfología computacional, en el que los métodos de formulación de reglas, para hacer operativo el conocimiento morfológico, tienen un papel esencial. A pesar de la diversidad de estos enfoques y opiniones sobre su validez como investigaciones lingüísticas, estos esquemas se han aplicado con relativo éxito sobre todo en el desarrollo de las tecnologías del lenguaje. Sin embargo, es indudable que los métodos basados en reglas no son las únicas formas de investigar los sistemas lingüísticos. Están también las investigaciones cuantitativas de los corpus que, si bien no capturan mediante reglas el carácter ilimitado del lenguaje, sí permiten hacer inferencias muy valiosas sobre los fenómenos lingüísticos retratados en los corpus. De hecho, aunque en este trabajo no se busca combinar estos métodos, vale la pena resaltar que los enfoques de reglas y los estadísticos se pueden combinar para desarrollar mejores sistemas de análisis de la morfología de las lenguas.

En esencia, en este capítulo, después del apartado de morfología computacional, se presentaron algunos trabajos y enfoques que suelen

[47] Las descripciones de los algoritmos que compitieron y de los métodos de evaluación que se aplicaron aparecen en Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen y Mikko Kurimo, "Empirical Comparison of Evaluation Methods for Unsupervised Learning of Morphology", *Traitement automatique des langues*, 52, 2 (2011, 45-90).

aplicarse al reconocimiento y descubrimiento de morfemas mediante computadoras. Específicamente, se revisaron algunos procedimientos para la segmentación morfológica de las palabras, los más tempranos y quizá más productivos en los desarrollos enfocados al descubrimiento automático de morfemas.

En el próximo capítulo, se describe el método de descubrimiento de afijos mediante mediciones de afijalidad, que se basa principalmente en la aplicación de la teoría de la información y el principio de economía. La idea, como se dijo al principio, es determinar o descubrir automáticamente conjuntos de signos afijales de las palabras gráficas de corpus de lenguas con morfología concatenativa.

CAPÍTULO 2
EL SIGNO AFIJAL

El objetivo de este capítulo es describir cómo obtener de manera no supervisada un conjunto de signos del nivel morfológico, de tipo afijal, a partir de un corpus electrónico[1]. En la primera sección, se examinan algunos conceptos lingüísticos, como los de morfema, morfo y afijo. Enseguida se abordan algunas cuestiones de carácter formal, para simplificar la exposición del procedimiento de descubrimiento de afijos. Luego se analizan y comparan diferentes técnicas de segmentación de palabras para finalmente describir un método de compilación de catálogos de prefijos y sufijos de lenguas de morfología concatenativa como el español.

2.1 Sobre las unidades morfológicas

Uno de los problemas de la investigación no supervisada de la morfología de una lengua, a partir de un corpus, es que los morfemas son unidades de significado y no hay, aún hoy en día, métodos exactos que descubran propiamente el significado de los morfemas. De hecho, muchos criterios de carácter semántico para determinar morfemas son relativos y varían de lingüista a lingüista, lo que dificulta que cualquier método automático los tome en cuenta.

Como sea, el morfema es una abstracción que puede entenderse como un conjunto de una o varias formas mínimas que comparten un

[1] Este capítulo está basado en Medina Urrea, "Automatic Discovery of Affixes by Means of a Corpus: A Catalog of Spanish Affixes", *Journal of Quantitative Linguistics*, 7:2 (2000, 97-114) y en *Investigación cuantitativa de afijos y clíticos del español de México*, Tesis doctoral, El Colegio de México (2003).

contenido y que Eugene Nida llamó *alomorfos*[2]. Específicamente, los alomorfos son las ocurrencias del morfema en contextos determinados. Así que el concepto de morfema implica, a grandes rasgos, que una misma unidad significativa (de la lengua) puede exhibir más de una forma, según el lugar donde ocurra (en el habla). Es decir, varias formas pueden compartir un significado: uno o más alomorfos constituyen un morfema.

De esta manera, el morfema para construir el plural de sustantivos en español puede manifestarse mediante los alomorfos *~s* y *~es*. Similarmente, el morfema representado por *~idad* puede manifestarse mediante *~edad* (*brev**edad***) o *~dad* (*bel**dad***), etc. Claro está que la distribución de los alomorfos depende de las bases a las que se adhieren. Así que el descubrimiento no supervisado de morfemas requiere de un procedimiento automático para encontrar fronteras entre los alomorfos, afijos y bases, de las palabras que conforman un corpus. Cuando un morfema comparte con otro alguno de sus alomorfos, tenemos un caso de homonimia o polisemia. Se trata de la situación en que dos signos morfológicos comparten un significante. Por ejemplo, la forma *~a* es una marca de género femenino en sustantivos y una de persona y modo, indicativo o subjuntivo, en formas verbales.

Por otra parte, los alomorfos también son conocidos como *variantes formales* o simplemente *morfos*[3]. Estos signos mínimos son unidades cuya búsqueda es relativamente fácil de llevar a cabo mediante computadoras, porque el morfo es una unidad del habla directamente observable en una muestra textual. En cambio, para buscar un morfema en un corpus, hay que conocer sus alomorfos de antemano, lo cual es un problema más complejo. Incluso en los casos de homonimia, el morfo

[2] Véase Nida, "The Identification of Morphemes", *Language* 24 (1948, 420). Zellig Harris ya antes los había llamado *alternantes de morfema*, véase "Morpheme Alternants in Linguistic Analysis", *Language* 18 (1942, 170). Compárense las definiciones de *morfema*, *alomorfo* y *morfo* en Glück, *op. cit.* (2000) y en Bußman, *Lexikon der Sprachwissenschaft*, Kröner, Stuttgart (1990), s.v. Allomorph y Morph.

[3] Charles Hockett, "Linguistic Elements and their Relations", *Language*, 37 (1961, 29-53).

compartido por dos morfemas es observable en la cadena hablada y, en la mayoría de los casos, su contexto de aparición permite desambiguar el morfema al que pertenece. De hecho, determinar a qué morfemas pertenece un morfo es un paso posterior en el procedimiento de descubrimiento o aprendizaje de la morfología. Como se verá, los experimentos de este trabajo se ocupan del morfo y no directamente del morfema.

De entre todos los fenómenos que se han concebido en la morfología[4], el de afijación ha sido muy estudiado. 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 (raíces + afijos). Como se dijo antes, las unidades afijales son el objetivo de este capítulo. Muchos morfos de tipo afijal pueden concebirse como elementos desgastados fonética y semánticamente, por lo que, para motivos de descubrimiento de los morfos afijales, este desgaste semántico puede verse como una ventaja, puesto que disminuye la dependencia en criterios cualitativos y subjetivos sobre su significado.

La mayor ventaja de los afijos es que, como se ha venido diciendo desde la introducción, éstos se prestan a una formalización lo suficientemente precisa como para descubrirlos automáticamente. Esto no excluye que otros tipos de signos, como los involucrados en el proceso de composición, también puedan descubrirse computacionalmente. Así, por ejemplo, aquí se caracterizará al afijo (y por exclusión a la base), como un signo que aparece adherido a la derecha (sufijo) o izquierda (prefijo) de otro signo (la base o lexema) y que[5]:

[4] Para diversas descripciones de tipos de morfemas, según se conciban como objetos, reglas, procesos, etc., véanse Nida, *Morphology. The Descriptive Analysis of Words*, The University of Michigan Press, Ann Arbor (1967 [1949], 68-77); Sapir, *El Lenguaje*, *op. cit.* (1992 [1921], 77-94); Spencer, *Morphological Theory. An Introduction to Word Structure in Generative Grammar*, Basil Blackwell, Cambridge (1991, 4-20); Anderson, *op. cit.* (1994, 48-66); Bergenholtz y Mugdan, *Einführung in die Morphologie*, Kohlhammer, Stuttgart (1979, 58-73).

[5] Por comodidad, se tratarán solamente sufijos y prefijos. Queda pendiente el estudio de otros tipos de afijos, como los infijos, circunfijos y los cambios vocálicos de algunas raíces verbales en español.

1. no ocurren aislados (son parte de las palabras),
2. ocurren en muchos vocablos de relativamente baja frecuencia,
3. tienen significados más generales (por ejemplo, gramaticales) que modifican al significado central de la palabra[6],
4. tienden a ser cortos y limitados en cuanto a repertorio fonológico.

Esta caracterización del afijo es la definición de trabajo para este capítulo[7]. Podemos verla como una premisa que sirve de base para formalizar y cuantificar la unidad afijo. La idea es aplicar las medidas involucradas en el cálculo cuantitativo de la propiedad abstracta de ser afijo. Por ejemplo, el punto 1 puede formalizarse mediante la noción tradicional de *cuadro*, que veremos más adelante. Se trata de un mecanismo para constatar, entre otras cosas, que los afijos siempre ocurren acompañados. Luego, el punto 2 se puede medir mediante el principio de economía. Por último, aunque la entropía no es una medida de significado, como medida de información, es un acercamiento interesante para el punto 3.

Respecto al punto 4, en este trabajo no se toma en cuenta. Es cierto que los afijos tienden a ser cortos, pero veremos que suelen ocurrir encadenados y las cadenas afijales no necesariamente son tan cortas (*~ié.ra.mos*, *~aba.n*, *~a.d.a.mente*, *~ié.ndo.se.las*, etc.). Además, el desgaste fonológico no ocurre ni instantánea ni inmediatamente después de que un segmento adquiere carácter afijal (*~torio*, *~miento* y *~mente*). Su tamaño reducido es más una consecuencia a largo plazo y no necesariamente una característica medular de estas unidades morfológicas.

[6] Siguiendo la distinción sapireana de carácter difuso entre contenido material y contenido relacional, los afijos tienden al contenido relacional (hay que notar que lo relacional opuesto a lo material puede variar de lengua a lengua); véase Bybee, *Morphology. A Study of the Relation between Meaning and Form*, John Benjamins, Amsterdam (1985, 7).

[7] Por simplicidad, se excluyen fenómenos como, además del problema de los alomorfos que tratamos arriba, la fusión de bases y afijos, *op. cit.* (Bybee 1985, 4-7); Anderson pone especial énfasis en el problema que este fenómeno implica para el descubrimiento automático de morfemas, *op. cit.* (1994, 389-391).

Una distinción interesante relacionada con el concepto de afijo, que también se puede investigar formalmente, es aquella entre afijos de derivación y de flexión. Greenberg plasmó en sus universales que la flexión aparece al exterior de la palabra y que los afijos de derivación aparecen entre los de flexión y la raíz. También se puede sacar provecho de la idea de que los de derivación son más léxicos y los de flexión más sintácticos. De hecho, Joan Bybee arguye que hay una escala continua entre los polos de derivación y flexión, siendo un extremo más léxico y el otro más sintáctico (1985, 81-109).

Como sea, valdría la pena una investigación automática y cuantitativa de los paradigmas, sobre todo de flexión. Lo importante es notar que las ideas utilizadas en este trabajo son apenas algunos ejemplos de criterios cuya aplicación automática se puede investigar.

2.2 Nociones formales preliminares

En este apartado se examina el aparato notacional en que se basan los apartados siguientes, que versan sobre los métodos de segmentación de vocablos. La idea es formalizar los procedimientos de extracción de afijos para, por un lado, facilitar su implementación y, por el otro, simplificar su exposición y entendimiento.

Sea Ψ una secuencia de ocurrencias de palabras gráficas (por ejemplo, un corpus) de tamaño ξ,

$$\Psi = \langle o_1, o_2, o_3, \ldots, o_x \rangle,$$

donde cada palabra o_i puede ser idéntica a cualquier otra, o_j, excepto en su posición en la secuencia. Así que cada tipo u ocurrencia de palabra tiene una frecuencia de aparición f_i. Sea Ω el número de tipos de palabras encontrados en Ψ. En este contexto, se pueden concebir dos conjuntos: uno de tipos, $V = \{v_1, v_2, v_3, \ldots, v_\Omega\}$, y otro de sus frecuencias, $F = \{f_1, f_2, f_3, \ldots, f_\Omega\}$.

Sea Φ el conjunto de Ω pares ordenados:

$$\Phi = \{\langle v_1 \mid f_1\rangle, \langle v_2 \mid f_2\rangle, \langle v_3 \mid f_3\rangle, \ldots, \langle v_\Omega \mid f_\Omega\rangle\},$$

donde cada v_i es miembro también del conjunto V y cada f_i del conjunto F, de tal manera que la frecuencia del tipo de palabra v_i es f_i. Lógicamente, la suma de estas frecuencias es igual al tamaño de la secuencia Ψ:

$$\sum_{i=1}^{n} f_i = |\Psi| = \xi$$

Por otra parte, cada corte posible de cada vocablo v_i se puede representar mediante dos signos de dos puntos, '::'. Así, si dividimos un vocablo en dos segmentos, $a :: b$ será la representación de esa división, donde a corresponde al inicio de la palabra, la parte izquierda, y b al final, la parte derecha. Para recordar que cada uno de estos fragmentos o segmentos de palabra es parte de una palabra particular, incluiremos el índice i, que corresponde al vocablo examinado, de tal manera que $a_i :: b_i \equiv v_i$. Evidentemente, cualquiera de los dos segmentos podrá ser una base o un afijo, según la información cuantitativa que se pueda obtener del corpus.

Además, sea j otro índice que indique cada corte posible del vocablo v_i (es decir, la columna donde se dividen los segmentos de ese vocablo). Así, si v_i tiene m_i caracteres de longitud, entonces contiene $m_i - 1$ cortes posibles, que podemos simbolizar $a_{ij} :: b_{ij}$, donde $j = \{1, 2, 3, \ldots, m_i - 1\}$. Por ejemplo, si tenemos el vocablo v_x, sus cortes se representarán como en la Figura 5:

$$a_{x,1} :: b_{x,1} \text{ (e::jemplo)}$$
$$a_{x,2} :: b_{x,2} \text{ (ej::emplo)}$$
$$a_{x,3} :: b_{x,3} \text{ (eje::mplo)}$$
$$\vdots$$
$$a_{x,\,m_x-1} :: b_{x,\,m_x-1} \text{ (ejempl::o)}$$

Figura 5. Representación de los cortes posibles de un vocablo x (v_x)

En la expresión $a_{ij} :: b_{ij}$, los segmentos a_{ij} y b_{ij} están en relación *sintagmática*, porque ocurren juntos en el corpus, de manera consecutiva,

uno después del otro. Sin embargo, suele ser que no siempre que ocurre uno, ocurre el otro. De hecho, varias palabras podrán compartir un mismo inicio, a_{ij}, mientras que otras podrán compartir un mismo final, b_{ij}. En otras palabras, a_{ij} y b_{ij} guardan relaciones *paradigmáticas* con otras palabras del corpus.

La variedad de segmentos que ocurre en un corpus después de a_{ij} puede representarse mediante el conjunto B_{ij} (así que $b_{ij} \in B_{ij}$). Similarmente, el conjunto de todos los segmentos que ocurren antes de b_{ij} puede ser A_{ij} (donde $a_{ij} \in A_{ij}$). Se puede construir un programa que recorra todo el vocabulario V para encontrar los conjuntos A_{ij} y B_{ij}, de cada par de segmentos en cada corte j de cada v_i, para analizar cuáles son más morfológicos.

Definamos el proceso de *alternancia* como aquel en el que, dado un vocablo v_i dividido en dos segmentos, se buscan todos los vocablos en el conjunto V que compartan con v_i uno de esos segmentos. Por ejemplo, si partimos del segmento a_{ij} de la palabra v_i, podemos recorrer todo el conjunto V para encontrar todas las palabras que inician con ese segmento y registrar todas las segundas partes de esas palabras, entre las cuales encontraremos b_{ij} que, valga la redundancia, es la segunda parte de v_i. Diremos entonces que b_{ij} *alterna* paradigmáticamente con todas las segundas partes que ocurren después de a_{ij}; esto es, los elementos de B_{ij}.

De manera similar, si partimos del segmento b_{ij}, podemos recorrer todo el conjunto V para encontrar las palabras que terminan con ese segmento y registrar todas las primeras partes de esas palabras, entre las cuales encontraremos a_{ij}, que es la parte inicial de v_i y *alterna* paradigmáticamente con todas esas primeras partes, a la izquierda de b_i. Así, al decir que uno de los extremos del vocablo *alterna*, estará implícito que el otro permanece estático o fijo[8].

[8] De Kock no define explícitamente el término *alternar* pero hace uso de él con un significado similar al que aquí se define (de Kock y Bossaert 1978, 17). Sin embargo, el término *alternación* (*alternation*) tiene en morfología varios sentidos, véase Matthews, *Morphology*, Cambridge University Press, Cambridge (1991, 114-119). También Rulon Wells, usa este término en el marco de la

En resumen, A_{ij} será el conjunto de segmentos que alternan a la izquierda del segmento b_{ij} de v_i, es decir, el conjunto de segmentos encontrados al fijar b_{ij} y dejar alternar a_{ij} ($a_{ij} \in A_{ij}$). También, B_{ij} será el conjunto que alterna a la derecha de a_{ij} (por lo que $b_{ij} \in B_{ij}$). Finalmente, $|A_{ij}|$ será el número de miembros en el conjunto A_{ij} y $|B_{ij}|$ el de B_{ij}.

2.3 Técnicas para cuantificar la afijalidad

En este apartado examinaremos la construcción automática de inventarios o catálogos de afijos. En esencia, utilizaremos algunas técnicas de descubrimiento morfológico no supervisado que, como se dijo, no requieren conocimiento previo de los límites entre morfemas. Específicamente, se describirán las cuentas de cuadros, los cálculos de entropía y las medidas de economía.

Para explorar estos diferentes métodos de segmentación morfológica, se calculan diferentes índices en cada corte de cada vocablo del conjunto *V*, con el objeto de compararlos y construir con ellos una herramienta que sirva para determinar si un corte dado corresponde o no a la frontera entre dos morfos.

2.3.1 Número de cuadros

En la lingüística estructural, la idea de *cuadro* sirve para validar el carácter morfológico de cualquier corte que se haga en el interior de un voca-

---

morfofonémica, en "Automatic Alternation", *Language* 25 (1949, 99-116); es interesante que se pueden utilizar sus términos *communis* (la parte que un conjunto de formas comparten) y *propria* (la parte propia a cada forma, que no comparte con las demás) para describir lo que hemos de entender por alternancia en este trabajo: los segmentos alternantes son el conjunto de partes propias (únicas) que están en uno de los extremos de un conjunto de vocablos, mientras que el segmento fijo es el otro extremo, que es la parte *communis* de todos ellos (104).

blo. Greenberg (1967 [1957], 20) la caracteriza como un conjunto de palabras que

> exists when there are four expressions in a language which take the form AC, BC, AD, BD. An example is English *eating:walking::eats:walks*, where A is *eat-*, B is *walk-*, C is *-ing*, and D is *-s*. One of the four members may be zero, as in *king:kingdom::duke::dukedom*, where C is zero.

Así que un cuadro es un conjunto de cuatro segmentos de vocablos, dos de la izquierda (a_1 y a_2) y dos de la derecha (b_1 y b_2) que combinados (los de la izquierda con los de la derecha) resultan en vocablos (a_1::b_1, a_1::b_2, a_2::b_1, a_2::a_2) presentes en el conjunto *V*. Uno de estos segmentos puede ser la cadena nula de símbolos, lo que se puede caracterizar como un morfema nulo, ø, para permitir cuadros tales como {*in*::*káuto, in*::*felís*, ø::*káuto*, ø::*felís*}.

Esta estructura combinatoria puede variar según el número de elementos que se requieran. Por ejemplo, se puede requerir seis, de tal manera que en lugar de cuadro se tiene un hexágono (seis segmentos contenidos en nueve palabras), o para corpus muy pequeños, se puede relajar el requisito para aceptar cuadros incompletos[9]. En la Tabla 5 se ilustran estas combinaciones:

Tabla 5. Estructuras combinatorias

cuadro	**cuadro incompleto**	**hexágono**		
A::a	A::a	A::a	A::b	A::c
A::b	A::b	B::a	B::b	B::c
B::a	B::a	C::a	C::b	C::c
B::b	B::c			

[9] "Un inventario restringido pide una regla elástica, un inventario ampliado admite una regla severa. La diferencia radica en que en un vocabulario limitado todas las posibilidades de realización, efectivas en el conjunto de la lengua, y que precisamente motivan la segmentación morfológica y la autorizan, no se hallan siempre representadas" (de Kock y Bossaert 1974, 195).

Con una computadora, se pueden contar los cuadros posibles de cada corte *j* de cada vocablo v_i de un corpus. Como se estableció, la idea es detectar todas las palabras que empiezan de la misma manera, a_{ij}, y cuyas segundas partes, B_{ij}, se combinan con otros inicios de palabra que se prefijan a b_{ij}. Es decir, cada corte posible de cada vocablo v_i se examina para contar el número de cuadros documentables allí, al buscar las posibles combinaciones en el conjunto *V*. Llamemos a este número c_{ij}, el número de cuadros encontrados en el corte *j* del vocablo v_i.

Evidentemente, no cualquier cuadro es morfológicamente aceptable; por ejemplo, {*k*::*apasidád, r̄*::*apasidád, k*::*apás, r̄*::*apás*}. Así que se vuelve necesario aplicar alguna prueba de correspondencia de significado (Greenberg 1967 [1957], 23) para detectar una verdadera frontera morfológica. La naturaleza de este experimento, sin embargo, impide la aplicación manual de pruebas de este tipo. De todas maneras, como en este tipo de trabajo los errores son inevitables, pero detectables gracias a sus bajas frecuencias (en comparación con los fenómenos sistemáticos de la lengua), tiene sentido asumir que serán muy pocos los cuadros correspondientes a un corte no morfológico, en comparación con la cantidad de cuadros que atestigüen uno verdaderamente morfológico.

Por último, en este trabajo se aplicó una restricción importante en el conteo de cuadros. Puesto que se trata de descubrir afijos, se contaron solamente aquellos cuadros cuyos segmentos alternantes guardaban una relación de pocos y muy frecuentes con respecto al segmento fijo o, viceversa, cuando el segmento fijo pertenece a un conjunto más pequeño de segmentos más numerosos que el conjunto de los segmentos alternantes.

2.3.2 Índice de entropía

Las intuiciones detrás del método de Harris —en el sentido de tomar el número de fonemas o letras que preceden o siguen una segmentación para descubrir una frontera morfológica— tienen mucho sentido

al considerar el fenómeno de afijación. Considerémoslo en términos de la teoría de la información (Shannon y Weaver 1964 [1949]); es decir, en términos de entropía.

Greenberg presenta esta idea con mucha claridad: "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" (1967 [1957], 91). De esta manera, se puede esperar que las bases o lexemas contengan más información que los afijos, porque la ocurrencia de una base nos debe *sorprender* más que la de un afijo. De hecho, si suponemos que un afijo contiene información predominantemente gramatical, entonces tiene sentido esperar que un pico de entropía dentro de una palabra señale el principio de una base, mientras que el lugar donde ocurra la entropía más baja será el inicio de un sufijo.

Tómese, por ejemplo, el conjunto Φ de pares ordenados de vocablos y sus frecuencias, $\langle v_i, f_i \rangle$. Existen cuando menos dos maneras de calcular la entropía en cada corte de cada vocablo, dependiendo de qué frecuencias son las que se toman en cuenta. Específicamente, los vocablos tienen una frecuencia en el corpus, pero sus terminaciones se repiten en otros vocablos. Así que tienen una frecuencia de aparición en los vocablos y otra en el corpus (como se verá en la sección 2.5.2, esta diferencia permite calcular dos tipos de probabilidades para los afijos). La frecuencia en el corpus es la suma de las frecuencias de los vocablos en los que aparece la terminación. En este trabajo se encontró que tomar en cuenta el primer tipo de frecuencia, su aparición en los vocablos, da mejores resultados.

Considérese que al examinar un corte de palabra y determinar el conjunto de segmentos que alternan en alguno de los extremos de esa palabra y que cada uno de esos segmentos alternantes tiene una frecuencia dada, se pueden calcular probabilidades para cada uno de estos segmentos. Es decir, si imaginamos un depósito hipotético de segmentos que alternan, podemos calcular las probabilidades que tiene cada uno de ser escogido al azar.

Así, a partir de un vocablo a_{ij}::b_{ij}, podemos imaginar un conjunto B_{ij} como un depósito de segmentos con posibilidades de ser seleccionados. La probabilidad de cada segmento en ese conjunto se dará de la siguiente manera:

$$0 \leq p(b_{kj}|a_{ij}) = \frac{f(b_{kj})}{f(a_{ij})} \leq 1,\ b_{kj} \in B_{ij},\ \mathrm{k} = 1, 2, 3, \ldots |B_{ij}|$$

donde la suma de estas probabilidades será 1:

$$\sum_{k=1}^{|B_{ij}|} \leq p\ (b_{kj} \mid a_{ij}) = 1$$

Por ejemplo, tomemos el vocablo *previamente* y examinemos el primer corte posible (*/p::rebiaménte/*). En la lista de vocablos del CEMC, hay 7 206 que empiezan con *p~*, es decir, $f(a_{ij})$ = 7 206, que es también el número de elementos del conjunto B_{i1} (a partir de v_i = */prebiaménte/*). En la última celda de la segunda columna de la Tabla 6 se muestra el total de formas en B_{i1} que empiezan con la letra 'p'.

Nótese que la mayoría de las formas que aparecen después de *p~* empiezan con vocal o consonante laminal. Un buen número de palabras empiezan con 'ps' (*ps-icología*, *ps-íquico*, etc.). Cuando después de *p~* ocurren otras consonantes (*pb~*..., *pc~*..., *pd~*..., etc.) podemos asumir que se trata de siglas, abreviaturas, palabras extranjeras, símbolos matemáticos o químicos o errores de dedo en el corpus.

El total de la última columna de la Tabla 6 (en negritas) nos muestra la entropía en ese corte. Como se ve, se mide en bits. Cada renglón de la última columna contiene la probabilidad de lo que ocurra después de *p~* multiplicada por el logaritmo de base 2 de la misma probabilidad: $p \times log_2\ (p)$. Esto es la aplicación de la fórmula presentada en la sección 1.3.4, que se puede actualizar de la manera siguiente:

$$\mathrm{H}(ij)^{izq} = -\sum_{x=1}^{f(a_{ij})} p(b_{xj}|a_{ij}) \times \log_2\left(p(b_{xj}|a_{ij})\right)$$

con la que se calcula la entropía en el corte *j* del vocablo *i* después del segmento a_{ij}.

Tabla 6. Entropía de la segmentación $p::B_{i,1}$

$a_{ij}::B_{ij}$	formas	$p(b_{kj}\|a_{ij})$	$-p \times \log_2(p)$
p::a	1 365	0.18943	0.45468
p::b	1	0.00014	0.00178
p::c	2	0.00028	0.00328
p::d	1	0.00014	0.00178
p::e	1 396	0.19373	0.45873
p::f	1	0.00014	0.00178
p::g	1	0.00014	0.00178
p::h	3	0.00042	0.00468
p::i	511	0.07091	0.27073
p::j	1	0.00014	0.00178
p::k	1	0.00014	0.00178
p::l	384	0.05329	0.22541
p::m	2	0.00028	0.00328
p::n	2	0.00028	0.00328
p::o	835	0.11588	0.36030
p::p	3	0.00042	0.00468
p::r	2 184	0.30308	0.52197
p::s	73	0.01013	0.06712
p::t	7	0.00097	0.00972
p::u	407	0.05648	0.23417
p::v	2	0.00028	0.00328
p::x	1	0.00014	0.00178
p::y	1	0.00014	0.00178
p::z	2	0.00028	0.00328
p::-	13	0.00180	0.01644
p::’	6	0.00083	0.00852
p::ø	1	0.00014	0.00178
total	7 206	1.00000	**2.66955 bits**

En esencia, la entropía mide la información, en el sentido técnico del término, que la segunda parte del vocablo debe proporcionar para reducir la incertidumbre que resulta de las alternativas posibles. Como ya se ha dicho, si definimos a un afijo como una unidad con relativamente poca información, se puede intuir que una medida baja de información, en relación con los otros cortes posibles de la palabra, indica el inicio de un sufijo, mientras que la medida más alta indicaría el inicio de una base.

Por otra parte, se puede calcular la entropía del mismo corte, pero en sentido contrario; esto es, mediante las probabilidades de cada segmento contenido en algún conjunto A_{ij}, que ocurre antes de un segmento particular b_{ij}. En este caso, los valores altos corresponden a fronteras de sufijos y los bajos a las de prefijos. Esto no es ni tan extraño ni tan novedoso; por ejemplo, Harris —aunque no habla de entropía— encuentra que las cuentas de fonemas anteriores a una segmentación son tan buenos indicios de frontera morfológica como las de los posteriores. En cuanto a la entropía concretamente, también Hafer y Weiss (1974) notaron que "en reversa" es tan buen indicador morfológico como de izquierda a derecha.

De hecho, puesto que los picos de entropía señalan el principio de bases después de un prefijo, los picos de entropía "al revés" marcan el inicio de un sufijo. Es más, en este experimento, se puede ver que las entropías en reversa son mejores indicadores de sufijos que los valores más bajos de la entropía de izquierda a derecha. Similarmente, los valores máximos de ésta última son mejores indicadores de fronteras entre prefijos y bases.

Por ejemplo, en la Tabla 7 se muestran los valores de entropía calculados en cada corte del vocablo *aparecer* (/*aparesér*/) en ambas direcciones. Así, el valor de entropía entre el prefijo *a~* y la base es de 2.792 (en negritas) y el valor de entropía entre *apares~* y el sufijo *~er* es 0.950 (renglón de izquierda a derecha), mientras que el valor de entropía entre este sufijo y la base es de 2.516 (en negritas) y la entropía entre la base y el prefijo *a~* es de 1.277 (renglón de derecha a izquierda). Nótese que los valores más altos corresponden con los cortes morfológicos, los de izquierda a derecha para prefijos y los de derecha a izquierda para sufijos.

Nótese también que los valores mínimos de entropía no necesariamente corresponden a los cortes morfológicos.

Tabla 7. Valores de entropía en cada segmentación del vocablo *aparecer*

	a	**p**	**a**	**r**	**e**	**s**	**e**	**r**
izq.-der.		**2.792**	1.818	1.630	1.298	1.270	0.950	1.303
der.-izq.		1.277	0.802	1.619	2.125	1.560	**2.516**	1.193

Una posible explicación es que los afijos, que cargan menos información que las bases, no dejan de cargarla: un signo sin contenido informativo no tiene mucho sentido. De hecho, puede haber entropías más bajas en el interior de los significantes de los signos. Así, la entropía más baja de derecha a izquierda (0.802) no marca el final de un prefijo, aunque la más alta sí marque el inicio del sufijo *~er*. De allí que solamente se pueda decir que el prefijo no termina donde está la entropía más baja. Por otra parte, saber dónde empieza o termina una base permite suponer el final o principio de un afijo. De todo esto, es importante recalcar que el afijo no es la unidad con menos información, sino que la base o lexema es el elemento con el más alto contenido de ésta. De esta manera, en los experimentos de este trabajo, la medida de entropía se refiere al promedio de información de los contextos adyacentes al supuesto afijo, que son las supuestas bases. Así, aunque el afijo contiene menos información, la entropía que se le asocia, es en realidad, la entropía de las bases, que necesariamente sería mayor que el del afijo.

Uno se podrá preguntar si es posible combinar los valores de una dirección con los de la otra, por ejemplo, substrayendo uno del otro. Harris propuso el uso de cuentas de fonemas anteriores principalmente para cotejar los resultados derivados de las cuentas de los fonemas posteriores, pero Hafer y Weiss combinaron estas cuentas de diferentes maneras sin obtener resultados alentadores. De hecho, los resultados fueron muy malos, por ejemplo, al sumarlos. Además, señalaron que "the union of

the measures produces too many incorrect cuts, while the intersection of the methods is too restrictive" (Hafer y Weiss 1974, 378). Como sea, veremos más adelante que los valores de entropía de cualquiera de las dos direcciones son mejores pistas por separado que, por ejemplo, la substracción del valor de una dirección menos el de la contraria.

2.3.3 Principio de economía

Como se mencionó en el capítulo anterior, Josse de Kock y Walter Bossaert desarrollaron en los años setenta un método para determinar, sin supervisión, las fronteras morfológicas entre bases y afijos. Su procedimiento está basado, como se dijo, en el principio de economía: el número de signos en todos los niveles del lenguaje tiende a ser menor que el número de cosas nombradas; de tal manera que cada signo puede utilizarse en diferentes contextos sin crear ambigüedad. Así, mientras menos signos de más frecuencia existan en el nivel morfológico, que den lugar a más signos (de baja frecuencia) del nivel sintáctico, la lengua será más económica. Este mecanismo sencillo nos proporciona un criterio para determinar los signos morfológicos.

Si dividimos un vocablo v_i en dos segmentos, $a_i::b_i$, y uno de éstos ocurre en muchos otros vocablos, mientras que el otro ocurre en unos pocos y, si el primero pertenece a un conjunto pequeño de segmentos muy frecuentes, mientras que el segundo pertenece a un conjunto muy grande, potencialmente infinito, de segmentos de baja frecuencia, se puede proponer un corte morfológico entre esos dos segmentos. Es más, el primero tendrá que ser un afijo y el segundo una base.

Por ejemplo, tómense los segmentos de la Figura 6. Cada segmento de la izquierda se combina con cada segmento de la derecha para formar verbos flexionados (*compra, comprada, comprado, comprando, … compró;… canta, cantada,… cantó;… controló*; …, etc.). Nótese que las formas de la derecha constituyen un conjunto *B* más bien pequeño de formas muy frecuentes y las de la izquierda el conjunto *A* de muchísimos

más miembros (un número potencialmente infinito) que son relativamente menos frecuentes (esto es, que aparecen en menos vocablos que las primeras). Esto es una pista de que los segmentos del conjunto *B* son afijos, específicamente sufijos. De esta manera, al comparar los tamaños de estos conjuntos, se puede argüir que el corte examinado es morfológico.

A	***B***
compr	a
cant	ada
alivi	ado
rest	ando
ray	ar
sum	aron
seleccion	aste
...	…
arrest	es
elabor	é
nad	o
anhel	ó
contrat	
am	
apel	
mand	
colabor	
control	
...	
∞	

Figura 6. Combinaciones de segmentos de la izquierda y de la derecha

Recuérdese que A_{ij} es el conjunto de segmentos que alternan a la izquierda del segmento b_{ij}. Es decir, dado un vocablo $a_{ij}::b_{ij}$, es el conjunto de segmentos encontrados al dejar alternar a_{ij} ($a_{ij} \in A_{ij}$). Similarmente, B_{ij} es el conjunto que alterna a la derecha de a_{ij} (por lo que $b_{ij} \in B_{ij}$). Entonces, $|A_{ij}|$ es el tamaño del conjunto A_{ij} y $|B_{ij}|$ el de B_{ij}.

2.3.3.1 Algunas restricciones

Para afinar los resultados, se pueden aplicar algunas restricciones. Específicamente, si se busca descubrir los afijos, se pueden eliminar de los conjuntos de ambos lados cualquier segmento que, según su frecuencia, sea potencialmente una base o lexema. Así, se puede prescindir desde un principio de aquellos segmentos que sean menos frecuentes que sus acompañantes, porque las bases son menos frecuentes que los afijos que las acompañan.

Sea A^p_{ij} el conjunto de segmentos que hipotéticamente funcionan como prefijos y B^s_{ij} el de sufijos hipotéticos. Entonces, A^p_{ij} es miembro de A_{ij} ($A^p_{ij} \in A_{ij}$) y contiene aquellos miembros de A_{ij} con frecuencias que comparadas con las de los miembros de B_{ij} se comportan como prefijos. Similarmente, los miembros de B^s_{ij} son también miembros de B_{ij} ($B^s_{ij} \in B_{ij}$) y funcionan como sufijos. Además, sea $|A^p_{ij}|$ el número de miembros de A^p_{ij} y $|B^s_{ij}|$ el número de miembros de B^s_{ij}.

Otro ejercicio útil para afinar las cuentas es el de mirar los caracteres que colindan con el segmento opuesto. Por ejemplo, dado un vocablo formado por los segmentos a_{ij} y b_{ij}, hay un conjunto de formas alternantes en el extremo opuesto de cada uno de éstos: $a_{ij}::B_{ij}$ y $A_{ij}::b_{ij}$. Con frecuencia se dará la situación en que B_{ij} contenga varios segmentos que empiezan con una letra en común (*~ra*, *~rada*, *~res*, etc.) o que A_{ij} contenga varios elementos que comparten la misma letra final (*sufre~*, *sube~*, *bate~*, etc.). Esas letras en común probablemente son parte del segmento acompañante, puesto que colindan con él. En otras palabras, si dentro del conjunto de afijos hipotéticos, hay alguno que aparece con varias bases que comparten el fonema adyacente a éste, se puede suponer que ese fonema es parte del supuesto afijo y no de las supuestas bases.

De manera similar, si dentro del conjunto de bases hipotéticas, hay un supuesto lexema cuyos afijos acompañantes comparten fonemas adyacentes a éste, se puede sospechar que esos fonemas pertenecen al lexema (de Kock y Bossaert 1978, 21). Así, los casos de fonemas adyacentes

a la segmentación, que se repiten en los segmentos que alternan, deben eliminarse de las cuentas. Por ejemplo, en la Figura 6, las formas *~a, ~ada, ~ado, ~ando, ~ar, ~aron, ~aste* cuentan sólo una vez; igualmente, *compr~, elabor~ y colabor~* cuentan también sólo una vez.

Una última restricción importante[10] consiste en requerir que ambos conjuntos, A_{ij} y B_{ij}, contengan miembros que aparezcan en más de un vocablo; es decir, cada segmento debe ocurrir en por lo menos dos vocablos diferentes, para garantizar la presencia de por lo menos un cuadro. Por otra parte, si queremos determinar un número mínimo de cuadros, no es fácil decidir cuántos son suficientes para garantizar que el corte sea morfológico[11]. Así que, para aceptar un corte como candidato, en este trabajo se requirió la presencia de únicamente un cuadro, puesto que la ausencia de cuadros es un indicio muy fuerte de que no hay corte morfológico.

2.3.3.2 Índice de economía

Por otra parte, la economía de un corte de palabra se puede calcular al comparar los tamaños de los conjuntos modificados con estas restricciones: mientras más grande sea la diferencia en número de segmentos considerados como bases con respecto al número de aquellos asumidos como afijos, más económica será el corte[12].

[10] De Kock y Bossaert aplicaron otras restricciones que resultaron tener poco impacto (1978, 23). Por ejemplo: requerir que las bases contengan por lo menos una vocal; eliminar afijos conocidos de entre los segmentos examinados como bases; etcétera.

[11] No es fácil determinar un límite mínimo de cuadros que atestigüen fronteras morfológicas (en parte porque, como se apuntó arriba, no hay una prueba automática de significado y con frecuencia se observan cuadros semánticamente inaceptables en cortes no morfológicos). Por otra parte, este requisito no es raro, además del de Kock y Bossaert, el procedimiento de Andreev para detectar paradigmas, por ejemplo, también requiere la presencia de cuadros (Cromm 1996, 8).

[12] En 1996, se llevó a cabo una primera aplicación de este método a la nomenclatura del Diccionario del español de México, cuyos resultados se presentaron en la ponencia "Un experimento cuantitativo de determinación de fronteras morfológicas del español de México", en el IV Encuentro Internacional de Lingüística en el Noroeste en Hermosillo, Sonora (noviembre de 1996).

Si alternan a la izquierda más segmentos de tipo base ($|A_{ij}| - |A^p_{ij}|$) que segmentos de tipo afijo a la derecha ($|B^s_{ij}|$), tiene sentido considerar sufijo al segmento de la derecha b_{ij}; y al revés, si más segmentos de tipo base alternan a la derecha ($|B_{ij}| - |B^s_{ij}|$) que segmentos de tipo afijo a la izquierda ($|A^p_{ij}|$), tiene sentido suponer que el segmento de la izquierda a_{ij} es un prefijo. De esta manera, tenemos esencialmente dos medidas de economía asociadas a un corte, dependiendo del tipo de afijo que se hipotetice:

$$k^p_{ij} = \frac{|B_{ij}| - |B^s_{ij}|}{|A^p_{ij}|}$$

donde k^p_{ij} medirá la economía de la segmentación j en el vocablo v_i y tendrá un valor mayor a la unidad cuando su primer segmento, a_{ij}, sea un prefijo o será una fracción cuando el segmento de la derecha b_{ij} sea un sufijo. Similarmente,

$$k^s_{ij} = \frac{|A_{ij}| - |A^p_{ij}|}{|B^s_{ij}|}$$

donde k^s_{ij} medirá la economía del corte j en el vocablo v_i y tendrá un valor mayor a la unidad cuando b_{ij} sea un sufijo o será menor a ésta cuando a_{ij} sea un prefijo.

Podemos ilustrar todo esto con el sufijo derivativo *~ura* (por ejemplo, en *fritura*) que, por un lado, alterna con los morfemas *~o* y *~a* (en *frito* y *frita*), con los grupos sufijales de plural *~o.s* y *~a.s* (en *fritos* y *fritas*) y con la secuencia de dos sufijos *~ura.s* (*frituras*); y, por el otro, se sufija a otras bases para formar sustantivos, que constituyen una clase abierta (*calentura*, *cordura*, *abreviatura*, etc., sumando *n* palabras). En este contexto, la medida de economía correspondería al número de palabras de un corpus con el sufijo *~ura* (*n* = el número de bases que lo acompañan: *calent~*, *cord~*, *abreviat~*, etc.) dividido entre el número de signos con que alterna en el corpus (en este ejemplo cinco): $k^s = n/5$. Así que, mientras mayor sea *n*, más carácter de afijo tendrá *~ura*.

En dirección contraria, diríamos que el fragmento *frit~* alterna con un gran número de formas (*calent*ura, *cord*ura, *loc*ura, *ric*ura, *amarg*ura, *blanc*ura, etc.) y que se le adhieren algunos pocos elementos: además de *~ura* y la secuencia *~ura.s*, están los sufijos *~o*, *~a*, *~o.s*, *~a.s*. La medida k^p para *frit~* es muy reducida. Resulta de dividir un número relativamente pequeño de acompañantes (sólo 6, contando *~ura*) entre uno mucho mayor de alternantes (el número de formas que terminan en *~ura*, descontando *fritura*): $k_s = 6/(n - 1)$. Así que, mientras mayor sea *n*, menos probable es que *frit~* sea un prefijo. De todo esto tenemos que concluir que *frit~* (en *fritura*) no es un afijo.

Por último, en los experimentos de este trabajo se aplicaron a cada corte de cada vocablo versiones normalizadas de estos índices, para obtener valores entre cero y la unidad, [0, 1]:

$$k^p_{ij} = 1 - \frac{|A_{ij}| - |A^p_{ij}|}{|B^s_{ij}|}$$

para el segmento de la izquierda como prefijo, y

$$k^s_{ij} = 1 - \frac{|B_{ij}| - |B^s_{ij}|}{|A^p_{ij}|}$$

para el de la derecha como sufijo.

2.4 Un experimento con el CEMC

Las técnicas examinadas en los apartados anteriores se aplicaron a una muestra aleatoria de vocablos del CEMC con el objeto de comparar su utilidad como criterios para descubrir afijos del español. Este corpus es una colección de casi mil textos, de alrededor de 2 000 palabras gráficas cada uno, agrupadas en párrafos escogidos al azar[13]. El corpus cuenta

[13] Los textos se originaron en la República Mexicana entre 1921 y 1974. Se trata tanto de obras escritas como de transcripciones de entrevistas grabadas. Están agrupados en 14 géneros textuales clasificados como lengua culta (literatura, periodismo, ciencias, técnicas, discursos políticos,

con 1 891 045 palabras gráficas ($|\Psi| = \xi$) que constituyen más de 79 000 tipos de palabras ($|V| = \Omega$) y que se usan para calcular los índices descritos arriba. Como se mencionó, el CEMC se compiló para seleccionar los vocablos de la nomenclatura inicial del DEM, así que parece una buena fuente de información para descubrir los afijos del español de México.

La muestra aleatoria de palabras se seleccionó del conjunto de tipos *V* del CEMC. Se eliminaron algunos vocablos como *convoy*, *nocaut*, *jueves*, etc., porque cualquier corte morfológico en estas palabras no parece muy afortunado, y todas las formas con menos de cinco letras de longitud. Así que quedaron 851 tipos de palabra para los que se calcularon los índices de número de cuadros, de entropía y de economía en cada uno de sus cortes (tomando en cuenta la información de todo el conjunto de tipos de palabra, *V*). Para cada vocablo, se almacenaron las formas cuyos valores fueron los más altos y los más bajos de cada índice. También, se determinó mediante inspección si el corte morfológico propuesto por cada índice (el valor más alto para cada vocablo) era o no válida. Para simplificar el experimento se contaron solamente los sufijos. Esto es, no se tomaron en cuenta los aciertos en la predicción de prefijos.

Como era de esperarse, determinar si eran correctos los cortes relacionados con afijos de flexión no presentó mayores dificultades. Sin embargo, no fue tan sencillo con muchas que involucraban afijos de derivación. Por ejemplo, algunos vocablos aparentan tener o exhiben una estructura que no se corresponde con su significado ni con su etimología: el vocablo *almohada* (del árabe: *al-muhádda*), para el que los índices propusieron una frontera entre *almoh~* y un supuesto sufijo *~ada*. Sin embargo, esta propuesta no se sostiene ni semántica ni eti-

---

religión y habla culta), subculta (literatura popular, habla media, lírica popular) y no-estándar (textos dialectales, documentos antropológicos, jergas y habla popular). Esta subdivisión en géneros contribuye a la diversidad de textos y la representatividad estadística del CEMC. Para una descripción más detallada de este corpus, véase Lara, Ham y García, *op. cit.* (1979) y el "Apéndice A" en Medina, *Investigación cuantitativa de afijos y clíticos del español de México*, *op. cit.* (2003, 350-373).

mológicamente. De todas maneras, la propuesta es compatible con la morfología del español. Aunque en este experimento se contó como un corte incorrecto, sería interesante ver qué cortes morfológicos proponen varios hablantes del español que desconozcan el origen etimológico del vocablo y averiguar si coinciden con la propuesta de estos índices. Si es así, sería interesante averiguar cómo saben que allí ocurre el sufijo participial *~ada*.

La mayoría de los errores tuvieron que ver con cortes propuestos en medio de lexemas, incluso produciendo hiatos (por ejemplo **su::élta-me*), o cortando afijos de sobra conocidos (por ejemplo **ekibokasióne::s*). Además, aunque el mejor corte propuesto por un índice sea morfológicamente equivocado, el segundo o tercero mejor bien pueden cortar correctamente la palabra, puesto que la mayoría de ellas tienen más de un corte morfológico correcto. De todas maneras, en este experimento sólo se tomaron en cuenta las mejores propuestas de corte de cada índice (su valor más alto) para cada vocablo.

Tabla 8. Comparación de índices: segmentaciones correctas en una muestra de 836 vocablos

índice	**aciertos**	**porcentaje**
cuadros, economía o entropía	764	90.41%
cuadros	737	87.22%
entropía	730	86.39%
economía	669	79.17%
substracción de entropías	272	32.19%

En la Tabla 8 se exhibe la comparación de los resultados de los índices. Como se dijo en la sección del índice de entropía, la peor de las medidas es la que toma en cuenta la diferencia entre las entropías de ambas direcciones (“substracción de entropías”, con sólo 32% de acier-

tos); esto es, la substracción de los valores de entropía con respecto a los sucesores menos aquellos con respecto a los predecesores.

Un aspecto interesante del experimento es que, al combinar los resultados de los índices de cuadros, economía y de entropía, el porcentaje de aciertos mejoró al 90.41%. Además, al comparar estas técnicas con otras, como las estadísticas de digramas, éstas siguieron obteniendo mejores resultados (Medina Urrea 2000, 97-114). Podemos conjeturar que tienen mejores resultados porque miden propiedades que hemos asociado directamente al fenómeno de la afijación y que toman en cuenta la estructura subyacente del inventario completo de vocablos. Es decir, se basan en nociones lingüísticas y miden las características determinantes de los afijos. Así que tiene sentido combinarlas para calcular la propiedad, hasta ahora cualitativa, que tienen algunos fragmentos de palabra de ser afijos[14].

2.5 Los catálogos de afijos

En este apartado se describen las bases para construir inventarios o catálogos de segmentos de palabras que actúen como afijos. Se ha establecido que un corte en el interior de una palabra será más morfológico o más económico mientras más cuadros tenga. Además, uno de los segmentos producidos por el corte será más afijal mientras mayor sea la diferencia entre su frecuencia y la frecuencia de las bases que lo acompañen.

Sin embargo, no hemos examinado con precisión las nociones de *cuadros suficientes*, *formas frecuentes*, *mayor diferencia*, etc. Evidentemente, hay cierto grado de arbitrariedad en decidir cuánto es suficiente, mucho o poco. Así que, en este trabajo, se pospusieron estas decisiones lo más posible. De hecho, en lugar de rechazar el carácter afijal de tal o cual segmento a partir solamente de los datos calculados en un vocablo

[14] En experimentos previos, incluso con una muestra menor de 217 formas, se han obtenido resultados similares: los índices de cuadros, entropía y economía obtuvieron entre 10 y 30 puntos porcentuales más que las de estadísticas de digramas.

aislado, resultó mejor hacerlo con base en todas sus ocurrencias en los vocablos del conjunto *V* del CEMC, sin importar si ese segmento verdaderamente representa o no un afijo en todos los contextos. De esta manera, se procedió a examinar cada corte de cada vocablo del conjunto *V* y, según las medidas calculadas, a agregar cada candidato a afijo a una estructura de datos. Esta estructura puede verse como un catálogo de afijos y la definiremos de la siguiente manera:

2.5.1 Definición formal de un catálogo de afijos

Sea Γ un catálogo de afijos descrito por el séxtuplo (*S, C, K, H, F′, F″*), donde *S* es el conjunto de segmentos afijales de γ elementos, $\{s_1, s_2, s_3, \ldots, s_\gamma\}$, extraídos de un corpus Ψ.

Sea entonces *C* el conjunto de promedios de cantidades de cuadros asociadas a cada ocurrencia de estos segmentos, $\{\bar{c}_1, \bar{c}_2, \bar{c}_3, \ldots, \bar{c}_\gamma\}$. Sea *K* el conjunto de promedios de índices de economía, $\{\bar{k}_1, \bar{k}_2, \bar{k}_3, \ldots, \bar{k}_\gamma\}$. Sea *H* el conjunto de promedios de entropía, $\{\bar{h}_1, \bar{h}_2, \bar{h}_3, \ldots, \bar{h}_\gamma\}$. Sea *F′* el conjunto de las frecuencias absolutas de cada segmento, $\{\Omega'_1, \Omega'_2, \Omega'_3, \ldots, \Omega'_\gamma\}$; y *F″* el conjunto de frecuencias de los segmentos como afijos entre los vocablos $\{\Omega''_1, \Omega''_2, \Omega''_3, \ldots, \Omega''_\gamma\}$. De esta manera, Γ también puede describirse como el conjunto de γ relaciones ordenadas o tuplas:

$$\begin{aligned}\Gamma = \{&\langle s_1, \bar{c}_1, \bar{k}_1, \bar{h}_1, \Omega'_1, \Omega''_1\rangle,\\ &\langle s_2, \bar{c}_2, \bar{k}_2, \bar{h}_2, \Omega'_2, \Omega''_2\rangle,\\ &\langle s_3, \bar{c}_3, \bar{k}_3, \bar{h}_3, \Omega'_3, \Omega''_3\rangle,\\ &\ldots \langle s_\gamma, \bar{c}_\gamma, \bar{k}_\gamma, \bar{h}_\gamma, \Omega'_\gamma, \Omega''_\gamma\rangle\}\end{aligned}$$

Además, se construyen dos catálogos separados; Γ^p, que contiene los prefijos, y Γ^s, que contiene los sufijos. En esta notación, los catálogos son conjuntos de tuplas sin un orden específico, pero veremos más adelante que el ordenamiento será de gran importancia. Las tuplas se

pueden ordenar por los valores, de mayor a menor o viceversa, de cualquiera de sus elementos o por una función de ellos.

2.5.2 Probabilidades de los afijos

A continuación, examinaremos dos tipos de frecuencias relativas o probabilidades asociadas a los afijos con respecto a su pertenencia a un catálogo formal como el definido arriba. Como se mencionó, podemos estimar dos tipos de probabilidades para cada afijo, una contando sus ocurrencias en los vocablos del corpus (esto es, en el conjunto *V*, sin tomar en cuenta las frecuencias de los vocablos) y otra en las palabras gráficas del corpus, que sí toma en cuenta las frecuencias de los vocablos.

La ocurrencia de una cadena de caracteres en un corpus no garantiza que esa cadena represente a un afijo. Por ejemplo, en los vocablos *comente*, *aumente*, *argumente*, etc., la cadena 'mente' no representa al afijo *~mente* ni la cadena 'te' al pronombre enclítico *~te*. Sin embargo, como hemos visto, los índices presentados arriba permiten determinar con cierta seguridad cuándo un segmento es un afijo y cuándo no. Así que cada fragmento de palabra podrá tener una frecuencia de aparición como forma y otro como afijo.

En la teoría clásica de la probabilidad[15], las probabilidades se pueden estimar directamente de los datos recolectados empíricamente simplemente al dividir el número de ocurrencias de un hecho imprevisto entre la suma de las ocurrencias del total de los hechos. De esta manera, podemos calcular la probabilidad de aparición de un vocablo v_i, que

[15] Para una presentación de la teoría de la probabilidad en el marco de la lingüística, véanse Manning y Schütze (1999, 40-58); Altmann, *Statistik für Linguisten*, Wissenschaftlicher Verlag Trier, Tréveris (1995, 61-91); Piotrowski, Lesohin y Lukjanenkov, *Introduction of Elements of Mathematics to Linguistics*, Brockmeyer, Bochum (1990, 162-216); Woods, Fletcher y Hughes (1986, 59-75); Piotrowski, Bektaev y Piotrowskaja, *Mathematische Linguistik*, Brockmeyer, Bochum (1985, 153-165).

ocurre en el corpus Ψ con ξ ocurrencias de palabras dividiendo la frecuencia del vocablo entre ξ:

$$0 \leq p(v_i) = \frac{f_i}{\xi} \leq 1,\ i = 1, 2, 3, \ldots, \Omega$$

donde, como quedó establecido arriba, f_i es la frecuencia de v_i, ξ es el tamaño del corpus Ψ y Ω es el total de vocablos, esto es, el tamaño del conjunto *V*.

De manera similar, al contar el número de veces que un segmento cumple su papel de afijo, se puede calcular la probabilidad de que el segmento sea en efecto un afijo en el vocabulario *V*. Si en el catálogo Γ hay un conjunto $F' = \{\Omega'_1, \Omega'_2, \Omega'_3, \ldots, \Omega'_\gamma\}$ de las frecuencias absolutas de los segmentos y un conjunto $F'' = \{\Omega''_1, \Omega''_2, \Omega''_3, \ldots, \Omega''_\gamma\}$ de las frecuencias de cada forma como afijo, $\Omega''_k \leq \Omega'_k \leq \Omega$ (para cada k = 1, 2, 3, …, γ), entonces, podemos estimar la probabilidad, **prob1**, de que un segmento s_k sea un afijo en el vocabulario *V* de la siguiente manera:

$$0 \leq p(s_k) = \frac{\Omega''_k}{\Omega'_k} \leq 1,\ k = 1, 2, 3, \ldots, \gamma$$

También es posible tomar en cuenta que cada vocablo en que aparece un segmento tiene su propia frecuencia de ocurrencia en el corpus. De hecho, un subconjunto de esas ocurrencias corresponderá a sus apariciones con carácter de afijo. Recuérdese que f_i es la frecuencia del vocablo *i* en el corpus. En el catálogo Γ hay un conjunto de segmentos s_k cuyas frecuencias están almacenadas en f'_{ki} y en f''_{ki}. La primera es igual a f_i y la segunda es igual a su frecuencia cuando s_k es afijo, pero es igual a 0 cuando no es afijo:

$$f'_{ki} = f_i, \qquad f''_{ki} = \begin{cases} f_i,\ s_k \in A_i^p,\ s_k \in B_i^s \\ 0,\ s_k \notin A_i^p,\ s_k \in B_i^s \end{cases}$$

Con esto se puede calcular una segunda probabilidad, **prob2**, que toma en cuenta las frecuencias de los vocablos en el corpus Ψ:

$$0 \leq p^{\Psi}(s_k) = \frac{\sum_{q=1}^{\Omega''_k} f''_{ki}}{\sum_{r=1}^{\Omega'_k} f'_{ki}} \leq 1,\ k = 1, 2, 3, \ldots, \gamma$$

Esto quiere decir que al recorrer una cadena de ocurrencias de palabras Ψ, $p^{\Psi}(s_k)$ es la probabilidad de que el segmento s_k sea un afijo en esa secuencia de ocurrencias[16].

2.6 Hacia un índice de afijalidad

A continuación, se presenta una manera de calcular un índice general de afijalidad, es decir, una medida del carácter de afijo que pueda tener una cadena de caracteres que forme parte de un vocablo. La ventaja de las medidas de economía, entropía y número de cuadros en la predicción de fronteras morfológicas se puede explicar en el hecho de que corresponden al concepto de afijo: son menos, más frecuentes y contienen menos información que otros tipos de signos. Así que tiene sentido caracterizar formalmente la cualidad que un segmento de palabra pueda tener de ser un afijo en términos de estos tres índices. Se puede proponer la siguiente fórmula para medirla:

$$AF(s_x) = \frac{f_x\, c_x\, k_x}{h_x}$$

[16] Estas probabilidades pueden contrastarse con las sugeridas por Meya para medir la probabilidad de ocurrencia de un morfema de alguna lengua a partir de una cadena de Markov (cosa, como establece la investigadora, útil en procedimientos de síntesis automática del habla, ya que las fronteras de morfemas proporcionan información sobre la prosodia (Meya 1986, 142). No es difícil imaginar cómo estas probabilidades se pueden afinar mediante la construcción de cadenas de Markov. La diferencia principal es que Meya presupone las reglas morfológicas (su red de hipótesis específicas al español y confeccionadas manualmente) para segmentar palabras en morfemas, mientras que aquí se propone un esquema de segmentación que no presupone las estructura morfológica.

donde f_x, c_x, k_x y h_x representan respectivamente la frecuencia, la cuenta de cuadros, la economía y la entropía del segmento s_x calculadas a partir de los vocablos de un corpus Ψ. Como ya se explicó, para que un segmento s_x sea un afijo, se espera, por un lado, que ocurra con una gran frecuencia en el corpus, que esté involucrado en un gran número de cuadros c_x, y que tenga asociada una gran cantidad de economía k_x. Por el otro, un afijo deberá contener una cantidad mínima de información h_x. En esta fórmula, mientras mayores sean las primeras cantidades y menor sea la cantidad de información, mucho mayor será la afijalidad $AF(s_x)$.

Sin embargo, hay dos cosas que reconsiderar. Primero, es obvio que la frecuencia del segmento es una señal de su afijalidad, pero también es cierto que la frecuencia depende de las muestras textuales de donde salgan las palabras. De hecho, las frecuencias varían entre diferentes corpus. Por ejemplo, hay una tendencia de ciertos afijos a ocurrir en ciertos tipos de textos y no en otros. La frecuencia misma puede concebirse como una manifestación o resultado de la economía, las estructuras combinatorias y el contenido de información inherentes al afijo. Así que aquí se prescindirá de ella.

Segundo, como se mencionó arriba, un contenido de información mínimo como criterio para determinar afijos no funciona tan bien como el pico de mayor información que caracteriza a las bases. En esencia, los afijos no son necesariamente los segmentos con menor información; sólo contienen menos que las bases. Por eso, no se tomará en cuenta la cantidad de información que contiene un afijo, sino la entropía de su contexto adyacente, esto es, de las bases que lo acompañan, que es directamente proporcional a la afijalidad. Al tomar estas consideraciones en cuenta, podemos redefinir a la afijalidad de la siguiente manera:

$$AF(s_x) = k_x\, c_x\, h_x$$

De esta manera, la cualidad que tiene s_x de ser un afijo es directamente proporcional al producto de alguna medida de economía (k) por el número de cuadros (c), por una medida (h) de la sorpresa inherente a la transición de ese segmento al siguiente —todas estas cantidades calculadas a partir de la frontera de un segmento de palabra y sus posibles segmentos adyacentes (supuestas bases) y del conjunto V de vocablos.

Esta generalización funciona para el cálculo de afijalidad de segmentos afijales de palabras aisladas, donde el supuesto afijo ocurre una vez. Sin embargo, la misma relación se sostiene si calculamos los promedios de los valores de afijalidad de varios afijos en varios vocablos. Esto resulta en un índice de afijalidad para las formas que pertenecen al catálogo de afijos Γ y que toman en cuenta todas las ocurrencias del segmento afijal:

$$AF(s_x) = \bar{k}_x \, \bar{c}_x \, \bar{h}_x$$

De manera intuitiva, estos índices se pueden combinar para representar otros tipos de fenómenos morfológicos. De hecho, un índice bajo de economía, un número pequeño de cuadros o poca entropía disminuirían la afijalidad de un segmento, pero el que uno de estos índices desfavorezca la afijalidad general, no significa que los otros dos no signifiquen nada: puede que sean indicios de algún otro tipo de morfo.

De hecho, sería pertinente explorar si —y hasta qué punto— un índice bajo de economía y un número alto de cuadros estarían relacionados con el fenómeno de composición. De manera similar, una medida baja de entropía seguramente estaría relacionada con algún tipo de unidad morfológica de contenido, porque la medida no mide su contenido de información, sino el de lo que le es adyacente en el vocablo. Recuérdese que, en los experimentos de este trabajo, la medida de entropía se refiere al pico que marca el principio de la supuesta base adyacente del segmento revisado. Así, aunque el afijo contiene menos

información, la entropía que se le asocia, es en realidad, la de las bases, que necesariamente será mayor que la del afijo.

Además, aparte de las ventajas de combinar estas medidas, conviene normalizarlas; es decir, ajustarlas en el intervalo [0, 1] para hacer comparables sus magnitudes. Hay varias maneras de proceder para hacer esto. Una posibilidad es la fórmula siguiente:

$$AF^{n}(s_x) = \frac{\frac{c_x}{\max c_i} + \frac{h_x}{\max h_i} + \frac{k_x}{\max k_i}}{3}$$

Ecuación 2. Índice normalizado de afijalidad

Así, cada promedio de cada índice puede normalizarse dividiéndolo entre el valor máximo obtenido para ese índice. Cuando no se conozca el máximo global, se puede utilizar el máximo del vocablo. Luego, para evitar números pequeños, se puede calcular el promedio de los índices normalizados, en lugar de multiplicarlos.

Este índice de afijalidad se calculó para los vocablos de la muestra aleatoria sacada del CEMC que se mencionó arriba. 764 de los 836 vocablos fueron segmentados correctamente, lo que significó que el 90.41% de esos vocablos se segmentaron correctamente. Es el mismo índice que se calculó para compilar los catálogos de afijos que se presentan en los siguientes capítulos.

2.7 Catálogos de afijos a partir del CEMC

En este apartado, se presentan los resultados de la aplicación al CEMC de los índices descritos en los apartados anteriores. Concretamente, se examinan los resultados de la construcción de dos catálogos de afijos del español de México, uno de sufijos y otro de prefijos.

Para construir los catálogos Γ^{s} y Γ^{p} a partir del CEMC, se llevaron a cabo varios experimentos que permitieron explorar los diferentes ca-

minos posibles en la recolección de los afijos. Un aspecto importante de esto es la aplicación de reglas de reescritura para convertir los grafemas en signos representativos de los fonemas del español de México. Así que se aplicaron las reglas de reescritura de la Tabla 9 para acercar la ortografía de los vocablos a su pronunciación.

Tabla 9. Reglas de reescritura de caracteres
para reflejar correspondencia entre grafemas y fonemas

reglas	**fonema**	**contextos**
‘v’ → ‘b’	/b/	todos
‘z’, ‘c’ → ‘s’	/s/	‘z’; ‘ce’, ‘ci’
‘c’, ‘qu’ → ‘k’	/k/	‘ca’, ‘que’, ‘qui’, ‘co’, ‘cu’
‘ch’ → ‘ʧ’	/ʧ/	todos
‘gu’ → ‘g’	/g/	‘gue’, ‘gui’
‘g’ → ‘g’	/g/	‘ga’, ‘go’, ‘gu’, ‘gü’
‘g’ → ‘j’	/x/	‘ge’, ‘gi’
‘h’ → ε	-	todos
‘y’ → ‘i’	/i/	fin de sílaba, después de vocal (‘ay’, ‘ey’, ‘oy’, ‘uy’).
‘y’, ‘ll’ → ‘y’	/ǰ/	principio de sílaba, antes de vocal
‘rr’ → ‘r̄ ’	/r̄ /	todos
‘r’ → ‘r̄ ’	/r̄ /	inicio de palabra; o después de sílaba que termina en ‘n’, ‘l’ o ‘s’

Otro aspecto importante de las palabras gráficas es su acentuación. En español escrito las sílabas tónicas se representan gráficamente

mediante una tilde sobre la vocal según reglas bien conocidas. Puesto que varios morfemas gráficos se distinguen entre sí por la aplicación de esta tilde (de tal manera que *~o* ≠ *~ó*, *~aras* ≠ *~arás*, and *~ás* ≠ *~as*), se pueden mantener los acentos de las palabras gráficamente acentuadas. Por otra parte, como los vocablos graves terminados en 'n', 's' o vocal no se acentúan en español escrito, es necesario introducir gráficamente estos acentos en las sílabas tónicas. Sin embargo, esto requeriría presuponer la estructura de la sílaba, lo cual implica supervisar los datos. Una solución posible es conservar solamente las tildes de las últimas sílabas, esto es, de las últimas vocales. Finalmente, se puede agregar el acento gráfico a la última vocal de las palabras que carecían de tilde y cuya terminación no fuera ni 'n' ni 's' ni vocal (lo que indicaba que son agudas): *coronel* → /koron**é**l/, *vejez* → /bex**é**s/, *codorniz* → /kodorn**í**s/.

El paso siguiente es examinar automáticamente cada posible corte *j* de cada vocablo v_i para determinar sus índices pertinentes: c^p_{ij} (número de cuadros al asumir un prefijo), c^s_{ij} (número de cuadros al asumir un sufijo)), k^p_{ij} (índice de economía al asumir prefijo), k^s_{ij} (índice de economía al asumir sufijo), h^p_{ij} (entropía de prefijo a base), h^s_{ij} (entropía de sufijo a base), etc. Estos valores proporcionan los criterios para determinar qué tan morfológico es cada corte *j* de v_i. Luego, se calcularon dos índices de afijalidad para cada corte: AF^p_{ij} y AF^p_{ij}, uno hipotetizando un prefijo y el otro suponiendo un sufijo.

Compárense los cortes posibles de los vocablos en las Tablas 10, 11 y 12. Cabría esperar que la frecuencia alta del sufijo *~mente*, que es un sufijo muy productivo, señalara cortes incorrectos dentro de lexemas como *aument~* y *coment~*, pero los mejores cortes según todos los índices (en negritas) para ambas palabras, *comente* y *aumente*, proponen un sufijo *~e*, marca de subjuntivo de 3ª persona singular de la primera conjugación. En estas tablas tenemos una puntuación perfecta de afijalidad porque, para normalizar los valores, se tomó en cuenta el valor máximo de cada palabra.

Tabla 10. Medidas de segmentación sufijal del vocablo *aumente [auménte]*

	a	**u**	**m**	**e**	**n**	**t**	**e**
cuadros, *c*^{s}		0	1021	20	0	0	**8348**
entropía, *h*^{s}		0	1.046	1.066	0.618	1.351	**2.018**
economía, *k*^{s}		0	0.921	0.55	0	0	**0.925**
***afijalidad, AF*^{s}**		0	0.545	0.375	0.102	0.223	**1.000**

Tabla 11. Medidas de segmentación sufijal del vocablo *comente [koménte]*

	k	**o**	**m**	**e**	**n**	**t**	**e**
cuadros, *c*^{s}		0	0	0	1009	602	**6505**
entropía, *h*^{s}		1.099	1.046	1.066	0.618	1.351	**2.018**
economía, *k*^{s}		0	0	0	0.476	0.94	**0.945**
***afijalidad, AF*^{s}**		0.1815	0.173	0.176	0.322	0.586	**1.000**

Tabla 12. Medidas de segmentación sufijal del vocablo *previamente [prebiaménte]*

	p	**r**	**e**	**b**	**i**	**a**	**m**	**e**	**n**	**t**	**e**
cuadros, c^s	0	0	0	6	**1050**	468	0	0	0	0	
entropía, h^s	0	0	1.609	1.022	**2.233**	1.046	1.066	0.618	1.351	2.018	
economía, k^s	0	0	0	0	0.966	**0.992**	0	0	0	0	
afijalidad, AF^s	0	0	0.24	0.153	**0.991**	0.638	0.159	0.092	0.202	0.301	

Tabla 13. Medidas de segmentación del vocablo *nacionalidad [nasionalidád]*

	n	**a**	**s**	**i**	**o**	**n**	**a**	**l**	**i**	**d**	**a**	**d**
cuadros, c^s	0	0	0	0	0	15	0	**135**	1.000	0	0	
entropía, h^s	0.562	1.792	0	0.628	0.959	**2.133**	0.915	1.782	0.537	0.315	0.863	
economía, k^s	0	0	0	0	0	0.4	0	**0.933**	0	0	0	
afijalidad, AF^s	0.088	0.28	0	0.098	0.15	0.513	0.143	**0.945**	0.084	0.049	0.135	

Nótese además que *~mente* no es el único segmento compitiendo como afijo: el segmento *~te* que se asemeja al pronombre enclítico, que normalmente acompaña gerundios, infinitivos e imperativos, también obtuvo un valor alto en el vocablo *comente*, mayor que 0.5 (esto es 0.586). Además, en las palabras donde sí ocurre el sufijo *~mente* (Tabla 12), otros cortes pueden exhibir una afijalidad más alta (véase *~amente*). Así, aunque *~mente* en *previamente* tiene la medida más alta de economía (0.992), el mejor valor de afijalidad (0.991) favorece la secuencia de sufijos *~a.mente*. Como sea, *~mente* tiene el segundo valor de afijalidad más alto (0.638).

En estos ejemplos se ve que puede ser deseable determinar un umbral para evitar aceptar segmentos con afijalidad demasiado baja. Así que, en la compilación de los catálogos, cada vez que se obtuvo un valor de afijalidad mayor al umbral de 0.5, el segmento en cuestión se insertó en el catálogo apropiado: si era un prefijo, el segmento izquierdo se insertó en Γ^p y, si era un sufijo, se agregó a Γ^s. Si dicho segmento ya estaba presente en el catálogo pertinente, se actualizaron los valores correspondientes a cada índice.

Como se ve en los ejemplos, muchas palabras contienen más de un afijo. De allí que fuera tan común encontrar más de un corte válido en cada palabra. Con respecto a esto, se puede seguir alguna de varias alternativas, según se acepten uno o varios de los cortes:

1) tomar todos los segmentos con afijalidad mayor a 0
2) tomar sólo el segmento con el mejor índice de afijalidad *(~amente* en la Tabla 12 e *~idad* en la Tabla 13) o,
3) tomar todos los segmentos con índices mayores a un valor umbral (al requerir 0.5 de afijalidad, se aceptarían *~amente* y *~mente* en la misma Tabla 12) o,
4) aplicar algún algoritmo; algunos ejemplos son:
    a) tomar el segmento con la afijalidad más alta y, recursivamente, tomar el próximo más alto entre ese segmento y la raíz del vocablo (hacia la izquierda cuando se trate de un sufijo); esto es, seleccionar el mejor, luego ignorar los valo-

res dentro del supuesto afijo (o cadena de afijos) y buscar el siguiente mejor valor en lo que queda del vocablo (en la Tabla 13, tomar el afijo *~idad* y luego *~alidad*),

b) tomar el segmento con la afijalidad más alta y, recursivamente, tomar el próximo más alto dentro de ese segmento, hacia el exterior del vocablo (hacia la derecha cuando se trate de un sufijo, hacia la izquierda cuando se hipotetice un prefijo); esto es, seleccionar el mejor, luego buscar los cortes entre éste y el exterior del vocablo (en la Tabla 12, tomar *~amente* y luego *~mente*),

c) etcétera.

La elección depende del tipo de catálogo que se quiera conseguir. Por ejemplo, para un conjunto de afijos de flexión podrán seleccionarse los afijos (o cadenas de afijos) entre el mejor corte y el exterior del vocablo (algoritmo b). Por otra parte, la primera alternativa (tomar todos los valores mayores a 0) resultaría en un catálogo enorme, el más grande posible, que contendría el número mayor de errores y afijos falsos.

Evidentemente, con este tipo de métodos siempre habrá errores de tipos muy diversos, desde los errores de dedo, en la escritura o transcripción de los documentos que forman parte de la muestra textual, hasta los errores de lógica en las técnicas aplicadas, pasando por los supuestos errores que involucran segmentos cuantitativamente afijales, pero que contravienen las mejores intuiciones de los morfólogos. Nos referiremos a este conjunto de segmentos de carácter afijal dudoso como material residual o simplemente residuos.

En los experimentos presentados en este trabajo, se aplicó la segunda opción: para cada palabra examinada se tomó únicamente el segmento con el valor mayor de afijalidad. Sin embargo, se mantuvieron las cuentas para registrar si los segmentos correspondían a los mejores cortes y si se trataba de las cadenas de caracteres más exteriores o más interiores.

Una vez que fueron examinados todos y cada uno de los cortes posibles de cada vocablo del conjunto V y se insertaron los segmentos

correspondientes en los catálogos Γ^s y Γ^p, se observaron varias cuestiones interesantes. Muchas de las cadenas de afijos que se recolectaron combinan derivación y flexión (*~ador.es*) o flexión verbal y enclítico (*~ar.lo*). Estas secuencias son comunes en español y muchos otros idiomas y, por tanto, no debe sorprender que aparezcan en inventarios de este tipo. Son secuencias de afijos típicas de muchas lenguas y resultan ser más afijales que muchos afijos aislados u que otras secuencias de otros tipos de morfos. Además, si se quiere investigar la afitáctica de la lengua, son estructuras que merecen atención especial.

Véase la Tabla 14, que muestra los primeros 50 sufijos del catálogo Γ^s en el que se reunieron 2 287 formas. Las más afijales contienen un solo morfo y tienen rangos menores; *~ó*, *~o*, *~s*, *~a* (rangos 1, 2, 3 y 4). En general, los segmentos con los promedios más altos de afijalidad también son grupos de afijos muy frecuentes en español (*~o.s*, *~a.s*, *~e.n*, *~a.r*, *~a.do*, etc.).

Los renglones de esta tabla están ordenados según la afijalidad que se despliega en la última columna. La primera columna muestra el rango de ordenamiento. Los segmentos más afijales tienen un rango menor y aparecen al principio de la tabla y los menos afijales tienen rangos mayores y se despliegan hacia el final. La tercera columna exhibe las frecuencias, que no son las frecuencias de aparición del segmento en el corpus. Cada frecuencia se refiere al número de vocablos en que el segmento obtuvo el valor de afijalidad más alto, con respecto a los otros cortes posibles del vocablo examinado. Nótese, además, que dicha frecuencia no necesariamente implica mayor afijalidad. Si bien los afijos más frecuentes tienden a ser los más afijales, ni el más frecuente es el más afijal (*~es* con frecuencia 2 479 tiene rango 36), ni los menos frecuentes ocurren necesariamente al final del catálogo (*~arán* con frecuencia 256 tiene rango 27).

En las siguientes tres columnas se despliegan los promedios normalizados de cuadros, economía y entropía. Luego están las columnas que muestran las probabilidades de los afijos, **prob1** y **prob2**, ya sea en el vocabulario V o en el corpus Ψ.

Tabla 14. Selección de sufijos del español según el CEMC en orden de afijalidad

	sufijo	frecuencia	cuadros	economía	entropía	prob1	prob2	afijalidad
1	~ó	1428	0.7371	0.9192	0.8720	0.8745	0.9003	0.8428
2	~o	6314	0.6860	0.9788	0.8017	0.4695	0.6291	0.8222
3	~s	12013	1.0000	0.9968	0.4609	0.5378	0.5125	0.8192
4	~a	7687	0.5753	0.9818	0.8888	0.5153	0.4431	0.8153
5	~os	4554	0.4775	0.9754	0.8235	0.5162	0.5639	0.7588
6	~as	4324	0.4216	0.9779	0.8645	0.6075	0.5965	0.7547
7	~en	945	0.4107	0.8991	0.9060	0.8630	0.2368	0.7386
8	~ar	1633	0.2178	0.9621	0.9149	0.7346	0.8928	0.6982
9	~ado	1429	0.2061	0.9619	0.9070	0.7099	0.9231	0.6917
10	~ando	976	0.1836	0.9544	0.9162	0.8399	0.9708	0.6847
11	~e	2363	0.4200	0.9482	0.6817	0.2738	0.2295	0.6833
12	~é	639	0.4104	0.8198	0.8153	0.8925	0.4090	0.6818
13	~aba	828	0.1821	0.9565	0.9024	0.8894	0.9564	0.6803
14	~aron	736	0.1779	0.9604	0.8935	0.8943	0.9726	0.6773
15	~ada	1135	0.1654	0.9491	0.9159	0.7385	0.9227	0.6768
16	~arse	665	0.1462	0.9541	0.9072	0.8428	0.9521	0.6692
17	~ados	941	0.1477	0.9549	0.9008	0.7189	0.8582	0.6678
18	~aban	551	0.1434	0.9395	0.9002	0.9062	0.9578	0.6610
19	~adas	813	0.1316	0.9449	0.9041	0.7670	0.8687	0.6602
20	~an	1775	0.1950	0.9434	0.8354	0.6187	0.6729	0.6579
21	~ara	370	0.1098	0.9151	0.9151	0.8916	0.9848	0.6467
22	~ará	387	0.1210	0.9295	0.8739	0.9214	0.9021	0.6415
23	~arlo	316	0.0927	0.9291	0.8849	0.9159	0.9588	0.6356
24	~arla	270	0.0795	0.9185	0.9071	0.9310	0.9650	0.6350
25	~arme	244	0.0868	0.9134	0.8916	0.9313	0.9537	0.6306

Tabla 14 (continuación).
Selección de sufijos del español según el CEMC en orden de afijalidad

	sufijo	frecuencia	cuadros	economía	entropía	prob1	prob2	afijalidad
26	~andose	260	0.0795	0.9136	0.8855	0.8966	0.9067	0.6262
27	~arán	256	0.0900	0.9112	0.8759	0.9242	0.9118	0.6257
28	~ido	445	0.1038	0.8567	0.9140	0.7672	0.8516	0.6248
29	~ita	453	0.0963	0.8965	0.8729	0.7639	0.8077	0.6219
30	~aría	231	0.0806	0.8869	0.8920	0.9240	0.9059	0.6198
31	~amos	645	0.1345	0.8801	0.8415	0.7380	0.8804	0.6187
32	~amente	624	0.1189	0.9784	0.7534	0.8607	0.9793	0.6169
33	~arlos	201	0.0646	0.8959	0.8853	0.9095	0.9235	0.6153
34	~ador	268	0.0549	0.8927	0.8965	0.7768	0.7696	0.6147
35	~aran	196	0.0745	0.8688	0.8857	0.9245	0.9145	0.6097
36	~es	2479	0.1876	0.9529	0.6885	0.4846	0.6193	0.6097
37	~ito	421	0.0932	0.8752	0.8594	0.7360	0.7269	0.6093
38	~aste	136	0.0573	0.8344	0.9237	0.8662	0.8713	0.6051
39	~arte	144	0.0553	0.8446	0.9136	0.9057	0.8802	0.6045
40	~antes	187	0.0365	0.8802	0.8944	0.6404	0.4992	0.6037
41	~adores	196	0.0417	0.8653	0.9029	0.7717	0.8551	0.6033
42	~idos	269	0.0727	0.8321	0.9042	0.7270	0.8084	0.6030
43	~ida	304	0.0831	0.8372	0.8864	0.7221	0.9108	0.6022
44	~arlas	139	0.0535	0.8775	0.8755	0.8910	0.9136	0.6021
45	~asión	540	0.0634	0.9278	0.8111	0.5273	0.8398	0.6007
46	~amiento	141	0.0209	0.8871	0.8935	0.6104	0.8256	0.6005
47	~ante	240	0.0340	0.8769	0.8883	0.6383	0.7293	0.5998
48	~arle	176	0.0657	0.8618	0.8716	0.9514	0.9866	0.5997
49	~asiones	263	0.0455	0.9155	0.8365	0.6726	0.8612	0.5992
50	~aremos	104	0.0420	0.8431	0.9103	0.9369	0.8746	0.5985

Como se dijo, los rangos mayores implican menor afijalidad, la que nos hace dudar sobre el carácter afijal de los segmentos que ocurren hacia el final del catálogo. No están claras las fronteras entre los afijos bien conocidos, afijos dudosos y residuos no morfológicos. De hecho, algunos afijos conocidos aparecen con rangos similares a los de errores evidentes. Por ejemplo, el sufijo de flexión verbal *~áis*, que en México no es productivo, obtuvo el rango 2085 con un índice de afijalidad de 0.3198 (que evidentemente queda fuera de la tabla) y aparece muy cerca de segmentos que no parecen sufijos, como *~gún* de *algún* o *ningún*.

Las frecuencias registradas permiten el cálculo de las probabilidades descritas en la sección 2.5.2. Su relación con las afijalidades de los segmentos es interesante: los valores mayores para cualquiera de los tipos de probabilidad no necesariamente implican un mayor índice de afijalidad y los valores menores no corresponden a índices menores. Véanse las columnas **prob1** y **prob2** de la Tabla 14. Las probabilidades más altas implican mayor certidumbre de afijalidad, ya sea en el vocabulario *V* o en el corpus, pero no mayor afijalidad.

Por ejemplo, considérese el sufijo *~a*, que obtuvo un alto índice de afijalidad (0.8153), pero bajas probabilidades (0.5153 y 0.4431 en el renglón 4 de la Tabla 14). La letra 'a' ocurre al final de sufijos como *~aba*, *~iera* o *~ería* en palabras como *compraba*, *sufriera* y *tortillería*. Así que esa letra en ese lugar no representa un afijo. Por eso, las posibilidades de que *~a* aparezca como sufijo resultan menores que aquellas que tienen los sufijos como *~aba* y en general los sufijos largos como *~mente* o cadenas sufijales como *~ándoselas*.

Considérese también el sufijo *~ó*, que obtuvo probabilidades más bien altas (0.8745 y 0.9003). Esto significa que al encontrarnos casualmente con la palabra gráfica *buscó*, podemos casi afirmar que *~ó* es un sufijo, con mucha más seguridad que si nos encontráramos con *busca* y esperáramos que *~a* fuera también un afijo. Así que, al observar una cadena de ocurrencias de palabras en contexto, se pueden utilizar estas probabilidades como criterios de confianza en la determinación de morfemas.

Como se mencionó, mientras más largos son los segmentos, mayor es la probabilidad de que sean afijos o una secuencia de ellos. Consecuentemente, los sufijos aislados de flexión, segmentos más cortos, tienden a tener probabilidades menores. Además, obsérvense las diferencias entre los dos tipos de probabilidad. Mientras mayor es la segunda, **prob2**, con respecto a la primera, **prob1**, más frecuentes son las formas en las que el afijo aparece. Esto significa que el afijo es muy frecuente en el corpus, por ejemplo, *~ado*, *~ada*, *~asión*, *~asiones*, *~amiento*. Por otra parte, si la **prob1** es mayor que **prob2**, el afijo en cuestión tiende a ocurrir en palabras de baja frecuencia. Así, los pronombres enclíticos que aparecen en gerundios, infinitivos e imperativos pertenecen a este grupo: *~lo*, *~la*, *~se*, *~los*, *~las*, *~le*, *~me* y *~nos*.

Al principio de esta tabla aparece la mayoría de los sufijos de flexión verbal. Muchos incluyen la vocal temática [a] de la primera conjugación, que es la más numerosa. Algunas de las cadenas de afijos terminan con algún pronombre enclítico (*~te*, *~se*, *~la*, etc.). También aparecen algunos sufijos exclusivamente de derivación (*~asión*, *~ador*) y a veces ocurren también sus formas en plural (*~antes* de *~ante*; *~asiones* de *~asión*).

Como se comentó, muchos otros segmentos afijales contienen también más de un afijo (*~ura.s*, *~idad.es*, *~a.r.á.s*, *~aba.s*, etc.). Su presencia en el catálogo implica que obtuvieron los valores más altos de afijalidad en las palabras examinadas y que los valores de afijalidad de los segmentos interiores fueron menores.

Algunas formas se distinguen entre sí solamente por el acento gráfico (*~ará*, 3ª persona singular del futuro; *~ara*, 1ª y 3ª personas singulares del subjuntivo pasado). Estas distinciones se habrían perdido si no se hubieran conservado las tildes de las últimas sílabas. Además, como se ve, la ausencia de acentos gráficos en otras sílabas no causó desajustes en este experimento: *~andose* no se distingue de *-ándose*.

En general, el catálogo de sufijos es una lista relativamente limpia de formas reconocibles donde los índices más bajos de afijalidad corresponden a formas poco productivas, a casos dudosos y a segmentos que no pueden considerarse sufijos tomando en cuenta criterios cualitativos.

Tabla 15. Selección de prefijos del español según el CEMC en orden de afijalidad

	prefijo	frecuencia	cuadros	economía	entropía	prob1	prob2	afijalidad
1	a~	2074	0.0519	0.9511	0.9630	0.2175	0.2454	0.6553
2	r̆e~	1866	0.0773	0.9639	0.9108	0.4999	0.6156	0.6507
3	semi~	63	0.0048	0.9430	0.9617	0.7000	0.4078	0.6365
4	des~	1041	0.0577	0.9700	0.8470	0.4813	0.7355	0.6249
5	auto~	82	0.0145	0.9076	0.9144	0.4767	0.5803	0.6122
6	pro~	481	0.0209	0.9166	0.8892	0.4546	0.7168	0.6089
7	kontra~	92	0.0157	0.8883	0.9219	0.4532	0.1738	0.6086
8	anti~	66	0.0042	0.9108	0.8976	0.4125	0.1791	0.6042
9	in~	1066	0.0345	0.9551	0.8176	0.3731	0.4243	0.6024
10	sub~	198	0.0101	0.8532	0.9334	0.6187	0.6009	0.5989
11	r̆adio~	31	0.0212	0.8803	0.8728	0.7209	0.3871	0.5914
12	porta~	16	0.0462	0.9062	0.8216	0.3556	0.2830	0.5913
13	sali~	15	0.0619	0.9171	0.7908	0.3488	0.6266	0.5899
14	intra~	21	0.0018	0.9113	0.8444	0.4286	0.5000	0.5858
15	sobre~	126	0.0169	0.9298	0.8063	0.8182	0.1228	0.5844
16	trai~	17	0.1323	0.9348	0.6688	0.3469	0.3522	0.5786
17	idro~	32	0.0025	0.9122	0.8157	0.5246	0.6757	0.5768
18	pre~	425	0.0210	0.9115	0.7917	0.4516	0.4459	0.5748
19	per~	239	0.0133	0.8465	0.8512	0.3469	0.2619	0.5703
20	r̆i~	44	0.0078	0.7299	0.9549	0.2178	0.2388	0.5642
21	inter~	139	0.0100	0.8475	0.8276	0.5055	0.4259	0.5617
22	kon~	866	0.0295	0.9429	0.7018	0.3991	0.3115	0.5581
23	electro~	21	0.0130	0.8139	0.8398	0.4468	0.6010	0.5556
24	mono~	30	0.0215	0.8125	0.8317	0.5085	0.5187	0.5552
25	dis~	217	0.0113	0.8777	0.7739	0.3344	0.2682	0.5543

Por otra parte, el catálogo de prefijos, Γ^{p}, reunió más de 3566 segmentos, muchos más que el de sufijos. Para evitar falsos prefijos en la Tabla 15, se eliminaron aquellos con un valor de afijalidad menor a 0.45.

Tabla 15 (continuación).
Selección de prefijos del español según el CEMC en orden de afijalidad

	prefijo	frecuencia	cuadros	economía	entropía	prob1	prob2	afijalidad
26	super~	43	0.0046	0.7956	0.8605	0.3333	0.1255	0.5536
27	pi~	165	0.0187	0.7508	0.8870	0.3167	0.2764	0.5522
28	neo~	16	0.0010	0.8473	0.8064	0.4706	0.5000	0.5516
29	psiko~	20	0.0015	0.9270	0.7183	0.5128	0.5209	0.5490
30	mikro~	27	0.0018	0.8225	0.8189	0.7105	0.7653	0.5477
31	jeo~	24	0.0014	0.9156	0.6920	0.6316	0.7895	0.5363
32	laba~	15	0.0557	0.8490	0.6858	0.3191	0.3232	0.5302
33	ʧi~	80	0.0171	0.7360	0.8308	0.3137	0.5873	0.5280
34	foto~	20	0.0311	0.8615	0.6772	0.5128	0.5219	0.5233
35	poli~	29	0.0017	0.7397	0.8176	0.3053	0.1858	0.5196
36	trans~	148	0.0080	0.7800	0.7639	0.6435	0.6991	0.5173
37	tene~	16	0.0185	0.9063	0.6147	0.5000	0.4771	0.5132
38	ɾete~	15	0.0016	0.8626	0.6711	0.3846	0.1633	0.5118
39	deja~	20	0.0607	0.8517	0.6206	0.3077	0.1728	0.5110
40	media~	17	0.0514	0.8208	0.6560	0.5000	0.1437	0.5094
41	ante~	26	0.0361	0.7771	0.7003	0.4194	0.4858	0.5045
42	tras~	57	0.0065	0.7230	0.7817	0.3373	0.2800	0.5038
43	tele~	33	0.0047	0.7202	0.7831	0.4714	0.8265	0.5027
44	bio~	29	0.0089	0.9022	0.5937	0.3053	0.1541	0.5016
45	ex~	417	0.0125	0.8421	0.6433	0.3949	0.4459	0.4993
46	ob~	122	0.0055	0.7562	0.7120	0.3333	0.5117	0.4912
47	ɾetro~	16	0.0025	0.8491	0.5994	0.4103	0.2683	0.4836
48	krea~	19	0.0201	0.7444	0.6709	0.4872	0.3019	0.4784
49	multi~	26	0.0018	0.8323	0.6002	0.4815	0.1934	0.4781
50	eki~	34	0.0023	0.8385	0.5672	0.3953	0.6393	0.4693

También se requirió que exhibieran la mejor afijalidad en por lo menos 15 de los vocablos en que ocurren y que fueran el prefijo más lejano de la base en por lo menos 10 vocablos. Además, se incluyeron sólo aquellos con una probabilidad de ocurrir como prefijos en los vocablos de por lo menos 30% (**prob1**), y en la cadena de palabras Ψ (el corpus) de al menos 10% (**prob2**). A todos se les requirió un índice normalizado de economía de por lo menos 0.7 y de entropía (también normalizada) de 0.5. Por último, todas las formas ocurren más de 10 veces. De todas maneras y con todas estas restricciones, aparecen formas inciertas que permanecieron por no poderse filtrar automáticamente. Es evidente que el nivel de ruido entre los prefijos es más alto que aquél que se observa entre los sufijos.

Esto tal vez se pueda explicar por la enorme cantidad de préstamos de otras lenguas y las muchas y variadas abreviaturas (siglas, símbolos químicos, etc.) que ocurren en el CEMC. Esto mismo puede explicar los aparentes errores dentro los mejores 50 prefijos de la Tabla 15, como r̄ i~ (rango 20) y ʧi~ (rango 33) de vocablos como *rico*, *risa*, *chico*, *chiva*, etc. Vale la pena notar que estos supuestos prefijos tienen valores de cuadros muy bajos. De hecho, el número de cuadros es el índice más bajo en todas las formas de la tabla.

Es interesante que muchos segmentos no sean propiamente prefijos del español: además de los pseudoprefijos tradicionales (*electro~*, *psiko~*, *mikro~*, etc.), hay verbos conjugados en presente de la 3ª persona singular típicos de composiciones, como *laba~* (rango 32), *krea~* (rango 48) y muchos otros distribuidos por todo el catálogo. Cabe señalar que estos últimos no habrían sido aceptados en el catálogo si se hubiera evitado la entrada de segmentos que pudieran tener en el interior una frontera de sufijo. De esta manera, *laba~* de *lavamanos* y *lavaplatos* no habría sido seleccionado como prefijo, porque *~a* se puede identificar como sufijo en *laba~*, mientras que, con este método, *~amanos* y *~aplatos* podrían haber sido considerados sufijos —aunque después de todo no merezcan aparecer entre los sufijos, porque ambos son bases y *lavamanos* y *lavaplatos* son ejemplos de composición.

Por otra parte, estos catálogos tal vez ilustren las diferencias entre sufijos y prefijos en la lengua española. Ya se ha observado que en español los sufijos poseen una mayor *capacidad* gramatical que los prefijos. Además, el hecho de que varios prefijos tengan la misma forma que algunas preposiciones (*con~*, *en~*, *a~*, etc.) tal vez acerque la prefijación del español a la composición. Sin embargo, la sufijación y la prefijación siguen siendo más parecidas entre sí que entre esta última y la composición[17].

Greenberg caracteriza la diferencia entre sufijos y prefijos mediante la tendencia del hablante a anticipar los sonidos (1967 [1957], 90-93). Los prefijos tienden a fundirse con la base porque el hablante tiende a anticipar todo aquello que contenga su mensaje, mientras que los sufijos tienden a permanecer relativamente estables porque el hablante matiza con ellos lo ya dicho. Así se explica que los prefijos y las bases se fundan rápidamente —mediante los consiguientes cambios fonológicos—, ocasionando que las bases tiendan a desarrollar muchas más irregularidades que los sufijos, que permanecen relativamente estables.

Esto también está relacionado con el hecho de que, por lo menos en español, tenemos un sistema de sufijos compacto y muy organizado que codifica información léxica y morfosintáctica, mientras que los prefijos no cargan información de tipo gramatical, sino de contenido. De hecho, la entropía al inicio de las palabras (2.55716 bits) es considerablemente más alta que la entropía con que empiezan al revés (1.83899 bits), lo que significa que es menos predecible un prefijo que un sufijo, al menos en el CEMC. Así que en lenguas que no dependan de sus estructuras morfológicas para codificar la información gramatical —o que simplemente no cuenten con un sistema de sufijos— un intento de descubrir sufijos con este método probablemente resultará en un catálogo parecido al de los prefijos del español: voluminoso y poblado de segmentos dudosos.

[17] Véase Moreno de Alba, *La prefijación en el español mexicano*, UNAM, México (1996, 15-17).

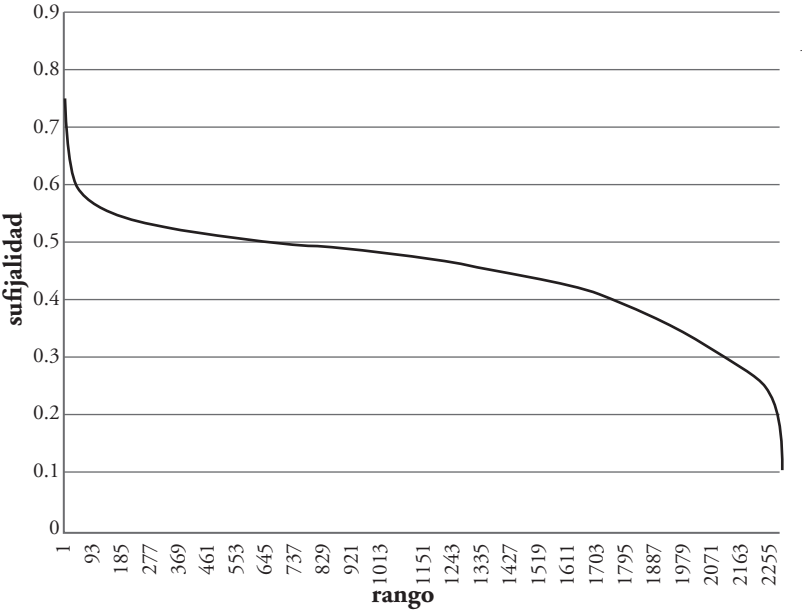


Figura 7. Distribución de los valores de afijalidad (sufijalidad) de todos los segmentos recogidos en el catálogo de sufijos del CEMC

En la curva de la Figura 7 podemos ver los valores de afijalidad de todos los segmentos acumulados en el catálogo Γ^s en orden de mayor a menor, incluidos los segmentos no afijales que, como se dijo arriba, tienen valores menores y, por lo tanto, aparecen hacia la derecha de la curva. De manera similar, la Figura 8 muestra la curva de valores de afijalidad de los prefijos también en orden de rango. Como puede verse, la diferencia principal está en el número de segmentos. A ambos grupos de datos se les aplicó el ajustador de Altmann (*Altmann-Fitter*)[18] y se observó que ambos se ajustan a la distribución hipergeométrica negativa.

[18] *Iterative Fitting of Probability Distributions*, RAM-Verlag, Lüdenscheid (2020 [1997]), que se encuentra disponible en https://www.ram-verlag.eu/software-neu/software/.

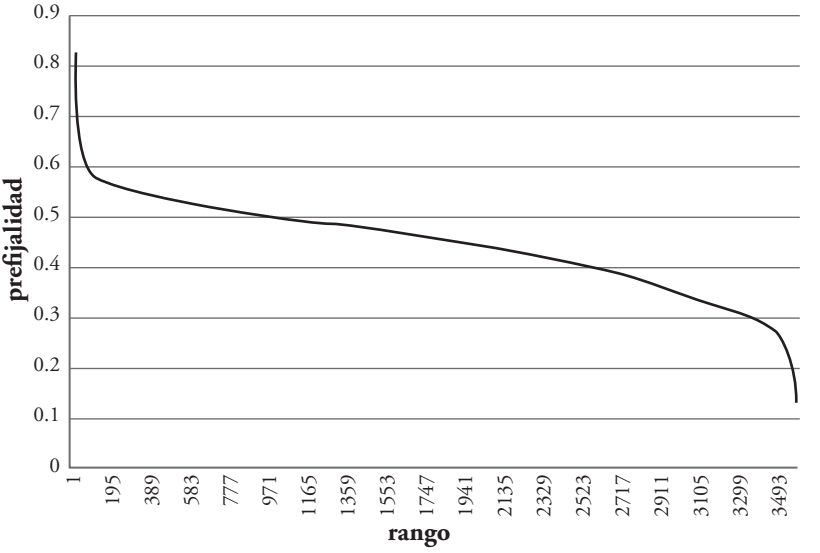


Figura 8. Distribución de los valores de afijalidad (prefijalidad) de todos los segmentos recogidos en el catálogo de prefijos del CEMC

Como se ve, la construcción de estos catálogos es flexible. Según las necesidades de investigación, pueden variar los criterios de inclusión de segmentos afijales y se pueden aplicar diferentes tipos de filtros para eliminar ciertos tipos de segmentos. En el siguiente capítulo, se revisarán los resultados de otro experimento de extracción de afijos a partir del CEMC. Específicamente, se examinarán los 749 segmentos sufijales para determinar si corresponden a los sufijos de flexión y derivación conocidos del español.

2.8 Hacia la evaluación del cálculo de la afijalidad

En este capítulo, se describió un camino para la construcción de catálogos de afijos para lenguas como la española. Los miembros de este tipo de catálogos son afijos porque exhiben ciertas propiedades lingüístico-formales que se han asociado al concepto de afijo, no porque

alguien en particular haya decidido que deberían ser tratados como afijos.

Específicamente, los afijos son pocos, pero muy frecuentes y se adhieren a numerosos segmentos de baja frecuencia (bases) para formar muchos signos del nivel léxico, satisfaciéndose así la necesidad de un inventario económico de signos. Los signos léxicos se relacionan unos con otros porque comparten afijos. De hecho, un afijo se combina con muchas bases que también se combinan con otros afijos para formar todavía más signos léxicos. También, cada afijo participa en un número significativo de cuadros. Finalmente, los afijos contienen menos información que los lexemas o bases. Por eso, la entropía de los contextos a los que se adhieren es mayor.

Sin duda, el método presentado puede mejorarse. Por ejemplo, valdría la pena explorar las maneras posibles de combinar los valores de cuadros, economía y entropía. De hecho, se pudo ver que la entropía es un buen marcador de fronteras entre las raíces y las secuencias de afijos, mientras que las medidas de economía funcionan muy bien para separar los afijos flexivos de las bases. Por otra parte, valdría la pena investigar otras técnicas de segmentación basadas en las características de lo que llamamos afijos. Finalmente, sería necesario examinar la naturaleza de los errores en este tipo de experimentos: ¿cómo se deben tratar?, ¿qué es propiamente un error?

Por lo pronto, podemos decir que se pudo determinar un conjunto de signos prefijales de la lengua española y otro de signos sufijales, muchos de los cuales tienen funciones en el nivel morfosintáctico. Como se verá adelante, este método se puede aplicar a corpus y muestras textuales de otras lenguas. La presencia en ellas de este tipo de signos merece ser estudiada con éstas y otras herramientas que puedan concebirse para describir la unidad lingüística llamada afijo.

En este contexto, es necesario explorar las maneras en que este método de cálculo de afijalidad se puede aplicar para investigar los afijos de otras lenguas y para evaluar su utilidad en otros proyectos de investigación. En particular, sus resultados se pueden comparar con los de otros

métodos de segmentación de palabras. También, se puede comprobar su aplicabilidad en el desarrollo de otros programas y se pueden examinar sus resultados en diferentes lenguas para determinar su valor e integridad.

Al final de este libro, en el único apéndice, se presentan algunas rutinas en Python, un lenguaje de programación muy popular, para hacer operativos los cálculos de entropía, economía y número de cuadros.

CAPÍTULO 3
APLICACIONES DEL DESCUBRIMIENTO DE AFIJOS

En los campos del procesamiento del lenguaje natural, la lingüística computacional, la lingüística de corpus y las tecnologías del lenguaje, los sistemas que se desarrollan deben evaluarse para determinar el grado de sus logros y comparar sus resultados con los de otros sistemas y experimentos. Hay varias maneras de hacerlo: por ejemplo, comparando los resultados de varios métodos e implementaciones, comprobando su utilidad en el desarrollo de otros programas, examinando sus resultados para determinar su valor e integridad, etcétera.

En particular, los métodos de segmentación morfológica también pueden ser evaluados de diversas maneras. Por ejemplo, se pueden comparar unos con otros para determinar cuáles tienen mejores resultados. Al respecto, Méndez Cruz *et al.* (2016) diseñó, basándose en mediciones de afijalidad, una serie de experimentos para segmentar morfológicamente el español de México y los comparó con los resultados de los métodos de Morfessor y ParaMor[1]. Los experimentos de Méndez muestran cómo y en qué aspectos y condiciones las mediciones de afijalidad presentan mejores resultados que estos métodos de segmentación.

Otra manera de evaluar estas técnicas es observar su utilidad en el desarrollo de tecnologías del lenguaje y programas de cómputo empleados en el tratamiento de corpus electrónicos, como lematizadores y lexematizadores (*lemmatizers* o *stemmers*) y etiquetadores gramaticales o de categorías gramaticales (*Part-Of-Speech or POS taggers*) y de aplicaciones de utilidad potencial a los hablantes de lenguas de bajos recursos, como

[1] Los resultados se publicaron en Méndez-Cruz, Medina-Urrea y Sierra, "Unsupervised morphological segmentation based on affixality measurements", *Pattern Recognition Letters*, 84 (2016, 127-133).

sintetizadores de habla o de voz (*speech synthesizers*). En este capítulo, nos ocuparemos de estos desarrollos en la primera sección.

Una manera más de valorar estos métodos es examinar sus resultados para apreciar cuánto de los paradigmas de flexión y derivación de una lengua se logran descubrir en un corpus. En la segunda sección de este capítulo, revisaremos los sufijos y cadenas de sufijos del español, extraídos a partir del CEMC, para corroborar que en efecto se descubrió la mayoría de los sufijos derivativos y flexivos conocidos de esta lengua.

Por otra parte, también se pueden aplicar estos métodos a corpus y muestras textuales de otras lenguas de morfología concatenativa para corroborar que se detectan sus sistemas afijales. Así como se pueden determinar los afijos y cadenas de afijos de una lengua como la española, a partir de un corpus, mediante la medición de los rasgos que se han considerado característicos de los afijos en las lenguas del mundo, es interesante que estos mismos criterios sirvan para llevar a cabo experimentos de descubrimiento de afijos en otras lenguas. Este es el tema del tercer apartado del presente capítulo.

En resumen, a continuación se presentan y comentan varios desarrollos y experimentos en los que la compilación automática de catálogos de afijos ha demostrado su utilidad. Luego, examinaremos con detalle los sufijos del español, que se presentaron en el capítulo anterior, y mostraremos los resultados de experimentos de extracción de afijos del checo, el ralámuli y el chuj. Por último, se presentará un ejercicio de evaluación en términos de las medidas de precisión y *recall* (recuperación comprensiva o exhaustividad), que son técnicas estándar de evaluación en los campos de recuperación de información y minería de textos.

3.1 Algunos desarrollos basados en la extracción de afijos

En este apartado, se presentan algunos usos de la segmentación morfológica automática no supervisada en el desarrollo de programas de

cómputo. Primero, se muestra una posible aplicación para la construcción de programas que *lexematizan* palabras flexionadas, esto es, que las desnudan de sus afijos de flexión y posiblemente de derivación. Luego, se ejemplifica su posible aplicación en programas que generan reglas para aplicar etiquetas de categorías gramaticales a las palabras gráficas de un corpus. Por último, se presenta un uso posible en el desarrollo de sintetizadores de voz.

3.1.1 Lematización y lexematización automáticas

Muchas herramientas para la explotación y el análisis de los corpus textuales requieren un procedimiento de lematización, que a menudo se reduce al truncamiento automático de las palabras gráficas para eliminar flexiones. Típicamente, se aplican técnicas sencillas, como el algoritmo de Porter, que se describió en la sección 1.2.5 y que se ha implementado para muchas lenguas esencialmente europeas. Como se vio, la desventaja de un método como el de Porter es que requiere del conocimiento *a priori* de la morfología.

En particular, muchos desarrollos de carácter lingüístico deben tratar directa o indirectamente con fenómenos morfológicos. De allí que en los sistemas de extracción y recuperación de información, de minería de textos, de procesamiento de diccionarios electrónicos, etc. sean populares los lematizadores (*lemmatizers*), lexematizadores (*stemmers*), analizadores morfológicos, etc. Si han de desarrollarse aplicaciones como éstas para lenguas que cuenten con pocos recursos electrónicos, que son la mayoría de las lenguas de México y del mundo (muchas de las cuales exhiben morfologías muy complejas), el descubrimiento no supervisado de morfemas será una etapa importante de su desarrollo.

Como sabemos, el lema o forma canónica de una palabra se usa para consultar su significado en los diccionarios. Normalmente, nos encontramos las palabras en sus contextos y, para buscarlas en el diccio-

nario, debemos determinar sus lemas. Esto requiere la eliminación de los morfemas de flexión que puedan tener en sus contextos de uso. Por ejemplo, los lemas de las palabras flexionadas *olvidaríamos* y *pequeñitas* son *olvidar* y *pequeño* y son las formas que usamos para encontrar sus entradas en los diccionarios.

En disciplinas como el procesamiento del lenguaje natural y la lingüística de corpus, la lematización es un procedimiento automático que busca asignarle a cada palabra gráfica una forma única sin flexiones, esto es, busca desnudarla de sus afijos para quedarse con la base o lexema[2]. En la lengua inglesa, suele ser que al eliminar los afijos en efecto nos quedemos con los lemas (*contacted* → *contact*, *hearts* → *heart*), lo que no es el caso en español (*olvidaríamos* → *olvid*, y *pequeñitas* → *pequeñ*), donde más bien llegamos al lexema, que no suele ocurrir libre en el discurso y que no corresponde con la entrada canónica del diccionario. Así que en lugar de referirnos a verdaderos lematizadores, aquí hablaremos de *lexematizadores* que en esencia separan el lexema o la raíz de los afijos.

Para esto, el algoritmo de Porter sigue siendo muy popular. Sin embargo, hay que repetirlo: para poder implementarlo se requiere del conocimiento previo de la morfología sufijal de la lengua a la que se le quiere aplicar. Así que, para desarrollar un lexematizador basado en este algoritmo, hay que conocer la lengua *a priori*. La lengua española ya cuenta con varias implementaciones de este método[3].

Sin embargo, si nos enfrentamos con la tarea de encontrar los lexemas de un corpus del español del siglo XVI, podemos dudar de los resultados de aplicar una implementación desarrollada para el siglo XXI. De allí que tenga sentido determinar los sufijos y grupos sufijales a partir de un corpus del siglo XVI para construir un lexematizador específicamente diseñado para ese estado de lengua.

[2] Véanse, por ejemplo, Jurafsky y Martin, *op. cit.* (2009, 611), McEnery y Wilson, *op. cit.* (2001, 53), Manning y Schütze, *op. cit.* (1999, 132).

[3] Algunas implementaciones del algoritmo de Porter para varias lenguas europeas se encuentran en http://snowballstem.org/.

Al respecto, dos de los problemas que surgen al desarrollar un lexematizador para estados de lengua poco documentados son: 1) los sistemas morfológicos de las lenguas pueden variar a lo largo del tiempo y, típicamente, estos cambios no están suficientemente documentados; y 2) un mismo idioma puede presentar características ortográficas diferentes en cada estado de lengua.

En primer lugar, la fonología española sufrió algunos cambios alrededor del siglo XVI, por lo que se requieren reglas de transcripción específicas para esa época. Algunas reglas han sido propuestas para la transcripción automática de los documentos del siglo XVI, pero todavía son provisionales y el consenso parece estar lejos de ser alcanzado. Por otra parte, las irregularidades ortográficas de los documentos antiguos hacen que las transcripciones fonológicas automáticas tengan muchos errores. De allí que este experimento sea básicamente un ejercicio de lengua escrita.

En comparación con otras lenguas, la morfología española ha cambiado relativamente poco durante los últimos cinco siglos. Así que se puede suponer que un lexematizador de tipo Porter para el español de hoy podría ser aplicado a esos siglos con el fin de eliminar la flexión. Sin embargo, dado que existen técnicas para descubrir afijos de manera no supervisada que pueden utilizarse para la lexematización, conviene comparar los resultados de una lexematización basada en medidas de afijalidad con los de una implementación del algoritmo de Porter (del siglo XXI).

Eso es lo que se hizo en el experimento reportado en Medina Urrea (2006)[4]. Específicamente, se compararon los resultados de segmentar, mediante medidas de afijalidad, una muestra textual del siglo XVI para obtener los lexemas, con los resultados de aplicarle a esa muestra una

[4] "Towards the Automatic Lemmatization of 16th Century Mexican Spanish: A Stemming Scheme for the CHEM", *Lecture Notes in Computer Science*, 3878 (Medina Urrea 2006, 101-104).

versión del algoritmo de Porter para el español contemporáneo[5]. El lexematizador basado en afijalidad simplemente elimina los supuestos sufijos.

El corpus meta para el experimento estuvo constituido por 95 documentos[6] del siglo XVI. Estos documentos comprenden alrededor de 257 385 palabras gráficas, que corresponden aproximadamente a 15 834 vocablos. No se tomaron en cuenta los nombres propios, que suelen ocurrir en mayúsculas. Tampoco se segmentaron las palabras con menos de cuatro caracteres.

Al examinar estos documentos, se pueden observar algunas idiosincrasias ortográficas del siglo XVI que hacen que muchos referentes tengan varias formas gráficas (por ejemplo, *admynistración, admynystracjon, administraçjon, adminystracjón*, etc.). Así que el texto debe normalizarse para reducir el número de palabras gráficas, esto es, formas ortográficas que se refieren a lo mismo (*merced, merçed* → *merced*; *cantava, cantaba* → *cantaba*; *yndio, jndio* → *indio*, etc.), aunque se pueda causar cierta homofonía entre palabras cortas. Esto se logra con la aplicación de un conjunto de reglas para modificar algunos caracteres y mejorar la correspondencia grafema-fonema[7]. Estas reglas se muestran en la Tabla 16.

[5] Para este experimento, se utilizó una versión para el español contemporáneo que se desarrolló en el Grupo de Ingeniería Lingüística del Instituto de Ingeniería de la UNAM (http://grupos.iingen.unam.mx/iling/es-mx/Paginas/default.aspx).

[6] Se trata de muestras textuales de los *Documentos Lingüísticos de la Nueva España. Altiplano Central* (Company Company 1994), de los *Indígenas de la Inquisici*ón Apostólica de fray Juan de Zumárraga (Buelna Serrano 2009) y del *El habla de Diego de Ordaz. Contribución a la historia del español americano* (Lope Blanch 1985).

[7] Para la formulación de reglas de reescritura del siglo XVI, se pueden consultar trabajos muy variados, que van de obras generales y abarcadoras, como la de Lara, *Historia mínima de la lengua española*, El Colegio de México (2013), a tesis académicas muy específicas, como la de Reyes Careaga, *Reglas de correspondencia entre sonido y grafía en el español hablado en México en el siglo XVI: Una aportación al Corpus histórico del español en México*, UNAM, Tesis de licenciatura, México (2008).

Tabla 16. Reglas de reescritura de caracteres para reflejar las correspondencias entre grafemas y fonemas en textos en español del siglo XVI

reglas	fonema	contextos
'h' → ε	-	todos
'v' → 'b'	/b/	todos
'ch' → 'ʧ'	/ʧ/	todos
'rr' → 'r̄'	/r̄/	todos
'ç', 'c', 'z' → 's'[a]	/s/	'çe', 'çi'; 'ce', 'ci'; en toda 'z'
'c', 'qu' → 'k'	/k/	'ca', 'que', 'qui', 'co', 'cu'
'g' → 'j'	/x/	'ge', 'gi'
'gu' → 'g'	/g/	'gue', 'gui'
'y' → 'i'	/i/	fin de sílaba, después de vocal ('ay', 'ey', 'oy', 'uy'); o al inicio de palabra, antes de consonante ('yn', 'yd', etc.)
'j' → 'i'	/i/	entre consonantes o entre consonante y vocal; o al inicio de palabra antes de consonante ('jn', 'jd', etc.)
'r' → 'r̄'	/r̄/	inicio de palabra; o después de sílaba que termina en 'n', 'l' o 's'

[a]En el siglo XVI, el sistema de estridentes del español se colapsaba. Así que el fono [θ] probablemente nunca existió como fonema en América (J. W. Harris 1969, 189-206).

Afortunadamente, cuando se examinan las muestras textuales y los conjuntos de sufijos extraídos, es posible observar que la gran variabilidad ortográfica en los documentos antiguos se produce principalmente en las raíces y las bases de las palabras gráficas. Además, como los documentos fueron editados siguiendo la costumbre de los filólogos de reconstruir palabras gráficas abreviadas, cuando reconstruyeron afijos y cadenas de afijos, lo hicieron siguiendo criterios estándar, limitando la variabilidad ortográfica de los afijos en las abreviaturas reconstruidas.

Para truncar las palabras el lexematizador determina las secuencias de sufijos a partir de la muestra textual mediante, como se dijo, el método el cálculo de afijalidades. Sin la aplicación de las reglas de correspondencia grafema-fonema de la Tabla 16, el método produjo 565 cadenas sufijales. Con la aplicación de estas reglas se obtuvieron 487. Finalmente, al eliminar los acentos gráficos (excepto los del final de las palabras que distinguen a morfemas de flexión verbal), se obtuvieron 470 sufijos. Luego, las listas de secuencias de sufijos descubiertas de esta manera se utilizaron en el experimento de lexematización, el cual, como se mencionó, consistió en eliminar del final de cada palabra de la muestra textual el sufijo o cadena de sufijos de esa lista.

Para evaluar los resultados, se tomó al azar uno de los documentos del siglo XVI, de 12 424 palabras gráficas, al que se le aplicaron las dos lexematizaciones. Luego, por inspección, se evaluó si las segmentaciones propuestas coincidían con cortes morfológicos correctos. El porcentaje de aciertos del lexematizador basado en medidas de afijalidad fue de 99.32%. En cambio, el segmentador de Porter del siglo XXI obtuvo el 93.28%, que mejoró a 95.97% al contar los aciertos por los tipos de palabra (en el conjunto *V*), en lugar de sus ocurrencias gráficas (en el corpus).

Es interesante que una implementación del algoritmo de Porter para el español contemporáneo obtenga una calificación tan alta al aplicarla a un documento de un estado más temprano de la lengua. Esto corrobora la observación de que la morfología del español ha cambiado relativamente poco en los últimos siglos. En cualquier caso, parece mejor aplicar un método no supervisado de descubrimiento de sufijos para lexematizar un corpus, que desarrollar un lematizador de Porter específicamente diseñado para el siglo XVI.

Como sea, uno debe preguntarse, ¿hasta qué punto se puede aplicar un método desarrollado, para una etapa de una lengua, a estados anteriores, o posteriores, de esa lengua? Evidentemente, la respuesta depende del idioma meta. Por ejemplo, aquellas lenguas relativamente más innovadoras, como el francés o el inglés, sufrieron más cambios en menos tiempo. Así que, seguramente, un lematizador tipo Porter

basado en sus estados actuales será menos adecuado de lo que parece ser para el español.

3.1.2 Etiquetado de categorías gramaticales basado en la transformación de reglas

Para el análisis sintáctico y morfosintáctico automático de lenguas como la española, tiene sentido tomar en cuenta la morfología sufijal. Un tipo de análisis muy necesario en lingüística computacional y lingüística de corpus es el que sirve para etiquetar cada palabra gráfica de una muestra textual con su categoría gramatical (*POS tagging*)[8]. Para esto, las técnicas de estadísticas de n-gramas son métodos muy populares; véase Allen (1995). Aunque son efectivas, generalmente requieren de corpus de millones de palabras gráficas que permitan calcular las probabilidades de las etiquetas que se busca aplicar a las palabras de los corpus.

Otro de los métodos de etiquetado más conocidos es el de Eric Brill[9] y se puede entrenar con textos previamente etiquetados de mucho menor tamaño. Su implementación cuenta con dos módulos para tratar separadamente la información sintáctica de la morfológica. En

[8] Para las generalidades del etiquetado de categorías gramaticales, se pueden consultar Jurafsky y Martin, *op. cit.* (2009, 123-163) y Manning y Schütze, *op. cit.* (1999, 341-377). Hoy en día, existen muchos métodos de etiquetado de categorías gramaticales para el español; véanse, por ejemplo, Jiménez y Morales, "SEPE: A POS Tagger for Spanish", *Lecture Notes in Computer Science* 2276 (2002, 250-259) y Morales Carrasco y Gelbukh, "Evaluation of TnT Tagger for Spanish", *Proceedings of the Fourth Mexican International Conference on Computer Science*, ENC 2003, Tlaxcala (2003, 18-25).

[9] Véanse Brill, "A simple Rule-Based Part of Speech Tagger", ACL, Trento (1992), *A Corpus-Based Approach to Language Learning*. Tesis doctoral, Universidad de Pensilvania, Filadelfia (1993), "Some Advances in Transformation-Based Part of Speech Tagging", AAAI (1994), "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging", *Computational Linguistics*, 21:4 (1995). Para una introducción al método de Brill véase Jurafsky y Martin, *op. cit.* (2009, 151-153). En el Laboratorio de Lenguaje Natural del Centro de Investigación en Computación del Instituto Politécnico Nacional se entrenó un etiquetador Brill para el español ([2004] 2007).

esencia, este método capta información gramatical y la codifica automáticamente en reglas, sin intervención humana. Si bien las reglas no son propiamente lingüísticas, sí representan información sobre la secuencia de las categorías gramaticales en el discurso. Como sea, en el módulo morfológico se puede introducir una lista de afijos, como los descubiertos mediante el cálculo de afijalidad, para observar si el método tiene o no mejores resultados.

Carlos Méndez Cruz[10] diseñó una serie de experimentos basados en una pequeña muestra textual de español del siglo XVI, similar a la del experimento de lexematización. Esta muestra fue etiquetada manualmente con categorías gramaticales siguiendo los criterios de codificación del grupo EAGLES[11]. A pesar de lo reducido de la muestra, el método de Brill fue capaz de generar reglas útiles para etiquetar las categorías gramaticales del español del siglo XVI, con una precisión de un poco más del 81%, lo cual es excelente, sobre todo al considerar que el corpus es antiguo.

En esencia, el objetivo principal del método de Brill es generar reglas *productivas* para etiquetar las palabras con su categoría gramatical. Una regla productiva es aquella que etiqueta correctamente un conjunto de palabras y cuyo efecto no se vea duplicado ni inhibido por la aplicación de otras reglas generadas por el mismo sistema. Así, la productividad no se trata solamente de etiquetar muchas palabras, sino también de que su etiquetado sea el más correcto posible desde un principio.

Brill estableció que su método no requiere de información morfológica explícita, porque revisa automáticamente un número previamente definido de letras finales e iniciales de cada palabra para hacer inferencias sobre la etiqueta gramatical que le corresponde. Logra su cometido, pero este acercamiento es más bien una técnica rudimentaria de fuerza

[10] *Identificación automática de categorías gramaticales en español del siglo XVI*, México, UNAM. Tesis de maestría (2009).

[11] EAGLES (Expert Advice Group for Language Engineering Standards) es una iniciativa de la Unión Europea, que trabaja desde 1994 en la definición de lineamientos de etiquetado de categorías gramaticales para las lenguas europeas, incluida la española.

bruta. Así que Méndez Cruz se propuso averiguar si el método pudiera beneficiarse con las técnicas de descubrimiento de la morfología mediante medidas de afijalidad.

De esta manera, diseñó varios experimentos y obtuvo, en casi todos, una precisión similar de alrededor de 81%. Sin embargo, el número de reglas se redujo con la inclusión de los afijos descubiertos previamente en el módulo morfológico. Lo que significa que las reglas resultaron más productivas. Esto quiere decir que, en términos de precisión, el método de Brill no mejoró significativamente, pero produjo un efecto de economía en la generación de reglas: menos reglas lograron un etiquetado gramatical similar. Esta mejora muestra cómo una técnica que busca medir fenómenos lingüísticos como el de la afijalidad compite con un método de fuerza bruta que debe recorrer todo el espacio de reglas generables para lograr alguna tarea relacionada con la estructura de la lengua.

3.1.3 Sintetizadores de voz

La síntesis de habla o de voz se refiere a la reproducción del sonido del habla mediante dispositivos electrónicos[12]. Quizá el tipo de sintetizadores de voz más conocido es el que recibe de entrada un texto y emite de salida el sonido de esa cadena textual. Esto es, el que genera una voz artificial que reproduce lo dicho en ese texto (*Text-To-Speech Synthesis*).

Un método muy popular de síntesis de voz es el conocido como de selección de unidades (*Unit-Selection Synthesis*)[13] que consiste en pregrabar ciertas porciones del habla y almacenarlas en una base de datos. Luego, al recibir un texto de entrada, el sintetizador combina o concatena estas grabaciones para *armar* el audio de ese texto y reproducirlo. Por ejemplo, si se graba el sonido de todas las letras aisladas de

[12] Para una descripción general de los métodos de síntesis del habla véase Jurafsky y Martin, *op. cit.* (2009, 249-284).

[13] Sobre síntesis de selección de unidades, véase *ibid.* (2009, 276-280).

un idioma, un sintetizador puede leer cualquier texto reproduciendo cada sonido de cada letra según se las vaya encontrando en la cadena textual. Evidentemente, el efecto de esto es burdo y poco natural, por lo que las unidades seleccionadas suelen ser de mayor tamaño que el sonido de una letra aislada: como difonos, trifonos, sílabas, palabras función, palabras de contenido comunes e, incluso, expresiones muy frecuentes.

Así, mientras más cortas sean las unidades del habla grabadas, más burda y mecánica será la voz que se escucha y, mientras más largas sean estas unidades, más inteligible y natural será el habla producida. Esto se debe a que en las fronteras entre las unidades grabadas ocurren los defectos o irregularidades que le quitan naturalidad a la reproducción del texto. De allí que, aunque existen técnicas para suavizar estos defectos, uno de los mayores problemas de los sistemas de síntesis concatenativa es determinar cómo combinar estas unidades para minimizar el número de fronteras entre ellas.

Entre los tipos de unidades a seleccionar, tiene sentido tomar en cuenta los afijos y las cadenas de afijos, porque, como hemos visto, son unidades que ocurren mucho, en combinación con las bases, que son secuencias de sonidos que ocurren poco. Al grabar el sonido del repertorio de afijos de una lengua, se evitará tener que suavizar las fronteras entre los fonos dentro de esas estructuras afijales.

Esta fue la idea de un experimento de construcción de un sintetizador de voz para la lengua ralámuli o tarahumara de San Luis Majimachi, Bocoyna, Chihuahua[14]. El experimento se basó en el catálogo de sufijos que resultó del procedimiento de extracción aplicado a las narraciones compiladas por Patricio Parra (2003), cuyos resultados se pueden observar en la Tabla 30 que aparece en la sección 3.3.2.

Para lograr el discurso más natural posible, se aplicó el enfoque de selección de unidades basado en palabras función, secuencias de sufijos

[14] Los detalles técnicos de este experimento se publicaron en Medina Urrea, Herrera Camacho y Alvarado García, "Towards the Speech Synthesis of Raramuri: a Unit Selection Approach based on Unsupervised Extraction of Suffix Sequences", *Research in Computing Science*, 41 (2009).

(derivacionales y flexivos) y difonos de la lengua. De esta manera, se grabaron estas unidades para construir la base de datos de audio que fue el corazón de este sistema de síntesis de texto a voz.

Maribel Alvarado, especialista de la lengua en este experimento, observó que hubo 10 sufijos derivativos importantes que no aparecieron en el catálogo. Se trató esencialmente de formas derivativas verbales, o modificadores de la transitividad o alguna característica semántica de las formas verbales (*~lo*, *~ni*, *~pu* [*~bu*], *~tu*, *~to*, *~pu*, *~tu*, *~repu*, *~bu*, *~bona*; las formas repetidas son homófonos). Esto podría significar que la pequeña muestra utilizada es más representativa de las estructuras nominales que de las verbales. Como sea, estos sufijos faltantes se agregaron al conjunto de unidades a ser procesadas por el sintetizador.

Cabe señalar que el sintetizador resultante no ofreció la naturalidad de voz que los sintetizadores del estado del arte de lenguas como el inglés y el alemán. Sin embargo, este sistema requirió de muy poca memoria y sirvió para trazar una ruta posible para el desarrollo de sintetizadores para lenguas de pocos recursos electrónicos como el ralámuli.

3.2 Los sufijos del español de México

En el capítulo anterior, los sufijos se perfilaron como los miembros de un conjunto de morfemas compacto y muy organizado de la morfología española, que, como sabemos, se emplea para codificar información léxica y gramatical.

En esta sección, nos enfocamos en los 749 segmentos más sufijales del CEMC, que se reproducen en la Tabla 17. La idea es agrupar aquellos que son reconocibles en varias tablas, para clasificarlos en sufijos de flexión, derivativos, pronombres enclíticos (que son gráficamente afijales) y combinaciones de éstos. Esas tablas aparecen a partir de la página 115.

Tabla 17. Sufijos y grupos sufijales del español de México (CEMC)

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
1	~ó	1428	0.737	0.919	0.872	0.875	0.900	0.843
2	~o	6314	0.686	0.979	0.802	0.470	0.629	0.822
3	~s	12013	1.000	0.997	0.461	0.538	0.513	0.819
4	~a	7687	0.575	0.982	0.889	0.515	0.443	0.815
5	~os	4554	0.478	0.975	0.824	0.516	0.564	0.759
6	~as	4324	0.422	0.978	0.865	0.608	0.597	0.755
7	~en	945	0.411	0.899	0.906	0.863	0.237	0.739
8	~ar	1633	0.218	0.962	0.915	0.735	0.893	0.698
9	~ado	1429	0.206	0.962	0.907	0.710	0.923	0.692
10	~ando	976	0.184	0.954	0.916	0.840	0.971	0.685
11	~e	2363	0.420	0.948	0.682	0.274	0.230	0.683
12	~é	639	0.410	0.820	0.815	0.893	0.409	0.682
13	~aba	828	0.182	0.957	0.902	0.889	0.956	0.680
14	~aron	736	0.178	0.960	0.894	0.894	0.973	0.677
15	~ada	1135	0.165	0.949	0.916	0.739	0.923	0.677
16	~arse	665	0.146	0.954	0.907	0.843	0.952	0.669
17	~ados	941	0.148	0.955	0.901	0.719	0.858	0.668
18	~aban	551	0.143	0.940	0.900	0.906	0.958	0.661
19	~adas	813	0.132	0.945	0.904	0.767	0.869	0.660
20	~an	1775	0.195	0.943	0.835	0.619	0.673	0.658
21	~ara	370	0.110	0.915	0.915	0.892	0.985	0.647
22	~ará	387	0.121	0.930	0.874	0.921	0.902	0.642
23	~arlo	316	0.093	0.929	0.885	0.916	0.959	0.636
24	~arla	270	0.080	0.919	0.907	0.931	0.965	0.635
25	~arme	244	0.087	0.913	0.892	0.931	0.954	0.631
26	~andose	260	0.080	0.914	0.886	0.897	0.907	0.626
27	~arán	256	0.090	0.911	0.876	0.924	0.912	0.626
28	~ido	445	0.104	0.857	0.914	0.767	0.852	0.625
29	~ita	453	0.096	0.897	0.873	0.764	0.808	0.622
30	~aría	231	0.081	0.887	0.892	0.924	0.906	0.620
31	~amos	645	0.135	0.880	0.842	0.738	0.880	0.619
32	~amente	624	0.119	0.978	0.753	0.861	0.979	0.617

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
33	~arlos	201	0.065	0.896	0.885	0.910	0.924	0.615
34	~ador	268	0.055	0.893	0.897	0.777	0.770	0.615
35	~aran	196	0.075	0.869	0.886	0.925	0.915	0.610
36	~es	2479	0.188	0.953	0.689	0.485	0.619	0.610
37	~ito	421	0.093	0.875	0.859	0.736	0.727	0.609
38	~aste	136	0.057	0.834	0.924	0.866	0.871	0.605
39	~arte	144	0.055	0.845	0.914	0.906	0.880	0.605
40	~antes	187	0.037	0.880	0.894	0.640	0.499	0.604
41	~adores	196	0.042	0.865	0.903	0.772	0.855	0.603
42	~idos	269	0.073	0.832	0.904	0.727	0.808	0.603
43	~ida	304	0.083	0.837	0.886	0.722	0.911	0.602
44	~arlas	139	0.054	0.878	0.876	0.891	0.914	0.602
45	~asión	540	0.063	0.928	0.811	0.527	0.840	0.601
46	~amiento	141	0.021	0.887	0.894	0.610	0.826	0.601
47	~ante	240	0.034	0.877	0.888	0.638	0.729	0.600
48	~arle	176	0.066	0.862	0.872	0.951	0.987	0.600
49	~asiones	263	0.046	0.916	0.837	0.673	0.861	0.599
50	~aremos	104	0.042	0.843	0.910	0.937	0.875	0.599
51	~itos	271	0.059	0.866	0.870	0.719	0.681	0.598
52	~n	3586	0.277	0.963	0.553	0.445	0.213	0.598
53	~asas	9	0.019	0.935	0.838	0.300	0.338	0.597
54	~itas	210	0.053	0.872	0.863	0.707	0.616	0.596
55	~aré	127	0.062	0.827	0.899	0.894	0.828	0.596
56	~ero	292	0.042	0.846	0.898	0.684	0.863	0.595
57	~ir	209	0.051	0.862	0.868	0.688	0.709	0.594
58	~aja	10	0.016	0.915	0.851	0.313	0.461	0.594
59	~ase	114	0.035	0.836	0.906	0.671	0.404	0.593
60	~arnos	139	0.054	0.863	0.858	0.885	0.939	0.592
61	~oro	15	0.022	0.909	0.838	0.333	0.481	0.590
62	~eros	200	0.034	0.827	0.899	0.627	0.771	0.587
63	~años	7	0.018	0.947	0.792	0.350	0.036	0.586
64	~eso	17	0.028	0.871	0.850	0.270	0.239	0.583
65	~eras	137	0.025	0.790	0.932	0.548	0.580	0.582
66	~alas	14	0.015	0.862	0.868	0.400	0.676	0.582

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
67	~adora	124	0.030	0.830	0.883	0.709	0.748	0.581
68	~idas	218	0.069	0.814	0.860	0.710	0.778	0.581
69	~osa	178	0.023	0.834	0.883	0.522	0.718	0.580
70	~usa	10	0.012	0.861	0.863	0.333	0.171	0.578
71	~ilo	14	0.019	0.859	0.852	0.269	0.282	0.576
72	~abamos	115	0.051	0.807	0.866	0.920	0.913	0.575
73	~iendo	276	0.103	0.860	0.758	0.899	0.920	0.574
74	~anta	16	0.012	0.845	0.863	0.552	0.734	0.573
75	~oso	191	0.027	0.812	0.880	0.569	0.580	0.573
76	~ala	30	0.017	0.808	0.893	0.492	0.737	0.573
77	~orar	6	0.006	0.842	0.868	0.200	0.346	0.572
78	~adoras	41	0.010	0.828	0.877	0.569	0.555	0.572
79	~arían	81	0.036	0.805	0.873	0.921	0.911	0.572
80	~amo	7	0.010	0.903	0.798	0.259	0.292	0.570
81	~ió	303	0.096	0.869	0.744	0.777	0.888	0.570
82	~ame	74	0.032	0.751	0.923	0.748	0.799	0.569
83	~ana	58	0.006	0.826	0.873	0.261	0.258	0.569
84	~ija	7	0.010	0.859	0.836	0.292	0.077	0.568
85	~ayo	9	0.016	0.862	0.826	0.310	0.462	0.568
86	~i	115	0.018	0.686	1.000	0.285	0.032	0.568
87	~able	115	0.022	0.814	0.867	0.442	0.582	0.568
88	~osos	112	0.019	0.817	0.866	0.519	0.478	0.567
89	~esos	8	0.027	0.966	0.709	0.229	0.449	0.567
90	~alado	6	0.023	0.859	0.819	0.286	0.078	0.567
91	~ijo	10	0.017	0.875	0.809	0.370	0.754	0.567
92	~ases	19	0.009	0.877	0.813	0.475	0.675	0.566
93	~er	264	0.072	0.775	0.851	0.624	0.546	0.566
94	~apa	9	0.015	0.919	0.764	0.474	0.500	0.566
95	~irse	108	0.035	0.818	0.846	0.777	0.735	0.566
96	~ata	32	0.019	0.770	0.908	0.299	0.720	0.565
97	~osas	91	0.011	0.825	0.860	0.474	0.796	0.565
98	~istas	181	0.023	0.843	0.830	0.670	0.732	0.565
99	~andolo	78	0.026	0.815	0.851	0.848	0.864	0.564
100	~alo	45	0.015	0.800	0.877	0.563	0.805	0.564

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
101	~ate	107	0.036	0.759	0.897	0.713	0.758	0.564
102	~ismo	213	0.023	0.876	0.792	0.603	0.845	0.563
103	~eto	10	0.004	0.838	0.846	0.161	0.085	0.563
104	~osar	8	0.008	0.920	0.758	0.615	0.821	0.562
105	~olar	6	0.001	0.873	0.811	0.240	0.547	0.562
106	~eses	26	0.012	0.811	0.861	0.456	0.943	0.561
107	~ilado	6	0.027	0.807	0.850	0.273	0.304	0.561
108	~ables	88	0.024	0.828	0.831	0.440	0.492	0.561
109	~asta	9	0.013	0.870	0.798	0.450	0.059	0.561
110	~ika	288	0.016	0.910	0.750	0.375	0.380	0.559
111	~andola	58	0.021	0.761	0.893	0.817	0.828	0.558
112	~iko	341	0.018	0.893	0.763	0.423	0.563	0.558
113	~anes	29	0.003	0.881	0.790	0.592	0.878	0.558
114	~asa	21	0.015	0.829	0.831	0.280	0.878	0.558
115	~aya	13	0.013	0.767	0.894	0.317	0.490	0.558
116	~una	10	0.020	0.795	0.859	0.303	0.015	0.558
117	~anas	18	0.004	0.850	0.819	0.158	0.203	0.558
118	~era	335	0.037	0.797	0.837	0.603	0.533	0.557
119	~ista	231	0.025	0.841	0.801	0.683	0.889	0.556
120	~arás	46	0.023	0.775	0.869	0.780	0.685	0.555
121	~eje	10	0.021	0.812	0.833	0.476	0.495	0.555
122	~iso	28	0.020	0.757	0.889	0.400	0.226	0.555
123	~esas	24	0.015	0.786	0.865	0.333	0.090	0.555
124	~ikas	167	0.009	0.909	0.747	0.390	0.360	0.555
125	~orado	6	0.004	0.826	0.835	0.150	0.172	0.555
126	~abas	30	0.016	0.784	0.863	0.517	0.548	0.554
127	~ikos	193	0.012	0.917	0.735	0.362	0.436	0.554
128	~ieron	238	0.083	0.835	0.745	0.919	0.974	0.554
129	~oja	7	0.005	0.884	0.773	0.280	0.183	0.554
130	~ansas	14	0.004	0.783	0.871	0.452	0.360	0.553
131	~al	375	0.028	0.819	0.807	0.458	0.352	0.551
132	~ee	9	0.016	0.783	0.856	0.214	0.539	0.552
133	~andome	48	0.021	0.744	0.888	0.762	0.798	0.551
134	~ales	281	0.025	0.843	0.785	0.497	0.649	0.551

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
135	~amientos	47	0.010	0.792	0.850	0.681	0.616	0.551
136	~arles	72	0.032	0.818	0.802	0.935	0.957	0.551
137	~ino	48	0.006	0.791	0.853	0.381	0.786	0.550
138	~inas	38	0.004	0.772	0.873	0.250	0.448	0.550
139	~asen	27	0.015	0.736	0.899	0.587	0.106	0.550
140	~aso	94	0.012	0.756	0.881	0.653	0.843	0.550
141	~us	41	0.008	0.729	0.912	0.318	0.014	0.549
142	~ano	67	0.009	0.789	0.850	0.274	0.450	0.549
143	~ale	48	0.020	0.710	0.917	0.658	0.889	0.549
144	~udo	21	0.007	0.725	0.915	0.344	0.489	0.549
145	~idor	12	0.004	0.797	0.846	0.387	0.217	0.549
146	~ina	115	0.014	0.759	0.873	0.312	0.396	0.548
147	~irte	8	0.022	0.774	0.843	0.286	0.495	0.547
148	~etos	6	0.014	0.804	0.821	0.136	0.111	0.546
149	~ieras	7	0.003	0.927	0.708	0.226	0.622	0.546
150	~alos	32	0.013	0.758	0.866	0.533	0.755	0.546
151	~adamente	53	0.010	0.813	0.814	0.616	0.783	0.546
152	~ona	95	0.013	0.796	0.827	0.540	0.758	0.545
153	~ila	16	0.002	0.731	0.903	0.302	0.276	0.545
154	~ako	10	0.003	0.843	0.789	0.227	0.445	0.545
155	~aras	28	0.019	0.703	0.912	0.412	0.632	0.545
156	~adero	16	0.006	0.739	0.887	0.432	0.135	0.544
157	~atibo	55	0.011	0.858	0.758	0.696	0.850	0.542
158	~andole	75	0.032	0.753	0.840	0.904	0.889	0.542
159	~año	9	0.015	0.811	0.799	0.346	0.147	0.542
160	~eno	15	0.014	0.814	0.795	0.146	0.067	0.541
161	~igo	12	0.012	0.833	0.779	0.279	0.652	0.541
162	~uto	9	0.003	0.738	0.882	0.237	0.078	0.541
163	~ojo	6	0.005	0.820	0.798	0.250	0.124	0.541
164	~ear	75	0.018	0.784	0.820	0.481	0.451	0.541
165	~isos	11	0.016	0.765	0.841	0.306	0.441	0.541
166	~etas	31	0.013	0.755	0.852	0.301	0.274	0.540
167	~ete	59	0.016	0.670	0.933	0.444	0.418	0.540
168	~anos	53	0.008	0.783	0.828	0.301	0.462	0.540

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
169	~rando	10	0.003	0.876	0.740	0.079	0.061	0.539
170	~uro	16	0.015	0.757	0.845	0.327	0.594	0.539
171	~ten	31	0.005	0.849	0.763	0.199	0.235	0.539
172	~urar	6	0.003	0.836	0.778	0.162	0.554	0.539
173	~tado	54	0.007	0.890	0.720	0.182	0.350	0.539
174	~lo	792	0.087	0.932	0.596	0.708	0.276	0.538
175	~atiba	44	0.009	0.843	0.763	0.647	0.864	0.538
176	~laba	7	0.021	0.821	0.772	0.143	0.342	0.538
177	~ota	24	0.008	0.708	0.897	0.358	0.521	0.538
178	~ía	970	0.103	0.820	0.688	0.760	0.789	0.537
179	~acho	6	0.002	0.874	0.736	0.300	0.100	0.537
180	~ato	60	0.009	0.749	0.853	0.423	0.482	0.537
181	~esa	51	0.020	0.681	0.910	0.381	0.174	0.537
182	~atos	20	0.009	0.806	0.795	0.253	0.644	0.537
183	~ería	121	0.028	0.700	0.881	0.617	0.873	0.536
184	~enas	10	0.005	0.802	0.801	0.217	0.069	0.536
185	~és	89	0.015	0.723	0.869	0.536	0.574	0.536
186	~andolas	22	0.008	0.764	0.835	0.611	0.590	0.536
187	~inos	34	0.005	0.782	0.820	0.370	0.300	0.536
188	~eo	99	0.017	0.745	0.841	0.425	0.628	0.534
189	~ajes	20	0.004	0.730	0.868	0.345	0.249	0.534
190	~iera	165	0.054	0.779	0.767	0.897	0.938	0.533
191	~se	1619	0.149	0.941	0.510	0.762	0.237	0.533
192	~istes	6	0.004	0.831	0.765	0.273	0.536	0.533
193	~andolos	44	0.015	0.772	0.810	0.786	0.763	0.532
194	~asón	8	0.003	0.784	0.810	0.308	0.046	0.532
195	~taron	29	0.009	0.847	0.741	0.185	0.165	0.532
196	~irme	23	0.009	0.759	0.829	0.500	0.406	0.532
197	~tando	29	0.007	0.860	0.729	0.177	0.264	0.532
198	~imos	151	0.048	0.780	0.766	0.668	0.758	0.532
199	~adito	10	0.003	0.757	0.835	0.333	0.305	0.532
200	~iando	21	0.002	0.789	0.803	0.309	0.159	0.531
201	~lados	7	0.021	0.867	0.705	0.071	0.047	0.531
202	~taban	20	0.007	0.862	0.725	0.172	0.643	0.531

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
203	~eta	64	0.018	0.702	0.874	0.448	0.386	0.531
204	~aka	19	0.003	0.745	0.845	0.380	0.748	0.531
205	~isa	118	0.016	0.800	0.776	0.667	0.701	0.531
206	~la	633	0.068	0.914	0.610	0.635	0.044	0.531
207	~adita	21	0.010	0.708	0.873	0.467	0.283	0.530
208	~rados	9	0.003	0.855	0.733	0.057	0.191	0.530
209	~elos	31	0.024	0.714	0.853	0.290	0.429	0.530
210	~raba	8	0.003	0.871	0.716	0.076	0.078	0.530
211	~ela	48	0.017	0.750	0.822	0.276	0.589	0.530
212	~ego	7	0.017	0.883	0.688	0.184	0.024	0.530
213	~etes	20	0.008	0.635	0.945	0.303	0.246	0.529
214	~isimo	98	0.030	0.843	0.714	0.790	0.773	0.529
215	~anse	11	0.003	0.789	0.796	0.282	0.156	0.529
216	~iados	16	0.002	0.827	0.758	0.216	0.238	0.529
217	~taba	42	0.007	0.853	0.726	0.246	0.699	0.529
218	~edo	7	0.002	0.792	0.792	0.269	0.141	0.529
219	~atibas	32	0.008	0.828	0.751	0.561	0.753	0.529
220	~ena	18	0.005	0.750	0.831	0.202	0.249	0.529
221	~ono	14	0.002	0.831	0.752	0.318	0.411	0.529
222	~erían	8	0.026	0.888	0.671	0.267	0.403	0.528
223	~esía	8	0.003	0.806	0.776	0.157	0.474	0.528
224	~tó	53	0.005	0.823	0.756	0.256	0.295	0.528
225	~lada	9	0.004	0.863	0.717	0.086	0.084	0.528
226	~í	138	0.103	0.592	0.889	0.664	0.035	0.528
227	~ean	25	0.007	0.743	0.834	0.347	0.771	0.528
228	~isas	22	0.005	0.731	0.846	0.393	0.364	0.528
229	~tar	75	0.010	0.839	0.734	0.240	0.344	0.528
230	~radas	12	0.002	0.838	0.743	0.095	0.123	0.527
231	~irán	52	0.023	0.763	0.796	0.703	0.553	0.527
232	~onas	20	0.006	0.768	0.806	0.318	0.917	0.527
233	~ol	14	0.002	0.741	0.836	0.141	0.332	0.527
234	~le	613	0.063	0.902	0.613	0.576	0.236	0.526
235	~tan	51	0.008	0.837	0.733	0.210	0.153	0.526
236	~iyas	54	0.020	0.761	0.797	0.495	0.245	0.526

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
237	~ese	143	0.102	0.603	0.873	0.638	0.391	0.526
238	~isima	63	0.025	0.830	0.722	0.716	0.609	0.526
239	~aña	8	0.003	0.831	0.743	0.258	0.247	0.525
240	~atas	6	0.002	0.738	0.837	0.111	0.042	0.525
241	~tados	35	0.004	0.865	0.706	0.178	0.223	0.525
242	~eada	21	0.007	0.768	0.799	0.300	0.261	0.525
243	~ías	120	0.017	0.856	0.700	0.436	0.719	0.524
244	~tara	15	0.006	0.855	0.712	0.203	0.172	0.524
245	~irlas	8	0.004	0.794	0.774	0.308	0.250	0.524
246	~iya	107	0.030	0.729	0.813	0.695	0.612	0.524
247	~is	115	0.011	0.795	0.765	0.270	0.199	0.524
248	~otes	33	0.005	0.721	0.845	0.446	0.389	0.524
249	~ren	12	0.003	0.870	0.696	0.117	0.138	0.523
250	~eña	15	0.002	0.770	0.797	0.366	0.204	0.523
251	~tada	28	0.004	0.877	0.687	0.138	0.165	0.523
252	~ían	336	0.134	0.724	0.710	0.794	0.690	0.523
253	~oneros	8	0.002	0.905	0.661	0.471	0.702	0.522
254	~atibos	30	0.008	0.838	0.721	0.566	0.703	0.522
255	~aderos	6	0.003	0.745	0.817	0.231	0.053	0.522
256	~idad	334	0.042	0.921	0.602	0.633	0.734	0.521
257	~itan	6	0.002	0.873	0.689	0.150	0.195	0.521
258	~rar	21	0.002	0.832	0.731	0.103	0.096	0.521
259	~otas	15	0.005	0.717	0.841	0.375	0.577	0.521
260	~tamos	25	0.008	0.823	0.733	0.245	0.717	0.521
261	~enos	6	0.021	0.764	0.778	0.100	0.720	0.521
262	~iga	9	0.003	0.783	0.776	0.220	0.278	0.521
263	~irla	23	0.014	0.740	0.809	0.404	0.322	0.521
264	~eka	7	0.004	0.901	0.656	0.140	0.290	0.521
265	~atoria	22	0.004	0.779	0.778	0.550	0.376	0.520
266	~ejo	9	0.004	0.798	0.759	0.250	0.387	0.520
267	~tadas	22	0.003	0.844	0.714	0.153	0.181	0.520
268	~eó	21	0.004	0.770	0.786	0.313	0.237	0.520
269	~uros	8	0.002	0.700	0.859	0.216	0.555	0.520
270	~t	36	0.004	0.692	0.865	0.196	0.139	0.520

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
271	~atorios	8	0.003	0.852	0.704	0.242	0.171	0.520
272	~elo	60	0.026	0.766	0.766	0.291	0.368	0.519
273	~iada	14	0.002	0.849	0.707	0.163	0.213	0.519
274	~eos	18	0.002	0.805	0.750	0.190	0.215	0.519
275	~laban	6	0.012	0.768	0.777	0.171	0.216	0.519
276	~eando	41	0.010	0.740	0.806	0.432	0.314	0.519
277	~omas	6	0.002	0.822	0.732	0.177	0.182	0.519
278	~eja	12	0.004	0.752	0.798	0.286	0.481	0.518
279	~los	410	0.055	0.937	0.563	0.594	0.030	0.518
280	~in	11	0.001	0.683	0.870	0.212	0.004	0.518
281	~ines	36	0.006	0.700	0.848	0.514	0.773	0.518
282	~irlos	16	0.010	0.732	0.812	0.390	0.328	0.518
283	~eres	15	0.014	0.723	0.816	0.254	0.373	0.518
284	~irá	78	0.031	0.755	0.768	0.804	0.695	0.518
285	~iría	43	0.022	0.688	0.842	0.672	0.679	0.518
286	~erla	22	0.021	0.802	0.729	0.367	0.375	0.517
287	~oya	6	0.008	0.784	0.759	0.273	0.046	0.517
288	~ea	79	0.017	0.728	0.808	0.348	0.251	0.517
289	~almente	74	0.007	0.809	0.735	0.468	0.434	0.517
290	~las	262	0.036	0.925	0.590	0.478	0.028	0.517
291	~um	22	0.005	0.743	0.803	0.324	0.320	0.517
292	~onero	10	0.002	0.794	0.754	0.526	0.826	0.517
293	~esito	20	0.005	0.642	0.902	0.435	0.209	0.517
294	~gos	8	0.002	0.864	0.682	0.064	0.049	0.516
295	~esta	13	0.012	0.822	0.715	0.245	0.019	0.516
296	~r	2587	0.191	0.948	0.410	0.698	0.584	0.516
297	~osamente	34	0.006	0.801	0.742	0.347	0.412	0.516
298	~iyo	75	0.017	0.728	0.802	0.625	0.512	0.516
299	~ras	179	0.021	0.937	0.589	0.256	0.410	0.516
300	~aje	60	0.008	0.693	0.846	0.566	0.559	0.515
301	~osidad	11	0.002	0.800	0.744	0.262	0.051	0.515
302	~eko	6	0.011	0.848	0.686	0.231	0.486	0.515
303	~entes	70	0.012	0.782	0.750	0.251	0.358	0.515
304	~inado	7	0.002	0.800	0.743	0.113	0.080	0.515

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
305	~iyos	36	0.011	0.710	0.822	0.480	0.257	0.514
306	~ra	917	0.078	0.877	0.588	0.537	0.681	0.514
307	~ine	6	0.002	0.762	0.779	0.140	0.206	0.514
308	~esen	16	0.005	0.789	0.748	0.203	0.186	0.514
309	~ansa	23	0.004	0.713	0.824	0.469	0.489	0.514
310	~tarse	12	0.002	0.841	0.697	0.101	0.065	0.514
311	~akas	7	0.002	0.782	0.756	0.233	0.704	0.514
312	~rado	20	0.003	0.814	0.723	0.085	0.262	0.513
313	~aria	44	0.007	0.823	0.710	0.306	0.156	0.513
314	~irlo	39	0.021	0.713	0.804	0.629	0.479	0.513
315	~ula	23	0.005	0.758	0.775	0.299	0.259	0.512
316	~ikamente	63	0.004	0.838	0.695	0.463	0.316	0.512
317	~rada	14	0.002	0.810	0.725	0.081	0.157	0.512
318	~ró	12	0.002	0.826	0.709	0.087	0.073	0.512
319	~erías	24	0.007	0.719	0.809	0.369	0.368	0.512
320	~iese	9	0.003	0.885	0.647	0.273	0.256	0.512
321	~ulas	8	0.003	0.823	0.708	0.160	0.032	0.512
322	~ite	8	0.002	0.732	0.800	0.129	0.127	0.511
323	~ele	12	0.021	0.800	0.713	0.146	0.090	0.511
324	~use	9	0.010	0.852	0.671	0.281	0.641	0.511
325	~il	42	0.007	0.684	0.840	0.321	0.197	0.510
326	~tikas	27	0.003	0.861	0.668	0.206	0.063	0.510
327	~imientos	9	0.006	0.909	0.616	0.200	0.138	0.510
328	~udos	8	0.005	0.724	0.802	0.267	0.342	0.510
329	~lado	14	0.015	0.811	0.704	0.100	0.062	0.510
330	~esido	10	0.004	0.785	0.741	0.167	0.210	0.510
331	~itado	8	0.002	0.843	0.685	0.174	0.258	0.510
332	~eado	23	0.009	0.765	0.756	0.307	0.287	0.510
333	~par	6	0.020	0.741	0.767	0.162	0.123	0.510
334	~oma	11	0.003	0.792	0.734	0.190	0.652	0.509
335	~do	2437	0.175	0.947	0.406	0.560	0.681	0.509
336	~iles	24	0.004	0.722	0.800	0.293	0.406	0.509
337	~ritas	6	0.002	0.786	0.739	0.158	0.476	0.509
338	~ates	10	0.003	0.744	0.779	0.250	0.109	0.509

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
339	~ocho	7	0.018	0.847	0.661	0.368	0.128	0.508
340	~ense	30	0.026	0.615	0.884	0.357	0.286	0.508
341	~asos	37	0.004	0.748	0.771	0.430	0.688	0.508
342	~abos	7	0.001	0.757	0.764	0.368	0.158	0.507
343	~tores	47	0.007	0.869	0.645	0.505	0.669	0.507
344	~tará	18	0.002	0.789	0.729	0.200	0.141	0.507
345	~aro	9	0.004	0.750	0.766	0.225	0.653	0.507
346	~ieran	75	0.038	0.685	0.798	0.743	0.785	0.507
347	~tibos	37	0.014	0.894	0.611	0.296	0.386	0.507
348	~iar	37	0.005	0.794	0.721	0.285	0.250	0.506
349	~tarán	7	0.002	0.818	0.699	0.123	0.113	0.506
350	~tas	254	0.013	0.890	0.616	0.253	0.390	0.506
351	~tos	176	0.012	0.887	0.619	0.178	0.383	0.506
352	~ta	604	0.028	0.850	0.641	0.351	0.601	0.506
353	~arios	51	0.006	0.824	0.688	0.311	0.399	0.506
354	~ensias	19	0.007	0.742	0.769	0.181	0.172	0.506
355	~ises	17	0.002	0.826	0.689	0.354	0.280	0.506
356	~eño	14	0.002	0.710	0.805	0.275	0.135	0.506
357	~ene	8	0.018	0.863	0.636	0.178	0.007	0.506
358	~eme	21	0.045	0.644	0.827	0.318	0.564	0.505
359	~esita	10	0.004	0.713	0.798	0.303	0.800	0.505
360	~ariamos	36	0.023	0.630	0.862	0.878	0.831	0.505
361	~ientes	26	0.006	0.834	0.674	0.280	0.178	0.505
362	~ke	26	0.015	0.764	0.735	0.176	0.001	0.505
363	~aditas	7	0.004	0.746	0.764	0.250	0.152	0.505
364	~sas	124	0.013	0.927	0.573	0.278	0.134	0.504
365	~tarla	7	0.002	0.795	0.716	0.137	0.083	0.504
366	~ote	55	0.011	0.728	0.773	0.377	0.474	0.504
367	~table	7	0.017	0.725	0.768	0.137	0.292	0.504
368	~alidades	8	0.002	0.758	0.751	0.267	0.097	0.504
369	~té	20	0.008	0.760	0.742	0.204	0.271	0.503
370	~eas	15	0.004	0.810	0.694	0.161	0.346	0.503
371	~tika	46	0.005	0.841	0.663	0.197	0.088	0.503
372	~alisar	6	0.001	0.748	0.759	0.167	0.049	0.503

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
373	~aduras	11	0.003	0.657	0.847	0.306	0.379	0.502
374	~istika	9	0.003	0.782	0.722	0.257	0.380	0.503
375	~erse	105	0.055	0.716	0.736	0.827	0.857	0.502
376	~ao	14	0.006	0.606	0.894	0.326	0.351	0.502
377	~arias	17	0.006	0.818	0.681	0.191	0.490	0.502
378	~iste	70	0.031	0.610	0.864	0.693	0.892	0.502
379	~oniko	11	0.001	0.728	0.776	0.268	0.346	0.502
380	~ua	6	0.003	0.881	0.620	0.087	0.006	0.501
381	~isadas	12	0.004	0.865	0.632	0.207	0.112	0.501
382	~ise	25	0.013	0.729	0.760	0.410	0.849	0.501
383	~atibamente	8	0.003	0.800	0.699	0.400	0.748	0.501
384	~iado	31	0.004	0.791	0.707	0.252	0.140	0.500
385	~oche	7	0.015	0.855	0.630	0.500	0.894	0.500
386	~onika	11	0.002	0.744	0.755	0.306	0.237	0.500
387	~onsito	14	0.004	0.704	0.792	0.560	0.477	0.500
388	~ablemente	6	0.002	0.758	0.738	0.143	0.042	0.500
389	~akos	6	0.001	0.763	0.733	0.214	0.197	0.499
390	~ismos	15	0.010	0.760	0.727	0.300	0.705	0.499
391	~tiko	45	0.004	0.839	0.654	0.196	0.092	0.499
392	~turas	7	0.001	0.747	0.749	0.123	0.069	0.499
393	~mas	25	0.010	0.849	0.635	0.122	0.064	0.498
394	~alisa	6	0.002	0.677	0.813	0.200	0.065	0.498
395	~andote	14	0.006	0.710	0.776	0.500	0.246	0.497
396	~les	455	0.039	0.907	0.546	0.364	0.372	0.497
397	~tero	7	0.001	0.768	0.723	0.105	0.034	0.497
398	~ueba	6	0.008	0.932	0.552	0.462	0.888	0.497
399	~uso	6	0.004	0.806	0.680	0.194	0.077	0.496
400	~sos	128	0.010	0.936	0.543	0.288	0.114	0.496
401	~so	257	0.020	0.896	0.573	0.347	0.156	0.496
402	~isimas	19	0.009	0.695	0.784	0.576	0.551	0.496
403	~ieramos	11	0.006	0.789	0.693	0.379	0.211	0.496
404	~ama	12	0.011	0.816	0.659	0.214	0.583	0.495
405	~enso	6	0.001	0.807	0.678	0.273	0.179	0.495
406	~tamente	17	0.002	0.763	0.721	0.218	0.320	0.495

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
407	~h	18	0.002	0.673	0.811	0.209	0.010	0.495
408	~iano	9	0.001	0.745	0.740	0.130	0.381	0.495
409	~isiones	10	0.002	0.807	0.677	0.156	0.515	0.495
410	~ia	204	0.018	0.872	0.595	0.198	0.374	0.495
411	~toras	10	0.002	0.797	0.685	0.435	0.515	0.495
412	~eser	10	0.011	0.753	0.718	0.167	0.331	0.494
413	~sa	281	0.019	0.879	0.582	0.317	0.249	0.494
414	~ario	76	0.006	0.765	0.709	0.362	0.304	0.494
415	~ín	55	0.008	0.623	0.848	0.519	0.533	0.493
416	~itar	7	0.002	0.773	0.704	0.130	0.444	0.493
417	~roso	7	0.001	0.864	0.614	0.111	0.151	0.493
418	~añas	8	0.002	0.776	0.701	0.286	0.124	0.493
419	~enses	6	0.001	0.744	0.733	0.200	0.233	0.493
420	~re	60	0.004	0.733	0.741	0.181	0.310	0.493
421	~isimos	33	0.015	0.733	0.729	0.674	0.443	0.492
422	~riko	7	0.001	0.862	0.613	0.063	0.032	0.492
423	~san	11	0.002	0.876	0.599	0.076	0.046	0.492
424	~ansia	22	0.005	0.731	0.740	0.310	0.336	0.492
425	~enes	10	0.002	0.709	0.764	0.204	0.212	0.492
426	~men	6	0.026	0.817	0.631	0.100	0.157	0.491
427	~nas	77	0.011	0.905	0.559	0.181	0.243	0.491
428	~lote	6	0.012	0.752	0.709	0.240	0.414	0.491
429	~oh	6	0.013	0.702	0.758	0.429	0.064	0.491
430	~tido	7	0.002	0.827	0.643	0.119	0.016	0.491
431	~andonos	18	0.011	0.651	0.809	0.529	0.512	0.490
432	~alisasión	18	0.003	0.776	0.692	0.340	0.271	0.490
433	~isar	69	0.008	0.832	0.631	0.496	0.320	0.490
434	~ie	13	0.002	0.738	0.731	0.186	0.024	0.490
435	~tarlo	8	0.001	0.755	0.713	0.125	0.078	0.490
436	~gas	10	0.011	0.732	0.726	0.096	0.076	0.490
437	~ses	33	0.006	0.816	0.647	0.140	0.113	0.490
438	~ere	12	0.016	0.813	0.638	0.188	0.126	0.489
439	~emos	360	0.147	0.650	0.669	0.898	0.581	0.489
440	~tiba	54	0.018	0.843	0.606	0.335	0.316	0.489

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
441	~lan	9	0.002	0.710	0.754	0.102	0.299	0.489
442	~edor	12	0.014	0.707	0.744	0.279	0.647	0.488
443	~to	366	0.023	0.837	0.605	0.215	0.442	0.488
444	~ares	46	0.007	0.837	0.620	0.305	0.435	0.488
445	~día	8	0.013	0.857	0.593	0.105	0.038	0.487
446	~ran	141	0.035	0.889	0.538	0.298	0.203	0.487
447	~isan	16	0.005	0.793	0.664	0.276	0.219	0.487
448	~rosa	9	0.001	0.781	0.679	0.158	0.330	0.487
449	~alismo	22	0.003	0.716	0.743	0.393	0.413	0.487
450	~ernos	11	0.013	0.717	0.730	0.220	0.186	0.487
451	~tibo	63	0.019	0.866	0.576	0.366	0.335	0.487
452	~tibas	39	0.015	0.893	0.551	0.312	0.272	0.487
453	~idades	86	0.014	0.840	0.605	0.677	0.583	0.487
454	~ios	64	0.006	0.886	0.567	0.138	0.177	0.487
455	~uela	16	0.019	0.678	0.761	0.500	0.831	0.486
456	~isado	35	0.010	0.828	0.621	0.343	0.203	0.486
457	~idores	11	0.005	0.711	0.742	0.500	0.433	0.486
458	~ajo	16	0.007	0.698	0.752	0.348	0.289	0.486
459	~ago	8	0.011	0.667	0.778	0.267	0.076	0.485
460	~enta	18	0.004	0.743	0.708	0.165	0.120	0.485
461	~bar	10	0.003	0.754	0.698	0.111	0.222	0.485
462	~iendose	68	0.035	0.700	0.720	0.701	0.616	0.485
463	~taría	7	0.001	0.784	0.670	0.137	0.046	0.485
464	~ien	7	0.002	0.765	0.687	0.206	0.984	0.485
465	~pe	9	0.005	0.709	0.739	0.173	0.073	0.485
466	~ben	9	0.015	0.773	0.665	0.132	0.269	0.484
467	~c	11	0.002	0.621	0.830	0.131	0.020	0.484
468	~ias	60	0.006	0.909	0.537	0.135	0.221	0.484
469	~tora	7	0.004	0.823	0.623	0.189	0.084	0.484
470	~sita	18	0.009	0.783	0.658	0.184	0.093	0.484
471	~tor	51	0.006	0.856	0.588	0.573	0.655	0.483
472	~enado	6	0.003	0.670	0.776	0.162	0.356	0.483
473	~nan	6	0.001	0.867	0.579	0.048	0.014	0.482
474	~erlos	15	0.022	0.730	0.695	0.300	0.289	0.482

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
475	~ramos	31	0.006	0.834	0.607	0.231	0.260	0.482
476	~iente	51	0.011	0.713	0.723	0.384	0.199	0.482
477	~iere	11	0.003	0.720	0.723	0.333	0.324	0.482
478	~elas	20	0.012	0.672	0.762	0.182	0.514	0.482
479	~ide	6	0.002	0.789	0.655	0.115	0.425	0.482
480	~mos	1103	0.182	0.857	0.406	0.691	0.653	0.482
481	~erlas	6	0.012	0.775	0.658	0.171	0.337	0.482
482	~irle	12	0.014	0.626	0.802	0.343	0.072	0.481
483	~onada	6	0.004	0.960	0.477	0.107	0.040	0.481
484	~isó	15	0.005	0.823	0.614	0.268	0.116	0.481
485	~idamente	9	0.005	0.650	0.785	0.257	0.065	0.480
486	~aramos	26	0.015	0.578	0.847	0.684	0.714	0.480
487	~tikos	34	0.002	0.788	0.651	0.197	0.130	0.480
488	~gó	7	0.003	0.743	0.694	0.093	0.032	0.480
489	~ora	146	0.012	0.908	0.519	0.447	0.104	0.479
490	~el	28	0.009	0.526	0.902	0.272	0.008	0.479
491	~na	185	0.019	0.871	0.546	0.194	0.140	0.479
492	~isión	39	0.005	0.700	0.730	0.302	0.336	0.478
493	~io	182	0.017	0.790	0.628	0.248	0.397	0.478
494	~ones	815	0.043	0.956	0.434	0.846	0.922	0.478
495	~alidad	50	0.006	0.705	0.721	0.476	0.385	0.477
496	~te	1072	0.090	0.884	0.457	0.359	0.490	0.477
497	~só	6	0.001	0.833	0.595	0.045	0.031	0.476
498	~tra	12	0.003	0.822	0.604	0.179	0.010	0.476
499	~ensia	83	0.018	0.612	0.799	0.343	0.534	0.476
500	~sitos	28	0.003	0.840	0.585	0.322	0.123	0.476
501	~are	9	0.003	0.663	0.762	0.360	0.385	0.476
502	~iales	10	0.002	0.821	0.604	0.103	0.164	0.476
503	~imiento	83	0.022	0.720	0.683	0.722	0.666	0.475
504	~me	565	0.082	0.856	0.485	0.787	0.124	0.474
505	~gado	9	0.002	0.723	0.697	0.100	0.079	0.474
506	~m	21	0.003	0.709	0.711	0.136	0.107	0.474
507	~adura	11	0.003	0.598	0.821	0.344	0.345	0.474
508	~iadas	10	0.003	0.708	0.710	0.164	0.103	0.474

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
509	~da	441	0.155	0.930	0.334	0.200	0.304	0.473
510	~erlo	35	0.035	0.680	0.704	0.493	0.300	0.473
511	~mente	981	0.089	0.976	0.354	0.881	0.976	0.473
512	~aderas	6	0.002	0.632	0.785	0.333	0.137	0.473
513	~tasión	23	0.002	0.700	0.717	0.147	0.193	0.473
514	~saba	7	0.001	0.801	0.615	0.072	0.060	0.472
515	~sen	31	0.010	0.824	0.583	0.158	0.094	0.472
516	~teros	6	0.001	0.707	0.709	0.120	0.073	0.472
517	~iaba	9	0.002	0.690	0.724	0.184	0.225	0.472
518	~ma	45	0.009	0.782	0.625	0.117	0.058	0.472
519	~onado	8	0.005	0.928	0.481	0.123	0.031	0.471
520	~das	174	0.173	0.929	0.311	0.116	0.097	0.471
521	~ibles	13	0.002	0.704	0.709	0.146	0.078	0.471
522	~res	334	0.024	0.955	0.434	0.412	0.352	0.471
523	~iré	14	0.009	0.687	0.716	0.400	0.094	0.471
524	~iendolo	12	0.012	0.684	0.717	0.333	0.279	0.471
525	~li	7	0.009	0.679	0.724	0.146	0.178	0.471
526	~ular	9	0.003	0.793	0.616	0.087	0.112	0.470
527	~istiko	12	0.003	0.696	0.709	0.375	0.400	0.469
528	~nar	6	0.002	0.848	0.557	0.034	0.015	0.469
529	~oles	14	0.013	0.736	0.656	0.175	0.400	0.468
530	~eles	11	0.006	0.646	0.750	0.183	0.061	0.467
531	~onadas	8	0.003	0.849	0.551	0.200	0.097	0.467
532	~ian	11	0.002	0.750	0.647	0.193	0.206	0.466
533	~erte	18	0.023	0.627	0.748	0.290	0.287	0.466
534	~no	108	0.010	0.823	0.565	0.179	0.040	0.466
535	~esitos	7	0.005	0.603	0.790	0.241	0.231	0.466
536	~ís	21	0.003	0.644	0.750	0.328	0.204	0.465
537	~ible	31	0.006	0.685	0.705	0.261	0.115	0.465
538	~isada	23	0.006	0.795	0.594	0.277	0.160	0.465
539	~oras	26	0.004	0.883	0.508	0.167	0.042	0.465
540	~esió	8	0.003	0.776	0.615	0.151	0.148	0.465
541	~nos	361	0.046	0.801	0.546	0.519	0.256	0.464
542	~jas	7	0.003	0.822	0.567	0.081	0.106	0.464

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
543	~ós	8	0.001	0.659	0.732	0.444	0.225	0.464
544	~lar	18	0.004	0.777	0.611	0.077	0.062	0.464
545	~sía	12	0.002	0.846	0.544	0.118	0.099	0.464
546	~gar	9	0.005	0.694	0.692	0.094	0.019	0.464
547	~andoles	13	0.009	0.607	0.775	0.565	0.341	0.464
548	~eaba	13	0.003	0.566	0.821	0.277	0.146	0.464
549	~tibamente	10	0.007	0.790	0.594	0.217	0.257	0.463
550	~uras	24	0.013	0.886	0.492	0.183	0.115	0.463
551	~sito	64	0.006	0.791	0.592	0.464	0.299	0.463
552	~ifikada	7	0.003	0.743	0.640	0.194	0.103	0.462
553	~dos	214	0.165	0.920	0.300	0.118	0.240	0.462
554	~go	56	0.014	0.676	0.695	0.221	0.245	0.462
555	~ai	8	0.002	0.594	0.788	0.250	0.010	0.461
556	~ro	155	0.010	0.855	0.519	0.209	0.093	0.461
557	~be	32	0.010	0.718	0.655	0.224	0.353	0.461
558	~uye	6	0.003	0.730	0.649	0.171	0.103	0.461
559	~eaban	12	0.005	0.601	0.776	0.293	0.215	0.461
560	~íos	7	0.005	0.686	0.691	0.212	0.290	0.460
561	~de	34	0.008	0.767	0.605	0.168	0.004	0.460
562	~atorias	8	0.003	0.688	0.688	0.400	0.139	0.460
563	~nales	6	0.001	0.893	0.484	0.049	0.161	0.459
564	~rías	23	0.006	0.849	0.522	0.190	0.324	0.459
565	~ga	42	0.011	0.666	0.700	0.180	0.085	0.459
566	~lasión	6	0.001	0.780	0.595	0.055	0.014	0.459
567	~ores	189	0.016	0.915	0.444	0.380	0.347	0.458
568	~tes	183	0.018	0.916	0.440	0.205	0.181	0.458
569	~isando	9	0.004	0.759	0.610	0.180	0.088	0.458
570	~ura	88	0.018	0.761	0.595	0.411	0.525	0.458
571	~irnos	18	0.012	0.582	0.779	0.462	0.323	0.458
572	~esko	7	0.002	0.750	0.621	0.206	0.051	0.458
573	~or	324	0.020	0.866	0.485	0.485	0.269	0.457
574	~nada	6	0.001	0.788	0.582	0.035	0.003	0.457
575	~rito	9	0.001	0.683	0.684	0.145	0.190	0.456
576	~arselo	10	0.007	0.623	0.735	0.455	0.462	0.455

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
577	~dió	12	0.004	0.833	0.528	0.164	0.151	0.455
578	~tibidad	10	0.003	0.791	0.570	0.313	0.280	0.455
579	~oide	9	0.002	0.650	0.712	0.429	0.500	0.455
580	~isia	9	0.003	0.667	0.693	0.243	0.140	0.454
581	~one	12	0.001	0.828	0.532	0.214	0.027	0.454
582	~tura	19	0.002	0.666	0.692	0.198	0.117	0.453
583	~tito	8	0.002	0.730	0.627	0.131	0.079	0.453
584	~ría	405	0.060	0.792	0.506	0.684	0.465	0.453
585	~sando	6	0.001	0.772	0.585	0.050	0.035	0.453
586	~isados	17	0.006	0.786	0.565	0.221	0.108	0.452
587	~alista	13	0.003	0.638	0.715	0.245	0.188	0.452
588	~jos	10	0.003	0.820	0.532	0.130	0.245	0.451
589	~ja	23	0.010	0.729	0.614	0.157	0.173	0.451
590	~ne	32	0.003	0.729	0.621	0.165	0.042	0.451
591	~arías	6	0.011	0.530	0.811	0.286	0.175	0.451
592	~che	11	0.009	0.663	0.678	0.149	0.104	0.450
593	~po	11	0.009	0.627	0.712	0.136	0.068	0.449
594	~irían	11	0.013	0.549	0.784	0.314	0.230	0.449
595	~isasión	106	0.010	0.861	0.475	0.663	0.447	0.449
596	~atorio	13	0.005	0.633	0.706	0.325	0.176	0.448
597	~diendo	6	0.004	0.828	0.510	0.111	0.050	0.447
598	~selo	7	0.014	0.819	0.509	0.115	0.077	0.447
599	~cha	11	0.002	0.672	0.666	0.112	0.006	0.447
600	~iana	6	0.001	0.658	0.680	0.107	0.148	0.447
601	~lón	8	0.005	0.683	0.652	0.191	0.052	0.447
602	~ulares	10	0.002	0.709	0.628	0.172	0.137	0.446
603	~osis	12	0.001	0.620	0.716	0.158	0.425	0.446
604	~dieron	8	0.004	0.811	0.520	0.143	0.042	0.445
605	~gan	10	0.005	0.678	0.652	0.104	0.038	0.445
606	~olojía	18	0.003	0.596	0.736	0.265	0.209	0.445
607	~uales	9	0.003	0.688	0.644	0.237	0.538	0.445
608	~abilidad	19	0.006	0.631	0.697	0.396	0.734	0.445
609	~ge	8	0.003	0.674	0.658	0.148	0.050	0.445
610	~ros	27	0.006	0.860	0.467	0.053	0.025	0.444

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
611	~rán	204	0.070	0.862	0.397	0.465	0.288	0.443
612	~ón	957	0.051	0.920	0.358	0.467	0.770	0.443
613	~niko	17	0.002	0.780	0.546	0.157	0.159	0.443
614	~iremos	9	0.011	0.645	0.672	0.310	0.103	0.443
615	~jo	20	0.004	0.739	0.583	0.119	0.015	0.442
616	~atura	9	0.002	0.531	0.791	0.346	0.109	0.441
617	~ié	8	0.002	0.624	0.698	0.229	0.202	0.441
618	~x	10	0.002	0.527	0.794	0.149	0.065	0.441
619	~tones	10	0.001	0.691	0.628	0.270	0.111	0.440
620	~ifikasión	23	0.010	0.686	0.623	0.383	0.253	0.440
621	~nado	18	0.003	0.731	0.584	0.083	0.135	0.439
622	~sado	14	0.003	0.739	0.574	0.062	0.039	0.439
623	~ulo	15	0.003	0.736	0.574	0.188	0.259	0.437
624	~ré	66	0.046	0.796	0.469	0.277	0.172	0.437
625	~ual	12	0.004	0.682	0.622	0.261	0.495	0.436
626	~tón	12	0.002	0.724	0.578	0.222	0.068	0.435
627	~des	29	0.003	0.802	0.496	0.090	0.209	0.433
628	~ioso	8	0.002	0.675	0.623	0.118	0.041	0.433
629	~eska	10	0.004	0.674	0.620	0.200	0.067	0.433
630	~emia	6	0.001	0.606	0.691	0.200	0.085	0.432
631	~l	358	0.018	0.804	0.474	0.278	0.086	0.432
632	~iamos	96	0.043	0.578	0.675	0.636	0.439	0.432
633	~ria	18	0.009	0.849	0.437	0.066	0.121	0.432
634	~rte	83	0.057	0.863	0.375	0.305	0.109	0.432
635	~tismo	10	0.001	0.587	0.702	0.185	0.283	0.430
636	~amento	15	0.004	0.504	0.780	0.441	0.368	0.429
637	~ienta	6	0.002	0.682	0.602	0.188	0.291	0.429
638	~én	9	0.002	0.568	0.716	0.209	0.776	0.429
639	~sio	11	0.001	0.681	0.599	0.104	0.172	0.427
640	~ial	32	0.004	0.716	0.561	0.208	0.388	0.427
641	~sidad	28	0.003	0.775	0.502	0.267	0.408	0.427
642	~yo	38	0.003	0.846	0.430	0.190	0.015	0.426
643	~kan	7	0.002	0.716	0.560	0.056	0.068	0.426
644	~ches	6	0.012	0.619	0.645	0.146	0.153	0.425

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
645	~erme	48	0.034	0.524	0.717	0.696	0.643	0.425
646	~rse	786	0.113	0.903	0.257	0.743	0.720	0.424
647	~ío	16	0.004	0.563	0.702	0.242	0.049	0.423
648	~rle	56	0.071	0.924	0.274	0.217	0.127	0.423
649	~siones	324	0.036	0.720	0.511	0.480	0.407	0.422
650	~rá	462	0.080	0.836	0.350	0.728	0.420	0.422
651	~iba	15	0.009	0.822	0.435	0.061	0.017	0.422
652	~mo	118	0.004	0.839	0.422	0.184	0.033	0.422
653	~tario	6	0.001	0.685	0.578	0.107	0.062	0.421
654	~sión	863	0.041	0.784	0.440	0.558	0.559	0.421
655	~edad	24	0.014	0.617	0.633	0.500	0.361	0.421
656	~bo	15	0.005	0.800	0.452	0.043	0.010	0.419
657	~rían	76	0.028	0.760	0.467	0.437	0.197	0.418
658	~sis	45	0.003	0.690	0.561	0.266	0.493	0.418
659	~rás	32	0.018	0.741	0.494	0.283	0.085	0.418
660	~itis	9	0.001	0.537	0.715	0.188	0.277	0.418
661	~torio	10	0.004	0.803	0.444	0.152	0.146	0.417
662	~ya	62	0.006	0.764	0.480	0.220	0.030	0.416
663	~remos	62	0.036	0.767	0.444	0.337	0.242	0.416
664	~ifika	12	0.009	0.604	0.633	0.240	0.093	0.416
665	~erá	68	0.042	0.592	0.611	0.840	0.959	0.415
666	~rme	202	0.060	0.846	0.337	0.509	0.434	0.414
667	~rlo	140	0.082	0.887	0.269	0.292	0.159	0.413
668	~ola	13	0.008	0.691	0.538	0.082	0.090	0.412
669	~’s	13	0.001	0.507	0.729	0.433	0.092	0.412
670	~ndo	345	0.092	0.902	0.240	0.219	0.187	0.411
671	~ás	36	0.009	0.699	0.521	0.205	0.154	0.409
672	~ron	317	0.083	0.939	0.206	0.281	0.446	0.409
673	~rio	35	0.007	0.817	0.398	0.103	0.068	0.407
674	~sina	7	0.003	0.594	0.623	0.125	0.251	0.407
675	~tro	7	0.002	0.625	0.593	0.071	0.002	0.406
676	~cho	7	0.003	0.570	0.640	0.075	0.003	0.404
677	~csión	6	0.003	0.734	0.476	0.048	0.018	0.404
678	~nal	10	0.002	0.755	0.455	0.064	0.224	0.404
679	~iensia	6	0.001	0.732	0.475	0.188	0.116	0.402

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
680	~dón	6	0.001	0.599	0.608	0.222	0.474	0.402
681	~ral	10	0.001	0.603	0.600	0.089	0.376	0.402
682	~rlas	18	0.040	0.877	0.283	0.082	0.050	0.400
683	~isarse	7	0.004	0.631	0.562	0.132	0.063	0.399
684	~ndole	14	0.019	0.945	0.233	0.118	0.060	0.399
685	~ndose	43	0.046	0.953	0.197	0.108	0.096	0.399
686	~riamos	12	0.022	0.739	0.432	0.177	0.110	0.398
687	~ba	148	0.105	0.772	0.314	0.115	0.044	0.397
688	~erán	21	0.024	0.614	0.550	0.404	0.070	0.396
689	~je	44	0.005	0.647	0.537	0.225	0.188	0.396
690	~sko	8	0.001	0.701	0.486	0.105	0.282	0.396
691	~sar	16	0.003	0.671	0.513	0.063	0.035	0.395
692	~nse	7	0.007	0.862	0.315	0.052	0.012	0.395
693	~ús	7	0.001	0.494	0.690	0.368	0.073	0.395
694	~rnos	115	0.046	0.777	0.357	0.449	0.249	0.393
695	~rla	93	0.055	0.838	0.284	0.227	0.103	0.393
696	~isidad	11	0.001	0.508	0.665	0.324	0.227	0.391
697	~emente	7	0.002	0.731	0.441	0.042	0.018	0.391
698	~dor	108	0.081	0.844	0.246	0.251	0.178	0.390
699	~ei	6	0.007	0.330	0.833	0.207	0.040	0.390
700	~rs	13	0.002	0.659	0.508	0.296	0.328	0.389
701	~ridad	9	0.001	0.629	0.528	0.173	0.300	0.386
702	~rios	9	0.007	0.782	0.367	0.036	0.006	0.385
703	~ste	118	0.063	0.708	0.382	0.389	0.061	0.384
704	~cto	6	0.001	0.659	0.490	0.069	0.011	0.383
705	~dero	11	0.005	0.682	0.463	0.164	0.131	0.383
706	~rles	14	0.025	0.810	0.312	0.117	0.058	0.382
707	~endo	14	0.011	0.875	0.252	0.038	0.053	0.379
708	~nidad	7	0.001	0.518	0.615	0.149	0.045	0.378
709	~án	52	0.024	0.794	0.307	0.097	0.090	0.375
710	~ko	90	0.015	0.768	0.339	0.085	0.141	0.374
711	~itud	14	0.004	0.563	0.555	0.483	0.500	0.374
712	~dores	77	0.068	0.829	0.224	0.248	0.160	0.374
713	~ente	84	0.013	0.743	0.360	0.057	0.131	0.372
714	~d	48	0.007	0.817	0.292	0.061	0.285	0.372

rango		frec.	cuad.	econ.	entr.	prob1	prob2	afijdad.
715	~rlos	42	0.041	0.801	0.272	0.135	0.084	0.371
716	~ka	78	0.020	0.711	0.374	0.074	0.132	0.368
717	~nte	212	0.028	0.864	0.206	0.113	0.143	0.366
718	~sia	154	0.004	0.717	0.374	0.342	0.228	0.365
719	~ño	6	0.001	0.651	0.431	0.061	0.008	0.361
720	~ntes	40	0.019	0.784	0.275	0.068	0.008	0.359
721	~sta	11	0.002	0.793	0.281	0.025	0.005	0.359
722	~ble	87	0.027	0.753	0.291	0.218	0.332	0.357
723	~nsia	32	0.006	0.830	0.223	0.100	0.041	0.353
724	~eron	11	0.013	0.847	0.196	0.036	0.012	0.352
725	~miento	248	0.039	0.773	0.239	0.709	0.704	0.350
726	~kos	20	0.007	0.700	0.326	0.029	0.008	0.344
727	~nes	31	0.010	0.791	0.227	0.027	0.006	0.343
728	~yas	6	0.003	0.639	0.387	0.035	0.005	0.343
729	~'l	9	0.001	0.613	0.409	0.409	0.227	0.341
730	~mento	7	0.005	0.465	0.554	0.090	0.090	0.341
731	~dora	32	0.051	0.757	0.215	0.154	0.104	0.341
732	~ña	6	0.004	0.581	0.437	0.065	0.018	0.340
733	~á	70	0.029	0.697	0.293	0.091	0.207	0.340
734	~on	65	0.031	0.826	0.155	0.053	0.004	0.337
735	~kas	14	0.008	0.639	0.364	0.024	0.006	0.337
736	~nta	7	0.002	0.618	0.381	0.041	0.052	0.334
737	~ska	8	0.001	0.550	0.433	0.090	0.065	0.328
738	~iento	8	0.006	0.797	0.162	0.021	0.002	0.322
739	~lidad	8	0.007	0.611	0.309	0.036	0.049	0.309
740	~mientos	12	0.011	0.653	0.256	0.103	0.019	0.307
741	~bles	16	0.023	0.607	0.261	0.054	0.014	0.297
742	~bilidad	7	0.002	0.541	0.344	0.081	0.045	0.296
743	~ión	41	0.025	0.745	0.111	0.025	0.032	0.294
744	~dad	18	0.004	0.662	0.182	0.030	0.086	0.283
745	~tis	7	0.001	0.496	0.321	0.115	0.114	0.273
746	~ban	22	0.046	0.588	0.148	0.033	0.007	0.261
747	~ad	22	0.005	0.644	0.107	0.034	0.133	0.252
748	~ilidad	8	0.002	0.439	0.254	0.076	0.013	0.232
749	~smo	6	0.007	0.475	0.061	0.016	0.005	0.181

Como se dijo, el material de esta tabla sirvió para construir otras tablas que se presentan en los siguientes apartados. Estas tablas permiten observar la naturaleza de los resultados del procedimiento aplicado al CEMC para extraer sufijos. En especial, permiten examinar lo que se logró identificar y lo que faltó descubrir. En esencia, las tablas presentan morfos y cadenas de morfos. Se agrupan, como se mencionó, en tres tablas: cadenas de sufijos flexivos, de sufijos derivativos y cadenas con enclíticos pronominales.

3.2.1 Sufijos flexivos

Una de las cosas que sobresalen de los sufijos de la Tabla 17 es que muchas formas representan más de un morfema, por lo que podemos caracterizarlas como homófonas. Por ejemplo, el sufijo *~s* (afjdad. 0.8192) marca el plural en sustantivos (como en *silla**s***, *oso**s***, etc.), pero también es una desinencia verbal de 2ª persona singular (*corre**s***, *compra**s***, etc.). Por otra parte, podemos ver el sufijo *~e* (afjdad. 0.6833) como uno derivativo (en *ajust**e***, *embragu**e***, etc.), pero también como una desinencia de subjuntivo (en *compr**e***, *sum**e***, etc.).

La Tabla 18 contiene formas sufijales que representan marcas de flexión nominal y que también representan desinencias verbales y sufijos derivativos (*~a*, *~e* y *~o*). Normalmente, esta homografía-homonimia no es un obstáculo para distinguir el tipo de sufijo, puesto que el contexto ayuda a desambiguarlo. De hecho, las palabras y las cadenas de afijos con que un afijo aparece son contextos que brindan mucha información porque disminuyen la ambigüedad. Por ejemplo, en la cadena *~alidades* no hay ambigüedad en cuanto a que *~es* es un alomorfo del morfema plural de sustantivos y adjetivos.

Tabla 18. Sufijos de flexión nominal

sufijos	frecuencia	afijalidad
~a	7 687	0.8153
~o	6 314	0.8222
~s	12 013	0.8192
~os	4 554	0.7588
~as	4 324	0.7547
~es	2 479	0.6097

En esencia, en esta tabla hay sufijos y cadenas afijales de flexión nominal (masculino, femenino y/o plural). Estas formas a menudo participan en fenómenos de concordancia: de sustantivos con artículos, determinadores y adjetivos y viceversa. También otros morfemas, como *~amiento* y *~ansia* (que son derivativos) participan en los fenómenos de concordancia, aunque propiamente no ocurra en ellos un sufijo de flexión nominal (*~a* u *~o*). Como sea, aproximadamente 82 de las formas del catálogo terminan en *~o* (con un promedio de afijalidad de 0.5089). Por los contextos, la mayoría son grupos donde sí ocurre el morfema de flexión (*~ad**o***, *~er**o***, *~ik**o***, *~iy**o***, etc.). Similarmente, aproximadamente 77 formas terminan en *~a* (con promedio de afijalidad de 0.50497), donde muchas sí son marcas de género femenino (*~it**a***, *~ador**a***, *~os**a***, *~an**a***, etc.).

El caso de la marca de plural nominal es menos claro. Hay algo más de 100 secuencias de afijos que contienen el sufijo de flexión nominal de plural. Se trata en su mayoría de algún sufijo derivativo seguido de uno de los alomorfos *~s* o *~es* (por ejemplo, *~ado**s***, *~ante**s***, *~ador**es***, *~ita**s***, etc.). La afijalidad promedio de esas secuencias es de 0.5211, cantidad poco despreciable, al considerar que más de la mitad de la lista tiene menos.

Como sea, los sufijos de flexión nominal son pocos. En cuanto a la flexión verbal, la mejor evidencia de que los segmentos que se des-

cubrieron son en verdad los más afijales se obtiene al corroborar que el sistema de flexión verbal ocurre completo dentro de las 740 formas más afijales de la Tabla 17. Estas formas verbales se agruparon en las siguientes tres tablas. De esta manera, la Tabla 19 contiene los paradigmas de conjugación verbal del modo indicativo, la Tabla 20 los del subjuntivo y la Tabla 21 muestra los sufijos de verboides.

Tabla 19. Sufijos de flexión verbal del modo indicativo

~ar			**~er**			**~ir**		
presente								
~o			6314			0.8222		
~eo	99	0.5342	~es	2479	0.6097			
~as	4324	0.7547						
~eas	15	0.5028						
~a	7687	0.8153	~e	2363	0.6833			
~ea	79	0.5173						
~amos	645	0.6187	~emos	360	0.4888	~imos	151	0.5317
~an	1775	0.6579	~en	945	0.7386			
~ean	25	0.5276						
pretérito								
~é	639	0.6818	~í	138	0.5279			
~aste	136	0.6051	~iste	70	0.5015			
~ó	1428	0.8428	~ió	303	0.5698			
~eó	21	0.5202						
~amos	645	0.6187	~imos	151	0.5317			
			~ieron	238	0.5542			
~aron	736	0.6773	~eron	11	0.3521			
~ron			317			0.409		

En la primera parte de la Tabla 19, se exhiben los paradigmas del presente y del pretérito. La segunda parte contiene los del futuro y la tercera las formas del pospretérito y del copretérito. Estas tablas están

divididas en tres columnas, una para cada conjugación. Como en la tabla anterior, cada forma aparece con su frecuencia (como sufijo) y su índice normalizado de afijalidad. Obsérvese que algunas formas no contienen la vocal temática que identifica la conjugación pertinente, por lo que se muestran abarcando las tres columnas. Otras formas son comunes a los paradigmas de la segunda y tercera conjugaciones, por lo que abarcan las dos columnas correspondientes. Nótese que se incluyeron algunas formas de paradigmas irregulares, esto es, formas con material adicional (por ejemplo *~eas* en formas como *caporaleas*, *bateas*, *tarareas* de los verbos *caporalear*, *batear* y *tararear*) o más cortas que las regulares (*~eron* de *fueron*, *trajeron*, *produjeron*).

Tabla 19 (continuación).
Sufijos de flexión verbal del modo indicativo

~ar			**~er**			**~ir**		
futuro								
~aré	127	0.596	[~eré]			~iré	14	0.471
	~ré		66				0.4370	
	~é		639				0.6818	
~arás	46	0.5554	[~erás]				[~irás]	
	~rás		32				0.4176	
	~ás		36				0.4093	
~ará	387	0.6415	~erá	68	0.4153	~irá	78	0.5177
	~rá		462				0.4221	
	~á		70				0.3396	
~aremos	104	0.5985	[~eremos]			~iremos	9	0.4426
	~remos		62				0.41560	
	~emos		366				0.43700	
~arán	256	0.6257	~erán	21	0.396	~irán	52	0.5272
	~rán		204				0.4431	
	~án		52				0.3749	

Tabla 19 (continuación).
Sufijos de flexión verbal del modo indicativo

~ar			~er			~ir		
pospretérito								
~aría	231	0.6198	[~ería]			~iría	43	0.5175
~ría			405				0.4527	
~ía			970				0.5371	
~arías	6	0.4505	~erías	24	0.5118	[~irías]		
~rías			23				0.4593	
~ías			120				0.5244	
~ariamos	36	0.5049	[~eríamos]			[~iríamos]		
~riamos			12				0.3977	
~iamos			96				0.4318	
~arían	81	0.5715	~erían	8	0.5284	~irían	11	0.4487
~rían			76				0.4183	
~ían			336				0.5226	
copretérito								
~aba	828	0.6803	~ía		970		0.5371	
~eaba	13	0.4636						
~abas	30	0.5544	~ías		120		0.5244	
~abamos	115	0.5746	~iamos		96		0.4318	
~aban	551	0.661	~ían		336		0.5226	
~eaban	12	0.4605						

Por otra parte, no se incluyen otras formas de verbos regulares que empiezan con semiconsonante que sí aparecen en el catálogo (*~iar*, *~iaba*, *~iado* e *~iando*). Tampoco se incluyen formas como *~ua* de *averigua* (rango 380 en la Tabla 17) porque otros miembros del paradigma (*~uan*, *~uamos*, etc.) no están en el catálogo. Tampoco las formas *~úa*, *~úan*, etc. de verbos como *actuar* y *perpetuar* (*actúa*, *perpetúa*) están en

el catálogo. Sin embargo, estos casos quedan cubiertos con los sufijos regulares de la primera conjugación.

Otras familias de sufijos irregulares que tampoco aparecen en la Tabla 19, pero que sí ocurren en el catálogo, contienen material que no corresponde propiamente ni a la base ni al sufijo (*~go*, *~ga*, *~gas* y *~gan* en verbos como *oír*, *traer*, *venir*, *tener*, etc.; así como *~ka*, *~kas*, *~kan* y *~ska* en los subjuntivos de *producir* y *conducir*).

Aunque también presentes en el catálogo, tampoco se incluyen en estas tablas las marcas de plural que ocurren al final de varias desinencias verbales y que pueden considerarse morfemas separados; esto es, *~mos*, marca de plural de 1ª persona (afjdad. de 0.4818) y *~n*, marca plural de 3ª persona y 2ª formal (0.5977).

También las flexiones del tiempo futuro, los morfemas correspondientes al verbo *haber* (*~é*, *~ás*, *~á*, *~emos* y *~án*), aparecen en el catálogo, por lo que se incluyen en la Tabla 19. La mayoría tiene una afijalidad comparativamente baja, pero destaca el hecho de que todo el paradigma esté presente.

Algunas formas se muestran entre corchetes. Son las que se echan de menos por no haber ocurrido dentro de las 749 más afijales. Muchas son formas con uno o más sufijos de flexión adheridos a una vocal temática (por ejemplo *~eré*, *~irás*, *~eríamos*, etc.). En cambio, los grupos de sufijos sin las vocales temáticas, que son comunes a las tres conjugaciones, sí están todos (*~ré*, *~rás*, *~remos*, *~ría*, *~rías*, y *~ríamos*).

En la Tabla 20 están los paradigmas del subjuntivo. De nuevo, las formas faltantes aparecen entre corchetes. Éstas pertenecen a los paradigmas del pretérito (*~ásemos*, *~ieses ~iésemos* e *~iesen*) y del futuro (*~aren*, *~iéremos* e *~ieren*). Los sufijos del presente de subjuntivo están completos en parte porque comparten formas con el presente de indicativo, aunque en diferentes conjugaciones (los de la primera en indicativo son las de segunda y tercera en subjuntivo). De hecho, la ausencia de formas subjuntivas del futuro y del segundo paradigma del pretérito es compatible con apreciaciones de que están cayendo en desuso.

Tabla 20. Flexiones del subjuntivo

~ar			~er/~ir		
presente					
~e	2363	0.6833	~a	7687	0.8153
~es	2479	0.6097	~as	4324	0.7547
~emos	360	0.4888	~amos	645	0.6187
~en	945	0.7386	~an	1775	0.6579
pretérito					
~ara	370	0.6467	~iera	165	0.5332
~ra		917		0.5144	
~aras	28	0.5446	~ieras	7	0.5462
~ras		179		0.5157	
~aramos	26	0.4800	~ieramos	11	0.4959
~ramos		31		0.4823	
~aran	196	0.6097	~ieran	75	0.5067
~ran		141		0.4873	
~ase	114	0.5925	~iese	9	0.5117
~ases	19	0.5664	[~ieses]		
			~eses	26	0.5613
[~ásemos]			[~iésemos/~ésemos]		
~asen	27	0.5497	[~iesen]		
			~esen	16	0.514
futuro					
~are	9	0.4759	~iere	11	0.4820
			~ere	12	0.4889
~ares	46	0.4878	~[i]eres	15	0.5178
~res		334		0.4711	
~aremos	104	0.5985	[~iéremos]		
[~aren]			[~ieren]		
~ren		12		0.5230	

Tabla 21. Sufijos de verboides

sufijos	frecuencia	afijalidad
~ar	1 633	0.6982
~ear	75	0.5405
~er	264	0.5661
~ir	209	0.5938
~r	2 587	0.5161
~ando	976	0.6847
~eando	41	0.5189
~iendo	276	0.5738
~ndo	345	0.4111
~ado	1 429	0.6917
~eado	23	0.5098
~ido	445	0.6248
~do	2 437	0.5092

Por último, los sufijos para formar verboides se muestran en la Tabla 21. Los participios se repiten en la Tabla 22, de sufijos derivativos, porque también sirven para formar adjetivos a partir de raíces verbales. Cabe notar que, en promedio, estos pocos sufijos tienen una afijalidad de alrededor de 0.56. Lo interesante es que no falta ninguno e incluso se incluyen formas sin vocales temáticas (*~r*, *~ndo* y *~do*) o con material adicional (*~ear*, *~eando*, *~eado*).

3.2.2 Sufijos derivativos

Como se verá a continuación, los sufijos derivativos y las cadenas de sufijos que los contienen son numerosos y tienden a ocurrir menos que los de flexión. En ese sentido son menos económicos y, por lo tanto, menos afijales. En esta subsección se organizan en tres grupos, los sufijos derivativos relacionados con el verbo (Tabla 22), ya sea porque convierten

raíces verbales en sustantivos o adjetivos, o porque convierten raíces no verbales (por ejemplo, adjetivos) en verbos; grupos de sufijos adverbiales (Tabla 23); y sufijos derivativos nominales (Tabla 24).

Tabla 22. Sufijos derivativos y verbales
(con y sin marcas de flexión nominal)

sufijos	frecuencia	afijalidad
~a	7687	0.8153
~as	4324	0.7547
~o	6314	0.8222
~os	4554	0.7588
~e	2363	0.6833
~es	2479	0.6097
~ado	1429	0.6917
~ao	14	0.5017
~eado	23	0.5098
~ados	941	0.6678
~ada	1135	0.6768
~eada	21	0.5247
~adas	813	0.6602
~ido	445	0.6248
~idos	269	0.6030
~ida	304	0.6022
~idas	218	0.5810
~da	441	0.4732
~das	174	0.4713
~do	2437	0.5092
~dos	214	0.4616
~ando	976	0.6847
~*isó*	15	0.4805
~*is*ar	69	0.4903
~*isa*rse	7	0.3989
~*isa*ndo	9	0.4577
~*isa*da	23	0.4651
~*isa*das	12	0.5005
~*isa*do	35	0.4861
~*isa*dos	17	0.4521
~*ifik*a	12	0.4155
~*ifika*da	7	0.4620

Respecto a la derivación que involucra verbos como bases o que resulta en la formación de verbos, en la Tabla 22 se exhiben, primero, algunos sufijos que se adhieren a verbos para formar sustantivos, *compr**a***, *logr**o***, *cort**e***. Luego, se muestran las marcas participiales para formar adjetivos y sustantivos a partir de bases verbales. Finalmente, aparecen los sufijos que se adhieren a sustantivos o adjetivos para formar verbos (*miti**ficar***, *esceni**ficar***... ; *neutral**izar***, *plural**izar***, etc.).

Algunas formas, *~itar*, *~itan* y el participio *~itado* obtuvieron un grado de afijalidad alrededor de 0.5 (0.493, 05214 y 0.5099, respectivamente), por lo que pudieron haber sido incluidas en esta tabla. Sin embargo, al examinar el tipo de vocablos de donde provienen (*editar*, *dinamitar*, *gravitar*, *evitar*, *meditar*, *necesitar*, etc.), sus reducidas frecuencias de ocurrir en éstos como afijos (7, 6 y 8 respectivamente), y no encontrar algún significado que los justifique como morfemas, se excluyeron de esta tabla.

Con respecto a la derivación de adverbios, en la Tabla 23 se reúnen las secuencias de sufijos que terminan en *~mente*. Lo que salta a la vista es el tipo de sufijos a los que se adhiere que, como era de esperarse, es el de los sufijos derivativos que forman adjetivos.

Tabla 23. Grupos de sufijos con marca adverbial

sufijos	**frecuencia**	**afijalidad**
~mente	981	0.4728
~ablemente	6	0.4996
~amente	624	0.6169
~adamente	53	0.5457
~almente	74	0.5171
~atibamente	8	0.5004
~tibamente	10	0.4634
~emente	7	0.3911
~idamente	9	0.4801
~ikamente	63	0.5123
~osamente	34	0.5160

En cuanto a la derivación nominal, el conjunto de secuencias de sufijos derivativos es enorme, como se ve en la Tabla 17. Para revisarlos de manera sistemática, se cotejan con uno de los inventarios compilados por Moreno de Alba, concretamente, el de sufijos ordenados por su for-

ma[15], que no consigna todos los sufijos analizados por ese investigador, pero sí algunos de los más importantes, agrupados por su semejanza formal y ordenados por porcentaje de ocurrencias.

La pertinencia de la semejanza formal es evidente, ya que el conjunto de sufijos extraídos del CEMC es un conjunto de formas (cadenas de morfos). Así que en la Tabla 24 varias secuencias de sufijos, que contienen por lo menos uno de derivación nominal, se agrupan por parecido formal. Estos sufijos de derivación, además de ocurrir en las secuencias de morfos obtenidas en el experimento del capítulo anterior, están documentados en alguno de los grupos de sufijos que inventarió Moreno de Alba.

Como se dijo, Moreno organiza los sufijos por semejanza formal. En ese sentido, los grupos son alomorfos de algún morfema, aunque como dice el investigador a veces sea difícil argumentar que conserven una "aceptable homogeneidad de sentido" (Moreno de Alba 1986, 183). Una diferencia importante entre los sufijos de Moreno y los de la Tabla 24 es que éstos últimos representan secuencias de sufijos, entre los que se encuentra el alomorfo en cuestión, mientras que Moreno los presenta aislados. Evidentemente, a partir de estos grupos se puede investigar de manera automática la afitáctica de todas y cada una de las secuencias y, por lo tanto, del español de México, tanto de morfos de flexión como de derivación[16].

[15] *Morfología derivativa nominal en el español de México*, UNAM, México (Moreno de Alba 1986, 183-205).

[16] Existen trabajos para investigar la morfotáctica de los afijos descubiertos de manera no supervisada mediante autómatas de estados finitos. Véase por ejemplo, para el español de México, la tesis doctoral de Carlos Méndez, *Generación automática de una gramática de estados finitos para la morfología del español*. UNAM, México (2013).

Tabla 24. Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	frecuencia	afijalidad
~(V)(C)ión	~asión	540	0.6007
	~*alis*asión	18	0.4903
	~isión	39	0.4782
	~tasión	23	0.4727
	~lasión	6	0.4585
	~*is*asión	106	0.4485
	~*ifik*asión	23	0.4395
	~sión	863	0.4213
	~csión	6	0.4042
	~ión	41	0.2938
~V	~a	7687	0.8153
	~e	2363	0.6833
	~o	6314	0.8222
	~as	4324	0.7547
	~os	4554	0.7588
	~es	2479	0.6097
	~ea	79	0.5173
	~eo	99	0.5342
	~eos	18	0.5192

El primer grupo[17], ~(V)(C)ión (*~ación*, *~ión*), que sirve para formar sustantivos de acción o efecto, fue el más frecuente de su material (12% de su total de vocablos). Esto, al considerar las frecuencias, no coincide con los datos obtenidos en este trabajo. De hecho, el orden utilizado por Moreno tampoco corresponde al de afijalidad. Por ejemplo, el promedio de afijalidad del primer grupo, ~(V)(C)ión, es bajo

[17] En la notación de Moreno de Alba, los paréntesis indican que el elemento puede estar ausente. Además, 'V' significa cualquier vocal, 'C' cualquier consonante y '-' flexión nominal de género.

(0.45077). En particular, al compararlo con el del grupo siguiente, ~V (vocales para formar sustantivos con sentido de acción o efecto: *~e*, *~o*, y con material adicional, *~eo*), que alcanza 0.7736, tomando en cuenta solamente los sufijos de una sola vocal y sin marca de plural. Es más, el grupo de alomorfos de diminutivo, ~(V)(C)it- (*~adito*, *~itito*, *~ita*), que sigue a los dos primeros muestra frecuencias menores y alcanza un promedio de afijalidad de 0.5155.

Tabla 24 (continuación).
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	**frecuencia**	**afijalidad**
(V)(C)it-	~ita	453	0.6219
	~ito	421	0.6093
	~*ad*ito	10	0.5316
	~*ad*ita	21	0.5303
	~*es*ito	20	0.5165
	~*es*ita	10	0.5051
	~*on*sito	14	0.5001
	~sita	18	0.4834
	~sito	64	0.4629
	~rito	9	0.4561
	~tito	8	0.4528
	~itos	271	0.5984
	~itas	210	0.5962
	~*ad*itas	7	0.5045
	~*es*itos	7	0.4658
	~sitos	28	0.4760

	sufijos	frecuencia	afijalidad
~(V)al	~al	375	0.5515
	~ual	12	0.4362
	~ial	32	0.4268
	~ales	281	0.5509
	~uales	9	0.4449
	~iales	10	0.4756

Además, el segundo grupo, ~V, comparte formas con el paradigma de flexión nominal de género, lo que explica el alto promedio de afijalidad, pero hay varios grupos con promedios de afijalidad más alta que éste. En suma, el orden de importancia de Moreno no concuerda ni con los datos de frecuencia ni con la medida de afijalidad. Algunos ejemplos de vocablos que exhiben los sufijos de los primeros tres grupos son: *aclar**ación***, *nacion**alización***, *admi**sión***; *siembr**a***, *enfoqu**e***, *muestr**eo***; y *cafec**ito***, *sent**adito***, *cabez**oncito***, etcétera. El grupo siguiente, representado por ~(V)al (*~al*, *~ual*), sirve típicamente para formar adjetivos que designan relación o caracterización. Algunos ejemplos son *experiment**al***, *mund**ial*** y *man**ual***.

Luego, los sufijos del grupo ~(V)(C)(C)ic- (*~ico*, *~ística*) forman adjetivos y sustantivos con sentido técnico; por ejemplo, *magn**ífico***, *electr**ónica*** y *caracter**ístico***. El conjunto de sufijos con el esquema ~(V)(C)(C)ad (*~idad*, *~dad*) suelen formar sustantivos abstractos. Algunos ejemplos son: 'carn**osidad**', 'nacion**alidad**', 'suav**idad**', 'cruel**dad**' y 'puber**tad**'.

Tabla 24 (continuación).
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	frecuencia	afijalidad
~(V)(C)(C)ic-	~ika	288	0.5585
	~iko	341	0.5581
	~ikas	167	0.5551
	~ikos	193	0.5544
	~tika	46	0.5028
	~*ist*ika	9	0.5024
	~*o*niko	11	0.5015
	~*o*nika	11	0.5001
	~tiko	45	0.4989
	~riko	7	0.4922
	~*ist*iko	12	0.4691
	~niko	17	0.4429
	~*if*ika	12	0.4155
	~tikas	27	0.5104
	~tikos	34	0.4800
~(V)(C)(C)ad	~idad	334	0.5214
	~*os*idad	11	0.5154
	~*al*idades	8	0.5035
	~idades	86	0.4865
	~*al*idad	50	0.4771
	~*abil*idad	19	0.4447
	~sidad	28	0.4268
	~edad	24	0.4211
	~*is*idad	11	0.3913
	~ridad	9	0.3861
	~nidad	7	0.3776
	~lidad	8	0.3090
	~*bil*idad	7	0.2957
	~dad	18	0.2826
	~ad	22	0.2517
	~*il*idad	8	0.2316

El grupo siguiente, representado por ~(V)Vnte (*~ante*, *~iente*), sirve para formar adjetivos a partir de verbos; por ejemplo, *ayud**ante***, *conven**iente*** y *absorb**ente***.

Tabla 24 (continuación).
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	**frecuencia**	**afijalidad**
~(V)Vnte	~ante	240	0.5998
	~antes	187	0.6037
	~entes	70	0.5146
	~ientes	26	0.5048
	~iente	51	0.4821
	~ente	84	0.3721
	~nte	212	0.3659
	~ntes	40	0.3591

El siguiente conjunto, ~Vd- (*~ado*, *~ida*), corresponde al grupo presentado en la Tabla 22. Se trata de varias cadenas de sufijos de origen participial, con vocal temática y marcas de flexión nominal. Algunos ejemplos de vocablos formados por este grupo son: *señ**alado***, *part**ida***, *subord**inado*** y *clas**ificada***.

Tabla 24 (continuación).
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	frecuencia	Afijalidad
	~ado	1 429	0.6917
	~ada	1 135	0.6768
	~ados	941	0.6678
	~adas	813	0.6602
	~ido	445	0.6248
	~idos	269	0.6030
	~ida	304	0.6022
	~idas	218	0.5810
	~alado	6	0.5669
	~orado	6	0.5549
	~iados	16	0.5289
	~eada	21	0.5247
	~iada	14	0.5193
~Vd-	~inado	7	0.5146
	~esido	10	0.5099
	~eado	23	0.5098
	~isadas	12	0.5005
	~iado	31	0.5004
	~isado	35	0.4861
	~enado	6	0.4827
	~onada	6	0.4806
	~iadas	10	0.4736
	~onado	8	0.4714
	~onadas	8	0.4673
	~isada	23	0.4651
	~ifikada	7	0.4620
	~isados	17	0.4521

El siguiente grupo de la tabla es el representado mediante ~Vncia o ~anza (*~ancia*, *~anza*). Se utiliza para formar sustantivos de acción o resultado de la acción; por ejemplo, *mat**anza***, *toler**ancia*** y *apar**iencia***.

Tabla 24 (continuación).
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	**frecuencia**	**afijalidad**
~Vncia ~anza	~ansas	14	0.5525
	~ansa	23	0.5138
	~ensias	19	0.5059
	~ansia	22	0.4916
	~ensia	83	0.4761
	~iensia	6	0.4024
	~nsia	32	0.3529
~(u)os-	~osa	178	0.5802
	~oso	191	0.5732
	~osos	112	0.5673
	~osas	91	0.5653
	~ioso	8	0.4332
~(Vd)er-	~ero	292	0.5953
	~era	335	0.5569
	~eros	200	0.5867
	~eras	137	0.5822
	~onero	10	0.5168
	~oneros	8	0.5224
	~adero	16	0.5439
	~aderos	6	0.5216
	~aderas	6	0.4727
	~dero	11	0.3831

Luego está el conjunto representado por ~(u)os- (*~oso*, *~uosa*). Este grupo sirve para formar adjetivos de cualidad o defecto. Algunos ejemplos son: *fam**oso***, *mald**oso***, *defect**uoso*** y *juic**ioso***. El grupo ~(Vd)er- (*~adero*, *~era*) sirve para formar sustantivos y adjetivos que designan algún agente, instrumento, objeto, alimento, etc., por ejemplo, *sal**ero***, *limon**ero*** y *panad**era***.

Tabla 24 (continuación).
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	**frecuencia**	**afijalidad**
~Vm(i)ent-	~amiento	141	0.6005
	~amientos	47	0.5506
	~imientos	9	0.5103
	~imiento	83	0.4751
	~amento	15	0.4292
	~miento	248	0.3500
	~mento	7	0.3412
	~mientos	12	0.3068
~(Vt)iv-	~atibo	55	0.5418
	~atiba	44	0.5382
	~atibas	32	0.5286
	~atibos	30	0.5219
	~tibo	63	0.4868
	~tiba	54	0.4888
	~tibos	37	0.5066
	~tibas	39	0.4866
	~iba	15	0.4220
~Vble	~able	115	0.5679
	~ables	88	0.5611
	~ibles	13	0.4713
	~ible	31	0.4652
	~ble	87	0.3569
	~bles	16	0.2973

El conjunto con la forma ~Vm(i)ent- (*~amento*, *~imienta*) forma sustantivos de acción o resultado. Por ejemplo, las voces *carg**amento***, *herr**amienta*** y *mov**imiento***. El siguiente grupo está representado por ~(Vt)iv- (*~ativo*, *~iva*). Sus formas sirven para formar adjetivos. Algunos ejemplos son: *signific**ativa***, *conflic**tivo***, *consec**utivo*** y *expres**ivo***. Luego está el conjunto de formas representado por ~Vble (*~able*, *~ible*) y que sirve para formar adjetivos que suponen capacidad y aptitud. Por ejemplo, *inaplaz**able***, *sens**ible*** e *insol**uble***.

El grupo ~(V)Cor- (*~ador*, *~idor*, *~tor*) sirve para formar adjetivos que designan oficios, profesiones y ocupaciones. Algunos adjetivos son: *pirograb**ador***, *profer**idora*** y *protec**tora***. El conjunto de formas representado mediante ~a(ta)ri- (*~aria*, *~atario*) se utiliza para formar sustantivos y adjetivos con significados típicamente colectivos, locativos, etc. Por ejemplo, *origin**ario***, *universi**taria*** y *prole**tario***. El grupo de la notación ~í- (*~ío*, *~ía*) se utiliza para formar sustantivos abstractos. Ejemplos de estos sustantivos son: *mejor**ía***, *arque**ología***, *burgu**esía*** y *jud**ío***.

Tabla 24 (continuación).
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	frecuencia	afijalidad
~(V)Cor(-)	~ador	268	0.6147
	~edor	12	0.4883
	~idor	12	0.5487
	~adora	124	0.5810
	~adores	196	0.6033
	~idores	11	0.4858
	~adoras	41	0.5716
	~tores	47	0.5069
	~toras	10	0.4947
	~tora	7	0.4837
	~tor	51	0.4832
	~dor	108	0.3902
	~dores	77	0.3736
	~dora	32	0.3407
	~or	324	0.4573
	~ora	146	0.4793
	~ores	189	0.4584
	~oras	26	0.4649
~a(ta)ri-	~aria	44	0.5133
	~arios	51	0.5059
	~arias	17	0.5017
	~ario	76	0.4935
	~tario	6	0.4214
~í-	~ía	970	0.5371
	~esía	8	0.5283
	~ías	120	0.5244
	~sía	12	0.4639
	~íos	7	0.4604
	~olojía	18	0.4450
	~ío	16	0.4232

El conjunto siguiente, ~(V)(C)ón(-) (*~ón*, *~ona*), se emplea para construir sustantivos y adjetivos aumentativos o de acción contundente. Algunos ejemplos son: *apret*ón, *pis**otón***, *saca**tones*** y *empuj**oncito***. El grupo representado mediante ~Vría (*~aría*, *~ería*) se utiliza en la formación de sustantivos. Por ejemplo, sustantivos como *secret**aría***, *not**aría***, *enferm**ería*** e *ingeni**ería***. Luego está en conjunto ~(V)(C)ura (*~ura*, *~adura*) para formar sustantivos. Algunos sustantivos formados con estos sufijos son: *pint**ura***, *criat**ura*** y *colgad**ura***.

Tabla 24 (continuación).
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	**frecuencia**	**afijalidad**
~(V)(C)ón(-)	~ón	957	0.4430
	~ones	815	0.4777
	~ona	95	0.5453
	~onas	20	0.5269
	~tón	12	0.4346
	~tones	10	0.4397
	~oneros	8	0.5224
	~onero	10	0.5168
	~onsito	14	0.5001
~Vría	~aría	231	0.6198
	~ería	121	0.5364
	~erías	24	0.5118
	~rías	23	0.4593
	~arías	6	0.4505
~(V)(C)ura	~ura	88	0.4577
	~uras	24	0.4632
	~adura	11	0.4739
	~aduras	11	0.5024
	~tura	19	0.4531
	~turas	7	0.4987
	~atura	9	0.4414

El conjunto ~(V)(C)ez(a) (*~ez*, *~aleza*) sirve para formar sustantivos a partir de adjetivos. Por ejemplo, los sustantivos *madur**ez***, *tris**teza*** y *estupid**eces***. El grupo siguiente es el representado por ~(V)(C) ori- (*~orio*, *~atoria*). Estas formas se utilizan para formar sustantivos y adjetivos. Algunos ejemplos son: *reclin**atorio***, *escap**atoria*** y *dormi**torio***. Enseguida está el conjunto ~Vdor(a) (*~ador*, *~edor*, *~idora*) para formar sustantivos que designan objetos, instrumentos o lugares. Por ejemplo, los sustantivos *incub**adora***, *corr**edor*** y *med**idor***. El conjunto de formas representado por ~in- (*~ino*, *~ina*) sirve para formar sustantivos muy diversos y adjetivos que señalan semejanzas y características. Algunos ejemplos son: *alcal**ino***, *cervant**ino*** y *estudiant**inas***.

Tabla 24 (continuación).
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	**frecuencia**	**afijalidad**
~(V)(C)ez(a)	~és	89	0.5357
	~esa	51	0.5368
	~eses	26	0.5613
	~esas	24	0.5551
~(V)(C)ori-	~atoria	22	0.5203
	~atorios	8	0.5195
	~atorias	8	0.4597
	~atorio	13	0.4477
	~torio	10	0.4172

	sufijos	frecuencia	afijalidad
~Vdor(a)	~ador	268	0.6147
	~edor	12	0.4883
	~idor	12	0.5487
	~adores	196	0.6033
	~idores	11	0.4858
	~adora	124	0.5810
	~adoras	41	0.5716
	~dor	108	0.3902
	~dores	77	0.3736
	~dora	32	0.3407
~in-	~ino	48	0.5500
	~inas	38	0.5498
	~ina	115	0.5483
	~inos	34	0.5356

El siguiente grupo es muy variado y está representado por la notación ~t- (*~te*, *~to*). Estas formas dan lugar a sustantivos y adjetivos abstractos de acción o efecto. Algunos vocablos con estos sufijos son: *muer**te***, *aten**to*** y *ven**ta***, *produc**to*** e *institu**to***. El conjunto representado mediante ~(i)(t)ud (*~itud*, *~ud*) sirve para formar sustantivos también abstractos que designan cualidad, acción o conducta. Por ejemplo, los sustantivos *exact**itud***, *inept**itud*** y *juvent**ud***. El antepenúltimo conjunto es el representado por ~(c)ill- (*~illo*, *~illa*). Este grupo se utiliza típicamente para formar sustantivos y adjetivos despectivos y a veces diminutivos. Algunos ejemplos son: *chiqu**illo***, *cortin**illa*** y *molin**illo***. El penúltimo grupo está constituido por ~(i)ci- (*~icio*, *~cia*). Estas formas se unen a raíces adjetivas para formar sustantivos abstractos; por ejemplo, las voces *inmund**icia***, *silen**cio*** e *infan**cia***. El último de los grupos organizados por forma es ~i- (*~ia*, *~ie*, *~io*) y es el menos documentado entre los materiales de Moreno de Alba (0.4% de su total). Las formas de este grupo dan lugar a sustantivos abstractos

relacionados con verbos o nombres. Algunos ejemplos son los sustantivos *molest**ia***, *domin**io*** y *progen**ie***.

Tabla 24 (continuación).
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	**frecuencia**	**afijalidad**
~t-	~to	366	0.4879
	~ta	604	0.5061
	~te	1,072	0.4769
	~tas	254	0.5062
	~tos	176	0.5062
	~tes	183	0.4582
	~rte	83	0.4315
	~ste	118	0.3839
	~cto	6	0.3832
	~sta	11	0.3585
	~nta	7	0.3336
	~uto	9	0.5410
	~ato	60	0.5369
~(i)(t)ud	~itud	14	0.3739
~(c)ill-	~iya	107	0.5240
	~iyo	75	0.5157
	~iyas	54	0.5258
	~iyos	36	0.5144
~(i)ci-	~sio	11	0.4271
	~sia	154	0.3650
	~isia	9	0.4544
~i-	~ia	204	0.4948
	~io	182	0.4780
	~ie	13	0.4902
	~ias	60	0.4842
	~ios	64	0.4865

Como se puede corroborar al examinar los 749 sufijos de la Tabla 17, las formas que Moreno de Alba consignó en su estudio no agotan todo sufijo o grupo de sufijos derivativos que están allí. Hasta aquí se ha mostrado que los segmentos de la Tabla 17 no son meros residuos. De hecho, todavía hay muchas que también se pueden agrupar por forma.

Por otra parte, si comparamos el promedio de cantidades normalizadas de afijalidad de los sufijos derivativos (Tablas 22 y 24) con aquellos flexivos que ocurren en el verbo (Tablas 19, 20 y 21), encontramos que los primeros (afjdad. 0.4864) son menos afijales que los segundos (afjdad. 0.5350). Estos promedios son burdos, pero a grandes rasgos van de acuerdo con la intuición de que los sufijos verbales, los más frecuentes, pero con menos tipos o formas, son los más afijales. Los derivativos son un conjunto de más formas (más de 200) con menos ocurrencias cada una. De hecho, mientras que en la Tabla 24 aparece una selección de sufijos derivativos larga e incompleta, los tipos de segmentos de flexión que faltan son pocos y, como se vio, están marcados en las tablas mediante corchetes cuadrados.

3.2.3 Enclíticos

Se mencionó que, al adherirse gráficamente a la palabra escrita, los enclíticos ocurren entre los sufijos gráficos. Al considerar las semejanzas entre clíticos y afijos, esto no debe causar extrañeza[18]. La Tabla 25 muestra aquellos segmentos que consisten únicamente de enclíticos y que aparecen dentro de las 600 formas más afijales.

[18] De hecho, no es raro que ya se hayan analizado los pronombres clíticos del español como marcas de flexión del verbo en concordancia con los complementos; véase el tercer capítulo de Rini, *Motives for Linguistic Change in the Formation of the Spanish Object Pronouns*, Juan de la Cuesta, Newark, Delaware (1992). También, véase la discusión sobre los pronombres personales del portugués como formas de flexión en Spencer, *op. cit.* (1991, 382): "if the pronoun forms really are inflections, it's hard to see what grammatical function they have".

Tabla 25. Enclíticos descubiertos como sufijos gráficos

enclíticos	**frecuencia**	**afijalidad**
~me	565	0.4744
~te	1,072	0.4769
~se	1,619	0.5332
~la	633	0.5306
~las	262	0.5171
~le	613	0.5260
~les	455	0.4973
~lo	792	0.5384
~los	410	0.5182
~nos	361	0.4643
~selo	7	0.4472

Nótese que, de todas las combinaciones de enclíticos posibles, solamente *~selo* (rango 598) ocurre en la Tabla 17. Lo importante es que todos los pronombres enclíticos están ahí. La forma correspondiente al pronombre *~os* del español de España no ocurre lo suficiente en el CEMC (*~os* ocurre como marca de nombre masculino en plural). Es interesante que los acusativos tengan en promedio mayor de afijalidad (0.52608) que los dativos (0.49535).

Como era de esperarse, la mayoría de los enclíticos en la tabla ocurrió adherida a sufijos que se utilizan para marcar gerundios, imperativos e infinitivos. La Tabla 26 muestra los sufijos del gerundio que ocurrieron con algún enclítico.

Tabla 26. Gerundio y enclíticos

gerundio + enclítico	frecuencia	afijalidad
~**a**ndome	48	0.5512
~**a**ndote	14	0.4973
~**a**ndose	260	0.6262
~**a**ndole	75	0.5417
~**a**ndoles	13	0.4637
~**a**ndolo	78	0.5640
~**a**ndolos	44	0.5324
~**a**ndola	58	0.5584
~**a**ndolas	22	0.5356
~**a**ndonos	18	0.4904
~i**e**ndose	68	0.4850
~i**e**ndolo	12	0.4709
~ndole	14	0.3988
~ndose	43	0.3986

El único paradigma completo es el de las formas con vocal temática de la primera conjugación. Destaca el gerundio con el pronombre *~se* (*~**a**ndose*) por ser el más afijal y más frecuente como cadena sufijada en toda la tabla. De la segunda y tercera conjugaciones solamente hay dos segmentos, a los que se adhieren los enclíticos *~se* y *~lo*. Las formas sin vocal temática también son sólo dos. Es significativo que los segmentos de los paradigmas incompletos tengan las afijalidades más bajas. También es de notarse que en general todas estas formas tengan frecuencias relativamente bajas como sufijos.

Tabla 27. Imperativo y enclíticos

~ar			**~er/~ir**		
~ame	74	0.5687	~eme	21	0.5052
~ate	107	0.5637	~ete	59	0.5398
~ese	143	0.5256	~ase	114	0.5925
~ala	30	0.5727	~ela	48	0.5296
~alas	14	0.5817	~elas	20	0.4819
~ale	48	0.5490	~ele	12	0.5112
~ales	281	0.5509	~eles	11	0.4673
~alo	45	0.5639	~elo	60	0.5194
~alos	32	0.5458	~elos	31	0.5301
~anos	53	0.5396	~enos	6	0.5210
plural					
~ense	30	0.5079	~anse	11	0.5291
~nse		7		0.3949	

En la Tabla 27 se listan los segmentos con sufijos de imperativo y enclíticos. Los paradigmas están completos. Así que se muestran sólo aquellos con un enclítico. Es de notarse, sin embargo, que algunas formas son homónimas de otros sufijos, especialmente derivativos; por ejemplo, la secuencia *~anos* de *campir**anos***.

Tabla 28. Infinitivo y enclíticos

~ar			**~er**			**~ir**			**~r**		
~arme	244	0.6306	~erme	48	0.4247	~irme	23	0.5322	~rme	202	0.4144
~arte	144	0.6045	~erte	18	0.4660	~irte	8	0.5465	~rte	83	0.4315
~arse	665	0.6692	~erse	105	0.5020	~irse	108	0.5659	~rse	786	0.4243
~arla	270	0.6350	~erla	22	0.5174	~irla	23	0.5209	~rla	93	0.3926
~arlas	139	0.6021	~erlas	6	0.4816	~irlas	8	0.5240	~rlas	18	0.3998
~arle	176	0.5997	[~erle]			~irle	12	0.4810	~rle	56	0.4228
~arles	72	0.5505	[~erles]			[~irles]			~rles	14	0.3824
~arlo	316	0.6356	~erlo	35	0.4729	~irlo	39	0.5127	~rlo	140	0.4128
~arlos	201	0.6153	~erlos	15	0.4823	~irlos	16	0.5180	~rlos	42	0.3711
~arnos	139	0.5917	~ernos	11	0.4868	~irnos	18	0.4576	~rnos	115	0.3932
~isar*se*	7	0.3989									
~**a**rse*lo*	10	0.4553									

Finalmente, la Tabla 28 contiene las ocurrencias de infinitivo con enclíticos. Como se ve, los paradigmas de la segunda y tercera conjugaciones no están completos. Curiosamente, las formas que faltan son del dativo. Si bien, la presencia del paradigma de las formas sin vocales temáticas cubre el espacio que dejan las formas faltantes, es de notarse que todos los paradigmas tienen valores bajos de afijalidad, especialmente aquellos con enclítico *~les*. De hecho, en las tablas anteriores se observa un poco de lo mismo, aunque mucho menos pronunciado. Por último, destaca la casi total ausencia de combinaciones de enclíticos. En la Tabla 25 sólo está la secuencia *~selo*, que es la misma que se encuentra en la cadena ***~arselo*** de la Tabla 28.

El promedio de afijalidad de los 87 segmentos que contienen algún enclítico es de 0.51. Esto es interesante al observar que las formas de

cerca de la mitad de los segmentos de la Tabla 17 tienen una afijalidad menor. Por otra parte, los sufijos de flexión tienen como promedio 0.53 y los derivativos 0.49. Entonces, según los criterios cuantitativos de afijalidad, los enclíticos del CEMC son menos afijales que los sufijos flexivos, pero más que los derivativos.

Esto no es un argumento para decir que los enclíticos sean algo intermedio, pero no hay razón para extrañarse de su promedio más bien alto de afijalidad. Si bien, en análisis previos ya han sido considerados marcas de flexión (aunque no exhiban otra función gramatical que la de servir de anáforas), la enclisis de pronombres se aleja mucho conceptualmente del fenómeno de derivación. Lo que indica este promedio es que los enclíticos, que gráficamente se adhieren a la palabra, son un grupo finito de pronombres con un número finito de maneras posibles de combinarse.

3.2.4 Hacia un catálogo de sufijos del español de México

Como se ve, hasta aquí se logró reunir la mayoría de los sufijos de flexión del español de México y un conjunto significativo de los derivativos. A partir de esto, se puede construir una caracterización completa y sistemática del conjunto de sufijos del español de México. Es importante notar que queda pendiente la tarea exhaustiva de examinar cada uno para determinar su comportamiento y su significado en cada contexto, mediante, por ejemplo, el uso de concordancias.

También resalta la necesidad de investigar el fenómeno de parasíntesis, es decir, la coocurrencia de ciertos prefijos y ciertos sufijos para formar vocablos nuevos[19]. Esto dependería naturalmente de una investigación completa del catálogo de prefijos cuantitativamente reconocibles a partir de un corpus.

[19] Véase una caracterización de construcciones parasintéticas en el español de México en Moreno de Alba, *op. cit.* (1996, 31-37).

Además, los sufijos derivativos descubiertos en el CEMC mediante el cálculo de afijalidades merecen examinarse todavía con más profundidad. A pesar de que la mayoría de las formas de la Tabla 17 se pueden reorganizar para mostrar su pertinencia dentro un subsistema morfológico flexivo y uno léxico derivativo, es necesario analizar lo que aparentemente no entra en ningún lado. Es decir, aunque se observan varios segmentos, que se reconocen como elementos con significado, que se adhieren a otros para formar nuevas palabras, hay muchos que sólo con un análisis más detallado, específicamente examinando sus contextos, podrán considerarse verdaderos morfemas del español.

3.3 Experimentos con corpus de otras lenguas

En esta sección se reportan algunas pruebas de extracción de afijos llevadas a cabo con muestras textuales del checo, ralámuri y chuj. En esencia, se examinan los resultados de algunos experimentos para descubrir grupos de afijos de estas lenguas, una lengua indoeuropea, de la rama eslava, y dos lenguas amerindias no emparentadas, de las familias yutoazteca y maya: el checo estándar, una variante de ralámuli o tarahumara y una variante del chuj[20]. Más importante aún, se busca comparar estos experimentos para evaluar sus resultados mediante medidas de utilizadas en disciplinas como extracción y recuperación

[20] Esta sección está basada en Medina-Urrea, "Affix Discovery by Means of Corpora: Experiments for Spanish, Czech, Ralámuli and Chuj" (2007, 277-299) en Mehler y Köhler, *Aspects of Automatic Text Analysis*, Springer (2007), en el que se reportan y actualizan los experimentos descritos en Medina y Buenrostro, "Características cuantitativas de la flexión verbal del chuj", *Estudios de Lingüística Aplicada*, 38 (2003, 15–31), Medina y Hlaváčová, "Automatic Recognition of Czech Derivational Prefixes", *Lecture Notes in Computer Science* 3406 (2005, 189–197) y Medina y Alvarado, "Un experimento de reconocimiento automático de la derivación léxica del ralámuli" (2009, 243-251), en Cuevas, ed., *La lengua y la antropología para un conocimiento global del hombre. Homenaje a Leonardo Manrique*, México: Instituto Nacional de Antropología e Historia.

de información, minería de datos y de textos, reconocimiento de patrones, etcétera.

De esta manera, en las próximas tres subsecciones, se presentan los catálogos de estas lenguas. Específicamente, se muestran catálogos de candidatos a prefijos del checo, sufijos derivativos del ralámuli y prefijos y sufijos de la flexión verbal del chuj. Por último, se describen los conceptos de precisión y recuperación comprensiva (*recall*) y se aplican en un ejercicio de evaluación comparativa de los resultados de estos experimentos.

3.3.1 Prefijos del checo

El método para medir afijalidad se aplicó a una lista de lemas extraída de un corpus de la lengua checa (Medina Urrea y Hlaváčová 2005). Específicamente, se utilizó la lista de 166,733 lemas, con frecuencia de ocurrencia mayor a cinco, del Corpus Nacional Checo (Český národní korpus 2005). Aunque la lengua checa es flexiva y el método se puede aplicar para descubrir tanto sufijos como prefijos, este experimento se enfocó en el sistema de prefijos, por ser el objetivo de investigación de Jaroslava Hlaváčová de la Universidad Carolina de Praga[21]. En la Tabla 29, aparecen los 35 prefijos con la mayor afijalidad de un total de 1,411 formas extraídas. Es significativo que sean los prefijos verbales los que encabecen la lista: *vy~, za~, od~, roz~, do~, pod~, u~*, entre los primeros 11. También se extrajeron otros prefijos verbales importantes (*o~*, *po~*, *pře~*, *před~*, *v~*, *z~*), pero por tener rangos mayores, quedaron fuera de esta tabla.

[21] De hecho, Hlaváčová estudia los patrones productivos de prefijos verbales de lenguas eslavas, como checo, eslovaco y ruso; véase "Productive Verb Prefixation Patterns", *The Prague Bulletin of Mathematical*, 101 (Hlaváčová y Nedoluzhko 2014, 111–122).

Tabla 29. Prefijos más prominentes del checo

	prefijo	frec.	cuadros	economía	entropía	afijalidad
1	vy~	2,391	1.0000	0.8845	0.8920	0.9255
2	za~	2,189	0.7820	0.8454	0.9102	0.8459
3	při~	1,361	0.7272	0.9103	0.8822	0.8399
4	od~	1,330	0.6319	0.8622	0.9270	0.8070
5	roz~	1,234	0.6114	0.8403	0.8634	0.7717
6	na~	2,129	0.6072	0.7568	0.9235	0.7625
7	do~	1,496	0.4379	0.7204	0.9331	0.6971
8	proti~	297	0.1315	0.9649	0.9851	0.6938
9	pod~	783	0.2589	0.7804	0.9871	0.6755
10	mimo~	85	0.0442	0.9970	0.9565	0.6659
11	u~	1,939	0.4492	0.6378	0.8977	0.6616
12	severo~	54	0.0296	1.0000	0.9072	0.6456
13	jiho~	52	0.0285	0.9906	0.9105	0.6432
14	osmi~	74	0.0776	0.9392	0.9014	0.6394
15	spolu~	151	0.0623	0.9441	0.9003	0.6356
16	ení~	241	0.1196	0.9260	0.8589	0.6348
17	mezi~	130	0.0561	0.9268	0.9154	0.6328
18	super~	200	0.0661	0.8615	0.9622	0.6299
19	nad~	224	0.0692	0.8408	0.9682	0.6261
20	pro~	2,205	0.3712	0.6071	0.8921	0.6235
21	troj~	87	0.0460	0.8589	0.9584	0.6211
22	video~	105	0.0441	0.9293	0.8889	0.6208
23	dvoj~	150	0.0603	0.8617	0.9399	0.6206
24	několika~	67	0.1017	0.9571	0.8029	0.6206
25	polo~	276	0.0831	0.8478	0.9289	0.6199
26	radio~	73	0.0274	0.9059	0.9240	0.6191
27	šesti~	113	0.0815	0.9299	0.8443	0.6186
28	pěti~	168	0.0848	0.8859	0.8845	0.6184
29	sebe~	198	0.0718	0.9017	0.8799	0.6178
30	dvou~	198	0.1137	0.8460	0.8885	0.6161

	prefijo	frec.	cuadros	economía	entropía	afijalidad
31	sedmi~	66	0.0655	0.9630	0.8165	0.6150
32	sky~	41	0.0298	0.9182	0.8920	0.6133
33	jedno~	167	0.0928	0.8353	0.9110	0.6130
34	více~	98	0.0761	0.8825	0.8800	0.6129
35	dvaceti~	41	0.0476	0.9602	0.8200	0.6093

Es de notarse que todas estas formas representan prefijos aislados; esto es, no son cadenas de prefijos. Hláváčova nota que dentro de las primeras 100 ocurren solamente dos que son secuencias de prefijos (*popo~* rango 83 y *zne~* rango 96). Además, no hay formas que no sean prefijos entre las primeras 100 entradas del catálogo, esto es el 100% de éstas se reconocen como prefijales. También buscamos en todo el catálogo de 1 411 elementos, los 45 prefijos más tradicionales del checo. Entre los primeros 100 ocurrió el 48% de ellos y entre los primeros 500 el 75%.

Cabe señalar que algunos candidatos no representan prefijos propiamente, sino que su adhesión a otras formas puede considerarse más bien un fenómeno de composición. Lo interesante es que, cuantitativamente, se comportan como prefijos: ocurren en muchas palabras gráficas modificando sus significados. Este es el caso, entre otros, de prefijos que pueden denominarse numéricos —un tipo de flexión que se aplica a los adjetivos (normalmente idéntica al genitivo). De hecho, todos los adjetivos numerales pueden prefijarse, aunque esto suele limitarse a números pequeños. Por ejemplo, en la expresión *sedmihlavý drak* "un dragón que tiene siete cabezas" el prefijo *sedmi~* (rango 31) corresponde a "siete".

Hlaváčová y Hrušecký (2008) desarrollaron una herramienta que llamaron *Affisix* especializada en el descubrimiento de prefijos y basa-

da en las medidas de entropía y de cuadros[22]. Estos mismos autores llevaron a cabo experimentos para comparar algunos métodos para el reconocimiento de prefijos del inglés y del checo donde observan que los resultados son mejores al utilizar listas de lemas y no las palabras gráficas de una muestra textual[23]. Además, tuvieron buenos resultados al buscar inicios de vocablos que, al ser eliminados, dejaran una palabra en la lista de lemas. Encontraron también que la entropía funcionó muy bien, pero que las medidas de economía y cuadros no tuvieron tan buenos resultados (Hlaváčová y Hrušecký 2011, 241). Sin duda, su trabajo contribuye a distinguir las diferencias entre prefijos y sufijos, en relación con las medidas aplicadas en este trabajo.

3.3.2 Sufijos derivativos del ralámuli

Como se anunció arriba, el método también se aplicó a la lengua ralámuli[24]. Esta lengua, también conocida como rarámuri o tarahumara, es una lengua yutoazteca que se habla en el norte de México. La muestra textual utilizada representa, como se dijo, la variante dialectal de San Luis Majimachi, Bocoyna, Chihuahua[25]. Para los estándares de hoy en día, se trata de una muestra minúscula, que contiene apenas 3 584 palabras gráficas y 934 tipos de palabras. No se puede asumir que esta mues-

[22] Los experimentos de la aplicación de esta herramienta al checo, así como de la descripción de la herramienta misma están en "Affisix: Tool for Prefix Recognition", en Sojka, Horák, Kopeček, Pala, eds., *Text, Speech and Dialogue. TSD 2008. Lecture Notes in Computer Science, 5246*, Springer (Hlaváčová y Hrušecký 2008, 85-92).

[23] Véase Hlaváčová y Hrušecký, "Prefix Recognition Experiments" en Habernal y Matoušek, eds., *Text, Speech and Dialogue. TSD 2011. Lecture Notes in Computer Science 6836*. Springer (2011, 235-242).

[24] Los resultados se presentaron en el Primer Coloquio Leonardo Manrique (septiembre, 2004) en la ponencia "Análisis cuantitativo y cualitativo de la derivación léxica en ralámuli" y en el artículo Medina y Alvarado, art. cit. (2009, 243-251).

[25] Se trata de diferentes escritos de Patricio Parra (2003) en los que muestra aspectos de la cultura ralámuli, que van desde reflexiones a manera de ensayos hasta cuentos, la mayoría narrados en pasado.

tra sea suficientemente representativa de la lengua, pero la aplicación de este método nos sirve para explorar qué tan apropiado es el método para muestras pequeñas.

En la Tabla 30 aparecen las 35 formas más afijales. A pesar de que el ralámuli tiene relativamente pocas formas flexivas, el catálogo completo exhibe más cadenas con material flexivo de lo esperado. Esto es porque en esta lengua, como en español, la formación de palabras se lleva a cabo principalmente mediante sufijación y los sufijos derivativos ocurren seguidos de los flexivos.

Tabla 30. Sufijos más prominentes del ralámuli

	sufijo	frec.	cuadros	economía	entropía	afijalidad
1	~ma	35	1.0000	1.0000	0.8873	0.9624
2	~re	72	0.7257	0.7150	0.8559	0.7655
3	~sa	33	0.5644	0.7705	0.7620	0.6990
4	~si	27	0.6944	0.5264	0.8444	0.6884
5	~ra	62	0.5927	0.5593	0.8576	0.6699
6	~na	25	0.3650	0.6165	0.8048	0.5954
7	~go	4	0.1875	0.7913	0.6545	0.5444
8	~é	46	0.1277	0.3356	1.0000	0.4878
9	~ga	49	0.2449	0.3653	0.8154	0.4752
10	~á	60	0.1333	0.3054	0.9418	0.4602
11	~ame	50	0.2250	0.2728	0.8720	0.4566
12	~gá	17	0.3824	0.3536	0.6152	0.4504
13	~ka	19	0.2368	0.2525	0.8481	0.4458
14	~a	279	0.0945	0.1626	0.9833	0.4135
15	~ba	8	0.2344	0.1846	0.7460	0.3883
16	~ire	13	0.2019	0.2029	0.7457	0.3835
17	~áame	11	0.1023	0.1279	0.8696	0.3666
18	~í	38	0.0724	0.1990	0.8239	0.3651

	sufijo	frec.	cuadros	economía	entropía	afijalidad
19	~či	39	0.0897	0.2402	0.7460	0.3586
20	~yá	17	0.3456	0.1758	0.5438	0.3551
21	~e	157	0.1346	0.2465	0.6645	0.3485
22	~ayá	6	0.1250	0.4396	0.4692	0.3446
23	~mi	4	0.0625	0.2638	0.7047	0.3437
24	~ré	9	0.1389	0.2032	0.6683	0.3368
25	~né	3	0.0833	0.3517	0.5418	0.3256
26	~o	41	0.0000	0.0000	0.9759	0.3253
27	~ira	11	0.0568	0.0959	0.8021	0.3183
28	~i	137	0.0274	0.0757	0.8515	0.3182
29	~ča	6	0.0625	0.1758	0.7108	0.3164
30	~ri	25	0.1200	0.1515	0.6603	0.3106
31	~wa	9	0.0417	0.1172	0.7460	0.3016
32	~ne	3	0.1250	0.3517	0.4228	0.2998
33	~áa	13	0.0385	0.0000	0.8543	0.2976
34	~bo	10	0.2125	0.0000	0.6802	0.2976
35	~íre	6	0.2500	0.0000	0.6303	0.2934

Con base en su experiencia de trabajo de campo y teniendo en cuenta el trabajo de otros expertos, Maribel Alvarado determinó los 35 sufijos derivativos nominales y verbales más destacados de esta lengua. 25 de éstos ocurrieron dentro de las primeras 100 entradas del catálogo (71%). Las otras entradas son cadenas de sufijos (que, como se dijo, incluyen secuencias de elementos derivativos y flexivos) y formas residuales.

Examinar las formas residuales fue difícil. Surgieron preguntas sobre la fosilización de ciertos segmentos y sobre la relación entre la estructura de la sílaba y la afijalidad de otras formas, cuestiones que expertos en ralámuli deben estudiar con detenimiento. Por ejemplo, al tomar en cuenta la estructura canónica silábica del ralámuli (CV), notamos que hay formas sufijales VCV que deben ser estudiadas por especialistas antes de ser aceptadas o rechazadas como verdaderos afijos. Además,

hace falta un análisis que nos permita establecer si la vocal inicial es un morfema, es parte de un sufijo o es parte de la raíz que la precede (*~ame*, *~ayá*, *~irá* e *~íre*, rangos 11, 22, 27 y 35). La información sincrónica con la que contamos no es suficiente para descartar rastros de elementos afijales lexicalizados de la diacronía. Por el momento, para fines de evaluación, las entradas con estructura silábica inesperada no se contaron propiamente como sufijos ni cadenas de ellos.

Hay 10 sufijos derivativos que no aparecieron representados en el catálogo (*~lo*, *~ni*, *~pu* [*~bu*], *~tu*, *~to*, *~pu*, *~tu*, *~repu*, *~bu*, *~bona*). Se trata de elementos de derivación verbal, o modificadores de transitividad o alguna característica semántica de las formas verbales. Esto significa que la pequeña muestra utilizada es más representativa de las estructuras nominales que de las verbales. Como sea, vale la pena enfatizar que una parte significativa del sistema de derivación conocido del ralámuli, esencialmente el subsistema nominal, se recuperó de un conjunto de textos muy pequeño, que apenas constituye una muestra textual de esta lengua.

Finalmente, aunque se trata de una lengua con relativamente poca flexión, encontramos en los resultados, como ya se dijo, mucho material flexivo afijado al derivativo. Esto se debe al hecho de que los textos de entrada son narraciones, por lo que las palabras aparecen necesariamente flexionadas. Así que las investigaciones enfocadas al descubrimiento de la derivación deberán utilizar las listas de lemas de un diccionario. También, sería interesante construir un lematizador para esta lengua que permita eliminar los afijos de flexión y así enfocarse en la derivación.

De todas maneras, si hemos de concebir los sufijos de flexión como unidades más afijales que los de derivación, encontramos que esta intuición se corrobora en la Tabla 30 con la presencia de los cuatro morfemas de flexión más importantes del tarahumara encabezando la lista: *~ma*, *~re*, *~sa*, y *~si* que, como apunta Alvarado, son marcas de flexión de tiempo, aspecto y modo (Medina Urrea, Herrera Camacho y Alvarado García 2009, 252).

3.3.3 Prefijos y sufijos de flexión verbal del chuj

El chuj es una lengua maya que se habla en ambos lados de la frontera entre México y Guatemala. El procedimiento de descubrimiento de afijos también se aplicó a una muestra textual muy pequeña de esta lengua con 15 485 palabras gráficas y poco más de 2 300 tipos de palabra. Esta muestra fue compilada por Elsa Cristina Buenrostro (2002) en diversas estancias de trabajo de campo en el estado de Chiapas. Se trata de una colección de cinco narraciones, todas con intercambios conversacionales. Aunque es una muestra más grande que la del ralámuli, tampoco se puede considerar un corpus equilibrado y representativo de la lengua. Sin embargo, los resultados son interesantes porque su sistema de flexión verbal está constituido tanto por prefijos como sufijos y porque Buenrostro, dados sus intereses gramaticales[26], puso especial énfasis en compilar una colección de textos representativos de las estructuras verbales.

Uno de los objetivos de este experimento fue determinar si, con tan pocos datos, por lo menos la morfología flexiva del chuj podía descubrirse. Los resultados fueron alentadores[27]. Alrededor del 86% de los sufijos y prefijos de flexión ocurrieron entre los más afijales de ese pequeño corpus: 3 prefijos de tiempo, 4 pronombres absolutivos prefijados, 11 ergativos también prefijados y 7 sufijos de voz pasiva y antipasiva, uno de modo, otro de tiempo y dos vocales que marcan final de frase. Además, todos los afijos de flexión identificados ocurrieron apretados en los primeros 200 lugares de los dos catálogos, uno de prefijos y el otro de sufijos.

Aunque no todas las formas son de flexión verbal, en la evaluación se consideraron aciertos todas las formas reconocidas como afijales. Se obtuvieron formas afijales de todos tipos (afijos y secuencias de afijos,

[26] Buenrostro es autora del volumen *Chuj de San Mateo Ixtatán*, Archivo de lenguas indígenas, El Colegio de México, México (2009); véase también *La voz en Chuj de San Mateo Ixtatán*, El Colegio de México. Tesis doctoral (2013).

[27] Estos resultados se presentaron en el VII Encuentro Internacional de Lingüística en el Noroeste (noviembre, 2002) en la ponencia "Características cuantitativas de la morfología flexiva del chuj" y en el artículo Medina y Buenrostro, art. cit. (2003).

nominales y verbales, derivativos y flexivos) y también formas residuales. Como ya se ha observado, no hay una frontera nítida entre las afijales y las residuales. Lo importante es que las formas no verbales y no flexivas no dejan de constituir partes del componente afijal del chuj.

3.3.3.1 Catálogo de prefijos del chuj

En la Tabla 31, se consignan los 30 fragmentos de palabra gráficas más prefijales según los métodos descritos arriba. Como antes, están ordenados por cantidad de afijalidad, de más a menos:

Tabla 31. Prefijos más prominentes del chuj

	prefijo	frec.	cuadros	economía	entropía	afijalidad
1	ix~	180	1.0000	0.8000	0.9198	0.9066
2	in~	94	0.5278	0.8278	1.0000	0.7852
3	tz~	358	0.7286	0.6143	0.8976	0.7468
4	s~	186	0.3871	0.6517	0.9774	0.6721
5	ko~	71	0.3064	0.6710	0.8742	0.6172
6	ol~	190	0.4844	0.5276	0.8163	0.6094
7	w~	74	0.4338	0.7204	0.5346	0.5629
8	xsčí'~	1	0.0429	1.0000	0.5202	0.5210
9	tzin~	47	0.1708	0.4278	0.9512	0.5166
10	a~	165	0.1687	0.4051	0.9594	0.5111
11	olin~	26	0.1775	0.4572	0.8982	0.5110
12	y~	128	0.3488	0.5739	0.5830	0.5019
13	ay~	31	0.2198	0.5414	0.7003	0.4872
14	olač~	26	0.1747	0.4944	0.7805	0.4832
15	ač~	11	0.0571	0.4247	0.8534	0.4451
16	tzs~	49	0.0776	0.4267	0.8252	0.4432
17	ixin~	28	0.1199	0.3443	0.8335	0.4326

	prefijo	frec.	cuadros	economía	entropía	afijalidad
18	tzonh~	17	0.0782	0.2754	0.8365	0.3967
19	ak'~	12	0.0321	0.3172	0.8027	0.3840
20	k'a~	11	0.0156	0.3636	0.7591	0.3794
21	al~	15	0.0352	0.2123	0.8902	0.3792
22	e~	63	0.0304	0.1663	0.9240	0.3736
23	ma~	31	0.0115	0.2202	0.8884	0.3734
24	x~	44	0.0601	0.2883	0.7708	0.3731
25	b'ati~	1	0.0286	0.7500	0.2601	0.3462
26	tzač~	11	0.0701	0.3610	0.6039	0.3450
27	ixs~	24	0.0274	0.1688	0.8306	0.3423
28	olonh~	5	0.0229	0.4250	0.5715	0.3398
29	yak'~	12	0.0714	0.1797	0.7667	0.3393
30	k'e~	3	0.0476	0.2500	0.7153	0.3376

Es de notarse que entre las primeras 6 entradas ocurren todos los prefijos temporales del paradigma verbal *ix~, tz~, ol~* (rangos 1, 3 y 6) y *x~* (rango 24, 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~, ač~* (rangos 10, 4, 2, 22, 15) y los alomorfos *ko~* y *ku~* (rangos 5 y 45, este último fuera de la lista). 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.ač~* y *tz.a~* (rangos 9, 11, 17, 27, 16, 18, 55, 14 y 38). De hecho, si los personales aparecen como prefijos aislados, es porque los temporales alternan con ø~. Un prefijo interesante es *ma~* (rango 23), el que sirve para negar oraciones; su 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~* (rango 37, fuera de la tabla) que también ocurre libre, sirve entre otras cosas para introducir oraciones subordinadas.

Otro grupo de formas de la Tabla 31 no menos importante es el de aquellas que no contaron como aciertos, porque no representan exactamente algún prefijo verbal, sino que muestran material adicional, probablemente parte de la raíz; por ejemplo, *xsči'~* (rango 8), de sólo una ocurrencia como mejor prefijo de una palabra, *k'a~* (rango 20), *yak'~* (rango 29) y *k'e~* (rango 30). Otras formas son probablemente parte de un prefijo o raíz nominales o son material residual: *ay~* (rango 13) y *b'ati~* (rango 25). Todas estas formas constituyen un reto interesante para el lingüista, sobre todo si los mismos ocurren en muestras de mayor tamaño. Por un lado, como los errores en este tipo de trabajo son inevitables, la naturaleza del error debe examinarse con cuidado. Por el otro, siempre habrá formas que parecen errores, pero que conviene estudiar para determinar si podrían considerarse, aunque fuera incipientemente, como un tipo de morfema.

En el caso de este experimento, el corpus es muy pequeño, por lo que la presencia de residuos no causa extrañeza. De todas maneras, al tomarlos como errores, se puede calcular la proporción de aciertos de la Tabla 31: 24 ÷ 30 = 0.8. Por lo que la proporción de formas residuales es de 20%. En un corpus tan pequeño como el utilizado, los residuos son inevitables debido a la escasez de datos.

3.3.3.2 Catálogo de sufijos del chuj

Con respecto a los sufijos y grupos sufijales reunidos en este experimento, en la Tabla 32 se muestran las 30 formas más importantes según los criterios del método aplicado. El primer grupo es el de las formas con una vocal temática *~a* e *~i* (rangos 17 y 6) que permiten distinguir los verbos transitivos de los intransitivos y que, además, indican el final de la frase. Otro grupo interesante es el de los sufijos *~ok* y *~nak* (rangos 5 y 16) que son respectivamente marcas de modo y tiempo. En el habla ocurren en distribución complementaria, 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 32 no aparece ninguno, pero en el catálogo aparecen, con rangos mayores de 30, *~čaj* y *~aj* (rangos 67 y 63). Los otros miembros de este paradigma, *~nax* (rango 1037) y *~b'il* (rango 886), no ocurren solos entre los primeros 290 grupos sufijales. El sufijo *~ji* aislado no fue detectado en este experimento. Sin embargo, estos sufijos sí ocurren en grupos sufijales tales como *~a.ji* y *~ak'.nax* (con rangos 105 y 256). Estas dos formas tienen baja frecuencia como mejores afijos (13 y 2 respectivamente), por lo que su carácter afijal es cuestionable; de hecho, como veremos, *~ak'* es una raíz que significa "dar".

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 32 (rango 15). Los otros dos ocurren después (*~wi* con rango 41 y *~waj* con rango 165 y frecuencia de 3). El sufijo *~in* (rango 21) 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 muy 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 debe sorprender que morfemas tan polisémicos obtengan valores altos de afijalidad. De hecho, es de esperarse que las formas más frecuentes sean las más polisémicas: a mayor número de contextos, más sentidos.

Con respecto a los sufijos *~al* e *~il* (rangos 8 y 9), 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*, *~b'at*, *~el*, *~em*, *~ek'* (rangos 1, 4, 13, 29, 30) y, fuera de la tabla, *~k'o*č (rango 31) y *~pax* (rango 38). 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.

Tabla 32. Sufijos más prominentes del chuj

	sufijo	frec.	cuadros	economía	entropía	afijalidad
1	~kan	68	1.0000	0.9516	0.8832	0.9449
2	~nhej	23	0.4138	1.0000	0.7550	0.7229
3	~ta’	70	0.6164	0.7297	0.7894	0.7118
4	~b’at	63	0.5738	0.6282	0.8279	0.6766
5	~ok	68	0.4973	0.5433	0.9224	0.6543
6	~i	198	0.6061	0.4731	0.8056	0.6283
7	~xi	37	0.5215	0.6840	0.6680	0.6245
8	~al	84	0.3175	0.5109	1.0000	0.6095
9	~il	63	0.3892	0.4812	0.9288	0.5997
10	~ač	18	0.3786	0.7299	0.6647	0.5911
11	~kot	48	0.4167	0.5348	0.7223	0.5579
12	~ab’	50	0.3148	0.4972	0.8552	0.5557
13	~el	69	0.2507	0.4901	0.8714	0.5374
14	~ik’	12	0.5154	0.4701	0.5884	0.5246
15	~an	234	0.3398	0.3592	0.8074	0.5021
16	~nak	18	0.3354	0.4047	0.7641	0.5014
17	~a	144	0.1908	0.3762	0.9103	0.4924
18	~ila	9	0.3374	0.5282	0.5584	0.4747
19	~tak	19	0.1598	0.3697	0.8864	0.4720
20	~ab’i	9	0.2551	0.5294	0.6292	0.4712
21	~in	47	0.1836	0.3297	0.8722	0.4618
22	~alan	13	0.2222	0.3566	0.8046	0.4611
23	~ilan	7	0.5344	0.1147	0.7073	0.4521
24	~ala	9	0.3045	0.3658	0.6807	0.4503
25	~kani	8	0.1620	0.4732	0.7010	0.4454
26	~ni’	7	0.1693	0.6023	0.5544	0.4420
27	~ak’kan	11	0.3300	0.3233	0.6649	0.4394
28	~ak’	43	0.1860	0.3544	0.7749	0.4384
29	~em	16	0.2708	0.2076	0.8101	0.4295
30	~ek’	23	0.1755	0.2228	0.8753	0.4245

En medio de la tabla está el sufijo *~nak* (rango 16), participio de verbos intransitivos; y fuera de ella están *~e* (rango 77), clasificador numeral de inanimados, y *~oj* (rango 71), marca de infinitivo en oraciones de complemento. Los sufijos con carácter adverbial son *~ta'* (rango 3) que significa "inmediata o recientemente", *~alan* (rango 22) "debajo" y *~nhej* (rango 2) "sólo".

Algunos resultados que no aparecen en las primeras entradas de la Tabla 32 no son propiamente morfemas del chuj, pero tampoco son propiamente errores, porque son formas que ocurren al final de los préstamos españoles muy abundantes y profusos en el corpus: *~o* (rango 42) en 'remedio', 'konejo', 'ciento', 'puro', 'ejersito', 'exodo', 'templo', 'cuatro', 'bueno', 'mismo', 'pero', 'San Pransisko', etc.; y *~es* (rango 62) en 'tres', 'tonces', 'entonces', 'despues', 'jues'. Lo interesante es que, desde el punto de vista cuantitativo, en varias palabras estas formas tienen carácter morfológico (sobre todo *~o*). Similarmente, la ocurrencia de los préstamos españoles terminados en *a* (como 'semana', 'pena', etc.) debe haber contribuido a que el sufijo de la vocal temática *~a* obtuviera el rango 17. Como sea, la discusión sobre si deben considerarse o no sufijos del chuj será seguramente de interés para los especialistas. Por lo pronto, aquí no los consideraremos afijos del chuj, pero tampoco los consideraremos errores: si tienen relaciones afijales con objetos de un corpus, no pueden descartarse como parte del sistema implícito en ese corpus.

También, se aprecian algunos verbos de movimiento en la Tabla 32 que funcionan como clasificadores verbales: *~kan*, *~b'at* y *~el* (rangos 1, 4 y 13). No está el paradigma completo, pero, como se trata de formas más derivativas que flexivas, deben ser sufijos muy productivos para competir con la flexión verbal en los primeros lugares del catálogo.

Otra cosa que se observa en la Tabla 32 es que varias formas contienen verbos. Así, el sufijo adverbial *~alan* (rango 22), arriba citado, tiene la misma forma que la secuencia *~al.an*, donde *~al* significa "decir". Por otra parte, *~ak'* (rango 28), que también ocurre en *~ak'.an* (rango 27), significa "dar". La forma *~čam* (rango 50) es la raíz de "matar" y forma

verbos compuestos con significados como "golpear" y "acabar". Con rango menor están *~tak* (rango 19) que es la raíz del verbo "aceptar" y *~ab'* (rango 12) que es la raíz del verbo "oír" y suele utilizarse como sufijo citativo. Por último, está la forma residual *~ek* (rango 86) sin un valor morfológico evidente.

Al identificar todos estos sufijos podemos calcular la proporción de aciertos en la Tabla 32: 29 aciertos ÷ 30 = 0.97, lo que corresponde a un porcentaje de residuos del 3%[28]. Además, al tomar en cuenta las dos tablas (Tablas 31 y 32), obtenemos un índice de aciertos de 0.88; esto es, 53 aciertos dentro de las 60 formas más afijales. Así, el porcentaje total de residuos es de 12%, cuestión nada desalentadora al considerar el tamaño del corpus.

3.3.3.3 Los paradigmas de la flexión verbal del chuj

Una manera de evaluar el carácter afijal de los fragmentos de palabras es identificar aquellos que pertenecen a los paradigmas de flexión verbal de la lengua estudiada, aunque no aparezcan dentro de los 30 más prefijales y los 30 más sufijales. En el caso del chuj, lo importante es verificar que lo seleccionado corresponda con sus morfemas más afijales. También es importante identificar lo que no se seleccionó, pero debió haber sido seleccionado e incluido entre el catálogo, por su ya conocido carácter morfológico.

Como se mencionó, en el chuj hay prefijos y sufijos de flexión verbal. En la Tabla 33 se exhiben los prefijos y en la Tabla 34 los sufijos. Ambas tablas son propuestas de Cristina Buenrostro y están basadas en su experiencia y en sus investigaciones sobre la flexión verbal de esta lengua. 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. En las dos últimas columnas están los pronombres ergativos.

[28] Si los sufijos no flexivos y las formas españolas se consideraran como errores, tendríamos 16 de 30 "aciertos" (esto es, un porcentaje de 53% y 47% de residuos). De todas maneras, este porcentaje de aciertos tampoco se ve mal, dadas las limitaciones del experimento.

La columna de la derecha contiene los pronombres que se adhieren a raíces con vocal inicial (*w~*, *y~*, *k~*, etc.) y la de la izquierda aquellos que se adhieren a las que inician con consonante (*in~*, *a~*, *s~*, etc.).

Los tres primeros renglones muestran las formas del singular y los tres últimos las del plural. Las formas *ko~* y *ku~* son alomorfos. Todos estos prefijos ocurrieron en los resultados de la extracción con el rango que se muestra antes de cada prefijo (el número entero a su izquierda); después de cada prefijo (a su derecha) aparece su valor de afijalidad. Los que no aparecen en la Tabla 31 ocurrieron con un rango mayor de 30 en el catálogo completo.

Tabla 33. Paradigma de prefijos de flexión verbal del chuj

tiempo			**persona**	**persona gramatical**								
				absolutiva			**ergativa**					
3	tz~	0.7468	**1ª**	2	in~	0.7852	2	in~	0.7852	7	w~	0.5629
1	ix~	0.9066	**2ª**	15	ač~	0.4451	10	a~	0.5111	—	ø~	—
24	x~	0.3731	**3ª**	—	ø~	—	4	s~	0.6721	12	y~	0.5019
6	ol~	0.6094										
—	ø~	—	**1ª**	36	onh~	0.2977	5	ko~	0.6172	49	k~	0.2601
							45	ku~	0.2707			
			2ª	369	ex~	0.0867	22	e~	0.3736	143	ey~	0.1779
			3ª	—	ø~ eb’	—	4	s~ eb’	0.6721	12	y~ eb’	0.5019

Por otra parte, el sistema de sufijos de flexión verbal del chuj, propuesto por Buenrostro, aparece en la Tabla 34. También ahí, antes de cada sufijo se muestra su rango en el catálogo completo de sufijos y, después, su valor de afijalidad. Los sufijos verbales del chuj marcan especialmente voz, modo y final del enunciado. Como se indicó arriba, las vocales temáticas de la última columna sirven para distinguir verbos transitivos de intransitivos y marcan el final de la frase. En esta tabla se

aprecia que faltó una de las marcas de voz pasiva (*~ji*), esto es, que no apareció en el catálogo. Esto significa que se descubrieron automáticamente 11 de las 12 formas sufijales posibles (91.67%).

Tabla 34. Paradigma de sufijos de flexión verbal del chuj

	voz			modal/temporal			vocal temática		
pasiva	63	~aj	0.2960						
	67	~čaj	0.2891	5	~ok	0.6543	6	~i	0.6283
	886	~b'il	0.0790	16	~nak	0.5014	17	~a	0.4924
	1037	~nax	0.0619						
	—	**~ji**	—						
antipasiva	15	~an	0.5021						
	41	~wi	0.3763						
	165	~waj	0.1715						

Finalmente, si tomamos los dos grupos de afijos de flexión como uno solo, vemos que el procedimiento automático permitió aislar 31 de 32. Esto indica que 96.88% de los afijos de flexión verbal esperados de la palabra chuj ocurrió entre los fragmentos más afijales de los catálogos extraídos del corpus.

3.4 Una evaluación con medidas de *precisión y recuperación comprensiva*

Evidentemente, es un reto evaluar los resultados de la aplicación de cualquier método para descubrir grupos de afijos tan diferentes, pertenecientes a lenguas tan heterogéneas, representadas en muestras textua-

les tan disparejas. De todas maneras, conviene considerar algún esquema para comparar y evaluar los resultados, a pesar del tamaño reducido de dos de estas muestras. Debido a la necesidad de simplicidad, la evaluación que se examina a continuación se enfoca en los conjuntos de afijos encontrados en los experimentos presentados hasta aquí, ya sean de flexión o de derivación, prefijales o sufijales.

Recuérdese que cada muestra textual o corpus utilizado para compilar los catálogos de afijos fue el objeto de investigación de experimentos específicos. Además, en cada experimento hay dos conjuntos de afijos: el grupo A y el grupo B que se describen de la siguiente manera:

A) el grupo de candidatos a afijos extraídos de un corpus —el catálogo de afijos— y
B) el grupo de afijos que se conocen *a priori* y que esperamos encontrar entre los resultados de la extracción —los afijos que conoce el hablante o especialista y que queremos que el método descubra.

Estos grupos están representados en la Figura 9. Evidentemente, la mejor evaluación requeriría que fueran iguales (A = B). Como eso raramente será el caso, se necesita que su intersección ($A \cap B = C$) sea lo más grande posible; esto es, que la mayor parte de B ocurra en A.

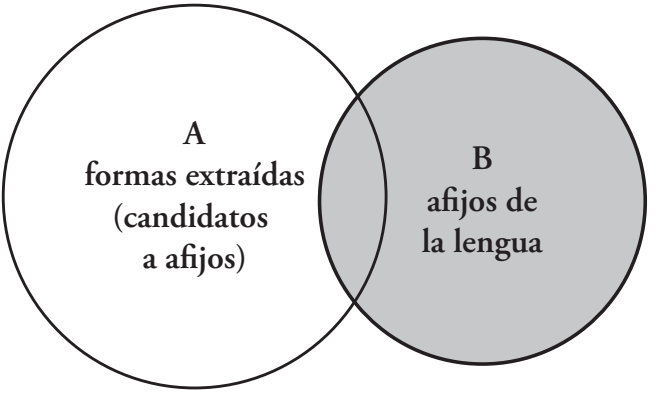


Figura 9. Las formas extraídas y los afijos de la lengua

En relación con el conjunto B, para conocerlo es necesario recurrir al conocimiento de la tradición lingüística —en los experimentos presentados, el conocimiento sobre la flexión verbal del español y sobre los prefijos tradicionales del checo— o apelar al conocimiento e intuición de hablantes, especialistas y sus investigaciones —los sufijos del ralámuli y los paradigmas de flexión verbal del chuj.

A partir de estos conjuntos, se pueden calcular las medidas de *precisión* y *recuperación comprensiva* o *recall*[29], de las que nos ocuparemos a continuación. Primero, la *precisión* representa la proporción de formas afijales detectadas automáticamente, respecto del total de candidatos extraídos del corpus. En otras palabras, toma en cuenta cuántos afijos "correctos" ocurren en el catálogo.

De esta manera, la precisión se calcula $\frac{|C|}{|A|}$, que es el tamaño del conjunto C (esto es, A ∩ B) dividido entre el tamaño del conjunto A. Es fácil ver que esto corresponde a una buena evaluación cuando la precisión se acerca a 1 y a una mala cuando se acerca a 0 (véase la Figura

[29] Se trata de medidas estándar para evaluar muy diversos métodos de reconocimiento de patrones, minería de textos, procesamiento del lenguaje natural, etc. (Jurafsky y Martin 2009, 455); véase también Manning y Schütze, *op. cit.* (1999, 268-269).

10). Además, los elementos de A que no están en C serán los residuos en el catálogo y su cantidad será |A| – |C|.

Figura 10. Proporción de afijos descubiertos en formas extraídas

Por otra parte, la segunda medida es la de *recall* o recuperación comprensiva. Ésta representa cuánto del conjunto de afijos buscado se encuentra entre los candidatos encontrados; en otras palabras, se trata de la proporción de afijos esperados (B) que ocurren en el catálogo (A). En la Figura 11, se pueden ver los tres conjuntos involucrados. Para una buena evaluación, queremos que el conjunto C (esto es, A ∩ B) incluya la mayor parte de los elementos del conjunto B. De esta manera, la recuperación comprensiva será en este trabajo la proporción de elementos de B que ocurrieron en C (esto es, $\frac{|C|}{|A|} = \frac{|A \cap B|}{|B|}$). Así, mientras que la precisión cuantifica aciertos versus errores, la recuperación comprensiva contrasta aciertos y omisiones.

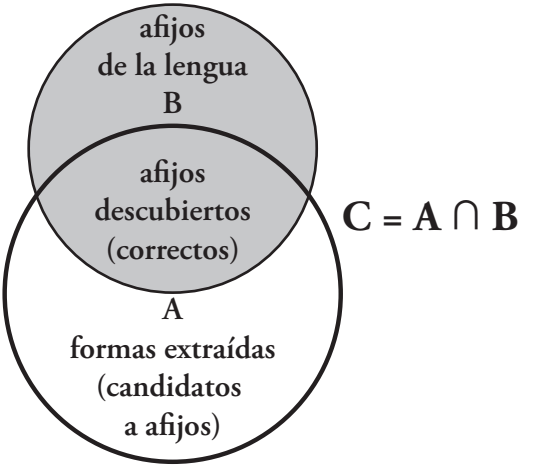


Figura 11. Afijos de la lengua y afijos descubiertos

En el contexto de una evaluación de los catálogos presentados, hay que notar que las muestras textuales son de dimensiones diferentes, el número esperado de afijos de cada lengua varía mucho y los tamaños de los catálogos de candidatos a afijos obtenidos son distintos. Además, son varios los tipos de afijos que compiten por aparecer en las primeras entradas de cada catálogo (nominales, verbales, etc.). Por ejemplo, si buscamos en el catálogo los afijos de flexión verbal, vemos que aparecen intrincados con los de flexión nominal y con los afijos derivativos más productivos. Así que en este experimento cuentan como aciertos, puesto que, desde el punto de vista de su afijalidad, no son propiamente errores.

Además, para medir la precisión utilizaremos una ventana de 50 entradas para cada catálogo. Se seleccionó 50 porque corresponde aproximadamente al promedio de tamaños de los 5 conjuntos de afijos enfocados en esta evaluación (véase el cuarto renglón de la Tabla 35). La proporción de aciertos se determina con esta ventana. De hecho, la precisión será el porcentaje de aciertos dentro de las primeras 50 entradas

de cada catálogo[30]. Con todo esto en mente, la precisión se calcula de la siguiente manera:

$$\text{Precisión} = \frac{\text{Número de afijos correctos dentro de las primeras 50 entradas}}{50}$$

Por otra parte, podemos caracterizar la medida de recuperación comprensiva con la siguiente fórmula:

$$\textit{Recall} = \frac{\text{Número de afijos conocidos } \textit{a priori} \text{ que aparecen en el catálogo}}{\text{Tamaño del conjunto de afijos conocidos } \textit{a priori}}$$

Respecto al tamaño del conjunto de afijos que se espera encontrar en cada experimento (|B|), cada uno de los especialistas de las lenguas examinadas tendrá que reflexionar sobre él, lo cual seguramente variará entre especialistas y entre hablantes. De hecho, para esta evaluación se les pidió a los especialistas contar separadamente alomorfos, homógrafos y formas polisémicas y no tomar en cuenta morfemas nulos. Además, en el caso del español, las flexiones en desuso (formas verbales informales de 2ª persona plural) y los grupos sufijales con enclíticos no se tomaron en cuenta.

En este contexto, examinemos la Tabla 35, que muestra datos generales sobre los corpus y las muestras textuales[31]. Las medidas de precisión y recuperación comprensiva aparecen al final, antes de las notas explicativas al pie de la tabla.

[30] Como se verá, el uso de ventanas de menor tamaño promueve valores de precisión más altos y valores de recuperación comprensiva más bajos, mientras que el uso de ventanas mayores corresponde con valores menores de precisión y valores mayores de recuperación comprensiva. Necesariamente, la recuperación comprensiva se reducirá al contar con una ventana de menor tamaño, porque los miembros más raros y menos productivos del conjunto de afijos examinado quedan fuera, en ventanas más pequeñas, mientras que los miembros más productivos de otros conjuntos de afijos también compiten por los primeros lugares del catálogo. Por otra parte, la precisión será menor al optar por una ventana de mayor tamaño porque, al crecer la ventana, las formas buscadas —las más raras y menos productivas— aparecerán junto con errores o formas residuales.

[31] Esta tabla es una actualización de aquella en Medina-Urrea, art. cit. (2007, 296).

Tabla 35. Resumen de los experimentos de segmentación afijal

	español	**checo**	**ralámuri**	**chuj**	
conjunto de afijos (B) que se busca en el catálogo	flexión verbal	prefijos	sufijos derivativos	flexión verbal prefijada	flexión verbal sufijada
tamaño de la muestra textual $\|\Psi\| = x$	1 891 045	166 733	3 584	15 485	
tipos de palabra en la muestra textual $\|V\| = \Omega$	79 000		934	2 300	
tamaño, *n*, del conjunto de afijos que se espera encontrar[a] (|B|)	163	45	35	20	12
afijos recuperados en el catálogo A (|C|)	156	45	28	20	11
errores, dudas o residuos[b] en las primeras 50 entradas del catálogo (|E|)	0	0	14	11	1
• precisión (1 – |E| / 50)	1.0000	1.0000	0.7200	0.7800	0.9800
• *recall* o recuperación comprensiva (|*A*| / |*B*|)	0.9571	1.0000	0.8000	1.0000	0.9167

[a] Los alomorfos, homógrafos y elementos polisémicos se contaron como formas distintas; los afijos nulos (ø) se excluyeron.

[b] Los errores, dudas o residuos se refieren a los candidatos rechazados como afijos o cadenas de afijos entre los primeros 50 candidatos del catálogo respectivo.

[c] Tamaño aproximado.

En el primer renglón se establece el nombre del tipo de afijos al que pertenecen las unidades afijales examinadas. El segundo y el tercero caracterizan los corpus utilizados: su tamaño en número de palabras gráficas y en número de tipos de palabra. Nótese que la muestra del checo es una muestra de lemas y, como cada lema ocurre una vez, el número de palabras es igual al número de tipos, que son los lemas. El cuarto renglón exhibe el tamaño del conjunto de afijos *a priori* (|B|) de acuerdo con la tradición o la experiencia de los especialistas. Los tamaños de estos conjuntos se mencionaron en cada descripción de cada experimento y, como se ve, son: 163 para la flexión verbal del español de

México, 45 para los prefijos del checo, 35 para los del ralámuli, 20 para los prefijos de flexión verbal del chuj y 12 para los sufijos de la flexión verbal de esa lengua. En el quinto renglón, aparece el número de afijos conocidos *a priori* que aparecieron en los catálogos (|C|). En el sexto, aparece el número de errores, dudas y formas no reconocidas dentro de las primeras 50 entradas del catálogo.

Como se dijo, los dos últimos renglones muestran las medidas de precisión y recuperación comprensiva. De nuevo, la medida de precisión es simplemente la proporción de aciertos entre los primeros 50 candidatos del catálogo y, por otra parte, la recuperación comprensiva corresponde a la proporción de miembros del conjunto de verdaderos afijos encontrados en todo el catálogo. Los resultados de ambas medidas de evaluación son alentadores. Sin duda, es necesario examinar más lenguas y otros corpus de estas mismas lenguas para obtener mejores comparaciones. Por lo pronto, este método de evaluación parece suficiente para acercarse al material lingüístico disponible en un corpus de alguna lengua concatenativa desconocida, similar en estructura a las presentadas aquí, para empezar a dilucidar su morfología afijal, aunque la muestra textual pueda no ser tan representativa como se pudiera desear.

3.5 Los catálogos de afijos como herramientas morfológicas

Hasta aquí, examinamos los resultados de aplicar un método no supervisado de descubrimiento de afijos a varias muestras textuales. Esto es, se presentaron y evaluaron los resultados de su aplicación a lenguas muy distintas, representadas en corpus muy desiguales. Hay, por supuesto, cuestiones puntuales que podrían mejorar la evaluación. Por ejemplo, sería mucho mejor si el método también tomara en cuenta los contextos de los afijos y de las palabras. Esto, por lo menos, facilitaría la evaluación de los residuos. Además, el examen mecánico y cuantitativo de las

secuencias de afijos es necesario para estudiar la afitáctica de las lenguas desde una perspectiva no supervisada.

Como sea, podemos ver que los resultados de revisar algunos desarrollos que se pueden construir con la aplicación de estas técnicas, de examinar y clasificar lo que contienen los catálogos de afijos y de ensayar el cálculo de medidas de evaluación como la de precisión y de recuperación comprensiva nos permiten construir un panorama valorativo que habla bien del método.

Dada la diversidad de estrategias morfológicas de las lenguas para formar nuevas palabras o flexionarlas, es evidente que el descubrimiento de morfemas va más allá de segmentarlas. Sin embargo, se ve que investigar cuantitativamente una de las estrategias más diseminadas, la afijación, tiene resultados interesantes. Esto se logra porque, aunque el método de cálculo de afijalidades no pueda pronunciarse categóricamente sobre la naturaleza afijal de alguna cadena de caracteres, esencialmente le asigna un valor a cada una, un valor que representa qué tanto podemos confiar en ella como representación de un afijo o secuencia de afijos en la lengua a la que pertenece.

Ciertamente, más que modelos morfológicos de las lenguas examinadas, los catálogos de afijos pueden verse como ventanas a fenómenos diversos que pueden describirse de diferentes maneras, de acuerdo con la perspectiva teórica preferida. Es decir, no constituyen una teoría morfológica, sino son herramientas para el descubrimiento de lo desconocido, más relacionadas con la minería de textos que con el diseño de formalismos basados en reglas. En el próximo capítulo veremos las bases de cómo utilizar estas herramientas para conocer la variación entre los registros de una lengua y entre lenguas emparentadas.

CAPÍTULO 4
HACIA EL CÁLCULO DE LA VARIACIÓN MORFOLÓGICA

Como hemos visto, a partir de las palabras gráficas de un corpus y mediante métodos no supervisados para descubrir afijos, podemos compilar catálogos de segmentos afijales en los que se consignan valores para estimar su facultad de actuar como afijos. Gracias a estos valores, un catálogo como éstos se puede ver como una especie de perfil de la lengua representada en la muestra, característico de su morfología afijal, un perfil que podría considerarse una verdadera huella dactilar.

Podemos comparar las extracciones morfológicas de diferentes corpus, representativas de diferentes registros, dialectos o estados diacrónicos, de una misma lengua o de varias lenguas emparentadas. Si cada elemento de un catálogo de afijos tiene asociado un valor de afijalidad, valdría la pena compararlo con el registrado para el mismo elemento en otro catálogo extraído de otra muestra textual. De hecho, como veremos, existen métodos para medir la similitud o distancia entre catálogos que toman en cuenta los valores de los afijos compartidos.

Cabe esperar que las diferencias entre perfiles de este tipo correspondan a las diferencias entre las morfologías de las lenguas retratadas en esos corpus. Así, la distancia entre dos catálogos de afijos o *perfiles morfológicos* se puede ver como una medida general de variación a nivel morfológico que puede servir para corroborar o descubrir la estructura dialectal o diacrónica representada mediante estos perfiles. Este es el tema de este capítulo.

En la primera sección, examinaremos un par de catálogos de segmentos sufijales de estudiantes de primaria cubanos, con la idea de observar las diferencias entre los sufijos que usan los niños de cuarto año y los que usan los de sexto; así que se comparan dos muestras de distintas etapas de desarrollo de una misma lengua. En la segunda sección,

se mostrarán algunos cambios, en mediciones de economía y entropía, en el interior de una estructura nominal del español que desapareció alrededor del siglo XVI, la frase nominal definida posesiva (*las nuestras provincias*, *el vuestro pecado*, etc.). En la tercera sección, se expone el concepto de cognado morfológico que nos permite relacionar formas afijales análogas, pertenecientes a diferentes lenguas emparentadas o a diferentes registros o dialectos de una lengua. Por último, en la tercera sección, se describe el concepto de distancia euclidiana y se muestran ejemplos de su aplicación para medir la similitud entre perfiles morfológicos de algunas lenguas mayas, en sincronía, y entre los de algunas muestras textuales de la lengua española de diferentes estados diacrónicos.

4.1 Diferencias entre perfiles de una misma lengua

El *Corpus electrónico de textos escritos por escolares de Guamá*, compilado por Celia Pérez Marqués (2003), es un conjunto de muestras textuales, recolectadas a principios de este siglo, representativas de la escritura de estudiantes de primaria, nacidos y criados en el municipio de Guamá, en Santiago de Cuba, por padres también nacidos y criados allí[1]. Se extrajeron sufijos y grupos sufijales de las muestras de cuarto y sexto grados y se almacenaron en catálogos separados para examinar la "competencia morfológica desplegada por estos escolares, la cual se pone de manifiesto en su vocabulario" (Pérez Marqués y Medina Urrea 2005).

El total de candidatos a sufijos de la muestra de cuarto año fue 904 y el de la muestra de sexto fue de 1696. En la Tabla 36, se pueden ver las formas sufijales con mayor afijalidad en cuarto año y en la Tabla 37 aquellas con mayor afijalidad de sexto. La diferencia entre la cantidad de informantes de sexto (129) y de cuarto (139), así como la producción de textos más extensos por parte de los primeros explican en parte por qué se

[1] Para construir este corpus se tomó el 20% de la población escolar existente en el municipio de Guamá en febrero de 2001. La muestra de cuarto año contiene textos de 127 informantes, mientras que la de sexto contiene 139 (C. M. Pérez Marqués 2004, 50).

descubrieron menos sufijos y grupos sufijales en los niños de cuarto. Sin embargo, como lo explica Pérez Marqués (2004, 109), también cabe esperar una competencia léxica y morfológica superior en los niños de sexto.

Lo importante es que las extracciones de estas muestras comparten 704 formas sufijales. Esto es, se detectaron 704 formas afijales usadas tanto en cuarto como en sexto; así que cada una tiene dos valores de afijalidad, uno para cada año escolar. Al restar estos dos valores, podemos distinguir los sufijos con mayor afijalidad en un grado, de aquellos con mayor afijalidad en el otro.

Tabla 36. Sufijos y grupos sufijales más prominentes de cuarto año

sufijo	frec.	cuadros	economía	entropía	afijalidad	diferencia	índice
1~emos	16	0.0799	0.8090	0.7513	0.5468	0.1439	0.0787
2~s	602	1.0000	0.8544	0.4287	0.7668	0.0881	0.0676
3~**a**	390	0.5426	0.6939	1.0000	0.7455	0.0862	0.0643
4~aba	13	0.1055	0.8205	0.7580	0.5614	0.0989	0.0555
5~**as**	147	0.2784	0.6853	0.9839	0.6492	0.0691	0.0449
6~ió	16	0.0741	0.5809	0.7902	0.4818	0.0576	0.0278
7~r	123	0.2469	0.9215	0.4323	0.5336	0.0412	0.0220
8~**o**	443	0.4836	0.6108	0.9335	0.6759	0.0318	0.0215
9~ía	29	0.0922	0.4993	0.7339	0.4418	0.0466	0.0206
10~iendo	12	0.1298	0.5388	0.6733	0.4473	0.0444	0.0199
11~íamos	11	0.0867	0.2787	0.6387	0.3347	0.0589	0.0197
12~n	199	0.2143	0.6694	0.5819	0.4885	0.0341	0.0167
13~**os**	318	0.2426	0.4620	0.8018	0.5021	0.0283	0.0142
14~**to**	64	0.0156	0.1793	0.7123	0.3024	0.0440	0.0133
15~en	40	0.1419	0.5083	0.9498	0.5333	0.0239	0.0127
16~an	86	0.2028	0.7487	0.9824	0.6446	0.0185	0.0119
17~er	22	0.0444	0.4970	0.8950	0.4788	0.0231	0.0111
18~**ita**	15	0.0496	0.3814	0.7068	0.3793	0.0220	0.0083
19~**e**	144	0.1045	0.2906	0.8615	0.4189	0.0154	0.0065
20~ar	74	0.2514	0.9415	0.9837	0.7256	0.0075	0.0054
21~mos	120	0.2533	1.0000	0.4056	0.5530	0.0004	0.0002

Tabla 37. Sufijos y grupos sufijales más prominentes de sexto año

sufijo	frec. cuadros	economía	entropía	afijalidad	diferencia	índice
1~ían	11 0.1129	0.6044	0.7163	0.4779	0.3527	0.1686
2~aron	29 0.2319	1.0000	0.9355	0.7224	0.2048	0.1479
3~ieron	13 0.0531	0.5307	0.8154	0.4664	0.2523	0.1177
4~iera	12 0.0560	0.4478	0.7608	0.4215	0.2507	0.1057
5~**ado**	59 0.0801	0.5307	0.9812	0.5307	0.1798	0.0954
6~amos	84 0.3056	0.9935	0.9894	0.7628	0.1250	0.0954
7~rse	15 0.0862	0.5273	0.3667	0.3267	0.2432	0.0795
8~rme	22 0.0823	0.4174	0.4399	0.3132	0.2529	0.0792
9~rnos	18 0.0364	0.5780	0.4453	0.3532	0.2036	0.0719
10~**eros**	16 0.0194	0.2420	0.7825	0.3480	0.1600	0.0557
11~**do**	163 0.1319	0.5655	0.5311	0.4095	0.1204	0.0493
12~arme	11 0.0470	0.3436	0.7163	0.3690	0.1327	0.0490
13~**ido**	26 0.0789	0.3604	0.7267	0.3887	0.1205	0.0468
14~ré	27 0.0505	0.2696	0.5941	0.3047	0.1248	0.0380
15~aré	12 0.0489	0.2633	0.6610	0.3244	0.1048	0.0340
16~ir	21 0.0476	0.3777	0.7747	0.4000	0.0790	0.0316
17~ría	25 0.0331	0.2937	0.6288	0.3185	0.0976	0.0311
18~ando	33 0.1432	0.9946	0.9264	0.6881	0.0406	0.0279
19~**oso**	12 0.0201	0.1807	0.7087	0.3032	0.0743	0.0225
20~aban	14 0.0419	0.3917	0.7515	0.3950	0.0527	0.0208
21~**itos**	22 0.0690	0.3665	0.7065	0.3807	0.0539	0.0205
22~**ada**	30 0.0063	0.0620	0.8528	0.3070	0.0638	0.0196
23~**í**	31 0.1496	0.4142	0.8771	0.4803	0.0360	0.0173
24~ó	61 0.4955	0.5705	0.8250	0.6303	0.0270	0.0170
25~**al**	25 0.0200	0.1884	0.7683	0.3256	0.0512	0.0167
26~**ito**	33 0.0637	0.4079	0.8291	0.4336	0.0290	0.0126
27~é	78 0.5389	0.6781	0.8365	0.6845	0.0151	0.0103
28~**ante**	14 0.0283	0.1549	0.7216	0.3016	0.0333	0.0100
29~**on**es	25 0.0235	0.3516	0.5275	0.3008	0.0301	0.0091
30~da	57 0.0578	0.3560	0.5528	0.3222	0.0261	0.0084
31~ero	22 0.0259	0.2082	0.8307	0.3549	0.0192	0.0068
32~imos	36 0.1475	0.7088	0.8715	0.5760	0.0049	0.0028
33~**es**	173 0.0967	0.4048	0.7768	0.4261	0.0039	0.0017

Se eliminaron todos los candidatos con frecuencia menor a 10 y cuya afijalidad fue menor de 0.3, por lo que quedaron solamente 54 formas: 21 de cuarto, que se despliegan en la Tabla 36; y 33 de sexto, en la Tabla 37. En ambas tablas, podemos ver que los sufijos no están ordenados por afijalidad, sino por un índice de ordenamiento que se describirá a continuación.

Las tablas tienen dos columnas nuevas a la derecha. En la penúltima columna está la diferencia entre la afijalidad de un grado menos la del otro. Nótese que en estas tablas esta diferencia siempre es positiva. Las entradas que aparecerían con diferencias negativas se muestran con valores positivos en la otra tabla. En la última columna aparece el índice de ordenamiento que toma en cuenta tanto los valores de afijalidad como las diferencias de la penúltima columna. El orden es de mayor a menor. Así que en los primeros lugares de cada catálogo aparecen las formas con más afijalidad y mayor diferencia (siempre positiva) entre los grados.

Los renglones con sufijos derivativos o grupos sufijales con material derivativo aparecen sombreados. Como se ve, algunas de estas formas contienen elementos sufijales que pueden ser tanto de flexión como de derivación, por ejemplo, el sufijo *~a* es marca de flexión en "compr***a*** los dulces" y de derivación en "hizo una compr***a*** de dulces". Aquellos que representan exclusivamente derivación aparecen subrayados y en negritas, aunque contengan material de flexión, como ***~ero.****s* *e* ***~ito.****s*, en la Tabla 37. En cuarto grado, se observa la secuencia ***~o.****s*, que ocurre al final de sustantivos masculinos en plural, y el grupo ***~it.****a*, que corresponde a los morfemas de diminutivo y femenino. El segmento *~to* es típico de palabras como *excepto*, *honesto*, *pronto*, *resto*, *cuarto*, etcétera.

Es interesante que la proporción de formas derivativas con respecto a las puramente flexivas sea mayor en sexto (45%) que en cuarto (33%). Por ejemplo, entre las formas derivativas más usuales de cuarto se puede observar el grupo sufijal diminutivo en femenino *~it.a*, mientras que, en sexto, se aprecia, además del grupo de diminutivos, el grupo sufijal

aumentativo ***~on.****es*. También en sexto, se emplea más el grupo ***~os.o*** para la formación de adjetivos y los sufijos y grupos sufijales ***~ero.****s*, ***~al*** y ***~ante*** para formar adjetivos y sustantivos.

Respecto a las formas flexivas, en cuarto grado predominan las que se refieren a la 1ª persona (singular o plural) del presente, pretérito y copretérito de indicativo, así como a la 1ª del presente del subjuntivo. En cambio, en sexto se observan entre las más usuales, además de aquellas, algunas formas correspondientes al futuro, al pospretérito de indicativo, y al pretérito de subjuntivo. Además, están presentes grupos sufijales que contienen la marca de infinitivo y un pronombre enclítico: *~r.se*, *~r.me* y *~r.nos*. Todo esto muestra que en sexto hay un uso de estructuras lingüísticas más complejas que en cuarto, lo cual se puede explicar por el avance en el proceso de maduración y escolarización de los alumnos.

Entre los 704 sufijos de menor frecuencia y menor afijalidad, que no aparecen en estas tablas, se observan los correspondientes a sustantivos abstractos (*~dad*, *~ción*, *~ería*), que también exhiben en sexto más afijalidad que en cuarto. Lo mismo ocurre con los sufijos y grupos sufijales que se emplean para formar adjetivos (*~ud.o*, *~os.a*) y con los que sirven para formar nombres de oficio u ocupación (*~ador*, *~ista*). De nuevo, todo esto concuerda con el incremento de la competencia morfológica en los estudiantes más avanzados.

4.2 Comparación entre frases nominales posesivas y frases definidas simples

Uno de los sentidos de la gramaticalidad de las estructuras lingüísticas es anterior a la noción chomskiana de aceptabilidad y se debe a Meillet, quien lo caracterizó como el estado de algunas palabras cuyo sentido y forma se han debilitado de tal manera que se vuelven cada vez más suplementarios con respecto a las unidades léxicas autónomas, convir-

tiéndose así en marcadores de roles gramaticales[2]. Meillet se refirió a esta idea como la evolución de las formas gramaticales (afijos, palabras gramaticales, etc.) a partir de palabras de contenido o formas léxicas[3]. Mientras más se desgasta fonológica y semánticamente una forma, más secundaria se hace y mayor es su fuerza de adhesión a las palabras de contenido (Meillet 1958 [1912], 139):

> L'affaiblissement du sens et l'affaiblissement de la forme des mots accessoires vont de pair ; quand l'un et l'autre sont assez avancés, le mot accessoire peut finir par ne plus être qu'un élément privé de sens propre, joint à un mot principal pour en marquer le rôle grammatical.

Como es bien sabido, este proceso de cambio ha sido denominado gramaticalización y, durante las últimas décadas, recibió especial atención de varios lingüistas como Dwight Bolinger, Bernd Heine, Elizabeth Traugott, Joan Bybee, Talmy Givón, entre otros, que han desarrollado ampliamente esta perspectiva. Se podría argumentar que este enfoque agrega poco a los métodos neogramáticos tradicionales. De hecho, algunos de sus seguidores han reconocido que todavía no existe un cuerpo de conocimiento que pueda llamarse una verdadera teoría de la gramaticalización (Fischer, Norde y Perridon 2004, 13)[4].

Como sea, este campo es muy interesante y muchas de sus observaciones merecen un tratamiento cuantitativo. De hecho, la frecuencia es

[2] Esta sección está basada en el artículo, "Towards the measurement of nominal phrase grammaticality: contrasting definite-possessive phrases with definite phrases of 13th to 19th century Spanish" en Grzybek y Köhler, *Exact Methods in the Study of Language and Text*, Mouton de Gruyter, Berlín (Medina-Urrea 2007, 427-437).

[3] Meillet, "L'évolution des formes grammaticales", en *Linguistique historique et linguistique générale*, Société de Linguistique de Paris, París (1958 [1912], 130-148).

[4] Existen diversas caracterizaciones de cómo se construyen las teorías. Una relevante para investigaciones cuantitativas del lenguaje es la que bosqueja Altmann en "Science and linguistics", publicada en Köhler y Rieger, eds., *Contribuitions to Quantitative linguistics*, Dordrecht, Kluwer (1993). También vale la pena conocer la propuesta de Bunge, *Philosophy of Science I* y *II*, Transaction Publishing, New Brunswick (1998 [1967]).

un criterio privilegiado para juzgar la gramaticalidad de una partícula, palabra o secuencia de palabras: cuanto más frecuente es algo, más *gramatical* se considera[5]. A continuación, examinaremos cómo podemos estimar una gramaticalidad cuantitativa mediante la adaptación de las medidas de economía y entropía, que se presentaron en el capítulo 2.

Para este propósito, nos enfocaremos en dos tipos de frases definidas del español. La idea es que, si con el paso del tiempo las palabras plenas se pueden convertir en auxiliares de algún tipo, que a su vez pueden transformarse en clíticos, los que tienen alguna probabilidad de convertirse en afijos, entonces las técnicas no supervisadas de descubrimiento de afijos, como la entropía y las relaciones económicas y entrópicas pueden servir para medir la gramaticalidad dentro de las frases. Como se verá, podemos llevar a cabo mediciones de entropía y economía en las frases definidas, las que proporcionan criterios cuantitativos para juzgar cuánto se adhieren o aglutinan los determinadores y los nombres dentro de la frase.

En este apartado, examinaremos la frase nominal definida posesiva (FN def-pos) que el milenio pasado fue común en español. Se trata de la frase que consiste en un sustantivo precedido por un artículo definido y un determinante posesivo. En esta lengua, esta construcción se volvió improductiva alrededor del siglo XVI. Algunos ejemplos son:

	determinadores		
	art. definido	**det. posesivo**	**sustantivo**
FN def-pos	*el*	*su*	*coraçon*
	la	*tu*	*onrra*
	los	*uuestros*	*prophetas*
	las	*nuestras*	*prouincias*

[5] La perspectiva de la gramaticalización propone ciertas etapas de cambio, a saber: palabra de contenido > palabra gramatical > clítico > afijo flexivo. Se dice que los elementos a la derecha de esta secuencia son más gramaticales.

Las frases definidas posesivas son un tipo de frase definida. Las contrastamos con la frase definida plana o regular, aquella constituida simplemente por un artículo definido y el sustantivo principal (FN def): *el coraçon*, *la onrra*, *los prophetas*, *las prouincias*. Para esto, se obtuvieron ejemplos de frases definidas posesivas y de frases definidas regulares de varios siglos del Corpus del español de Mark Davies (http://www.corpusdelespanol.org/).

La cantidad de frases definidas regulares que se pueden recuperar de todos los siglos del corpus es obviamente enorme, las frases definidas posesivas son comunes alrededor del siglo XIII, pero su número disminuye drásticamente hacia el siglo XVI. Por eso, se seleccionó un tamaño de muestra de 1 200 ejemplos para cada tipo de frase para cada siglo. De esos ejemplos, se examinaron y se eliminaron manualmente algunas formas que no correspondían a frases nominales definidas. La Tabla 38 muestra el número de ejemplos que se obtuvieron del corpus para cada siglo en cuestión.

Tabla 38. Total de ejemplos de frases nominales definidas

	1200	**1300**	**1400**	**1500**	**1600**	**1700**	**1800**
FN def	1 193	1 197	1 198	1 199	1 200	1 200	1 200
FN def-pos	1 194	1 174	1 152	645	136	312	182

Nótese que los tamaños de las muestras de las estructuras definidas posesivas muestran una reducción importante alrededor del siglo XVI, cuando perdió su productividad. De hecho, los pocos ejemplos de frases definidas posesivas en los siglos subsecuentes se produjeron principalmente en citas de textos antiguos.

Una forma interesante de contrastar estas estructuras, más allá de mirar su frecuencia, es examinar la variedad de sustantivos que ocurren dentro de estas estructuras. Para esto, se lematizaron manualmente los

sustantivos. Básicamente, la lematización consistió en eliminar la flexión del plural. Por lo tanto, frases como *el vuestro peccado* y *los nuestros pecados* corresponden a un mismo lema (*pecado*). En contraste, los sustantivos con flexión de género permanecieron como lemas separados. Así que frases como *el su rrey* y *la vuestra rreyna* representan dos lemas separados (*rey* y *reina*).

Después de la lematización, se calculó la entropía de los lemas de las frases nominales, esto es, la sorpresa que causan los lemas cuando han ocurrido los determinadores (la entropía de izquierda a derecha). En esencia, se trata de la medición de la variedad de lemas nominales en el conjunto de las frases. Como vimos en los capítulos anteriores, la entropía es un indicador de fronteras entre bases y afijos. Medirla entre los elementos en el interior de una frase puede dar indicios sobre la cohesión de los mismos en la frase. La Tabla 39 exhibe la variedad nominal en bits de cada tipo de estructura por siglo.

Tabla 39. Entropía de los sustantivos
en frases nominales definidas (bits)

	1200	**1300**	**1400**	**1500**	**1600**	**1700**	**1800**
FN def	9.277	9.327	9.388	9.603	9.662	9.673	9.659
FN def-pos	8.301	8.417	8.695	8.293	6.589	6.847	7.024

De nuevo, se observa un descenso en los valores de la frase definida posesiva a partir del siglo XVI. La Figura 12 muestra estos números gráficamente. Hay que tener en cuenta que dentro de los primeros tres siglos y a pesar de los tamaños de muestra similares, el conjunto de sustantivos pertenecientes a las frases definidas posesivas contiene entre medio bit y casi un bit menos de información que los de frases definidas simples.

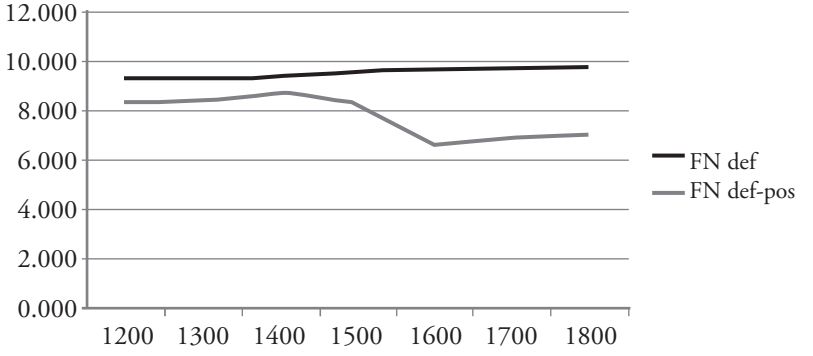


Figura 12. Entropía de los sustantivos en frases nominales definidas (bits)

Como se dijo, la entropía de los lemas en las frases definidas posesivas disminuye en el siglo XVI, lo que podría haberse esperado por los tamaños de las muestras, que van disminuyendo. En otras palabras, durante los primeros siglos, la variedad de elementos que ocurren dentro de estructuras definidas posesivas parece ser un poco menor que la variedad de elementos nombrados dentro de frases definidas regulares. Luego, los siglos siguientes muestran una disminución en la variedad de nombres de los primeros, mientras que la variedad de los sustantivos de las frases definidas simples permanece relativamente constante. Por un lado, a medida que pasa el tiempo, los artículos definidos parecen unirse a un número creciente de sustantivos, algunos de los cuales en siglos anteriores solían ocurrir sin ningún artículo. Por otro lado, la variedad de elementos nombrados en construcciones definidas posesivas disminuye, seguramente debido a la cantidad menor de ejemplos de este tipo de frase encontrados en los siglos subsecuentes.

Sin embargo, cabría haber esperado que las frases menos frecuentes exhibieran más información. ¿Cómo es posible entonces que la cantidad de información nominal no haya crecido cuando la estructura se volvió menos productiva (frecuente)? En realidad, hubo una reducción rápida de las ocurrencias disponibles junto con una reducción relativamente

lenta de la variedad de sustantivos. La Tabla 40 muestra la cantidad de información dividida entre el número de ejemplos de frase por siglo.

Tabla 40. Entropía de los lemas dividida entre número de frases nominales definidas (bits)

	1200	**1300**	**1400**	**1500**	**1600**	**1700**	**1800**
FN def	0.0078	0.0078	0.0078	0.0080	0.0081	0.0081	0.0081
FN def-pos	0.0070	0.0072	0.0076	**0.0129**	**0.0485**	**0.0219**	**0.0386**

Como se ve en los números en negritas, los valores de la FN def-posesiva aumentan ligeramente hacia el siglo XVI, especialmente con respecto a los de las frases definidas simples, mientras que el tamaño de las muestras de la FN def-posesiva decreció (ver Tabla 38). De allí que las frases menos frecuentes contengan más información.

De hecho, parece que toda la FN def-posesiva, como una unidad, empieza a parecerse más a una palabra de contenido que a una estructura analítica nominal, mientras que las frases definidas siguen siendo construcciones muy productivas, en las que unos pocos signos determinantes se combinan con una variedad ligeramente creciente de sustantivos[6].

Otro aspecto interesante que se puede tomar en cuenta, para medir el carácter gramatical de las frases, es el de las relaciones económicas. Como hemos visto, los signos morfológicos (signos económicos, como afijos o clíticos y signos de contenido, como bases o raíces), se combinan para formar signos del siguiente nivel lingüístico (el léxico).

[6] Nótese que en la frase definida se observa una tendencia estable, pero ligeramente al alza de la cantidad de información de los sustantivos. Los artículos definidos preceden una variedad de cosas nombradas que crece con los siglos. Es interesante que esta misma tendencia fuera argumentada en otros términos, sin evidencia cuantitativa, para los artículos definidos del francés en Epstein, "The Development of the Definite Article in French" en Pagliuca, ed., *Perspectives on Grammaticalization*, Amsterdam / Philadelphia, Benjamins (1994, 63-78).

De hecho, cada frase nominal se puede ver como un signo del nivel sintáctico (sujeto, complemento directo, indirecto o circunstancial) constituido por signos del nivel morfosintáctico (determinadores, adjetivos y sustantivos). La frecuencia ya es una medida de economía de signos. Así que los elementos más frecuentes dentro de las frases estructuran al sintagma y al discurso y lo hacen económico.

De hecho, el tamaño del conjunto de sustantivos lematizados que, de acuerdo con un corpus, ocurre en las frases nominales (independientemente de la presencia o ausencia de determinadores) se relaciona con la economía de la frase nominal, más que la mera frecuencia de la frase: mientras más tipos nominales lematizados haya dentro de una estructura, más económica será esa estructura. Además, los determinadores y sus combinaciones también pueden tenerse en cuenta para estimar la economía de cada tipo de frase. Como en el caso de los afijos en las palabras, mientras más determinadores tenga una frase nominal, menos económica podrá considerarse.

Para ilustrar esto, recordemos que las frases definidas simples en español tienen un paradigma de cuatro artículos en distribución complementaria, *el*, *la*, *los*, *las*. En cambio, las frases definidas posesivas tienen un paradigma más amplio que consiste en combinaciones de artículos definidos y determinadores posesivos: *el mi*, *el tu*, *el su*, *el nuestro*, *el vuestro*, *la mi*, *la tu*, *la su*, *la nuestra*, *la vuestra*, *los mis*, *los tus*, *los sus*, *los nuestros*, *los vuestros*, *las mis*, etc. De hecho, hay 56 combinaciones posibles (|{*el, la, los, las*} × {*mi, su, tu, mis, sus, tus, nuestro, nuestra, nuestras, nuestras, vuestro, vuestra, vuestros, vuestras*}|), pero sólo 20 de ellas son aceptables (**la nuestro*, **los tu*, **las tu*, etc. son combinaciones inaceptables).

Por lo tanto, una forma de estimar la economía de una frase nominal será comparar los tamaños de ambos conjuntos, el de nombres lematizados y el del paradigma de artículos o combinaciones de determinadores. Al dividir el tamaño del primero entre el tamaño del segundo, obtenemos una medida de economía de la frase nominal.

Además, aparte del hecho de que las frases definidas simples han sido siempre más frecuentes que las frases definidas posesivas, las primeras pue-

den considerarse más económicas que las segundas porque, dados conjuntos de ejemplos de tamaño similar (y suponiendo números similares de nombres lematizados, digamos *n*), el cociente de economía para frases definidas simples (*n* ÷ 4) sería mayor que el de las frases definidas posesivas (*n* ÷ *c*, donde *c* es el número de combinaciones de determinadores de acuerdo con una muestra de frases posesivas, que puede alcanzar 20). Obsérvense estos valores de economía a lo largo de los siglos en la Tabla 41:

Tabla 41. Valores de economía en la frase nominal definida

	1200	**1300**	**1400**	**1500**	**1600**	**1700**	**1800**
FN def	179.25	182.00	191.25	210.75	216.50	219.00	215.75
FN def-pos	20.36	21.73	19.82	16.43	5.35	7.85	5.50

Estos datos se muestran gráficamente en la Figura 13. Como era de esperarse, el carácter económico de las frases definidas simples ha sido siempre mayor que el de las frases definidas posesivas. También, se puede observar una disminución en los valores de esta segunda estructura, del siglo XVI al XVII. Si el carácter gramatical puede medirse por medio de la economía de frases, la FN def-posesiva simplemente pierde gran parte de su gramaticalidad, la que, para empezar, ya era considerablemente más baja que la de la FN def regular.

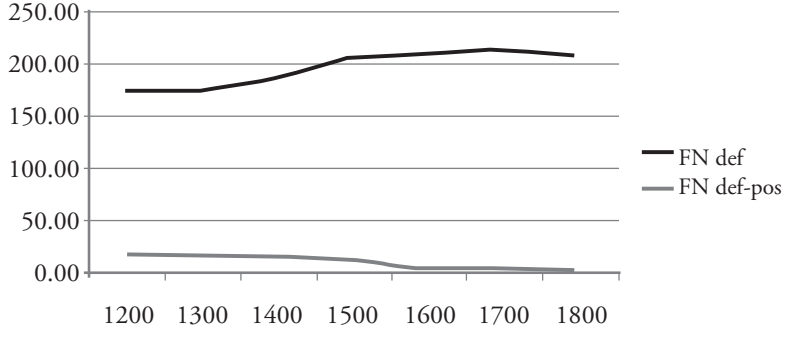


Figura 13. Valores de economía en la frase nominal definida

Es interesante que, al igual que las medidas de entropía (ver Figura 12), hay dos tendencias opuestas: a medida que pasan los siglos, las frases definidas simples llevan más información y se vuelven más económicas, mientras que las estructuras definidas posesivas se vuelven menos económicas y contienen menos información en términos absolutos, aunque más en términos relativos.

Si se trata de capturar la esencia de la gramaticalidad, las observaciones sobre el contenido de la información y las relaciones económicas en conjunto brindan una imagen más fina que la proporcionada por la observación de las frecuencias simples. De hecho, podemos estimar la gramaticalidad al multiplicar estos valores: $G = ik$, donde i es la magnitud que mide la información contenida en los sustantivos de las estructuras y k la magnitud que representa algún tipo de estructura económica. En el presente experimento, i es el número de bits y k el número de sustantivos lematizados dividido entre el número de miembros del paradigma del determinante. Para nuestro experimento, estos valores se muestran en la Tabla 42.

Tabla 42. Cantidad de información y estructura económica (G)

	1200	**1300**	**1400**	**1500**	**1600**	**1700**	**1800**
FN def	2766.80	2791.08	2811.65	2878.58	2898.64	2901.81	2897.77
FN def-pos	169.03	182.88	172.35	136.29	35.25	53.75	38.63
total	2935.83	2973.96	2984.00	3014.87	2933.89	2955.56	2936.40

Los datos aparecen gráficamente en la Figura 14. La Figura 15 muestra la misma información, pero sólo para las frases definidas posesivas.

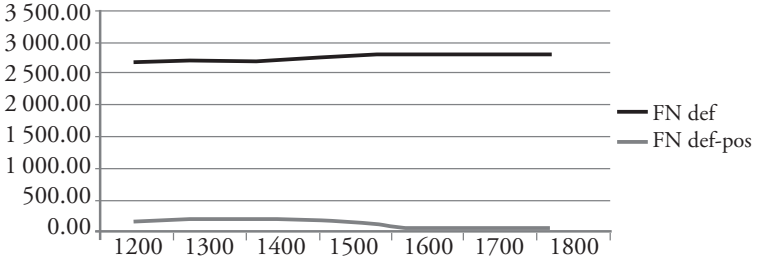


Figura 14. Cantidad de información × estructura económica

Como se dijo, los valores representan el número de bits en las estructuras económicas de las frases en cada siglo. Nótese que el valor correspondiente a las frases definidas regulares se queda alrededor de 2 850, aunque podría argumentarse que hay una tendencia débil al alza. En contraste, la FN def-posesiva exhibe una gramaticalidad decreciente, que se puede apreciar mejor en la Figura 15. Es interesante que, al menos para estos datos, las sumas de estos valores por siglo estén cerca de 3 000 (véase el último renglón de la Tabla 42). De hecho, sería interesante estimar este tipo de valores para todos los otros tipos de frase, incluyendo las verbales, para ver si al sumarlos hay alguna tendencia hacia algún valor constante a lo largo de los siglos.

Figura 15. Cantidad de información × estructura económica en la frase definida posesiva

En esta discusión exploratoria, se contrastaron dos tipos de frases nominales definidas españolas a lo largo del milenio pasado. Al observar sólo su frecuencia, fue posible inferir que las frases definidas posesivas perdieron su productividad en algún momento del siglo XVI. También, mediante medidas de entropía y economía, fue posible observar los cambios en gramaticalidad que sufrieron estos tipos de frases. Evidentemente, éstas son observaciones preliminares, considerando que las muestras son pequeñas y de representatividad limitada.

Ambos tipos de frases exhiben diferentes tendencias de cambio. Por un lado, las frases definidas simples muestran una cantidad de información y un valor de estructura económica más o menos constantes, tal vez con una ligera tendencia al alza. Por otro lado, las frases definidas posesivas muestran desde el principio menos estructura y menos variedad de sustantivos. Sería interesante averiguar si estas tendencias se pueden observar en otras estructuras lingüísticas. Por lo pronto, la medición de la gramaticalidad, mediante algo más que las frecuencias de las unidades examinadas, proporciona formas interesantes de examinar cómo cambian a lo largo del tiempo.

4.3 Sobre cognados y relaciones genéticas

Sheila Embleton (1986) describe la gran cantidad de trabajo cuantitativo que ya se había llevado a cabo en los años ochenta para evaluar el grado de relación genética entre lenguas con el objetivo de compararlas diacrónicamente[7]. La glotocronología de Morris Swadesh (1955) ya era entonces el enfoque más destacado[8]. Este método está inspirado en el fechado de carbono de material orgánico. En esencia, sirve para calcular un índice de cambio lingüístico por milenio, por lo que ha suscitado mucha controversia. Se basa en la compilación de bases léxicas de al me-

[7] *Statistics in historical linguistics*, Brockemeyer, Bochum (Embleton 1986).

[8] "Towards Greater Accuracy in Lexicostatistic Dating", *International Journal of American Linguistics* 21 (Swadesh 1955, 121-137).

nos 100 elementos (partes del cuerpo, fenómenos celestiales, números pequeños, pronombres personales, verbos de acción básicos, etc.) para comparar cognados de distintas lenguas y estimar su separación de la lengua ancestral en términos de milenios. La idea es calcular el porcentaje de cognados compartidos: mientras mayor sea, se presume que su separación como lenguas independientes es más reciente.

Los cognados son conjuntos de dos o más palabras con un origen común, esto es, que descienden de una misma voz perteneciente a la lengua ancestral. Pueden pertenecer a lenguas diferentes, como *starve* (inglés) y *sterben* (alemán), o a una misma lengua, como los vocablos españoles *delgado* y *delicado* y las palabras inglesas *shirt* y *skirt*. También los morfemas dentro de las palabras pueden ser cognados como *~idad*, *~ité* (francés), *~ität* (alemán), *~ity* (inglés), etc. y como los pronombres clíticos de las lenguas romances. Estos últimos tipos de cognados pueden constituir una base interesante para la comparación de lenguas de manera no supervisada.

Thomason y Kaufman[9] (1991 [1988], 5-8) recuerdan la controversia entre Frank Boas y Edward Sapir acerca de la posibilidad de distinguir similitudes genéticas entre lenguas con el propósito de clasificarlas. A grandes rasgos, Boas planteaba que la relación genética entre dos lenguas implica necesariamente correspondencias sistemáticas en todas las partes del lenguaje y que con el tiempo es imposible determinar si su similitud se debe a una genética compartida o al contacto entre sus hablantes. En cambio, Sapir consideraba que los elementos superficiales de la gramática podían cambiar por contacto con otras lenguas, mientras que los más profundos, como la morfología o el vocabulario, son más estables y se heredan a las lenguas hijas. Probablemente, como establecen Thomason y Kaufman, una verdadera relación genética entre lenguas se podrá plantear cuando "systematic correspondences can be

[9] *Language Contact, Creolization, and Genetic Linguistics*, University of California Press, Berkeley (Thomason y Kaufman 1991 [1988]).

found in all linguistic subsystems—vocabulary, phonology, morphology, and (we would add) syntax as well" (1991 [1988], 8).

Sin embargo, la posibilidad de reunir conjuntos de afijos característicos de varias lenguas para compararlos cuantitativamente nos permite explorar sus relaciones de similitud en el nivel de morfología afijal. En la siguiente sección, se compararán cuatro lenguas mayas, cuyo parentesco genético no está en entredicho, pero sí el carácter de cognados de las formas afijales que comparten. En cambio, en la última sección, examinaremos tres estados de lengua del español en México y España, en los que podemos asumir que sus afijos y cadenas de afijos sí son cognados.

Lo interesante es que, al contrastar lenguas emparentadas o registros diacrónicos de una misma lengua, pueden tomarse en cuenta los elementos morfológicos y no sólo los léxicos. Como hemos visto, las unidades léxicas, específicamente sus raíces o lexemas, transmiten más información sobre el mensaje de nuestro interlocutor que las otras unidades morfológicas (afijos, clíticos, partículas, modificadores), que expresan la información gramatical que estructura las frases y el discurso. De allí que una lista de unidades léxicas básicas, como las que se presupone en la glotocronología, sea más afín a la dimensión cultural de la vida de los hablantes que al sistema gramatical de comunicación de su lengua, que le da estructura lingüística a su pensamiento y a su cultura. En cambio, los grupos afijales, que exhiben información de carácter estructural, léxico y morfosintáctico, están más relacionados con la organización interna de las lenguas, que cualquier base léxica del tipo usado en la glotocronología.

La lista de voces que sirve como base glotocronológica puede incluir unidades gramaticales en tanto se desempeñen como pronombres, clasificadores, numerales, etc., que son funciones típicas de las unidades gramaticales. Si funcionan de esta manera, probablemente pertenezcan al conjunto de modificadores gramaticales o al de las partículas de naturaleza clítica o afijal; esto es, unidades de ocurrencia relativamente frecuente que se pueden haber desgastado fonológica y semánticamente y

suelen haber perdido su significado referencial. Lo importante es que las lenguas emparentadas genéticamente seguramente comparten cognados de tipo gramatical y no sólo léxico. Mientras más lejanas sean estas lenguas, genética y geográficamente, menos probable será que compartan unidades gramaticales.

Como lo plantea Swadesh, la compilación manual de bases léxicas, para comparar lenguas, será de mucho valor para estimar su lejanía en milenios de un ancestro común. Sin embargo, en la medida en que no se tomen en cuenta afijos y clíticos, tal comparación reflejará más el nivel de la cultura y el pensamiento, que el nivel íntimo de la estructura comunicativa de esas lenguas. Por supuesto que una lista de palabras no representa toda una cultura y menos un grupo de culturas. Sin embargo, el hecho de que un conjunto de lenguas pueda compartir referentes no significa que los signos, específicamente los significantes, sean independientes de sus culturas. Además, cuando estos significantes se refieren a conceptos supuestamente universales, representan más bien la intersección de los conjuntos de referentes importantes en esas culturas.

4.4 Variación entre perfiles morfológicos

Podemos comparar los valores de afijalidad de dos muestras diferentes, presuponiendo que sus afijos son cognados, para conocer cómo varían geográfica o temporalmente. Cuando son varias las formas que dos muestras comparten, podemos promediar las diferencias entre los valores de una y de la otra, para obtener un valor general que represente la cercanía o lejanía, similitud o distancia, entre estas muestras. En especial, si lo que tenemos es un conjunto de cognados afijales que pertenecen a dos lenguas distintas, cabe esperar que, al medir la cercanía entre estas muestras, obtengamos un valor abstracto que represente una distancia morfológica entre estas lenguas. En esta sección, examinaremos algunos experimentos para ejemplificar esto.

Primero, se describirá el método para medir distancias euclidianas. Esta técnica es muy conocida y se escogió para estimar distancias entre perfiles morfológicos por su sencillez y facilidad de aplicación. Luego, se presentarán los resultados de calcular estas distancias entre muestras textuales pequeñas de cuatro lenguas mayas, que son el chuj, el tojolabal, el yucateco y el huasteco, teenek o tének. Finalmente, examinaremos los resultados de un experimento de medición de distancias entre perfiles morfológicos del español provenientes de tres estados de lengua y dos continentes.

4.4.1 Distancias euclidianas

Como se dijo, la posibilidad de extraer conjuntos de afijos de diferentes muestras nos permite comparar los de una con los de otras. Existe una variedad de medidas que se pueden aplicar a la comparación de estos perfiles; por ejemplo, los coeficientes de distancia, como los de Minkowski, Manhattan o Canberra (Oakes 1998, 111-112). También se pueden usar otros coeficientes para medir el grado de correlación en lugar del de distancia, como el de *r* de Pearson o el de correlación de rango de Spearman (Woods, Fletcher y Hughes 1986, 169-74). Sería interesante, incluso, aplicar las similitudes de coseno que se utiliza en recuperación de información para comparar vectores; véanse, por ejemplo, Jurafsky y Martin, *op. cit.* (2009) y Manning y Schütze, *op. cit.* (1999, 540-541).

Como sea, entre estas medidas, la distancia euclidiana se escogió por su simplicidad y aplicabilidad. De esta manera, las distancias d_{jk} entre cada par de perfiles *j* y *k* se midieron mediante la siguiente fórmula:

$$d_{jk} = \sqrt{\sum_{i=1}^{n} (X_{ij} - X_{ik})^2}$$

donde X_{ij} representa el valor de afijalidad del elemento morfológico *i* en el perfil *j*, y X_{ik} representa el valor de afijalidad del mismo elemento en el perfil morfológico *k*; *n* es el número de elementos afijales que

comparten. Los valores promediados se obtienen aplicando la siguiente fórmula:

Ecuación 3. Distancia euclidiana

$$\delta_{jk} = \sqrt{\frac{\sum_{i=1}^{n} (X_{ij} - X_{ik})^2}{n}}$$

De esta manera, se miden las distancias euclidianas entre varios perfiles morfológicos, p_1, p_2, p_3, …, p_m, representando diferentes registros, dialectos o estados de lengua. Podemos colocar estas distancias en una matriz como la siguiente:

	p_1	p_2	p_3	…	p_m
p_1	δ_{11}	δ_{12}	δ_{13}	…	δ_{1m}
p_2	δ_{21}	δ_{22}	δ_{23}	…	δ_{2m}
p_3	δ_{31}	δ_{32}	δ_{33}	…	δ_{3m}
…	…	…	…	⋱	…
p_m	δ_{m1}	δ_{m2}	δ_{m3}	…	δ_{mm}

donde m es el número de perfiles morfológicos, $1 \leq i \leq m$, $1 \leq j \leq m$ y $1 \leq k \leq m$. Así, en el renglón del perfil p_3 y en la columna del perfil p_2 está la celda con el valor δ_{32} que corresponde a la distancia euclidiana que hay entre p_3 y p_2.

La diagonal de la matriz (celdas en gris) contiene las distancias de cada perfil hacia sí mismo. Evidentemente, éstas son siempre 0 ($\delta_{11} = \delta_{22} = \delta_{33} = \ldots = \delta_{mm} = 0$), porque la distancia de cada perfil morfológico a sí mismo es nula. Además, como δ_{32} es la misma distancia que δ_{23}, podemos prescindir de las celdas debajo de la diagonal. De esta manera, la matriz de distancias entre perfiles se verá como la de la Tabla 43:

Tabla 43. Matriz de distancias euclidianas (δ_{ij}) entre perfiles morfológicos (p_k)

	p_1	p_2	p_3	…	p_m
p_1	δ_{11}	δ_{12}	δ_{13}	…	δ_{1m}
p_2		δ_{22}	δ_{23}	…	δ_{2m}
p_3			δ_{33}	…	δ_{3m}
…				⋱	…
p_m					δ_{mm}

Finalmente, podemos utilizar estas distancias para estimar la similitud entre los perfiles. Como los valores de afijalidad están normalizados, todos los promedios, δ_{ij}, serán valores entre 0 y 1. Por lo que una manera de estimar la similitud puede ser simplemente calcular el valor absoluto de la diferencia entre 1 y la distancia. Por ejemplo, la similitud entre p_1 y p_2 será $1 - \delta_{12}$.

4.4.2 Distancias en sincronía entre las morfologías afijales de algunas lenguas mayas

El método más conocido y más antiguo para medir distancias entre lenguas genéticamente emparentadas probablemente es la glotocronología de Swadesh (1955, 121-137) que, como se dijo, sirve para medir la separación, en milenios, entre dos o más lenguas y su ancestro común, con base en sus vocabularios básicos. En los años sesenta y setenta, se llevó a cabo una gran cantidad de investigaciones cuantitativas para evaluar el grado de parentesco genético entre lenguas con el objeto de compararlas

diacrónicamente. Al respecto, Sheila Embleton (1986) resumió lo que se había publicado a mediados de los años ochenta. Recientemente, se han llevado a cabo trabajos que examinan ciertas medidas estadísticas de secuencias de palabras para comparar lenguas europeas: inglés, español, francés, ruso, alemán e italiano, entre 1855 y 2009 (Morales, y otros 2018). Un punto clave de estos métodos es la medición de la similitud o distancia entre formas.

En este apartado, se presentan los resultados de utilizar la estructura afijal de las palabras para indagar la relación de similitud morfológica entre cuatro lenguas mayas: chuj, tojolabal, yucateco y huasteco. Así que, a continuación, examinaremos la extracción no supervisada de afijos de estas lenguas, para luego medir distancias y similitudes entre ellas a partir de la afijalidad de sus formas prefijales y sufijales compartidas. Este tipo de trabajos podría contribuir a resolver controversias como aquella sobre la clasificación del chuj y del tojolabal en la que Kaufman (1974) las considera hermanas mientras que Robertson (1977) las coloca en ramas distintas de la familia maya.

En la primera sección, se dan los detalles de las muestras textuales que se utilizaron para este experimento. Enseguida, se revisan los resultados de la extracción de prefijos y sufijos de las lenguas enfocadas y del cálculo de las distancias euclidianas. Por último, se comenta la necesidad de ir más allá de esta similitud euclidiana y se presentan las observaciones finales.

4.4.2.1 Muestras textuales de cuatro lenguas mayas

En este trabajo se utilizan colecciones de textos del chuj, del tojolabal, del yucateco y del huasteco para determinar automáticamente los afijos y grupos afijales de estas lenguas[10]. La idea es que estos afijos las carac-

[10] Los resultados preliminares de este experimento se presentaron en el XI Encuentro Internacional de Lingüística en el Noroeste, en la Universidad de Sonora, Hermosillo, en noviembre de 2010 y en el Primer Encuentro de Estudios sobre el Chuj, Instituto de Investigaciones Antropológicas, Ciudad de México, en mayo de 2018.

terizan y pueden servir de material para el cálculo de distancias euclidianas. Sobra decir que, a pesar del tamaño reducido de estas muestras textuales, procedemos a utilizarlas presuponiendo cierto nivel de representatividad. Las muestras textuales son las siguientes:

1. Las narraciones del chuj compiladas por Elsa Cristina Buenrostro que se utilizaron en la extracción de prefijos y sufijos del chuj (2002) que se presentó en el capítulo anterior (alrededor de 15 500 palabras gráficas en 86 Kb).
2. La recopilación de cuentos del tojolabal, *Palabras de nuestro corazón. Mitos, fábulas y cuentos maravillosos de la narrativa tojolabal*, UNAM, Universidad Autónoma de Chiapas (Gómez Hernández, Palazón y Ruz 1999), que consta de 22 590 palabras (137 Kb).
3. Los cuentos del yucateco disponibles en la página electrónica *Yucatán, identidad y cultura maya* (Centro de Investigaciones Dr. Hideyo Noguchi s.f.)[11]. La muestra consta de 21 740 palabras en 129 Kb.
4. Las narraciones del huasteco compiladas por Lucero Meléndez del Instituto de Investigaciones Antropológicas de la UNAM (Meléndez Guadarrama 2010). Esta muestra consta apenas de 3 120 palabras gráficas en 17Kb. Como se ve, es la menor de todas, por mucho.

Respecto a las particularidades de la escritura en estas muestras, nos enfrentamos a una heterogeneidad en las convenciones de escritura, por ejemplo, en la representación de ciertos fonemas ("tz" *vs*. "ts", "x" *vs*. "sh", etc.). Así que se hicieron adaptaciones para unificar los criterios de escritura[12].

Además, si consideramos que la palabra gráfica es una secuencia de letras entre espacios, ciertas unidades lingüísticas pueden aparecer, en

[11] La página se encuentra en http://www.mayas.uady.mx/literatura/index_02.html.

[12] Gracias a Cristina Buenrostro Díaz por encargarse de la normalización de las grafías de las cuatro muestras textuales.

su escritura, pegadas a la palabra gráfica en alguna lengua, pero separadas de la misma en alguna otra, según hayan sido consideradas afijos o clíticos por quienes las transcribieron. En otras palabras, algo que aparece en una muestra como parte de la palabra gráfica, esto es, como afijo, en otra puede aparecer separado gráficamente. Por esta razón, en este experimento algunas formas clíticas en alguna lengua también se consideraron afijos, por la posibilidad de ser cognadas de formas afijales en las otras lenguas. Naturalmente, esta corrección no resuelve todos los problemas. La realidad de la escritura de las palabras del maya es más compleja. Si bien los artículos, demostrativos, etc. se escriben como palabras independientes, puede ser que aparezcan escritos adheridos a las palabras gráficas. Además, los predicados complejos o construcciones seriales (por ejemplo, dos verbos, verbo más otra clase de palabra, entre otros tipos) a veces se escriben juntos y a veces separados. Cabe esperar que todo esto repercuta en la exactitud de los datos obtenidos de las cuatro muestras.

En la siguiente sección, se presentan los resultados de la extracción automática de afijos a partir de cada una de las muestras. En la Tabla 44, se puede ver que se extrajeron 1 241 afijos del corpus del chuj, de los cuales 448 también se extrajeron del corpus de tojolabal y sólo 300 se encontraron del de yucateco. También se puede ver que el tojolabal y el yucateco comparten 420 grupos afijales.

Tabla 44. Cuentas simples de afijos compartidos entre cuatro lenguas mayas

	chuj	**tojolabal**	**yucateco**	**huasteco**
chuj	1 241	448	300	145
tojolabal		2 283	420	180
yucateco			1 592	187
huasteco				477

En la diagonal (en gris) está el total de formas encontradas para cada muestra textual. Es interesante que las lenguas que más afijos comparten sean el chuj y el tojolobal (448), mientras que las que menos comparten sean el chuj y el huasteco (145). De hecho, el huasteco es la lengua del conjunto que comparte menos formas afijales con las otras tres (145, 180 y 187), lo cual no es extraño puesto que es la lengua de la que se extrajeron menos afijos (477) y cuya muestra es la más pequeña.

4.4.2.2 Extracción de afijos

En la Tabla 45, podemos ver los resultados de la extracción de afijos del chuj. A la izquierda, se muestran las diez formas más prefijales de un total de 546 grupos prefijales extraídos. A la derecha, aparecen las diez más sufijales de un total de 1 065 (las primeras cinco ya se comentaron en el capítulo anterior). En la columna del centro aparece el rango de estas formas. De nuevo, como están ordenadas de mayor a menor afijalidad, las formas con rango 1 son las más afijales del corpus (*ix~* la más prefijal y *~kan* la más sufijal). A medida que crece el rango, las formas tienen una afijalidad menor y, por lo tanto, crece la probabilidad de que no sean elementos morfológicos de la lengua.

Como se mencionó arriba, para cada afijo se exhiben la frecuencia en que se detectó como segmento más afijal, los valores normalizados del número de cuadros en los que apareció, el grado de economía que exhibió y la entropía de las bases con las que apareció. El promedio de estos tres valores se muestra en la columna de afijalidad.

En el experimento previo de extracción de afijos del chuj (Medina y Buenrostro 2003), los resultados fueron un poco diferentes, porque no se tomó en cuenta la medida de cuadros. Una diferencia importante es que en aquel experimento no se detectó como prefijo la forma *w~* (marca de 1ª persona de ergativo), que en este nuevo experimento sí apareció con el rango 7. Lo importante es que, como se estableció entonces, todas estas formas son reconocibles como parte del paradigma de flexión que se prefija al verbo. En cuanto a los sufijos, estos primeros diez

Tabla 45. Afijos del chuj; primeros 10 prefijos del catálogo de 546 (derecha) y primeros 10 sufijos de 1 065 (izquierda)

prefijo	frec.	cuadros	economía	entropía	afijalidad		sufijo	frec.	cuadros	economía	entropía	afijalidad
ix~	180	1.0000	0.8000	0.9198	0.9066	1	~kan	68	1.0000	0.9516	0.8832	0.9449
in~	94	0.5278	0.8278	1.0000	0.7852	2	~nhej	23	0.4138	1.0000	0.7550	0.7229
tz~	358	0.7286	0.6143	0.8976	0.7468	3	~ta’	70	0.6164	0.7297	0.7894	0.7118
s~	186	0.3871	0.6517	0.9774	0.6721	4	~b’at	63	0.5738	0.6282	0.8279	0.6766
ko~	71	0.3064	0.6710	0.8742	0.6172	5	~ok	68	0.4973	0.5433	0.9224	0.6543
ol~	190	0.4844	0.5276	0.8163	0.6094	6	~i	198	0.6061	0.4731	0.8056	0.6283
w~	74	0.4338	0.7204	0.5346	0.5629	7	~xi	37	0.5215	0.6840	0.6680	0.6245
tzin~	47	0.1708	0.4278	0.9512	0.5166	8	~al	84	0.3175	0.5109	1.0000	0.6095
a~	165	0.1687	0.4051	0.9594	0.5111	9	~il	63	0.3892	0.4812	0.9288	0.5997
olin~	26	0.1775	0.4572	0.8982	0.5110	10	~ach	18	0.3786	0.7299	0.6647	0.5911

ocurrieron dentro de las 189 formas más sufijales de aquel experimento, y corresponden a marcas gramaticales de diversos usos, como verbos de movimiento que se usan como clasificadores verbales, participios de verbos intransitivos, adverbios, etcétera.

La Tabla 46 muestra los afijos del tojolabal, la Tabla 47 muestra las formas más afijales del yucateco y la Tabla 48 las del huasteco. De nuevo, las formas que aparecen al principio de cada tabla son las más afijales en el corpus correspondiente.

Conviene insistir que el ordenamiento de mayor a menor afijalidad de las formas contribuye a concentrar las formas más morfológicas hacia el principio de estas tablas, con los rangos menores. De nuevo, mientras más grande sea el rango y menor sea la frecuencia de las formas, más probable es que no sean verdaderos afijos de la lengua en cuestión.

Naturalmente, para determinar el carácter morfológico de todas estas formas es necesario revisar sus contextos y constatar que tienen una función y significado morfológicos. Sin embargo, como los experimentos de este trabajo han sido ejercicios no supervisados de extracción de afijos (un ejercicio en el que se evita la intervención del analista para descartar las formas sin carácter morfológico), esta revisión queda pendiente para trabajos futuros.

Como sea, vale la pena hacer algunas observaciones sobre los resultados del huasteco, que fueron extraídos de la muestra más pequeña, de apenas 3 120 palabras gráficas, por lo que podemos esperar muchos errores. Por ejemplo, según Meléndez Guadarrama[13] *k'apu~* no es verdaderamente un prefijo, sino la raíz del verbo *comer*, a la que se adhieren ciertos sufijos de flexión (2017, 183). Por otra parte, *k'i~* aparece al principio de palabras como *k'ima:ʔ* (casa), *k'itsa:ʔ* (día) y *k'itʃ'aʔ* (apretar), *k'ihla:b* (espíritu). Valdría la pena hacer pruebas para determinar si se trata verdaderamente de un prefijo en esas palabras. Por último,

[13] Véase *Huasteco de El Mamey San Gabriel, Tantoyuca, Veracruz*, Archivo de Lenguas Indígenas, El Colegio de México, México (Meléndez Guadarrama 2017).

Tabla 46. Afijos del tojolabal; primeros 10 prefijos del catálogo de 1 040 (izquierda) y primeros 10 sufijos de 2 039 (derecha)

prefijo	frec.	cuadros	economía	entropía	afijalidad		sufijo	frec.	cuadros	economía	entropía	afijalidad
ja~	246	0.4907	0.4538	0.9498	0.6314	1	~i’	573	1.0000	1.0000	0.8769	0.9590
oj~	43	0.1867	0.4314	1.0000	0.5394	2	~b’i	342	0.5806	0.9241	0.8571	0.7873
lek~	11	0.2072	0.4426	0.9009	0.5169	3	~a	980	0.6003	0.6247	0.9269	0.7173
yuj~	15	0.1941	0.4593	0.8779	0.5104	4	~ja	76	0.1814	0.9243	0.9171	0.6743
s~	650	0.2855	0.2564	0.9754	0.5058	5	~uk	133	0.3073	0.7699	0.8952	0.6575
oč~	21	0.2339	0.3160	0.9242	0.4914	6	~e’i’	45	0.1449	0.9453	0.8129	0.6344
ay~	26	0.1476	0.3705	0.9450	0.4877	7	~e’	220	0.2067	0.6816	0.8959	0.5947
el~	28	0.2237	0.4041	0.8189	0.4822	8	~xa	63	0.1846	0.6705	0.9236	0.5929
tan~	15	0.0980	0.5638	0.7713	0.4777	9	~ma	57	0.1622	0.6798	0.8840	0.5753
k’ot~	15	0.2157	0.4211	0.7885	0.4751	10	~b’a	49	0.1357	0.7420	0.8146	0.5641

Tabla 47. Afijos del yucateco; primeros 10 prefijos del catálogo de 646 (izquierda) y primeros 10 sufijos de 1 471 (derecha)

Prefijo	frec.	cuadros	economía	entropía	afijalidad		sufijo	frec.	cuadros	economía	entropía	afijalidad
al~	32	0.9397	0.5378	0.8783	0.7853	1	~o	403	1.0000	0.9747	0.9385	0.9711
j~	303	1.0000	0.3595	0.9722	0.7772	2	~e	370	0.9977	0.9901	0.8844	0.9574
a~	329	0.8418	0.3540	0.9990	0.7316	3	~ik	190	0.4371	0.9182	0.8590	0.7381
y~	225	0.9154	0.4470	0.5937	0.6520	4	~e’	84	0.3027	1.0000	0.8774	0.7267
aa~	94	0.6049	0.3594	0.9481	0.6375	5	~i	158	0.4026	0.8425	0.8731	0.7061
w~	130	0.7801	0.4838	0.6165	0.6268	6	~il	186	0.4022	0.7839	0.9213	0.7025
paa~	36	0.5307	0.3953	0.9255	0.6172	7	~a	225	0.3167	0.6338	1.0000	0.6502
o~	138	0.5682	0.2878	0.9112	0.5891	8	~ak	71	0.1713	0.6443	0.9750	0.5969
i~	123	0.4965	0.2485	1.0000	0.5817	9	~aak	16	0.1350	0.7212	0.9251	0.5938
an~	25	0.4455	0.2573	0.9225	0.5418	10	~en	111	0.2424	0.7110	0.8140	0.5891

parece que la forma *~ani*, con una afijalidad muy alta, es en realidad parte de la conjunción *ʔani*.

Otro criterio para constatar que estas formas son en verdad afijos es corroborar que aparecen como tales en más de una de las lenguas. Por ejemplo, la propiedad de *ta~* y *a~* de ser prefijos del huasteco crece al corroborar que también ocurren en los catálogos de las otras lenguas, con cierto grado de afijalidad. De hecho, si se corrobora que sus funciones y significados son similares en las cuatro, podemos descartar de que se trate de meros homónimos e, incluso, presumir que son cognados.

Lo interesante es que, a partir de estos cuatro catálogos, se detectaron 50 formas prefijales comunes a las cuatro lenguas, que se exhiben en la Tabla 49 y que, para este ejercicio no supervisado, nos servirán para establecer una similitud formal entre estas lenguas. Vale la pena aclarar que estas formas, que ocurren en las cuatro muestras, sólo comparten, valga la redundancia, la forma. Para determinar si tienen funciones análogas, es necesario analizar cada forma en cada uno de sus contextos en las cuatro lenguas, lo cual queda fuera del alcance de este trabajo. Similarmente, se encontraron 64 formas sufijales compartidas, que también nos permitirán medir la similitud morfológica. Estas se listan en la Tabla 50. La primera columna de estas Tablas muestra el rango de ordenamiento de los elementos afijales, la segunda los elementos afijales y las siguientes consignan los valores de afijalidad en cada una de las lenguas.

Estas formas se ordenaron de más a menos afijales. Esto es, aquellas con mayor promedio de afijalidad aparecen en los primeros lugares de la tabla. Como estos valores pueden variar de manera considerable entre una lengua y otra, también se tomó en cuenta el valor mínimo. Esto es, aquellas con el mayor promedio y con el valor mínimo más alto se consideraron como las más afijales. De esta manera, la forma *a~* es la más afijal del conjunto, porque su promedio de afijalidad (0.549) sumado al valor mínimo (0.44 del tojolabal) resultó en el criterio de ordenamiento (0.989), el más alto de los elementos de la tabla. Similarmente, el segundo prefijo más afijal es *al~*, que tiene un valor de ordenamiento de 0.703 (promedio de 0.4299 más 0.2731, que es su afijalidad mínima,

Tabla 48. Afijos del huasteco; primeros 10 prefijos del catálogo de 274 (izquierda) y primeros 10 sufijos de 337 (derecha)

prefijo	frec.	cuadros	economía	entropía	afijalidad		sufijo	frec.	cuadros	economía	entropía	afijalidad
ta~	23	0.3043	0.2816	1.0000	0.5286	1	~in	40	1.0000	0.9297	0.9930	0.9742
a~	67	0.4179	0.2766	0.8458	0.5134	2	~ani	26	0.9712	0.9431	0.9113	0.9419
k'apu~	12	0.3889	0.3145	0.8024	0.5019	3	~icj	35	0.5143	0.7598	0.7880	0.6874
ja~	31	0.0860	0.0660	0.8627	0.3382	4	~tam	14	0.2857	0.6600	0.7786	0.5748
te~	17	0.0980	0.0602	0.7716	0.3099	5	~tal	11	0.1250	0.5273	0.7579	0.4701
al~	11	0.0909	0.1240	0.6045	0.2731	6	~an	30	0.1917	0.3972	0.7906	0.4598
jun~	13	0.1026	0.0787	0.6083	0.2632	7	~u	17	0.2941	0.3016	0.7475	0.4477
t~	136	0.0417	0.0188	0.7015	0.2540	8	~'	38	0.2368	0.5854	0.4428	0.4217
k'i~	11	0.0606	0.0000	0.6864	0.2490	9	~a	66	0.0814	0.1448	1.0000	0.4087
x~	16	0.0833	0.0639	0.5805	0.2426	10	~al	75	0.0850	0.1550	0.8968	0.3789

en el huasteco), y la forma menos afijal de todas es *bo~* (rango 50) con un valor de ordenamiento de 0.206 (promedio 0.1251 + valor mínimo 0.0806). El mismo método se siguió para ordenar las formas sufijales de la Tabla 50.

Tabla 49. Formas prefijales compartidas y sus valores de afijalidad

	prefijo	chuj	tojolabal	yucateco	huasteco		prefijo	chuj	tojolabal	yucateco	huasteco
1	a~	0.5111	0.4400	0.7316	0.5134	26	to~	0.2952	0.1667	0.2918	0.1867
2	al~	0.3792	0.2819	0.7853	0.2731	27	k'~	0.2207	0.2316	0.2221	0.1852
3	ti~	0.2384	0.2677	0.4452	0.8962	28	pa~	0.2249	0.2915	0.2931	0.1571
4	ja~	0.2577	0.6314	0.3069	0.3382	29	o~	0.1206	0.2194	0.5891	0.1816
5	x~	0.3731	0.4520	0.5105	0.2426	30	wi~	0.2656	0.1472	0.4030	0.1540
6	ta~	0.2619	0.3282	0.2859	0.5286	31	at~	0.1952	0.1583	0.4399	0.1478
7	wa~	0.2482	0.4095	0.4571	0.2299	32	tz~	0.7468	0.1994	0.0843	0.1281
8	ma~	0.3734	0.2998	0.3613	0.2450	33	mu~	0.2051	0.2521	0.2748	0.1487
9	u~	0.2508	0.2309	0.5116	0.2422	34	an~	0.1956	0.2787	0.5418	0.0887
10	j~	0.1482	0.4130	0.7772	0.1996	35	t~	0.0978	0.2624	0.4354	0.2540
11	jun~	0.1877	0.4542	0.4400	0.2632	36	m~	0.1756	0.2259	0.3241	0.1410
12	ko~	0.6172	0.3429	0.3234	0.1610	37	te~	0.1881	0.1286	0.2867	0.3099
13	i~	0.1868	0.3175	0.5817	0.2282	38	na~	0.2533	0.1823	0.3053	0.1332
14	k~	0.2601	0.2890	0.3968	0.2159	39	ba~	0.1374	0.1922	0.3049	0.2061
15	ya~	0.2573	0.2763	0.3785	0.2222	40	l~	0.1984	0.2085	0.2946	0.1250
16	e~	0.3736	0.1953	0.4473	0.2207	41	ju~	0.1841	0.2347	0.2701	0.1212
17	ku~	0.2707	0.2343	0.4142	0.2187	42	b~	0.1121	0.1813	0.3282	0.2087
18	ka~	0.2888	0.2869	0.3521	0.1982	43	xa~	0.1543	0.3754	0.1741	0.0838
19	p~	0.2071	0.2481	0.3987	0.2329	44	k'i~	0.1172	0.1470	0.1122	0.2490
20	y~	0.5019	0.2747	0.6520	0.0892	45	k'u~	0.1734	0.1999	0.1036	0.1457
21	w~	0.5629	0.1222	0.6268	0.1081	46	ji~	0.0796	0.2212	0.1670	0.1837
22	k'a~	0.3794	0.1820	0.2381	0.2055	47	we~	0.0703	0.1849	0.1868	0.1867
23	pe~	0.3070	0.2024	0.2362	0.1863	48	pi~	0.0867	0.1009	0.2120	0.1505
24	ki~	0.1699	0.2613	0.2668	0.2450	49	bu~	0.0796	0.1317	0.2206	0.0892
25	n~	0.2082	0.1830	0.2774	0.2109	50	bo~	0.0842	0.0806	0.2463	0.0892

Tabla 50. Formas sufijales compartidas y sus valores de afijalidad

	sufijo	chuj	tojolabal	yucateco	huasteco		sufijo	chuj	tojolabal	yucateco	huasteco
1	~a	0.4924	0.7173	0.6502	0.4087	33	~s	0.2313	0.2110	0.2830	0.1372
2	~an	0.5021	0.4687	0.4546	0.4598	34	~ak	0.1693	0.1804	0.5969	0.0888
3	~al	0.6095	0.4000	0.5408	0.3789	35	~ax	0.1850	0.1544	0.2218	0.1994
4	~in	0.4618	0.2871	0.3700	0.9742	36	~no	0.1930	0.1752	0.2977	0.1407
5	~e	0.2661	0.4218	0.9574	0.3134	37	~ay	0.1583	0.1514	0.2878	0.1591
6	~i	0.6283	0.2948	0.7061	0.2506	38	~k	0.1585	0.1648	0.3480	0.1236
7	~il	0.5997	0.3996	0.7025	0.2346	39	~ke	0.0787	0.3187	0.3203	0.1928
8	~o	0.3629	0.4040	0.9711	0.2181	40	~nal	0.1648	0.1206	0.2001	0.2426
9	~ta	0.2795	0.4911	0.4931	0.3530	41	~ti	0.0806	0.2160	0.3374	0.2387
10	~’	0.2763	0.3707	0.4284	0.4217	42	~y	0.1612	0.1454	0.1440	0.1659
11	~ani	0.3248	0.2867	0.2055	0.9419	43	~at	0.0777	0.1765	0.3611	0.2525
12	~el	0.5374	0.4311	0.4485	0.1999	44	~es	0.2970	0.3347	0.3046	0.0416
13	~ul	0.2700	0.3772	0.4189	0.2061	45	~lan	0.4072	0.2518	0.0820	0.0720
14	~tal	0.1715	0.2076	0.5610	0.4701	46	~ni	0.0588	0.4529	0.2078	0.1323
15	~la	0.3022	0.2663	0.4607	0.1999	47	~ro	0.2216	0.2236	0.1996	0.0819
16	~a’	0.2156	0.4152	0.3099	0.1829	48	~om	0.2074	0.1034	0.1224	0.1706
17	~on	0.2107	0.4405	0.2933	0.1776	49	~t	0.1030	0.1735	0.4458	0.0568
18	~ol	0.3259	0.1618	0.4213	0.2350	50	~wal	0.3252	0.1330	0.0820	0.1332
19	~ik	0.2571	0.1693	0.7381	0.1187	51	~anil	0.1305	0.1718	0.2743	0.0815
20	~l	0.2761	0.2173	0.3289	0.1859	52	~ek	0.2502	0.0882	0.1664	0.1012
21	~na	0.1917	0.2388	0.2284	0.2526	53	~as	0.2162	0.2517	0.1147	0.0720
22	~x	0.1751	0.2042	0.2915	0.2521	54	~yal	0.3029	0.2011	0.0644	0.0815
23	~to	0.2658	0.2905	0.3800	0.1332	55	~im	0.0696	0.0821	0.2784	0.1776
24	~ala	0.4503	0.3333	0.3730	0.0888	56	~un	0.1421	0.1896	0.2324	0.0641
25	~xi	0.6245	0.2800	0.3116	0.0641	57	~mo	0.0787	0.0949	0.2520	0.0888
26	~ya	0.1715	0.2756	0.3812	0.1407	58	~ina	0.0696	0.1670	0.2204	0.0888
27	~k’	0.3826	0.1768	0.2371	0.1332	59	~lal	0.0787	0.1361	0.1781	0.0888
28	~te	0.1924	0.2261	0.3678	0.1332	60	~jan	0.0787	0.3819	0.1298	0.0330
29	~m	0.2325	0.1941	0.3627	0.1321	61	~j	0.1399	0.2085	0.2839	0.0236
30	~am	0.1416	0.2174	0.3601	0.1518	62	~mal	0.2081	0.2027	0.0580	0.0525
31	~ja	0.2358	0.6743	0.2320	0.0577	63	~kal	0.2936	0.1226	0.0449	0.0888
32	~n	0.1375	0.2083	0.3236	0.2070	64	~nto	0.1114	0.0949	0.0820	0.0888

Es de notarse que no es muy grande la diferencia entre el número de formas prefijales y sufijales que comparten las cuatro lenguas. Sin embargo, es significativo que haya más sufijos compartidos que prefijos. Sería interesante observar cuánto disminuyen estos conjuntos al agregar más lenguas mayas al conjunto examinado. También valdría la pena averiguar, mediante otros métodos y muestras textuales más amplias, si comparten más formas.

4.4.2.3 La matriz de distancias euclidianas

Con estos datos podemos medir las distancias euclidianas entre perfiles para estimar la similitud entre ellos. Como se dijo, las distancias mayores significan menor similitud, mientras que las menores implican mayor cercanía.

De esta manera, podemos calcular una distancia euclidiana para cada par de lenguas. Por ejemplo, la distancia entre el chuj y yucateco es la raíz cuadrada del promedio de diferencias entre los valores de afijalidad de cada afijo al cuadrado:

$$D\,(\text{chuj, yucateco}) = \sqrt{\frac{\sum_{i=1}^{n} (AF_{chuj}(i) - X_{yucateco}(i))^2}{n}}$$

donde n es el número de prefijos en la Tabla 49 (50) o de sufijos en la Tabla 50 (64), según se calcule la distancia euclidiana entre formas prefijales o sufijales. Por ejemplo, para calcular las distancias entre los sufijos del chuj y del yucateco, se saca la raíz cuadrada del promedio de las diferencias (al cuadrado) entre los valores de afijalidad de cada renglón de la Tabla 51. Esto es la raíz cuadrada del promedio de: $(0.4924 – 0.6502)^2$, $(0.4618 – 0.3700)^2$, …, $(0.1114 – 0.082)^2$.

Tabla 51. Aspecto de la Tabla 50

	sufijo	chuj	tojolabal	yucateco	huasteco
1	~a	0.4924	0.7173	0.6502	0.4087
2	~in	0.4618	0.2871	0.3700	0.9742
3	~e	0.2661	0.4218	0.9574	0.3134
4	~o	0.3629	0.4040	0.9711	0.2181
…	…	…	…	…	…
64	~nto	0.1114	0.0949	0.0820	0.0888

Mediante este procedimiento, se calcularon las matrices de distancias euclidianas de prefijos y de sufijos de las cuatro lenguas, que se muestran en la Tabla 52.

Tabla 52. Matrices de distancias euclidianas

	prefijos			
	chuj	**tojolabal**	**yucateco**	**huasteco**
chuj	0.0000	0.1585	0.2194	0.1892
tojolabal		0.0000	0.1947	0.1474
yucateco			0.0000	0.2405
huasteco				0.0000

	sufijos			
	chuj	**tojolabal**	**yucateco**	**huasteco**
chuj	0.0000	0.1486	0.2117	0.1938
tojolabal		0.0000	0.1983	0.1983
yucateco			0.0000	0.2625
huasteco				0.0000

A primera vista, podemos notar que las matrices de prefijos y de sufijos son parecidas. Si acaso, al considerar sólo los sufijos, el huasteco se aleja un poco más del grupo, mientras que el chuj y el tojolabal están un poco más cerca. Para unificar estas distancias, podemos sumar las matrices para obtener sólo una distancia entre cada par de lenguas. La nueva matriz aparece en la Tabla 53. Allí se puede ver que la menor distancia es la que hay entre el chuj y el tojolabal (0.3071, en negritas) y la mayor es la que hay entre el huasteco y el yucateco (0.5030, en cursivas).

Tabla 53. Distancias euclidianas entre perfiles morfológicos de cuatro lenguas mayas

	chuj	**tojolabal**	**yucateco**	**huasteco**
chuj	0.0000	**0.3071**	0.4311	0.3830
tojolabal		0.0000	0.3930	0.3457
yucateco			0.0000	***0.5030***
huasteco				0.0000

Por otra parte, al tomar en cuenta el ordenamiento de los elementos compartidos por las cuatro lenguas, que va de mayor a menor promedio de afijalidad (las Tablas 49 y 50), podemos eliminar del cálculo de distancias de aquellos elementos con menor afijalidad, porque tienen más probabilidad de ser errores. De hecho, se pueden calcular las distancias euclidianas solamente con aquellas con valor de ordenamiento (promedio + valor mínimo) mayor de 0.5. Esto es, con los primeros 17 prefijos de la Tabla 49 y los primeros 15 sufijos de la Tabla 50. Al medir las distancias entre los afijos de las cuatro lenguas, los valores de la matriz se actualizan en la Tabla 54:

Tabla 54. Distancias euclidianas entre afijos
más afijales de cuatro lenguas mayas

	chuj	**tojolabal**	**yucateco**	**huasteco**
chuj	0.0000	**0.3254**	0.5424	0.5058
tojolabal		0.0000	0.4961	0.4971
yucateco			0.0000	***0.7082***
huasteco				0.0000

Según estos valores, la distancia afijal entre el huasteco y el yucateco (0.7082, en cursivas) es mayor que aquella entre el huasteco y las otras dos lenguas. De hecho, el huasteco es el más alejado del grupo, lo cual puede deberse al tamaño del corpus y a los pocos afijos extraídos del mismo. Por otra parte, el chuj está un poco más lejos del yucateco (0.5424) que del huasteco (0.5058). Nótese que los cercanos, esto es, los más similares de todo el grupo siguen siendo el chuj y el tojolabal (0.3254, en negritas), los que, también geográficamente, se hablan en regiones más cercanas, como se puede ver en el mapa de la Figura 16:

Figura 16. Distribución geográfica

Es interesante que haya una analogía entre estas distancias euclidianas y la separación geográfica de los lugares donde se hablan estas lenguas. Naturalmente, estos datos no pueden ser definitivos. Como se advirtió arriba, las muestras textuales en las que se basó este experimento son muy pequeñas y su representatividad es cuestionable. Además, no sabemos cuánto de las morfologías de estas lenguas no se obtuvo con el método de descubrimiento de afijos; el experimento se basó en apenas 50 prefijos y 64 sufijos comunes y, para la última medición, sólo en 17 y 15 respectivamente.

4.4.2.4 De la coincidencia de afijos a la determinación de cognados afijales

La semejanza entre las formas más afijales de estas lenguas puede deberse a muchos factores que van desde la casualidad hasta la relación genética, pasando por las relaciones de contacto lingüístico entre sus hablantes. De todas maneras, puesto que sabemos que estas lenguas están emparentadas, cabe considerar la posibilidad de que las formas compartidas sean cognados, especialmente cuando forman palabras con significados análogos o funcionan de manera similar en las distintas lenguas. Mientras mayor sea el número de cognados que comparten dos lenguas, mayor será la probabilidad de que estén emparentadas genéticamente.

Para determinar si las formas afijales compartidas son cognados, es necesario analizar cada una. El objetivo sería averiguar si en efecto son afijos en las cuatro lenguas y si coinciden en su función, significado y contextos; por ejemplo, en la categoría gramatical de las palabras a las que se adhieren. Al respecto, en la opinión de por lo menos Martín Sobrino, especialista en maya yucateco, ni el prefijo *ja~* (rango 4, Tabla 49) ni el sufijo *~to* (rango 23, Tabla 50) parecen siquiera afijos del yucateco[14]. En cambio, en su opinión los prefijos *x~* y *j~* (rangos 5 y 10,

[14] Carlos Martín Sobrino Gómez, comunicación personal, 17 de octubre de 2018.

Tabla 49) y los sufijos *~tal*, *~nal* y *~es* (rangos 14, 40 y 44, Tabla 50) probablemente sí son cognados en estas lenguas.

Lo importante es que determinar el carácter de cognados de las formas afijales de estas lenguas bien podría contribuir a resolver controversias sobre, por ejemplo, aquella ampliamente debatida sobre la clasificación del chuj y del tojolabal dentro de alguna rama de la familia maya. Hay dos posiciones más o menos claras, la de Kaufmann (1974) que ubica al chuj como una lengua hermana del tojolabal y la de Robertson (1977, 105-120) que los ubica en subfamilias distintas; al tojolabal como una rama de las tzeltalanas y al chuj como una rama de las kanjobaleanas. Si bien en la clasificación del chuj y del tojolabal no se puede desestimar la similitud entre sus morfologías (según las medidas euclidianas), todavía no se puede afirmar que se trate de relaciones genéticas. Además, valdría la pena agregar a este experimento otras lenguas, tzeltalanas y kanjobaleanas, para ver si un mapa de distancias euclidianas más complejo puede contribuir a aclarar la relación entre el chuj y el tojolabal. También sería muy interesante incluir estados diacrónicos de estas lenguas para examinar sus distancias en el tiempo.

En resumen, en este apartado se calcularon distancias euclidianas entre los perfiles morfológicos de cuatro lenguas mayas. Estas distancias se vaciaron en matrices que permitieron visualizarlas y compararlas. Cabe notar que se tomaron en cuenta todos los prefijos y sufijos comunes a las cuatro lenguas. Al final, se calcularon las distancias sólo con aquellos afijos cuyos promedios de afijalidad sumados al valor mínimo fue mayor que 0.5, lo que acentuó las cercanías y lejanías entre los perfiles de cada par de lenguas.

Lo interesante fue que se pudo corroborar que el chuj y el tojolabal son los más próximos según su morfología afijal compartida y que la segunda lengua más próxima al tojolabal es el yucateco y la más lejana al yucateco es el huasteco. Como se vio, esto corresponde, a grandes rasgos, a la realidad geográfica de las comunidades donde se hablan estas

lenguas. Naturalmente, se puede argumentar que la correspondencia de los datos presentados con la realidad física es una casualidad y coincidencia. Más interesante sería que pudieran tomarse en cuenta sus relaciones comerciales o de contacto lingüístico a lo largo de los siglos y que se incluyeran otros tipos de morfemas, pero no deja de ser interesante que una coincidencia como ésta pueda suceder.

Muchas cuestiones quedan pendientes. Como se dijo, hace falta examinar las formas afijales detectadas para valorar su calidad de cognados. En particular, es necesario analizar los contextos en que ocurren en cada muestra textual. Por ahora, podemos apreciar que estas lenguas tienen morfologías afijales similares en su forma, pero no podemos establecer todavía su semejanza genética, ni estamos en posición de determinar si son similares por contacto o por mera casualidad.

4.4.3 Distancias en diacronía entre perfiles morfológicos del español

Como hemos visto, los métodos no supervisados para segmentar palabras gráficas pueden aplicarse a muestras textuales para recolectar conjuntos de afijos y grupos de afijos (es decir, perfiles morfológicos) que parecen caracterizar íntimamente a estas muestras. También hemos visto que, a partir de estas caracterizaciones, podemos medir distancias euclidianas para obtener una impresión de la similitud morfológica entre ellas.

A continuación, examinaremos la aplicación de estos métodos en la dimensión diacrónica[15]. En esencia, se presentan datos cuantitativos de tres siglos de la lengua española utilizada en lo que hoy es México (siglos XVI, XVIII y XX) y pequeñas muestras de español de España (siglos XVIII y XX). En particular se calculan y se examinan distancias euclidianas entre perfiles morfológicos de estos estados de lengua en estas zonas geográficas.

[15] Este capítulo está basado en Medina-Urrea, "Toward a comparison of unsupervised diachronic morphological profiles" en Gries, Wulff y Davies, eds., *Corpus-Linguistic Applications: Current Studies, New Directions*, Rodopi, Amsterdam (2010, 29-45).

4.4.3.1 Muestras textuales de tres estados de lengua

Como se dijo, los estados de lengua seleccionados para este experimento corresponden a los siglos XVI, XVIII y XX del español escrito en México[16]. Como con en el experimento de las lenguas mayas, es importante observar que estas muestras son también muy desiguales en cuanto a su tamaño y representatividad. De todas maneras, una vez aplicado el método, se obtuvieron sufijos y grupos sufijales de cada muestra y se almacenaron en catálogos para su posterior comparación. Los datos se ven interesantes, pero conviene ser precavidos con respecto a la naturaleza de las muestras utilizadas.

Por ejemplo, un aspecto importante de comparar textos de diferentes dialectos de una lengua es el de la transcripción fonológica. Como se vio para las lenguas mayas, esta tarea presenta dificultades. En primer lugar, la fonología española sufrió algunos cambios alrededor del siglo XVI, por lo que se requieren reglas de transcripción específicas para esa época. Algunas reglas han sido propuestas para la transcripción automática de los documentos del siglo XVI, como las de la Tabla 16 del experimento de lexematización, pero todavía son provisionales y el consenso parece estar lejos de ser alcanzado[17]. Además, las irregularidades ortográficas de los documentos antiguos hacen que las transcripciones fonológicas automáticas resulten en muchos errores. De allí que en este

[16] Para los dos primeros estados de lengua se utilizaron los documentos reunidos en el Corpus Histórico del Español en México (CHEM), del Instituto de Ingeniería de la UNAM, que comprenden versiones electrónicas de los *Documentos Lingüísticos de la Nueva España, Altiplano Central* (Company Company 1994), *Los procesos inquisitoriales contra indígenas* (Buelna Serrano 2009) y *El habla de Diego de Ordaz: contribución a la historia del español americano* (Lope Blanch 1985). Para el siglo XX se utilizaron los datos que se extrajeron del CEMC.

[17] Acerca de la correspondencia entre grafemas y sonidos del español del siglo XVI, se pueden consultar trabajos muy variados, que van de obras generales y abarcadoras, como de Lara, *Historia mínima de la lengua española*, El Colegio de México (2013), a tesis académicas muy específicas, como Reyes Careaga, *Reglas de correspondencia entre sonido y grafía en el español hablado en México en el siglo XVI: Una aportación al Corpus histórico del español en México*, UNAM, Tesis de licenciatura, México (2008).

experimento se decidiera no aplicar reglas de reescritura como las del experimento de lexematización. Así que éste es básicamente un ejercicio de lengua escrita.

Además, como este experimento se basa en textos escritos y no en transcripciones fonológicas, cabe esperar que los resultados muestren mayor similitud entre los estados de lengua de México y España de la que podía haberse esperado al aplicar las reglas de reescritura. Por ejemplo, los sufijos gráficos que contienen 'z' (como los utilizados en los patronímicos como *Gonzál~ez*, *Rodrígu~ez*, etc., o aquellos encontrados en otras derivaciones como *vej~ez*, *bell~eza*, etc.) tendrían diferentes transcripciones fonológicas dependiendo del lado del Atlántico (/gon. θá.leθ/ *vs*. /gon.sá.les/, /be.xéθ/ *vs*. /be.xés/). Sin embargo, las diversas pronunciaciones no producen diferencias morfológicas. En este experimento la información fonológica se pierde. De haberla conservado, los valores de similitud entre los perfiles morfológicos de ambos lados del Atlántico seguramente serían algo menores.

Por otra parte, cuando se examinan las muestras textuales y los conjuntos de sufijos extraídos, es posible observar que la gran variabilidad ortográfica en los documentos antiguos se produce principalmente en las raíces y las bases de las palabras gráficas y no en los sufijos. De hecho, como los documentos fueron editados siguiendo la práctica de los filólogos de reconstruir palabras gráficas abreviadas, cuando se reconstruyeron afijos y cadenas de afijos, se estandarizaron, limitando la variabilidad ortográfica de los afijos en las abreviaturas reconstruidas.

4.4.3.2 Extracción de sufijos

Los resultados de la extracción se muestran en las Tablas 55, 56 y 57, que exhiben los diez sufijos más afijales de cada siglo. De nuevo, las filas con rangos menores exhiben la mayor afijalidad.

Tabla 55. Grupos de sufijos del siglo XVI;
primeras 10 del catálogo de 760 entradas

	sufijo	frec.	cuadros	economía	entropía	afijalidad
1	~a	919	0.6410	0.9441	0.9505	0.8452
2	~s	1 288	1.0000	0.9968	0.4833	0.8268
3	~o	902	0.6676	0.9355	0.8164	0.8065
4	~ó	306	0.6420	0.8362	0.8720	0.7834
5	~as	403	0.3457	0.9325	0.9607	0.7463
6	~ar	259	0.3003	0.9447	0.9623	0.7358
7	~os	507	0.3721	0.9253	0.8883	0.7286
8	~ado	250	0.2681	0.9434	0.9631	0.7249
9	~e	534	0.4770	0.8904	0.7931	0.7202
10	~aba	137	0.1979	0.9093	0.9415	0.6829

La Tabla 55, correspondiente al siglo XVI, presenta los primeros elementos más sufijales de un catálogo de 760 entradas, extraído de una muestra de 151 966 palabras gráficas y 17 608 vocablos. La Tabla 56, correspondiente al siglo XVIII, presenta los resultados de un catálogo de 527 entradas extraídas de una muestra de 165 159 palabras gráficas y 15 916 vocablos.

Tabla 56. Grupos de sufijos de la Nueva España (siglo XVIII);
primeras 10 del catálogo de 527 entradas

	sufijo	frec.	cuadros	economía	entropía	afijalidad
1	~a	1 462	0.6558	0.9564	0.9371	0.8497
2	~o	1 418	0.7031	0.9574	0.8153	0.8253
3	~s	1 761	1.0000	1.0000	0.4558	0.8187
4	~as	593	0.3461	0.9503	0.9331	0.7432
5	~ó	287	0.4709	0.8447	0.9094	0.7417
6	~os	657	0.3799	0.9460	0.8863	0.7374

	sufijo	frec.	cuadros	economía	entropía	afijalidad
7	~ar	353	0.2279	0.9533	0.9658	0.7157
8	~ado	341	0.2184	0.9566	0.9555	0.7102
9	~e	629	0.3629	0.8919	0.7683	0.6744
10	~an	285	0.1799	0.9059	0.9235	0.6698

Las muestras de las que se extrajeron estos dos conjuntos de afijos son pequeñas. En cambio, la Tabla 57, que corresponde al siglo xx, presenta los 10 morfos más afijales de 749 afijos y grupos de afijos extraídos del CEMC, que revisamos en los capítulos anteriores y que cuenta con aproximadamente dos millones de palabras gráficas y casi 70 000 vocablos.

Tabla 57. Grupos de sufijos del México del siglo xx; primeras 10 del catálogo de 749 entradas

	sufijo	frec.	cuadros	economía	entropía	afijalidad
1	~ó	1 428	0.7371	0.9192	0.8720	0.8428
2	~o	6 314	0.6860	0.9788	0.8017	0.8222
3	~s	12 013	1.0000	0.9968	0.4609	0.8192
4	~a	7 687	0.5753	0.9818	0.8888	0.8153
5	~os	4 554	0.4775	0.9754	0.8235	0.7588
6	~as	4 324	0.4216	0.9779	0.8645	0.7547
7	~en	945	0.4107	0.8991	0.9060	0.7386
8	~ar	1 633	0.2178	0.9621	0.9149	0.6982
9	~ado	1 429	0.2061	0.9619	0.9070	0.6917
10	~ando	976	0.1836	0.9544	0.9162	0.6847

Como en otros experimentos, algunas formas son grupos de sufijos; por ejemplo, la forma *~o.s* (rango 5 en la Tabla 57) que encapsula dos morfemas flexivos: marcadores de género nominal y de número.

Por otra parte, los sufijos con valores de afijalidad altos, como los de estas tres tablas, no necesariamente son parecidos en función o comportamiento, lo que se refleja en las cantidades de cuadros, economía y entropía. Por ejemplo, el sufijo *~s* es alto en cuadros y economía, pero relativamente bajo en entropía, mientras que sufijo *~ar* tiene una puntuación alta en economía y entropía, pero baja en cuadros. Ambos aparecen entre las 10 primeras formas de las Tablas 55, 56 y 57.

Como se mencionó anteriormente, la entropía es una buena técnica para extraer las secuencias de afijos adheridas a las raíces, mientras que el índice de economía es mejor para extraer los afijos al exterior de las palabras, esto es, los de flexión. Así que, un afijo con altas puntuaciones de cuadros y economía, combinadas con baja entropía sería característico de las flexiones más externas, típicamente adheridas a otros afijos, inclusive flexivos, que es el caso del sufijo *~s*: como marca de plural, se puede unir a sustantivos ya flexionados (por ejemplo, *sólid~o.**s***, *erudit~a.**s***, *prend~id.it.o.**s***, *canast~it.a.**s***).

Por otro lado, un afijo con altos valores de entropía y economía, pero un bajo número de cuadros sería más probablemente una flexión que se une a las raíces, como el sufijo *~ar* que se adhiere directamente a las raíces verbales y que no es un morfema libre (por ejemplo, *compr~**ar***, *naveg~**ar***, *alegr~**ar***).

Ciertamente, se percibe la similitud entre los elementos de los perfiles en estas tres tablas. También se notan sus diferencias: las dos primeras (Tablas 55 y 56) comparten 9 de las 10 primeras formas con una ligera variación en el orden. La Tabla 57 comparte ocho artículos con las otras dos tablas y también exhibe alguna variación en el orden. Sin embargo, es difícil hacer juicios sobre su similitud sin una medida como la distancia euclidiana entre los valores de afijalidad.

4.4.3.3 Las matrices de distancias euclidianas

Las distancias euclidianas calculadas para todos los elementos de los tres perfiles morfológicos, parcialmente presentados en las tablas anteriores,

se exhiben en la Tabla 58. Cada celda contiene las distancias calculadas a partir de los afijos y grupos de afijos compartidos. De nuevo, la diagonal representa la distancia de la lengua de cada siglo a sí misma.

Tabla 58. Matriz de distancias euclidianas entre muestras diacrónicas del español en México

	siglo XVI	**siglo XVIII (Nueva España)**	**siglo XX (México)**
siglo XVI	0	0.0781	*0.0913*
siglo XVIII (Nueva España)		0	**0.0715**
siglo XX (México)			0

Estas distancias son promedios de diferencias entre valores normalizados de afijalidad, por lo que son valores que oscilan entre 0 y 1. En esta matriz estamos examinando etapas diacrónicas relativamente cercanas de una lengua conocida por su naturaleza conservadora. Lo que hay que notar de los valores de afijalidad que sirvieron para medir las distancias euclidianas es que se obtuvieron de un conjunto de muestras desiguales, dos relativamente pequeñas y una comparativamente enorme, el CEMC. Se podría esperar que esto interfiriera con los valores de la Tabla 58. Sin embargo, como podía haberse esperado, la distancia entre el siglo XVIII y el XX es la menor (en negritas) y la distancia entre éste y el XVI es la mayor (en cursivas).

Ciertamente, para hablar de la emergencia del español mexicano habría que tomar en cuenta otros fenómenos, como las condiciones sociohistóricas. Se puede señalar que una variedad propia del español pudo haber existido desde el principio del asentamiento y que se deben tomar en cuenta cuestiones ideológicas e identitarias para poder afirmar que el español de México emergió en tal o cual fecha. Lo que el cálculo de distancias euclidianas indica es que las variedades de los siglos XVIII y XX se parecen más entre sí (tienen una menor distancia) y que la va-

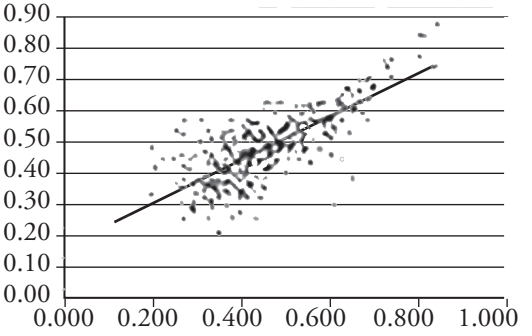
riedad del siglo XVI es menos similar a la del XVIII (son más distantes), como si hubiera habido un salto en el ritmo de cambio de la morfología afijal, porque en menos tiempo cambió más.

La Figura 17 muestra gráficas de dispersión para cada par de siglos. Cada elemento compartido por cada par de perfiles morfológicos aparece dibujado como un punto en el espacio cartesiano en la gráfica correspondiente. Así, los puntos en la Gráfica B, correspondiente a los siglos XVIII y XX, se ven más apretados que los de las otras gráficas (distancia euclidiana de 0.0715). La Gráfica C, correspondiente a los siglos XVI y XX, es la más dispersa (exhibe una distancia euclidiana mayor de 0.0913). La dispersión en la Gráfica A, correspondiente a los siglos XVI y XVIII, es visualmente mayor que la de B, lo que apunta, de nuevo, al posible salto que dio lugar a un sistema dialectal mexicano, en algún momento antes del siglo XVIII. Sin embargo, todas estas observaciones son necesariamente circunstanciales, porque está en entredicho la representatividad de las muestras pequeñas.

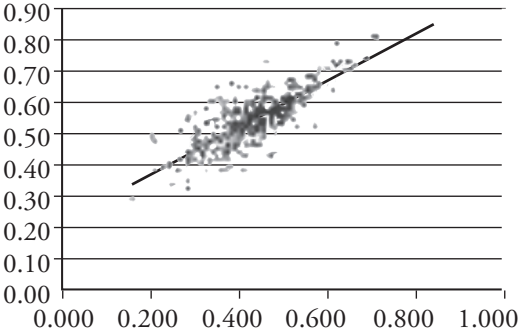
En este contexto, tiene sentido considerar otra variedad del español para contrastar las distancias o similitudes entre estos perfiles. Esto es, a los números de la matriz de distancias, les podemos agregar contexto mediante la inclusión de otros registros del español. Aunque esta lengua consta de varios dialectos nacionales prestigiosos, se escogió el español de España para este experimento, en parte por razones históricas, pero sobre todo por la mayor disponibilidad de muestras textuales.

De esta manera, también se obtuvieron pequeñas muestras textuales del español de España, de los siglos XVIII y XX[18]. Esto se llevó a cabo mediante varias búsquedas al Corpus de Referencia del Español Actual

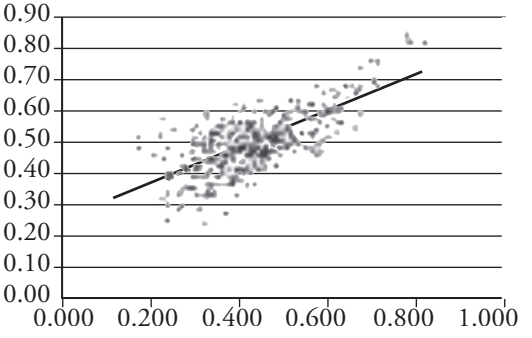
[18] Ferrasi y Bernardini (2010), en su reseña del libro de Gries, Wulff y Davies, eds., *op. cit.* (2010), señalan que las muestras de los siglos XVI y XVIII de este experimento no fueron descritas apropiadamente en Medina Urrea, art. cit. (2010) y sugieren que se pudo haber incluido también una muestra del siglo XVI del español de España. Ciertamente, valdría la pena incluir más datos de más estados de lengua y de más regiones en el experimento, pero es necesario aclarar que la muestra del siglo XVI utilizada no puede considerarse representativa del español de la Nueva España, principalmente porque todavía no había echado raíces en el Nuevo Mundo. De hecho, muchos de los documentos del siglo XVI fueron recolectados de archivos de México y España (del Archivo General



A (0.0781)



B (0.0715)



C (0.0913)

Figura 17. Gráficas de dispersión

Los datos de los siglos XVI y XVIII se comparan en A (374 entradas compartidas de afijos y grupos sufijales); los datos de los siglos XVIII y XX se comparan en B (362 entradas compartidas); y los datos de los siglos XVI y XX se comparan en C (335 entradas compartidas). Las distancias euclidianas aparecen entre paréntesis después de las etiquetas A, B y C.

(CREA) y al Corpus Diacrónico del Español (CORDE) de la Real Academia Española. Se buscaron palabras de contenido (por ejemplo, "elefante") y de españolismos (como "bañador") en documentos producidos en España para cada estado de lengua. Las concordancias resultantes se reunieron como muestras textuales. La Tabla 59 contiene los grupos sufijales castellanos del siglo XVIII y la Tabla 60 contiene los del siglo XX.

Tabla 59. Grupos de sufijos de la España (siglo XVIII); primeras 10 del catálogo de 429 entradas

	sufijo	frec.	cuadros	economía	Entropía	afijalidad
1	~s	1 546	1.0000	1.0000	0.5012	0.8338
2	~a	1 063	0.5585	0.9573	0.9628	0.8262
3	~o	1 050	0.5792	0.9448	0.8379	0.7873
4	~as	477	0.3156	0.9417	0.9360	0.7311
5	~os	586	0.3242	0.9278	0.9002	0.7174
6	~ar	245	0.2027	0.9449	0.9743	0.7073
7	~ado	221	0.1507	0.9393	0.9724	0.6875
8	~an	247	0.1676	0.9016	0.9422	0.6705
9	~ando	128	0.1147	0.9016	0.9641	0.6601
10	~ó	200	0.3371	0.7885	0.8369	0.6542

El catálogo presentado en la Tabla 59, correspondiente al siglo XVIII, tiene un total de 429 entradas, extraídas de una colección de 96 877 palabras gráficas y 13 882 vocablos. Similarmente, los afijos de la Tabla 60, correspondiente al siglo XX, son los 10 primeros de 551 formas, extraídas de una muestra de 125 969 palabras gráficas (17 509 vocablos). De nuevo, las muestras son pequeñas y los resultados se parecen mucho entre sí.

---

de la Nación y del Archivo de Indias), escritos en su mayoría por españoles, o son intercambios epistolares escritos también por españoles y enviados de España a América o viceversa.

Tabla 60. Grupos de sufijos de la España del siglo XX; primeras 10 del catálogo de 551 entradas

	sufijo	frec.	cuadros	economía	Entropía	afijalidad
1	~s	2,071	1.0000	0.9952	0.5179	0.8378
2	~a	1,470	0.5112	0.9617	0.9727	0.8152
3	~o	1,398	0.5961	0.9501	0.8692	0.8051
4	~as	760	0.3452	0.9486	0.9710	0.7549
5	~os	670	0.3196	0.9297	0.9221	0.7238
6	~ó	303	0.3869	0.8556	0.9064	0.7163
7	~ado	337	0.1668	0.9261	0.9786	0.6905
8	~ar	359	0.1830	0.9086	0.9676	0.6864
9	~aba	217	0.1423	0.9125	0.9587	0.6712
10	~ando	167	0.1149	0.8934	0.9913	0.6665

En la Tabla 61 se reúnen todos los valores de afijalidad de todas las formas comunes a las cinco muestras textuales. Con estos valores se calcularon las distancias euclidianas que aparecen en la Tabla 62, incluida la información de la Tabla 58.

Tabla 61. Sufijos del español en tres estados de lengua, en México y España

	sufijos	siglo XVI	siglo XVIII (Nueva España)	siglo XX (México)	siglo XVIII (España)	siglo XX (España)
1	~a	0.8450	0.8452	0.8153	0.8262	0.8040
2	~s	0.8153	0.8064	0.8192	0.8338	0.8271
3	~o	0.8040	0.8127	0.8222	0.7873	0.7959
4	~ó	0.8053	0.7405	0.8428	0.6542	0.7212
5	~as	0.7418	0.7338	0.7547	0.7311	0.7415
6	~os	0.7269	0.7276	0.7588	0.7174	0.7110

	sufijos	siglo XVI	siglo XVIII (Nueva España)	siglo XX (México)	siglo XVIII (España)	siglo XX (España)
7	~ar	0.7329	0.7122	0.6982	0.7073	0.6876
8	~ado	0.7314	0.7074	0.6917	0.6875	0.6882
9	~e	0.7415	0.6773	0.6833	0.6520	0.6228
10	~ando	0.6846	0.6645	0.6847	0.6601	0.6622
11	~an	0.6908	0.6732	0.6579	0.6705	0.6235
12	~en	0.7036	0.6442	0.7386	0.6154	0.6033
13	~aron	0.7030	0.6511	0.6773	0.6235	0.6477
14	~ada	0.6707	0.6673	0.6768	0.6230	0.6419
15	~ados	0.6961	0.6448	0.6678	0.6158	0.6408
16	~aba	0.6672	0.6355	0.6803	0.5957	0.6753
17	~arse	0.6389	0.6404	0.6692	0.6071	0.6273
18	~adas	0.6533	0.6227	0.6602	0.6065	0.6341
19	~ido	0.6709	0.6314	0.6248	0.5967	0.6168
20	~amente	0.6431	0.6360	0.6169	0.6119	0.6101
21	~aban	0.6207	0.6067	0.6610	0.5836	0.6245
22	~es	0.6319	0.6196	0.6097	0.6063	0.6162
23	~ara	0.6463	0.6377	0.6467	0.5286	0.5949
24	~amos	0.6203	0.5902	0.6187	0.6057	0.6025
25	~ase	0.7034	0.6369	0.5925	0.5531	0.5477
26	~n	0.6542	0.6011	0.5977	0.5891	0.5589
27	~ará	0.6715	0.6030	0.6415	0.5745	0.5105
28	~ir	0.6318	0.5918	0.5938	0.5804	0.5729
29	~ida	0.6257	0.5780	0.6022	0.5666	0.5915
30	~iendo	0.6398	0.5826	0.5738	0.5546	0.5736
31	~ero	0.5772	0.5908	0.5953	0.5650	0.5909
32	~i	0.6533	0.5703	0.5680	0.5649	0.5364
33	~arme	0.5631	0.5760	0.6306	0.5313	0.5838
34	~idos	0.6095	0.5502	0.6030	0.5489	0.5682
35	~er	0.6605	0.5739	0.5661	0.5191	0.5521
36	~andose	0.5876	0.5947	0.6262	0.4840	0.5390

	sufijos	siglo XVI	siglo XVIII (Nueva España)	siglo XX (México)	siglo XVIII (España)	siglo XX (España)
37	~ió	0.6105	0.5609	0.5698	0.5395	0.5458
38	~aré	0.6300	0.5430	0.5960	0.5293	0.5195
39	~idas	0.5826	0.5559	0.5810	0.5390	0.5547
40	~eros	0.5298	0.5816	0.5867	0.5572	0.5466
41	~ante	0.5809	0.5386	0.5998	0.5276	0.5459
42	~eras	0.5373	0.5581	0.5822	0.5485	0.5661
43	~ía	0.6482	0.5021	0.5371	0.5430	0.5608
44	~aría	0.5933	0.5480	0.6198	0.4982	0.5305
45	~arlo	0.5450	0.5830	0.6356	0.4789	0.5382
46	~osa	0.5418	0.5598	0.5802	0.5315	0.5666
47	~ieron	0.6285	0.5380	0.5542	0.5266	0.5324
48	~arla	0.4753	0.5909	0.6350	0.5213	0.5532
49	~arle	0.5673	0.5625	0.5997	0.4913	0.5501
50	~ito	0.5108	0.5496	0.6093	0.4838	0.6157
51	~era	0.5587	0.5456	0.5569	0.5405	0.5530
52	~é	0.6048	0.4840	0.6818	0.4647	0.5108
53	~lo	0.5999	0.5388	0.5384	0.5361	0.5197
54	~oso	0.5254	0.5327	0.5732	0.5336	0.5512
55	~arán	0.5935	0.5399	0.6257	0.4346	0.5068
56	~iera	0.5706	0.5407	0.5332	0.4990	0.5532
57	~ales	0.5376	0.5306	0.5509	0.5239	0.5466
58	~ita	0.4198	0.5383	0.6219	0.5169	0.5856
59	~al	0.5360	0.5290	0.5515	0.5058	0.5531
60	~antes	0.4870	0.5001	0.6037	0.5207	0.5612
61	~ador	0.4410	0.5324	0.6147	0.5273	0.5568
62	~la	0.5674	0.5284	0.5306	0.5175	0.5171
63	~ino	0.4783	0.5656	0.5500	0.5129	0.5437
64	~idad	0.5090	0.5293	0.5214	0.5308	0.5432
65	~r	0.5521	0.5316	0.5161	0.5216	0.5036
66	~andole	0.6048	0.5418	0.5417	0.4690	0.4650

	sufijos	siglo XVI	siglo XVIII (Nueva España)	siglo XX (México)	siglo XVIII (España)	siglo XX (España)
67	~osas	0.5199	0.5401	0.5653	0.4342	0.5575
68	~tar	0.5052	0.5194	0.5276	0.5159	0.5312
69	~los	0.5459	0.4947	0.5182	0.5115	0.5287
70	~le	0.5945	0.5129	0.5260	0.4951	0.4690
71	~tos	0.5085	0.5181	0.5062	0.5098	0.5358
72	~se	0.5901	0.5273	0.5332	0.4825	0.4440
73	~ta	0.5560	0.5122	0.5061	0.5054	0.4968
74	~tas	0.5228	0.4966	0.5062	0.5184	0.5293
75	~is	0.5151	0.4641	0.5236	0.5367	0.5273
76	~mas	0.5096	0.5461	0.4976	0.5119	0.4974
77	~ra	0.5561	0.5006	0.5144	0.4954	0.4944
78	~rado	0.4990	0.5170	0.5134	0.4948	0.5350
79	~to	0.5356	0.5137	0.4879	0.5039	0.5163
80	~ían	0.6310	0.4168	0.5226	0.4837	0.4871
81	~do	0.5258	0.5201	0.5092	0.4932	0.4925
82	~ano	0.4480	0.5013	0.5493	0.4836	0.5520
83	~imos	0.5446	0.4410	0.5317	0.4667	0.5485
84	~irse	0.4336	0.5305	0.5659	0.4781	0.5227
85	~iese	0.6157	0.5173	0.5117	0.4658	0.4201
86	~anos	0.5043	0.5110	0.5396	0.4840	0.4865
87	~las	0.4841	0.4996	0.5171	0.5054	0.5186
88	~ia	0.6060	0.5236	0.4948	0.4622	0.4123
89	~arlos	0.5123	0.4873	0.6153	0.4632	0.4162
90	~les	0.5641	0.4799	0.4973	0.4941	0.4577
91	~osos	0.4100	0.5212	0.5673	0.5041	0.4886
92	~io	0.5967	0.5552	0.4780	0.4390	0.4216
93	~ora	0.4600	0.5949	0.4793	0.5640	0.3864
94	~tado	0.4906	0.5103	0.5385	0.4235	0.5202
95	~ios	0.4725	0.4985	0.4865	0.5423	0.4817
96	~able	0.4303	0.4873	0.5679	0.4853	0.5047

	sufijos	siglo XVI	siglo XVIII (Nueva España)	siglo XX (México)	siglo XVIII (España)	siglo XX (España)
97	~í	0.5261	0.4380	0.5279	0.4836	0.4970
98	~ran	0.5210	0.4698	0.4873	0.5040	0.4872
99	~tan	0.4680	0.5106	0.5259	0.4680	0.4946
100	~itos	0.4348	0.4875	0.5984	0.4483	0.4928
101	~elo	0.4455	0.5071	0.5194	0.4489	0.5330
102	~iko	0.3949	0.4786	0.5581	0.4987	0.5230
103	~ras	0.4705	0.4426	0.5157	0.5158	0.5029
104	~emos	0.5794	0.4483	0.4888	0.4413	0.4887
105	~tó	0.4935	0.5007	0.5283	0.3921	0.5143
106	~re	0.5409	0.4620	0.4928	0.4544	0.4753
107	~eo	0.4095	0.5236	0.5342	0.4774	0.4806
108	~adores	0.4626	0.4246	0.6033	0.4116	0.5147
109	~ieran	0.5044	0.5297	0.5067	0.4149	0.4595
110	~in	0.4083	0.4862	0.5180	0.4748	0.5233
111	~so	0.4935	0.4805	0.4961	0.4528	0.4846
112	~arnos	0.5004	0.4612	0.5917	0.3543	0.4998
113	~arían	0.4493	0.4811	0.5715	0.4388	0.4582
114	~ores	0.4148	0.5045	0.4584	0.5125	0.5079
115	~isimo	0.4609	0.4696	0.5292	0.4260	0.5032
116	~ías	0.3303	0.5071	0.5244	0.5038	0.5207
117	~ros	0.4823	0.4776	0.4442	0.4769	0.5044
118	~or	0.4800	0.4835	0.4573	0.4847	0.4768
119	~ejo	0.3890	0.4993	0.5203	0.5328	0.4406
120	~na	0.4889	0.4649	0.4786	0.4914	0.4557
121	~mente	0.4795	0.4792	0.4728	0.4657	0.4753
122	~res	0.4904	0.4611	0.4711	0.4656	0.4790
123	~nos	0.5233	0.4633	0.4643	0.4607	0.4518
124	~amiento	0.5039	0.3756	0.6005	0.4184	0.4642
125	~enta	0.4184	0.4806	0.4852	0.4576	0.5125
126	~irme	0.4175	0.4812	0.5322	0.4644	0.4564
127	~t	0.4600	0.4603	0.5200	0.4412	0.4591

	sufijos	siglo XVI	siglo XVIII (Nueva España)	siglo XX (México)	siglo XVIII (España)	siglo XX (España)
128	~ro	0.4527	0.4834	0.4613	0.4473	0.4904
129	~il	0.4237	0.4965	0.5104	0.4439	0.4595
130	~erlo	0.4834	0.4158	0.4729	0.4569	0.5049
131	~iendose	0.5089	0.4198	0.4850	0.4716	0.4486
132	~ería	0.3662	0.4776	0.5364	0.4752	0.4769
133	~esta	0.4451	0.4063	0.5161	0.5374	0.4256
134	~ke	0.4669	0.4495	0.5047	0.4693	0.4305
135	~mos	0.5112	0.4485	0.4818	0.4292	0.4465
136	~ma	0.4819	0.4086	0.4718	0.4633	0.4891
137	~da	0.4992	0.4676	0.4732	0.4320	0.4427
138	~erse	0.5427	0.4733	0.5020	0.3731	0.4210
139	~de	0.4870	0.4761	0.4598	0.4462	0.4377
140	~ese	0.4377	0.4742	0.5256	0.3757	0.4858
141	~nas	0.4368	0.3635	0.4913	0.4836	0.5237
142	~sa	0.4418	0.4722	0.4936	0.4163	0.4688
143	~iente	0.4336	0.4753	0.4821	0.4411	0.4589
144	~imiento	0.5525	0.4533	0.4751	0.4118	0.3939
145	~me	0.5086	0.4354	0.4744	0.4196	0.4320
146	~ría	0.4953	0.4511	0.4527	0.4303	0.4362
147	~el	0.4825	0.4401	0.4790	0.4565	0.4072
148	~go	0.4913	0.4626	0.4616	0.4266	0.4217
149	~ones	0.4262	0.4240	0.4777	0.4804	0.4540
150	~no	0.4750	0.4627	0.4659	0.4257	0.4327
151	~és	0.3430	0.4127	0.5357	0.4557	0.5100
152	~das	0.4850	0.4154	0.4713	0.4419	0.4293
153	~iado	0.3375	0.4796	0.5004	0.3909	0.5336
154	~ias	0.4651	0.5202	0.4842	0.3223	0.4467
155	~ientes	0.3782	0.4523	0.5048	0.4258	0.4665
156	~ria	0.5140	0.4552	0.4317	0.4243	0.3982
157	~mo	0.4696	0.4551	0.4219	0.4723	0.3935

	sufijos	siglo XVI	siglo XVIII (Nueva España)	siglo XX (México)	siglo XVIII (España)	siglo XX (España)
158	~erme	0.4860	0.4422	0.4247	0.4484	0.3948
159	~ga	0.4132	0.4457	0.4590	0.4117	0.4636
160	~igo	0.3232	0.4404	0.5413	0.4861	0.3953
161	~ere	0.4290	0.4192	0.4889	0.4050	0.4276
162	~te	0.4638	0.3904	0.4769	0.4076	0.4242
163	~aje	0.3688	0.4172	0.5155	0.3683	0.4888
164	~ario	0.3841	0.4818	0.4935	0.4133	0.3849
165	~dos	0.4573	0.4155	0.4616	0.4058	0.3961
166	~uso	0.3619	0.4464	0.4965	0.4609	0.3684
167	~iré	0.3224	0.4410	0.4710	0.4713	0.4116
168	~ura	0.4156	0.4346	0.4577	0.4272	0.3754
169	~idades	0.2691	0.4302	0.4865	0.4538	0.4557
170	~rse	0.4467	0.4223	0.4243	0.3993	0.3899
171	~ré	0.4345	0.4310	0.4370	0.4042	0.3724
172	~edad	0.4295	0.4521	0.4211	0.3839	0.3775
173	~isima	0.2512	0.4210	0.5255	0.3775	0.4868
174	~ko	0.4643	0.4015	0.3740	0.3941	0.4128
175	~l	0.4626	0.3851	0.4321	0.3754	0.3867
176	~ón	0.3977	0.3732	0.4430	0.3922	0.4312
177	~rían	0.4418	0.4160	0.4183	0.3824	0.3701
178	~rme	0.4252	0.3965	0.4144	0.3829	0.4013
179	~rá	0.4269	0.3969	0.4221	0.4016	0.3656
180	~rán	0.4194	0.3976	0.4431	0.3771	0.3630
181	~ndo	0.4411	0.4016	0.4111	0.3856	0.3572
182	~ron	0.4539	0.3868	0.4090	0.3729	0.3719
183	~m	0.3694	0.3749	0.4742	0.3849	0.3909
184	~tes	0.4007	0.4199	0.4582	0.3081	0.4050
185	~rle	0.4054	0.3996	0.4228	0.3773	0.3868
186	~rla	0.4124	0.3660	0.3926	0.4182	0.3666
187	~remos	0.4218	0.3273	0.4156	0.3585	0.4245
188	~ka	0.4455	0.3558	0.3682	0.4019	0.3708

	sufijos	siglo XVI	siglo XVIII (Nueva España)	siglo XX (México)	siglo XVIII (España)	siglo XX (España)
189	~irá	0.2019	0.4553	0.5177	0.3837	0.3667
190	~endo	0.4117	0.3765	0.3792	0.3794	0.3619
191	~ente	0.4028	0.3865	0.3721	0.3682	0.3737
192	~rio	0.3925	0.4139	0.4073	0.3550	0.3237
193	~on	0.4247	0.4344	0.3372	0.4175	0.2639
194	~dor	0.4235	0.3416	0.3902	0.3740	0.3475
195	~án	0.3858	0.3691	0.3749	0.4407	0.3059
196	~eron	0.3576	0.4030	0.3521	0.3573	0.3845
197	~uras	0.3003	0.4261	0.4632	0.3772	0.2856
198	~tro	0.2750	0.3517	0.4062	0.4312	0.3685
199	~ja	0.3249	0.3303	0.4509	0.3400	0.3849
200	~sión	0.3377	0.3520	0.4213	0.3646	0.3466
201	~rlos	0.3703	0.3132	0.3711	0.3481	0.4005
202	~ás	0.3610	0.3715	0.4093	0.3297	0.2917
203	~d	0.4121	0.3083	0.3718	0.3456	0.3103
204	~rnos	0.3447	0.3998	0.3932	0.2569	0.3518
205	~ndose	0.3242	0.3404	0.3986	0.3370	0.3435
206	~nte	0.3644	0.3385	0.3659	0.3445	0.3285
207	~ste	0.2711	0.4942	0.3839	0.2749	0.3060
208	~dores	0.4143	0.2936	0.3736	0.2972	0.3086
209	~ntes	0.3191	0.3126	0.3591	0.3414	0.3329
210	~nta	0.3545	0.3515	0.3336	0.2998	0.3045
211	~á	0.3362	0.2876	0.3396	0.3491	0.2697
212	~ndole	0.3543	0.3108	0.3988	0.3086	0.2010
213	~miento	0.3714	0.2688	0.3500	0.2755	0.3041
214	~ba	0.3788	0.3110	0.3967	0.2216	0.2457
215	~nes	0.2948	0.2511	0.3427	0.3201	0.3435
216	~dad	0.3994	0.3086	0.2826	0.3025	0.2361
217	~ble	0.3009	0.3078	0.3569	0.2776	0.2771
218	~ión	0.2953	0.2793	0.2938	0.3000	0.2945
219	~ad	0.3531	0.1834	0.2517	0.3641	0.2914

Como antes, las celdas de la Tabla 62 albergan las distancias entre los perfiles morfológicos que encabezan los renglones con los que encabezan las columnas. Así, la distancia euclidiana entre la morfología afijal del español de México del siglo xx y la del español de España del mismo siglo es 0.715, que no es ni la más larga ni la más corta de la matriz.

Tabla 62. Matriz de distancias euclidianas entre algunas muestras diacrónicas del español en México y España

	siglo XVI	**siglo XVIII (Nueva España)**	**siglo XX (México)**	**siglo XVIII (España)**	**siglo XX (España)**
siglo XVI	0	0.0781	***0.0913***	***0.0833***	***0.0877***
siglo XVIII (Nueva España)		0	**0.0715**	**0.0591**	**0.0643**
siglo XX (México)			0	0.0804	**0.0715**
siglo XVIII (España)				0	**0.0576**
siglo XX (España)					0

De nuevo, la diagonal muestra ceros porque contiene las distancias de cada muestra a sí misma. Los datos muestran una pauta interesante. Las cinco distancias más cortas (en negritas con fondo gris claro) corresponden a las muestras de los siglos XVIII y XX. De hecho, la muestra del español mexicano del siglo XX se parece tanto a la del español del siglo XVIII en la Nueva España (0.0715) como a la del castellano del siglo XX (0.0715). Además, la muestra del español del siglo XVI es la más distante de todas (números blancos en cursivas), lo que apunta a que hubo un salto entre la morfología sufijal de ese siglo y la de los otros estados de lengua tanto de México como de España.

Los datos de la Tabla 62 se pueden visualizar mediante un dendograma generado mediante un análisis de agrupamiento o clúster jerárquico (*hierarchical cluster analysis*). La Figura 18 muestra los resultados de este análisis[19].

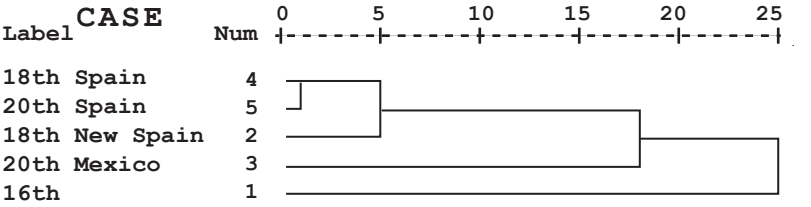


Figura 18. Dendograma de agrupamiento jerárquico de distancias euclidianas entre muestras diacrónicas del español de México y España
Tomada de Medina (2010, 42). Método del vecino más cercano (*Nearest Neighbor Method, Complete Linkage*); rescalamiento al rango [0-25].

En esta figura, las muestras españolas de los siglos XVIII y XX son las más cercanas. El siguiente perfil similar es el registro de la Nueva España del siglo XVIII. El siglo XX del español mexicano viene después. Finalmente, el perfil del siglo XVI es el más lejano de todos. El hecho de que este último registro sea el más distante morfológicamente señala un posible cambio general en el español de ambos lados del Atlántico, cuando las primeras oleadas de europeos llegaron a América. Como sea, estos datos parecen compatibles con lo que podíamos haber esperado.

En resumen, hasta aquí se presentaron datos cuantitativos de tres siglos de algunos registros de la lengua española. Los datos fueron extraídos de muestras de español de los siglos XVI, XVIII y XX provenientes de lo que hoy es México y España. Aunque estas muestras son peque-

[19] El dendograma es una representación visual de la distancia en la que se combinan los agrupamientos o clústers. Se examina de izquierda a derecha. Las líneas verticales se unen para mostrar los clústers. La posición de la línea en la escala indica la distancia en la que se unen. No se observan distancias reales. Las distancias observadas son reescaladas al rango 1 a 25 (la razón de las distancias reescaladas dentro del dendograma es la misma que la razón de las distancias originales).

ñas y no son realmente representativas (excepto en el caso del CEMC), pueden hacerse algunas observaciones de carácter preliminar relativas al nivel morfológico. Por un lado, el español de México parece haber emergido como un sistema dialectal antes del siglo XVIII, ya que las morfologías afijales de los siglos XVIII y XX están menos alejadas entre sí. Además, según los datos, el español contemporáneo de España parece estar más cerca del español del siglo XVIII de la Nueva España que del español mexicano del siglo XX. Esto posiblemente se debe a un sesgo en el registro escrito. Como sea, también parece un indicio de la naturaleza conservadora de estos dialectos del español en el nivel morfológico.

Como se vio hasta aquí, el método aplicado para descubrir afijos sirve para compilar, de manera no supervisada, catálogos de afijos y grupos de afijos de lenguas concatenativas como las mayas y el español. Como se ve, estos catálogos o perfiles morfológicos parecen caracterizar de manera íntima las muestras examinadas en este experimento. En este sentido, pueden considerarse verdaderas huellas dactilares. Además, se calcularon distancias euclidianas para medir la separación entre estos perfiles. Los resultados presentados son interesantes, aunque las muestras textuales utilizadas no sean verdaderamente representativas de los estados de lengua de los que provienen. De todas maneras, la medición de distancias y similitudes entre conjuntos de afijos y sus grupos afijales parece permitir la comparación de estados diacrónicos en periodos de tiempo relativamente cortos (en comparación a los de la glotocronología, que mide milenios), al menos en el caso del español escrito.

Sin duda, estos experimentos pueden mejorarse de muchas maneras, por ejemplo, probando otras técnicas de segmentación, con el fin de medir la afijalidad, teniendo en cuenta también valores para clíticos y otros modificadores y aplicando métodos alternativos para medir distancias y similitudes entre perfiles. Sería valioso aplicar estos métodos a los sistemas dialectales de otros idiomas en sus dimensiones diacrónica y sincrónica, así como geográfica y social. ¿Es posible corroborar que

el registro de clase media del mundo de habla hispana representa un conjunto de dialectos con mayor proximidad entre sí que con cualquier registro de las clases menos privilegiadas? ¿Podrían examinarse los cambios sobresalientes del purépecha de los siglos XVI y XIX, midiendo las distancias euclidianas entre sus perfiles morfológicos?

OBSERVACIONES FINALES

Sin duda, los experimentos presentados para aprender sobre la morfología afijal de lenguas concatenativas como el español, el chuj, el ralámuli, etc. pueden ser mejorados de muchas maneras. Sin embargo, es posible ver que con pocas herramientas conceptuales se logra mucho. Ojalá estos instrumentos sirvan también para reforzar o motivar el interés en los métodos cuantitativos y en el uso de muestras textuales para conocer éstos y otros fenómenos lingüísticos.

Conviene recapitular lo presentado hasta aquí. Al principio de este trabajo, se mostró un panorama general de la morfología computacional y de los métodos de reconocimiento supervisado y descubrimiento automático no supervisado de la morfología. Este panorama sirvió de marco para el desarrollo de los capítulos posteriores. Luego, se describieron las herramientas para llevar a cabo una investigación del nivel morfológico, destinada al descubrimiento automático de signos afijales del español, que utilizó como fuente de datos el CEMC. En el capítulo siguiente, se presentó un panorama de aplicaciones y recursos sencillos de evaluación, en el que se consideraron muestras de otras lenguas como el checo, el ralámuli y el chuj y se examinaron algunos desarrollos y aplicaciones, cuya ejecución puede beneficiarse de la información afijal extraída de las muestras textuales. Finalmente, en el último capítulo se reportaron algunos experimentos para examinar la variación en afijalidad en extracciones de afijos y cadenas afijales de diversas fuentes textuales de cuatro lenguas emparentadas (maya) y de tres estados de lengua del español.

Una cuestión clave en este trabajo, que también lo es en trabajos basados en corpus en general, es la dificultad de garantizar un nivel óptimo de representatividad de las muestras, además de que su recolección puede ser cara en términos de tiempo y recursos técnicos y humanos. Construir un corpus verdaderamente representativo de una lengua

requiere de mucho tiempo y dinero, por no hablar del gran esfuerzo humano involucrado. Aun así, los lingüistas a menudo recolectamos información de diversas fuentes para compilar pequeñas muestras textuales (algunas muy limitadas, otras menos), sin el rigor necesario para garantizar el nivel anhelado de representatividad de las lenguas o los fenómenos asociados a ellas. Sin embargo, también nos dan la oportunidad de aprender cosas sobre el lenguaje.

Por otra parte, un enfoque cuantitativo como el mostrado tiene muchas ventajas. En particular para los temas abordados aquí, al concebirse y cuantificarse el fenómeno de afijalidad, tenemos por lo menos las ventajas de la cuantificación que, como apunta Bunge, son: refinamiento conceptual y descripción y clasificación más precisas (Bunge 1967, 202). Además, resulta ser un recurso valioso para descubrir automáticamente las unidades más afijales de un corpus sin apoyarse en el conocimiento de la lengua allí representada. Naturalmente, esto no excluye la necesidad de aplicar otras técnicas, como las cualitativas, pero constituye un buen marco del cual se puede partir.

Otra ventaja es la posibilidad de conocer de una nueva manera las unidades lingüísticas de lenguas muy estudiadas. Aunque sean muy conocidas, como la española, al observar sus elementos y clasificarlos mediante representaciones numéricas abstractas que los caractericen como partes de un sistema, estas unidades se pueden continuar investigando y describiendo de manera cuantitativa. De hecho, como hemos visto, sus representaciones numéricas constituyen descripciones de los sistemas a los que pertenecen.

En este trabajo se ha visto que, a partir de una muestra, el índice de afijalidad resulta en valores altos para afijos y valores menores para todos los demás segmentos; es decir, que el carácter de ser elemento estructural de una palabra es una propiedad que en efecto se puede estimar mediante los cálculos de entropía, cuadros y economía. De esta manera, medimos algo que se cuela entre bases y afijos y que los caracteriza como algo más que sus significantes: sus relaciones con el resto de los segmentos del corpus.

En este contexto, hemos visto que los resultados de medir la afijalidad pueden variar según se trate de descubrir prefijos o afijos. Por ejemplo, en español los prefijos difieren de los sufijos en que los segundos forman parte de un aparato morfosintáctico denso que los hace más susceptibles de ser descubiertos mediante los índices propuestos. Como sea, las técnicas aplicadas no son especiales a los fenómenos del español. De hecho, la posibilidad de su aplicación a otras lenguas, a pesar de las diferencias tipológicas, hace a estas medidas en cierto grado universales. Es evidente que no todas las lenguas utilizan los mismos recursos de la misma manera (por ejemplo, los sufijos como elementos relevantes de la morfosintaxis), pero cabe esperar que todas sean sistemas entrópicos que hagan uso de diversas estrategias de economía, es decir, que entre sus segmentos más gramaticales haya algo de esa energía sapireana que los caracteriza con respecto al corpus donde ocurren. Así que es necesario seguir investigando otras técnicas que puedan medir estas propiedades. No se puede privilegiar uno o algunos métodos, porque sabemos que no son el fenómeno mismo que nos incumbe. No hay que confundirlos. Los métodos y los formalismos no son la lengua, son el andamiaje que nos acerca al edificio del lenguaje y nos ayudan a medirlo y conocerlo.

Mucho se ha hablado de la entropía y del principio de economía como fenómenos característicos de los universales del lenguaje. Por eso, es emocionante corroborar que el lenguaje es, en efecto, una estructura entrópica y económica, cuyas unidades pueden investigarse mediante métodos capaces de medir estas propiedades. El léxico tiene una estructura interna que no se puede pasar por alto al seleccionar el vocabulario estándar de una lengua. Como dice Josse de Kock, tal vez restringir el número de vocablos en un sistema puede *aliviar* la carga de memorización en los hablantes, pero puede afectar la economía interna de la estructura léxica, incluso en la dimensión diacrónica: "It even more interferes with historical conditioning as it ignores this internal organization" (de Kock y Bossaert 1978, 58). Ciertamente, los lexicógrafos deben tomar esto en cuenta para garantizar la inclusión equilibrada de los vocablos en los diccionarios.

A pesar de que las computadoras se empezaron a aplicar al estudio y procesamiento del lenguaje y la literatura desde el final de la segunda guerra mundial, aún en el siglo XXI no se han explorado todas las rutas posibles en el campo de la segmentación morfológica no supervisada, con todo y los avances que se han llevado a cabo con desarrollos como ParaMor y la familia de métodos Morfessor, entre otros, y con la emergencia de nuevas dinámicas para evaluarlos, como los esfuerzos del Morpho Challenge. Por otra parte, con el renovado entusiasmo en las redes neuronales de hoy en día, seguramente veremos pronto nuevos desarrollos, basados en el entrenamiento de estas redes, enfocados al análisis no supervisado de la morfología de las lenguas. Aun así, no deja de ser necesario entender los conceptos básicos de segmentación morfológica de entropía, cuadros y economía. El paso siguiente será pasar de los experimentos puntuales a investigaciones exhaustivas y a gran escala, así como a la exploración de otros fenómenos lingüísticos.

La reflexión lingüística tiene muchos caminos. Aquí apenas se exploraron algunos de carácter cuantitativo, en los que la curiosidad, sobre todo, ha servido de guía principal. De esta curiosidad por identificar objetos lingüísticos y explorar los métodos para hacerlo, surgió la necesidad de medir la fuerza o energía imaginada de la afijalidad. Se puede pensar que la medición de las propiedades lingüísticas implica la reducción y simplificación de la realidad al máximo. Sin embargo, para conocer verdaderamente estas propiedades, los métodos cuantitativos parecen imprescindibles, especialmente si optamos por seguir una de las reglas básicas de Galileo: medir todo lo medible e intentar hacer medible todo lo que aún no lo es[1].

[1] Citado por Bunge (1967, 203 [v. II]).

BIBLIOGRAFÍA

Český národní korpus. Praga: Instituto del Corpus Nacional Checo, 2005.

Abbagnano, Nicola. *Diccionario de filosofía.* 2a. Traducido por Alfredo N. Galletti. México: Fondo de Cultura Económica, 1991 [1961].

Aho, Alfred V., y Jeffrey D. Ullmann. *The Theory of Parsing, Translation, and Compiling.* Nueva York: Prentice-Hall, 1972.

Allen, James. *Natural Language Understanding.* 2a. Redwood, California: Benjamin / Cummings, 1995.

Altmann, Gabriel. "Michael Oakes, Statistics for Corpus Linguistics". *Journal of Quantitative Linguistics* 6, nº 3 (1990).

Altmann, Gabriel. "Science and linguistics". En *Contributions to Quantitative Linguistics*, editado por Reinhard Köhler, & Burghard Rieger, 3-10. Dordrecht: Kluwer, 1993.

——. *Statistik für Linguisten.* Tréveris: Wissenschaftlicher Verlag Trier, 1995.

Altmann-Fitter. Iterative Fitting of Probability Distributions. Prod. Gabriel Altmann. Lüdenscheid: RAM-Verlag, 2020 [1997].

Anderson, Stephen R. *A-Morphous Morphology.* Cambridge: Cambridge UP, 1994.

Antworth, Evan L. *PC-KIMMO: A Two-Level Processor for Morphological Analysis.* Dallas: Summer Institute of Linguistics, 1990.

Bergenholtz, Henning, y Joachim Mugdan. *Einführung in die Morphologie.* Stuttgart: Kohlhammer, 1979.

Bright, William, ed. *The International Encyclopedia of Linguistics.* Oxford: Oxford UP, 1992.

Brill, Eric. *A Corpus-Based Approach to Language Learning.* Filadelfia: Department of Computer and Information Science, Universidad de Pennsylvania. Tesis doctoral, 1993.

——. "A simple Rule-Based Part of Speech Tagger". *Proceedings of the Third Conference on Applied Natural Language Processing.* Trento: ACL, 1992. 112-116.

——. "Some Advances in Transformation-Based Part of Speech Tagging". *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI '94).* Seattle, 1994. 722-727.

Brill, Eric. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging". *Computational Linguistics* 21, nº 4 (1995): 543-565.

Buelna Serrano, María Elvira, recop. *Indígenas en la Inquisición Apostólica de fray Juan de Zumárraga.* México: UAM Azcapotzalco, 2009.

Buenrostro Díaz, Elsa Cristina, ed. *Corpus de la lengua chuj.* Recopilado por Elsa Cristina Buenrostro Díaz. México: Instituto de Investigaciones Antropológicas, UNAM, 2002.

——. *La voz en Chuj de San Mateo Ixtatán.* México: El Colegio de México. Tesis doctoral, 2013.

Buenrostro, Cristina. *Chuj de San Mateo Ixtatán.* México: El Colegio de México, 2009.

Bunge, Mario. *Philosophy of Science. I: From Problem to Theory, II: From Explanation to Justification.* 2 vols. New Brunswick: Transaction Publishing, 1998 [1967].

——. *Scientific Research. I: The Search for System, II: The Search for Truth.* 2 vols. Berlín / Heidelberg: Springer, 1967.

Bußmann, Hadumod. *Lexikon der Sprachwissenschaft.* 2a. Stuttgart: Kröner, 1990.

Bybee, Joan. *Morphology. A Study of the Relation between Meaning and Form.* Amsterdam: John Benjamins, 1985.

Centro de Investigaciones Dr. Hideyo Noguchi. *Yucatán, identidad y cultura maya.* Yucatán: Universidad Autónoma de Yucatán, s.f.

Charniak, Eugene. *Statistical Language Learning.* Cambridge (Mass.): MIT Press, 1993.

Church, Kenneth W., y Robert L. Mercer. "Introduction to the Special Issue on Computational Linguistics Using Large Corpora". *Computational Linguistics* 19 (1993): 1-24.

Company Company, Concepción, ed. *Documentos Lingüísticos de la Nueva España. Altiplano-Central.* Recopilado por Concepción Company Company. México: UNAM, 1994.

Creutz, Mathias, y Krista Lagus. "Inducing the morphological lexicon of a natural language from unannotated text". En *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 106–113. 2005.

——. "Inducing the morphological lexicon of a natural language from unannotated text". *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*. 2005. 106–113.

Creutz, Mathias, y Krista Lagus. "Induction of a simple Morphology for highly Iinflecting Languages". En *Proceedings of 7th Meeting of the ACL Special Interest Group in Computational Phonology*, 43–51. 2004.

Creutz, Mathias, y Krista Lagus. "Unsupervised discovery of morphemes". En *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, 21–30. Filadelfia, 2002.

Creutz, Mathias, y Krista Lagus. *Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0.* Helsinki University of Technology, 2005.

Cromm, Oliver. *Affixerkennung in deutschen Wortformen. Eine Untersuchung zum nicht-lexikalischen Segmentierungsverfahren von N. D. Andreev.* Fráncfort del Meno: Abschluß des Ergänzungsstudiums Linguistische Datenverarbeitung, 1996.

de Kock, Josse, y Walter Bossaert. "De la definición de estructuras lingüísticas con la ayuda de un ordenador. El morfema". En *Introducción a la lingüística automática en las lenguas románicas*, 181-227. Madrid: Gredos, 1974.

——. *Introducción a la lingüística automática en las lenguas románicas.* Madrid: Gredos, 1974.

——. *The Morpheme. An Experiment in Quantitative and Computational Linguistics.* Amsterdam: Van Gorcum, 1978.

Embleton, Sheila M. *Statistics in Historical Linguistics.* Bochum: Brockemeyer, 1986.

Epstein, Richard. "The Development of the Definite Article in French". En *Perspectives on Grammaticalization*, editado por William Pagliuca, 63-78. Amsterdam / Philadelphia: Benjamins, 1994.

Ferraresi, Adriano, y Silvia Bernardini. "Book Review of Stefan Thomas Gries, Stephanie Wulff, and Mark Davies (eds.), Corpus-linguistic applications. Current studies, new directions". *Empirical Language Research Journal* 4, nº 2 (2010).

Ferraresi, Adriano, y Silvia Bernardini. "Book Review of Stefan Thomas Gries, Stephanie Wulff, and Mark Davies, eds., Corpus-linguistic applications. Current studies, new directions". *Empirical Language Research Journal* 4, nº 2 (2010).

Fischer, Olga, Muriel Norde, y Harry Perridon. *Up and down the Cline - The Nature of Grammaticalization.* Amsterdam: Benjamins, 2004.

Flenner, Gudrun. "Ein quantitatives Morphsegmentierungssystem für spanische Wortformen". En *Computation Linguae II*, editado por Ursula Klenk, 31-62. Stuttgart: Franz Steiner, 1994.

Frakes, William. "Stemming Algorithms". En *Information Retrieval, Data Structures and Algorithms*, editado por William Frakes, & Ricardo Baeza-Yates, 131-160. New Jersey: Prentice Hall, 1992.

Gazdar, Gerald, y Chris Mellish. *Natural Language Processing in Prolog.* Wokingham, Reino Unido: Addison-Wesley, 1989.

Gelbukh, Alexander. "Brill's Tagger trained for Spanish". México, [2004] 2007.

Gelbukh, Alexander, Mikhail Alexandrov, y Sang Yong Han. "Detecting Inflection Patterns in Natural Language by Minimization of Morphological Model". En *CIARP 2004, Lecture Notes in Computer*

Science 3287, editado por A. Sanfeliu, J.F. Martínez Trinidad, & J.A. Carrasco Ochoa, 432–438. Springer, 2004.

Gelbukh, Alexander, y Grigori Sidorov. "Approach to Construction of Automatic Morphological Analysis Systems for Inflective Languages with Little Effort". En *Computational Linguistics and Intelligent Text Processing (CICLing-2003)*, 215-220. Springer, 2003.

Glück, Helmut, ed. *Metzler Lexikon Sprache.* 2a. Stuttgart: Verlag J.B. Metzler, 2000.

Goldsmith, John. "An Algorithm for the Unsupervised Learning of Morphology". *Natural Language Engineering* 12, nº 6 (2006): 353–371.

Goldsmith, John. "Segmentation and Morphology". En *Computational Linguistics and Natural Language Processing Handbook*, editado por Alex Clark, Chris Fox, & Shalom Lappin, 364-394. Blackwell, 2009.

Goldsmith, John. "Unsupervised Learning of the Morphology of a Natural Language". *Computational Linguistics* 27, nº 2 (2001): 153–198.

Gómez Hernández, Antonio, María Rosa Palazón, y Mario Humberto Ruz, *Palabras de nuestro corazón. Mitos, fábulas y cuentos maravillosos de la narrativa tojolabal.* México: UNAM / Universidad Autónoma de Chiapas, 1999.

Greenberg, Joseph H. *Essays in Linguistics.* Chicago: The University of Chicago Press, 1967 [1957].

Gries, Stefan Thomas, Stefanie Wulff, y Mark Davies, . *Corpus-Linguistic Applications: Current Studies, New Directions.* Amsterdam: Rodopi, 2010.

Hafer, Margaret A., y Stephen F. Weiss. "Word Segmentation by Letter Successor Varieties". *Information Storage and Retrieval* 10 (1974): 371-385.

Hammarström, Harald. "A Survey of Computational Morphological Resources for Low-Density Languages". *Journal of the Northern European Association for Language Technology*, 2009: 105-130.

Hammarström, Harald, y Lars Borin. "Unsupervised Learning of Morphology". *Computational Linguistics* 37, nº 2 (2010): 309-350.

Harris, James W. "Historical Excursus: Reflexes of the Medieval Stridents". En *Spanish Phonology*, 189-206. Cambridge (Mass.): MIT Press, 1969.

Harris, Zellig S. *A Theory of Language and Information.* Oxford: Clarendon, 1991.

——. "From Phoneme to Morpheme". *Language* 31, nº 2 (1955): 190-222.

——. "Morpheme Alternants in Linguistic Analysis". *Language* 18 (1942): 169-180.

Hlaváčová, Jaroslava, y Anna Nedoluzhko. "Productive Verb Prefixation Patterns". *The Prague Bulletin of Mathematical Linguistics*, nº 101 (abril 2014): 111–122.

Hlaváčová, Jaroslava, y Michal Hrušecký. "Affisix: Tool for Prefix Recognition". En *Text, Speech and Dialogue. TSD 2008, Lecture Notes in Computer Science 5246*, editado por Petr Sojka, Aleš Horák, Ivan Kopeček, & Karel Pala, 85-92. Springer, 2008.

Hlaváčová, Jaroslava, y Michal Hrušecký. "Prefix Recognition Experiments". En *Text, Speech and Dialogue. TSD 2011. Lecture Notes in Computer Science 6836*, editado por Ivan Habernal, & Václav Matoušek, 235-242. Springer, 2011.

Hockett, Charles F. "Linguistic Elements and their Relations". *Language* 37 (1961): 29-53.

Janßen, Axel. "Segmentierung französischer Wortformen in Morphe ohne Verwendung eines Lexikons". En *Computation Linguae I*, editado por Ursula Klenk, 74-95. Stuttgart: Franz Steiner, 1992.

Jäppinen, Harri. "Finite State Computational Morphology". En *Computation Linguae I*, editado por Ursula Klenk, 96-109. Stuttgart: Franz Steiner, 1992.

Jiménez Salazar, Héctor, y Guillermo Morales Luna. "SEPE: A POS Tagger for Spanish". *Lecture Notes in Computer Science* 2276 (2002): 250-259.

Johnson, Howard, y Joel Martin. "Unsupervised Learning of Morphology for English and Inuktitut". *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.* Edmonton, Canada: ACL, 2003.

Juilland, Alphonse, y E. Chang Rodríguez. *A Frequency Dictionary of Spanish Words.* La Haya: Mouton, 1965.

Jurafsky, Daniel, y James H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, NJ: Pearson Prentice Hall, 2009.

Kageura, Kyo. "Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences". *Journal of Quantitative Linguistics*, nº 6 (1999): 149-166.

Kaplan, Ronald M., y Martin Kay. "Phonological rules and finite-state transducers". *Annual Meeting of the Linguistics Society of America.* Nueva York, 1981.

Karttunen, Lauri. "KIMMO: a General Morphological Processor". *Texas Linguistic Forum* 22 (1983): 163-186.

Kaufman, Terrence. *Idiomas de Mesoamérica.* Guatemala: José de Pineda Ibarra, Ministerio de Educación, 1974.

Klenk, Ursula. "Automatische morphologische Analyse arabischer Verbformen". En *Computation Linguae II*, editado por Ursula Klenk, 84-101. Stuttgart: Franz Steiner, 1994.

Klenk, Ursula, ed. *Computation Linguae I.* Stuttgart: Franz Steiner (ZDL-Beiheft 73), 1992.

Klenk, Ursula, ed. *Computation Linguae II.* Stuttgart: Franz Steiner (ZDL-Beiheft 83), 1994.

Klenk, Ursula. "Verfahren morphologischer Segmentierung und die Wortstruktur des Spanischen". En *Computation Linguae I*, editado por Ursula Klenk, 110-124. Stuttgart: Franz Steiner, 1992.

Klenk, Ursula, y Hagen Langer. "Morphological Segmentation Without a Lexicon". *Literary and Linguistic Computing* 4, nº 4 (1989): 247-253.

Köhler, Reinhard. "Diversification of Coding Methods in Grammar". En *Diversification Processes in Language: Grammar*, editado por Ursula Rothe, 47-55. Hagen: Rottmann, 1991.

Koskenniemi, Kimmo. "Computational Morphology". En *International Encyclopedia of Linguistics*, de William Bright, 291-293. Oxford: Oxford UP, 1992.

Lara Reyes, Diego. *Sistema de segmentación automática de palabras en morfemas para el español.* México: CIC IPN. Tesis de maestría, 2008.

Lara, Luis Fernando, ed. *Diccionario del español de México.* México: El Colegio de México, 2010.

——. *Dimensiones de la lexicografía. A propósito del Diccionario del Español de México.* México: El Colegio de México, 1990.

——. *Historia mínima de la lengua española.* México: El Colegio de México, 2013.

Lara, Luis Fernando. "La cuantificación en el Diccionario del español de México". En *Dimensiones de la lexicografía. A propósito del Diccionario del español de México*, 56-68. México: El Colegio de México, 1990.

Lara, Luis Fernando, Roberto Ham Chande, y María Isabel García Hidalgo. *Investigaciones lingüísticas en lexicografía.* México: El Colegio de México, 1979.

Lope Blanch, Juan Manuel, ed. *El habla de Diego de Ordaz. Contribución a la historia del español americano.* México: UNAM, 1985.

Manning, Christopher D., y Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* Cambridge (Mass.): The MIT Press, 1999.

Matthews, Peter H. *Morphology.* 2a. Cambridge: Cambridge UP, 1991.

McEnery, Tony, y Andrew Wilson. *Corpus Linguistics.* 2a. Edinburgh: Edinburgh UP, 2001.

Medina Urrea, Alfonso. "Affix Discovery by Means of Corpora: Experiments for Spanish, Czech, Ralámuli and Chuj". En *Aspects of Automatic Text Analysis*, editado por Alexander Mehler, & Reinhard Köhler, 277-299. Berlín: Springer-Verlag, 2007.

——. "Automatic Discovery of Affixes by Means of a Corpus: A Catalog of Spanish Affixes". *Journal of Quantitative Linguistics* 7, nº 2 (2000): 97-114.

——. *Investigación cuantitativa de afijos y clíticos del español de México.* México: El Colegio de México. Tesis doctoral, 2003.

——. "Toward a comparison of unsupervised diachronic morphological profiles". En *Corpus-Linguistic Applications: Current Studies, New Directions*, editado por Stefan Thomas Gries, Stefanie Wulff, & Mark Davies, 29-45. Amsterdam: Rodopi, 2010.

Medina Urrea, Alfonso. "Towards the Automatic Lemmatization of 16th Century Mexican Spanish: A Stemming Scheme for the CHEM". *Lecture Notes in Computer Science* (Springer) 3878 (2006): 101–104.

Medina Urrea, Alfonso, José Abel Herrera Camacho, y Maribel Alvarado García. "Towards the Speech Synthesis of Raramuri: a Unit Selection Approach based on Unsupervised Extraction of Suffix Sequences". *Research in Computing Science* 41 (2009): 243-256.

Medina Urrea, Alfonso, y Carlos Francisco Méndez Cruz. "Arquitectura del Corpus Histórico del Español de México". En *Avances en la Ciencia de la Computación*, editado por Arturo Hernández Aguirre, & José Luis Zechinelli Martini, 248-253. México: Sociedad Mexicana de Ciencia de la Computación, 2006.

Medina Urrea, Alfonso, y Carlos Francisco Méndez Cruz. "El Corpus Histórico del Español en México". *Revista Digital Universitaria* 12, nº 7 (2011): 3-25.

Medina Urrea, Alfonso, y Jaroslava Hlaváčová. "Automatic Recognition of Czech Derivational Prefixes". *Lecture Notes in Computer Science* (Springer) 3406 (2005): 189–197.

Medina Urrea, Alfonso, y Maribel Alvarado García. "Un experimento de reconocimiento automático de la derivación léxica del ralámuli". En *La lengua y la antropología para un conocimiento global del hombre. Homenaje a Leonardo Manrique*, editado por Susana Cuevas Suárez, 243-251. México: Instituto Nacional de Antropología e Historia, 2009.

Medina, Alfonso, y Elsa Cristina Buenrostro. "Características cuantitativas de la flexión verbal del chuj". *Estudios de Lingüística Aplicada* 38 (2003): 15–31.

Medina-Urrea, Alfonso. "Towards the Measurement of Nominal Phrase Grammaticality: Contrasting Definite-Possessive Phrases with Definite Phrases of 13th to 19th Century Spanish". En *Exact Methods in the Study of Language and Text*, de Peter Grzybek, & Reinhard Köhler, 427-437. Berlín: Mouton de Gruyter, 2007.

Mehler, Alexander, y Reinhard Köhler. *Aspects of Automatic Text Analysis.* Berlín, Heidelberg: Springer, 2007.

Meillet, Antoine. "L'évolution des formes grammaticales". En *Linguistique historique et linguistique générale*, 130-148. París: Société de Linguistique de Paris, 1958 [1912].

Meléndez Guadarrama, Lucero, ed. *Narraciones del huasteco.* Recopilado por Lucero Meléndez Guadarrama. México: Instituto de Investigaciones Antropológicas, UNAM, 2010.

——. *Huasteco de El Mamey San Gabriel, Tantoyuca, Veracruz.* México: El Colegio de México, 2017.

Méndez Cruz, Carlos Francisco. *Generación automática de una gramática de estados finitos para la morfología del español.* México: UNAM. Tesis doctoral, 2013.

——. *Identificación automática de categorías gramaticales en español del siglo XVI.* México: UNAM. Tesis de maestría, 2009.

Méndez-Cruz, Carlos-Francisco, Alfonso Medina-Urrea, y Gerardo Sierra. "Unsupervised morphological segmentation based on affixality measurements". *Pattern Recognition Letters* 84 (2016): 127-133.

Meya, Montserrat. "Morphologische Analyse des Spanischen". En *Informationslinguistische Texterschließung*, editado por Christoph Schwarz, & Gregor Thurmair, 134-156. Zürich: Georg Olms, 1986.

Meyer-Eppler, Werner. *Grundlagen und Anwendungen der Informationstheorie.* 2a. Heidelberg: Springer, 1969.

Moliner, María. *Diccionario de uso del español.* Madrid: Gredos, 1992.

Monson, Christian. *ParaMor: From Paradigm Structure to Natural Language Morphology Induction.* Pittsburgh: Carnegie Mellon University. Doctoral Defense Draft, 2008.

Monson, Christian, Jaime Carbonell, Alon Lavie, y Lori Levin. "ParaMor: Finding Paradigms across Morphology". *Lecture Notes in Computer Science* 5152 (2008): 900-907.

Monson, Christian, Jaime Carbonell, Alon Lavie, y Lori Levin. "ParaMor: Minimally Supervised Induction of Paradigm Structure and Morphological Analysis". En *Computing and Historical Phonology: The Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*. Praga, 2007.

Morales Carrasco, Raúl, y Alexander Gelbukh. "Evaluation of TnT Tagger for Spanish". *Proceedings of the Fourth Mexican International Conference on Computer Science.* Tlaxcala: ENC 2003, 2003. 18-25.

Morales, José A., y otros. "Rank Dynamics of Word Usage at Multiple Scales". *Frontiers in Physics*, 2018: 1-43.

Moreno de Alba, José. *La prefijación en el español mexicano, UNAM, México, 1996.* México: UNAM, 1996.

——. *Morfología derivativa nominal en el español de México.* México: UNAM, 1986.

Naumann, Sven, y Hagen Langer. *Parsing: Eine Einführung in die maschinelle Analyse natürlicher Sprache.* Stuttgart: Teubner, 1994.

Nida, Eugene A. *Morphology. The Descriptive Analysis of Words.* Ann Arbor: The University of Michigan Press, 1967 [1949].

Nida, Eugene A. "The Identification of Morphemes". *Language* 24 (1948).

Oakes, Michael P. *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh UP, 1998.

Parra, Patricio, ed. *Memorias y textos ralámulis de la variante de la Sierra Tarahumara de San Luis Majimachi, Municipio de Bocoyna, Chihuahua.* 2003.

Pérez Marqués, Celia María, ed. *Corpus electrónico de textos escritos por escolares de Guamá.* Vol. CD. Santiago de Cuba: Centro de Lingüística Aplicada, 2003.

——. *Nuevo enfoque para un diagnóstico de desarrollo léxico.* Santiago de Cuba: Universidad de Oriente. Tesis doctoral, 2004.

Pérez Marqués, Celia, y Alfonso Medina Urrea. "Sufijos y grupos sufijales característicos en el vocabulario de escolares guamenses". *Actas I del IX Simposio Internacional de Comunicación Social.* Santiago de Cuba: Centro de Lingüística Aplicada del Ministerio de Ciencia, Tecnología y Medio Ambiente de Cuba, 2005. 125-129.

Piotrowski, R. G., K. B. Bektaev, y A. A. Piotrowskaja. *Mathematische Linguistik.* Traducido por A. Falk. Bochum: Brockmeyer, 1985.

Piotrowski, R. G., M. Lesohin, y K. Lukjanenkov. *Introduction of Elements of Mathematics to Linguistics.* Bochum: Brockmeyer, 1990.

Porter, M.F. "An Algorithm for Suffix Stripping". *Program* 14, nº 3 (1980): 130-137.

Rainer, Franz. *Spanische Wortbildungslehre.* Tübingen: Niemeyer, 1993.

Reyes Careaga, Teresita Adriana. *Reglas de correspondencia entre sonido y grafía en el español hablado en México en el siglo XVI: Una aportación al Corpus histórico del español en México.* México: UNAM. Tesis de licenciatura, 2008.

Rini, Joel. *Motives for Linguistic Change in the Formation of the Spanish Object Pronouns.* Newark, Delaware: Juan de la Cuesta, 1992.

Ritchie, Graeme D., Graham J. Russell, Alan W. Black, y Stephen G. Pulman. *Computational Morphology. Practical Mechanisms for the English Lexicon.* Cambridge (Mass.): MIT Press, 1992.

Robertson, John S. "A Proposed Revision in Mayan Subgrouping". *International Journal of American Linguistics* 43, nº 2 (1977): 105-120.

Rothe, Ursula, ed. *Diversification Processes in Language: Grammar.* Hagen: Rottmann, 1991.

Sapir, Edward. *El lenguaje.* Traducido por Margit Frenk, & Antonio Alatorre. México: Fondo de Cultura Económica, 1992 [1921].

Schwarz, Christoph, y Gregor Thurmair. *Informationslinguistische Texterschließung.* Zürich: Georg Olms, 1986.

Shannon, Claude E., y Warren Weaver. *The Mathematical Theory of Communication.* Urbana: University of Illinois Press, 1964 [1949].

Sierra Martínez, Gerardo E. *Introducción a los corpus lingüísticos.* México: Instituto de Ingeniería, UNAM, 2017.

Slabý, Rudolf, Rudolf Grossmann, y Carlos Illig. *Wörterbuch der spanischen und deutschen Sprache.* Wiesbaden: Brandstetter, 1975.

Spencer, Andrew. *Morphological Theory. An Introduction to Word Structure in Generative Grammar.* Cambridge: Basil Blackwell, 1991.

Spencer, Andrew, y Arnold M. Zwicky. *The Handbook of Morphology.* Oxford: Blackwell, 1998.

Sproat, Richard William. *Morphology and Computation.* Cambridge, Mass.: MIT Press, 1992.

Swadesh, Morris. "Towards Greater Accuracy in Lexicostatistic Dating". *International Journal of American Linguistics* 21, nº 2 (1955): 121-137.

Thomason, Sarah Grey, y Terrence Kaufman. *Language Contact, Creolization, and Genetic Linguistics.* Berkeley: University of California Press, 1991 [1988].

Thurmair, Gregor. "Ein Morphologisches Prozessorsegment zur Erzeugung von Grundformen mithilfe von Lernverfahren". En *Informationslinguistische Texterschließung*, editado por Christoph Schwarz, & Gregor Thurmair, 8-31. Zürich: Georg Olms, 1986.

Torres Moreno, Juan Manuel. "Beyond Stemming and Lemmatization: Ultra-stemming to Improve Automatic Text Summarization". *arXiv:1209.3126v1 [cs.IR]*, septiembre 2012.

Varro, Marcus Terentius. *On the Latin Language.* Editado por Roland G. Kent. Traducido por Roland G. Kent. Vol. II. Londres: Willian Heinemann, 1938 [47-45 a.C.].

Varrón, Marco Terencio. *De lingua Latina.* Traducido por Manuel Antonio Marcos Casquero. Madrid: Anthropos, 1990 [ca. 40 a.C.].

Virpioja, Sami, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, y Mikko Kurimo. "Empirical Comparison of Evaluation Methods for Unsupervised Learning of Morphology". *Traitement Automatique des Langues* 52, nº 2 (2011): 45-90.

Wall, Robert. *Introduction to Mathematical Linguistics.* Englewood Cliffs, Nueva Jersey: Prentice Hall, 1972.

Weaver, Warren. "Recent Contributions to the Mathematical Theory of Communication". En *The Mathematical Theory of Communication*, de Claude E. Shannon, & Warren Weaver, 3-28. Urbana: University of Illinois Press, 1964.

Wells, Rulon S. "Automatic Alternation". *Language* 25 (1949): 99-116.

Woods, Anthony, Paul Fletcher, y Arthur Hughes. *Statistics in Language Studies.* Cambridge: Cambridge UP, 1986.

APÉNDICE. CÓDIGO PYTHON

En este apéndice, se muestran algunos ejemplos de cómo implementar en Python los procedimientos para calcular la entropía y la economía y contar los cuadros en cada corte posible de una cadena de caracteres correspondiente a una palabra gráfica. Los ejemplos se presentan en tres cuadros divididos en código y resultados. El primer cuadro es para la entropía, el segundo para los cuadros y el tercero para la economía.

En el primero, se definen la *clase* ***arb*** y sus *métodos*. Esta *clase* permite almacenar las palabras de un corpus (en este ejemplo se usa el CEMC, sin transcripción fonológica) en dos estructuras arbóreas; una para alojar las palabras en su sentido de lectura, de inicio a fin (**derIzq**), y otra para alojarlas al revés, del final al principio (**izqDer**). La entropía se calcula en cada nodo de las estructuras arbóreas mediante el método ***calcularEntr*** (de la clase ***arb***). La invocación de este método aparece poco antes del ciclo de petición de palabras que despliega la entropía de cada corte posible.

En la segunda parte del cuadro, se muestra la salida del programa con los resultados de la petición del vocablo *contexto*. Nótese que el valor más alto del primer renglón (*der-izq*) corresponde con el corte entre el prefijo *con-* y la base *texto*, mientras que el valor más alto del segundo renglón (*izq-der*) corresponde con el corte entre la base *context-* y el sufijo *-o* (que alternaría con *-ual*).

```
import math

class arb:
	c= None
	arcos= None
	f= None
	entr= None
	alreves= None

	def __init__( self, caracter= '', alreves= False):
		self.c= caracter
		self.arcos= []
		self.f= 0
		self.entr= 0
		self.alreves= alreves

	def frecuencia( self, nodo):
		return nodo.f

	def entropia( self, nodo):
		return nodo.entr
```

```
	def encontrar( self, palabra):
		if palabra == '': return False
		if self.nodo( palabra + '.') == None: return False
		return True
	def insertar( self, palabra):
		if self.encontrar( palabra): return
		self.f+= 1
		actual= self
		arcs= actual.arcos
		for i in range( len( palabra)):
			encontrado= False
			for j in range( len( arcs)):
				if arcs[ j].c == palabra[ i]:
					encontrado= True
					actual= arcs[ j]
			if not encontrado:
				actual.arcos.append( arb( palabra[ i]))
				actual= actual.arcos[ -1]
			actual.f+= 1
			arcs= actual.arcos
		actual.arcos.append( arb( '.'))
		actual.arcos[ -1].f= 1
```

```
	def nodo( self, segmento):
		actual= self
		arcs= actual.arcos
		for i in range( len( segmento)):
			encontrado= False
			for j in range( len( arcs)):
				if arcs[ j].c == segmento[ i]:
					encontrado= True
					actual= arcs[ j]
			if not encontrado: return None
			arcs= actual.arcos
		return actual
	def calcularEntr( self):
		def proxCol( actual, p, arcs):
			if arcs == []:
				actual.entr= 0
			else:
				entropia= 0
				for i in range( len( arcs)):
					prob= ( arcs[ i].f) / actual.f
					entropia-= prob * math.log( prob, 2)
					proxCol( arcs[ i], p + arcs[ i].c, arcs[ i].arcos)
				actual.entr= entropia
		p= self.c
		arcs= self.arcos
		proxCol( self, p, arcs)
```

```
	def dameSegs( self, nodo, rev):
		def proxCol( p, arcs, bases, rev):
			for i in range( len( arcs)):
				if arcs[ i].arcos == []:
					bases.append( p)
					return
				proxCol( p + arcs[ i].c, arcs[ i].arcos, bases, rev)
		bases= []
		arcs= nodo.arcos
		for i in range( len( arcs)):
			proxCol( arcs[ i].c, arcs[ i].arcos, bases, rev)
		return bases

def limpiar (texto):
	punt= ".,;:@¡!¿?“”‘’)(][}{$%&/\\1234567890-=—_- — +*«»<>…^" + '"'
	texto= texto.lower()
	for i in punt:
		texto= texto.replace( i,' ')
	return texto
```

```
derIzq= arb()
izqDer= arb( alreves= True)
d= open( 'CEMC utf-8.txt', 'rt')
corpus= d.read()
corpus= limpiar( corpus)
texto= set( corpus.split())
d.close()
for i in texto:
	derIzq.insertar( i)
	izqDer.insertar( i[::-1])
print( 'corpus cargado')
derIzq.calcularEntr()
izqDer.calcularEntr()
while 1:
	palabra= input( 'Palabra: ')
	if palabra == '': break
	if derIzq.encontrar( palabra):
		entr1= []
		entr2= []
		for i in range( len( palabra)-1):
			A= palabra[:i+1]
			B= palabra[i+1:]
			entr1.append( derIzq.entropia( derIzq.nodo( A)))
			entr2.append( izqDer.entropia( izqDer.nodo( B[::-1])))
```

```
            print( '\t', end= '')
            for i in range( len( palabra)):
                  print( palabra[ i], '\t', end= '')
            print( '')
            print( 'der-izq\t', end= '')
            for i in range( len( entr1)):
                  print( '', "{:.3f}".format( entr1[ i]), '\t', end= '')
            print( '')
            print( 'izq-der\t', end= '')
            for i in range( len( entr2)):
                  print( '', "{:.3f}".format( entr2[ i]), '\t', end= '')
            print( '')
            print( '')
```

========== RESTART: /Users/.../entropía.py ==========
corpus cargado
Palabra: contexto

	c	o	n	t	e	x	t	o
der-izq	2.582	2.605	3.089	2.158	1.857	0.000	0.918	
izq-der	0.000	0.000	1.585	0.971	0.863	2.679	3.504	

En la segunda lámina, aparece el código de la función ***cuentaCdrs***, que recorre y cuenta los cuadros que conforman las dos mitades de una palabra con los segmentos de otras palabras. La función se invoca desde el ciclo de para desplegar resultados. En la segunda parte, la salida del programa muestra la lista de cuadros encontrados en el CEMC para la palabra *pesadumbre*. Vale la pena señalar que, para cortes muy productivos, la lista puede ser muy grande, especialmente para segmentaciones entre bases y sufijos de flexión. En el caso de *pesadumbre*, se sugieren dos cortes posibles, entre *pesa* y *dumbre* (con 9 cuadros) y entre *pesad-* y *-umbre*, siendo el segundo el que exhibe más cuadros (21). Nótese que algunas de estas estructuras parecen extrañas, porque están conformadas con artículos y clíticos (*las*, *lo*), nombres propios (*Servin*) o probables abreviaturas (*co*, *ca*). Cabe notar que estas estructuras raras son comunes en experimentos en los que no se eliminan previamente palabras función, siglas, etcétera.

```
def cuentaCdrs( palabra, der, rev):
       cuadros= []
       for i in range( len( palabra)-1):
              A= palabra[:i+1]
              B= palabra[i+1:]
              Arev= A[::-1]
              Brev= B[::-1]
```

```
n1= der.nodo( A)
n2= rev.nodo( Brev)
cuadros.append( 0)
primeras= rev.dameSegs( n2, der)
while '' in primeras: primeras.remove('')
while Arev in primeras: primeras.remove( Arev)
segundas= der.dameSegs( n1, rev)
while '' in segundas: segundas.remove('')
while B in segundas: segundas.remove( B)
Aind= der.encontrar( A)
Bind= der.encontrar( B)
if Aind:
      for a in primeras:
            r= a[::-1]
            if der.encontrar( r):
                  cuadros[ i]+= 1
                  c1= A + '::' + B
                  c2= A + '::' + '0'
                  c3= r + '::' + '0'
                  c4= r + '::' + B
                  print( c1, c2, c3, c4)
```

```
            for b in segundas:
                  if Bind and der.encontrar( b):
                        cuadros[ i]+= 1
                        c1= A + '::' + B
                        c2= '0' + '::' + B
                        c3= '0' + '::' + b
                        c4= A + '::' + b
                        print( c1, c2, c3, c4)
                  for a in primeras:
                        r= a[::-1]
                        if ( der.encontrar( r + b)
                        and der.encontrar( r + B)):
                              cuadros[ i]+= 1
                              c1= A + '::' + B
                              c2= A + '::' + b
                              c3= r + '::' + b
                              c4= r + '::' + B
                              print( c1, c2, c3, c4)
      return cuadros
```

```
derIzq= arb()
izqDer= arb( alreves= True)

d= open( 'CEMC utf-8.txt', 'rt')
corpus= d.read()
corpus= limpiar( corpus)
texto= set( corpus.split())
d.close()

for i in texto:
       derIzq.insertar( i)
       izqDer.insertar( i[::-1])

print( 'corpus cargado')
```

```
while 1:
       palabra= input( 'Palabra: ')
       if palabra == '': break
       if derIzq.encontrar( palabra):
              cuadros= cuentaCdrs( palabra, derIzq, izqDer)

              print( '\n=========================================')
              print( '\t', end= '')
              for i in range( len( palabra)):
                     print( palabra[ i], '\t', end= '')
              print( '\n-----------------------------------------')
              print( 'cuadros\t', end= '')
              for i in range( len( cuadros)):
                     print( ' ', cuadros[ i], '\t', end= '')
              print( '\n-----------------------------------------')
              print( '')
```

========== RESTART: /Users/.../cuadros.py ==========
corpus cargado
Palabra: pesadumbre
pesa::dumbre pesa::rme servi::rme servi::dumbre
pesa::dumbre pesa::r servi::r servi::dumbre
pesa::dumbre pesa::dos servi::dos servi::dumbre
pesa::dumbre pesa::do servi::do servi::dumbre

```
pesa::dumbre pesa::da servi::da servi::dumbre
pesa::dumbre pesa::n recie::n recie::dumbre
pesa::dumbre pesa::n servi::n servi::dumbre
pesa::dumbre pesa::mos podre::mos podre::dumbre
pesa::dumbre pesa::mos servi::mos servi::dumbre
pesad::umbre pesad::os servid::os servid::umbre
pesad::umbre pesad::os al::os al::umbre
pesad::umbre pesad::os l::os l::umbre
pesad::umbre pesad::os tech::os tech::umbre
pesad::umbre pesad::os cost::os cost::umbre
pesad::umbre pesad::os c::os c::umbre
pesad::umbre pesad::o servid::o servid::umbre
pesad::umbre pesad::o al::o al::umbre
pesad::umbre pesad::o l::o l::umbre
pesad::umbre pesad::o leg::o leg::umbre
pesad::umbre pesad::o tech::o tech::umbre
pesad::umbre pesad::o cost::o cost::umbre
pesad::umbre pesad::o az::o az::umbre
pesad::umbre pesad::o c::o c::umbre
pesad::umbre pesad::amente rel::amente rel::umbre
pesad::umbre pesad::a servid::a servid::umbre
pesad::umbre pesad::a al::a al::umbre
pesad::umbre pesad::a l::a l::umbre
pesad::umbre pesad::a cost::a cost::umbre
```

pesad::umbre pesad::a c::a c::umbre
pesad::umbre pesad::illa cost::illa cost::umbre

	p	e	s	a	d	u	m	b	r	e
cuadros	0	0	0	9	21	0	0	0	0	

En el tercero y último cuadro, se muestra el código de la función **calculaEcon**, que calcula el índice de economía. De nuevo, el ciclo de petición la invoca y despliega los resultados en dos renglones, der-izq para los prefijos, e izq-der para los sufijos. Al final del cuadro aparecen los resultados para la palabra temeridad. Como se ve, el valor más alto de economía, 24.83, corresponde al sufijo *~idad*.

```
def calculaEcon( palabra, der, rev):

      eco1= []
      eco2= []
      for i in range( len( palabra)-1):
            A= palabra[:i+1]
            B= palabra[i+1:]
            Arev= A[::-1]
            Brev= B[::-1]
            segA= [ A]
            segB= [ B]
            n1= der.nodo( A)
            n2= rev.nodo( Brev)
            eco1.append( 0)
            eco2.append( 0)
            colindante= []
```

```
primeras= rev.dameSegs( n2, der)
while '' in primeras: primeras.remove('')
while Arev in primeras: primeras.remove( Arev)

segundas= der.dameSegs( n1, rev)
while '' in segundas: segundas.remove('')
while B in segundas: segundas.remove( B)

Aind= der.encontrar( A)
Bind= der.encontrar( B)

if Aind:
      for a in primeras:
            r= a[::-1]
            if der.encontrar( r):
                  if not '' in segB: segB.append( '')
                  if not r in segA: segA.append( r)
```

```
            for b in segundas:
                  if not b[0] in colindante:
                        if Bind and der.encontrar( b):
                              if not '' in segA: segA.append( '')
                              if not b in segB: segB.append( b)

                  for a in primeras:
                        r= a[::-1]
                        if ( der.encontrar( r + b)
                        and der.encontrar( r + B)):
                              if not r in segA: segA.append( r)
                              if not b in segB: segB.append( b)
                              colindante.append( b[ 0])

            if len( segA) > 0: eco1[ i]= len( segB)/len( segA)
            if len( segB) > 0: eco2[ i]= len( segA)/len( segB)

      return( eco1, eco2)
```

```
derIzq= arb()
izqDer= arb( alreves= True)

d= open( 'CEMC utf-8.txt', 'rt')
corpus= d.read()
corpus= limpiar( corpus)
texto= set( corpus.split())
d.close()

for i in texto:
       derIzq.insertar( i)
       izqDer.insertar( i[::-1])

print( 'corpus cargado')
```

```
while 1:
       palabra= input( 'Palabra: ')
       if palabra == '': break
       if derIzq.encontrar( palabra):
              valores= calculaEcon( palabra, derIzq, izqDer)
              econ1= valores[ 0]
              econ2= valores[ 1]

              print( '\n=========================================')
              print( '\t', end= '')
              for i in range( len( palabra)):
                     print( palabra[ i], '\t', end= '')
              print( '\n-----------------------------------------')
              print( 'der-izq\t', end= '')
              for i in range( len( econ1)):
                     print( '', "{:.2f}".format( econ1[ i]), '\t', end= '')
              print( '')
              print( 'izq-der\t', end= '')
              for i in range( len( econ2)):
                     print( '', "{:.2f}".format( econ2[ i]), '\t', end= '')
              print( '\n-----------------------------------------')
              print( '')
```

```
================ RESTART: /Users/.../economía.py ================
corpus cargado
Palabra: temeridad

================================================
        t        e        m        e        r        i        d        a        d
------------------------------------------------
der-izq  1.00     1.00     1.25     0.31     0.04     1.00     1.00     1.00
izq-der  1.00     1.00     0.80     3.20     24.83    1.00     1.00     1.00
------------------------------------------------
```

El signo afijal en la muestra textual:
claves para entender el descubrimiento automático de morfemas
se terminó de imprimir en octubre de 2021, en los talleres de
Iniziativa Graphic, D.V., Alcanfores 45,
col. Valle del Sur, 09819, Ciudad de México.
Portada: Enedina Morales
La edición consta de 350 ejemplares.
Tipografía y formación: El Atril Tipográfico, S.A. de C.V.
Cuidó la edición la Dirección de Publicaciones de
El Colegio de México

CENTRO DE ESTUDIOS LINGÜÍSTICOS Y LITERARIOS

ESTUDIOS DE LINGÜÍSTICA
XXXVII

En el contexto del procesamiento automático de los corpus electrónicos, un método supervisado se refiere al enriquecimiento de estos recursos mediante la aplicación de etiquetas con información típicamente lingüística, de cualquier nivel del lenguaje. La ganancia adquirida contribuye al aprendizaje de cosas imprevisibles sobre los fenómenos lingüísticos y extralingüísticos que los corpus representan.

En cambio, un método no supervisado, como el que se describe en este libro, no presupone más que el corpus o muestra textual en su estado original, pero permite la generación de conocimiento lingüístico de maneras sorprendentes. Por ejemplo, el método descrito aquí busca identificar objetos lingüísticos mediante la medición de propiedades cuantitativas, como las relaciones combinatorias, entrópicas y económicas entre cadenas de caracteres.

Así, este método contribuye a descubrir la morfología afijal de lenguas concatenativas como el español, el maya o el ralámuli, sin conocimiento previo de sus morfologías, mediante medidas de una fuerza o energía imaginada de afijalidad que permiten determinar las fronteras entre lexemas y afijos.

Puede pensarse que la medición de propiedades lingüísticas implica la reducción y simplificación de la realidad al máximo. Sin embargo, para conocer verdaderamente estas propiedades, los métodos cuantitativos parecen imprescindibles, especialmente si optamos por seguir una de las reglas básicas de Galileo: *medir todo lo medible e intentar hacer medible todo lo que aún no lo es.*

ISBN: 978-607-564-307-6

EL COLEGIO DE MÉXICO