UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA MAESTRÍA Y DOCTORADO EN LINGÜÍSTICA

GENERACIÓN AUTOMÁTICA DE UNA GRAMÁTICA DE ESTADOS FINITOS PARA LA MORFOLOGÍA DEL ESPAÑOL

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN LINGÜÍSTICA

PRESENTA:

CARLOS FRANCISCO MÉNDEZ CRUZ

TUTORES PRINCIPALES

DR. ALFONSO MEDINA URREA

PROGRAMA MAESTRĉA Y DOCTORADO EN LING¦ ĉSTICA

DR. GERARDO SIERRA MARTÍNEZ

INSTITUTO DE INGENIERÍA, UNAM

COMITÉ TUTOR

DRA. CHANTAL MELIS VAN EERDEWEGH

INSTITUTO DE INVESTIGACIONES FILOLÓGICAS, UNAM

DR. GRIGORI SIDOROV

PROGRAMA MAESTRĉA Y DOCTORADO EN LING¦ ĉSTICA

MÉXICO, D. F. OCTUBRE 2013

A todas las personas que amablemente creen en mí

Resumen

El presente trabajo de investigación propone un método no supervisado de segmentación morfológica automática que infiere parte de la morfotáctica del español. Su principal interés ha sido el descubrimiento de patrones morfotácticos que describan el orden y secuencialidad de unidades morfológicas a partir de datos empíricos (corpus).

Ya que la morfotáctica de una lengua puede llegar a ser muy compleja, este trabajo se aboca únicamente al descubrimiento de bases y secuencias de sufijos (sufitáctica). Con estas unidades se crea un aparato de descripción formal que describe su orden y secuencialidad. Así, los objetivos planteados son:

1. Descubrir, a partir de corpus y mediante un método no supervisado de segmentación morfológica automática, los sufijos y sufitáctica de la lengua española.
2. Generar, a partir de los sufijos y sufitáctica descubiertos, una gramática de estados finitos que describa la morfotáctica del español.

Estos objetivos se llevaron a cabo mediante el procesamiento automático del Corpus del Español Mexicano Contemporáneo, que fue creado bajo criterios estadísticos como muestra representativa del léxico del español mexicano. Para lograr el primer objetivo, se modificó el método propuesto por Medina (2000; 2003), que cuantifica la afijalidad de segmentos al interior de una palabra. Este método propone que los valores más altos de afijalidad dan cuenta de fronteras morfológicas.

Después de un primer acercamiento – en el que se desarrolló un truncador morfológico que demostró su efectividad para la tarea de resumen automático de documentos – y diversos experimentos de segmentación morfológica, se determinó la mejor estrategia para

el descubrimiento de bases y sufijos de acuerdo a una evaluación hecha mediante un corpus segmentado manualmente.

Para el logro del segundo objetivo, se desarrolló un procedimiento que genera un autómata de estados finitos a partir de las unidades descubiertas. Como parte de la experimentación, dos autómatas fueron generados, uno a partir de la representación ortográfica del corpus y otro a partir de su representación fonológica. El autómata generado de esta última representación mostró patrones morfotácticos que no estaban presentes en el otro autómata, por lo que se tomó como mejor descripción de la morfotáctica del corpus.

Mediante la evaluación cualitativa de una muestra de patrones morfotácticos inmersos en el autómata, se observó que la gran mayoría son pertinentes y dan cuenta de distintas regularidades morfológicas del español. De esta manera, el método desarrollado incluye los siguientes pasos:

1. Cuantificar la afijalidad de segmentos al interior de la palabra.
2. Descubrir las bases y sufijos mediante una estrategia de segmentación basada en la afijalidad.
3. Descubrir los patrones morfotácticos mediante la generación de un autómata de estados finitos.

Este método ofrece distintas ventajas, como la posibilidad de describir automáticamente la morfotáctica (bases y sufijos) de lenguas predominantemente afijales. Además, permite sentar las bases para futuras investigaciones en el descubrimiento de la morfotáctica mediante la inferencia de su aparato de descripción y no mediante su construcción manual. Esto es importante porque se estudia la lengua sin presuponer sus unidades y su secuencialidad.

Índice general

Índice de tablas

Índice de figuras

Agradecimientos

Agradezco enormemente a mis tutores, no sólo por el apoyo para la realización de este trabajo, por sus críticas y aportaciones, sino también por ser una guían fundamental en mi formación como investigador. En especial, le doy gracias a Alfonso Medina por su entusiasmo y disposición por entablar innumerables charlas sobre la naturaleza del lenguaje, por su gran disposición a compartir su conocimiento y por incluirme en numerosos proyectos que me han permitido tener una visión de lo que es la lingüística y la manera de hacer ciencia. También, agradezco a Gerardo Sierra por su incansable preocupación por mantener a flote el gran proyecto llamado Grupo de Ingeniería Lingüística y permitirme ser parte de él.

Además, quiero dar gracias a Juan Manuel Torres por todas sus enseñanzas, que me han permitido entender más a fondo la labor de hacer ciencia, y por el gran apoyo recibido durante mi estancia en su grupo de investigación de la Universidad de Aviñón. Agradezco también a Chantal Melis, Grigori Sidorov y Ramón Zacarías por la lectura crítica y las recomendaciones realizadas a este trabajo de investigación.

Agradezco al proyecto del Diccionario del Español de México por el permiso para utilizar el Corpus del Español Mexicano Contemporáneo, pieza imprescindible de mi investigación. Además, agradezco al Dr. Grigori Sidorov del Laboratorio de Procesamiento de Texto y Lenguaje Natural del Centro de Investigación en Computación por el permiso para utilizar la lista de palabras del *System for automatic morphological analysis of Spanish*.

Finalmente, esta tesis se realizó gracias al apoyo otorgado por el Consejo Nacional de Ciencia y Tecnología (CONACYT) mediante las siguientes becas otorgadas: una de estudios de posgrado, una para realizar estancia en el extranjero (beca mixta) y una del pro-

yecto “Detección y medición automática de similitud textual” con clave 178248 para dar los últimos detalles a este trabajo de investigación. Además, gracias al apoyo del proyecto IN400312 “Análisis estilométrico para la detección de similitud textual” del Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la UNAM.

Introducción

De entre los distintos niveles de estudio del lenguaje, este trabajo de investigación se sitúa en el nivel morfológico. En términos generales, la morfología estudia cómo están constituidas las palabras de una lengua y los fenómenos involucrados en su formación[1], esto conlleva conocer las unidades que las forman y los fenómenos relacionados con la interacción de estas unidades[2].

El interés por estudiar el lenguaje, y en especial las palabras, es muy antiguo; piénsese por ejemplo en la gramática de Panini 350 años a. C. para el sánscrito. Aunque este interés nunca se perdió, fueron los estructuralistas los que pusieron énfasis en estudiar unidades morfológicas (Hockett, 1971; Bloomfield, 1961), gracias al concepto de signo lingüístico, y en determinar procedimientos sistemáticos para descubrirlas (Nida, 1949). Surgió entonces la propuesta de que las palabras están formadas de morfemas y se vio a éstos como las unidades mínimas del análisis morfológico.

[1] Dejaré de lado las discusiones de carácter lingüístico a propósito de la pertinencia del concepto palabra y de considerar al morfema como una unidad mínima con o sin significado. Para efectos de mi trabajo, adoptaré estos términos como las unidades mínimas del análisis morfológico, reconociendo que las palabras están formadas por morfemas. Para discusiones sobre el concepto de palabra véase Anderson (1985, págs. 150-156), González Calvo (1998, págs. 11-37), Lara (2004, págs. 401-408) y Pena (1999, págs. 4327-4328). Diversas definiciones de morfema como unidad con significado son las de Hockett (1971, pág. 125) y Nida (1949, pág. 1); definiciones distintas o que cuestionan su característica de unidad con significado son las de Aronoff (1976, pág. 15), Anderson (1985, pág. 161) y Lara (2006, pág. 62).

Por otro lado, en las descripciones y discusiones de corte computacional, así como en la experientación de mi trabajo, me referiré a la palabra como la cadena de caracteres separada por espacios o signos en un corpus (palabra gráfica).

[2] Pena establece como objetivos de la morfología: "a) delimitar, definir y clasificar las unidades del componente morfológico, b) describir cómo tales unidades se agrupan en sus respectivos paradigmas y c) explicar el modo en que las unidades integrantes de las palabras se combinan y constituyen conformando su estructura interna" (1999, pág. 4307).

Dividir las palabras en morfemas trajo como consecuencia la necesidad de describir su orden y secuencialidad (Hockett, 1971, pág. 131). Así, buena parte del análisis morfológico de una lengua ha consistido en la determinación de los morfemas y su morfotáctica[3]. Siguiendo a Lara (2006, pág. 65), la morfotáctica es la característica de las lenguas de ordenar sus morfemas en secuencias determinadas[4].

El nivel básico de una descripción morfotáctica suele ser el orden que presentan en la lengua los morfemas ligados con relación a la base[5] de la palabra. Éstos pueden ser prefijos (por ejemplo ***in***-*confesable*), sufijos (por ejemplo *blanc*-***o***) o circunfijos (por ejemplo ***en***-*roj*-***ecer***) dependiendo de que aparezcan antes, después, o de que rodeen la base. Todos estos segmentos son llamados afijos.

El siguiente nivel de descripción morfotáctica es la secuencialidad de estos afijos y las restricciones que hacen que ciertas secuencias no sean permitidas. Diversas han sido las propuestas para explicar la naturaleza de estas restricciones de ordenamiento, pero aún no parece haber consenso en cuál puede dar cuenta de todos los casos posibles[6].

Conocer la secuencialidad de prefijos y sufijos[7] permite estudiar los fenómenos de flexión y derivación, o el orden entre afijos que expresan categorías gramaticales o sintácti-

[3] La idea de morfotáctica como secuencialidad está muy ligada a la morfología concatenativa, donde los fenómenos morfológicos se basan en la adición de material fonológico o escrito. Si bien el presente trabajo de investigación se basa en este tipo de morfología, entiendo que la complejidad morfológica rebasa este tipo de fenómenos y las posibilidades son muy amplias en las distintas lenguas humanas (reduplicación, eliminación, cambio vocálico, etcétera.)

[4] Discutiré el concepto de morfotáctica al inicio del capítulo 1.

[5] Usaré el término *base* como el segmento de la palabra sobre el que se dan los procesos de flexión o derivación, véase Pena (1999, pág. 4318).

[6] Revisaré algunas de estas propuestas en la sección 1.2.

[7] Me referiré al orden y secuencialidad de los afijos de una lengua como *afitáctica* o *morfotáctica afijal*. Asimismo, a la secuencialidad de los sufijos de una lengua le llamaré *sufitáctica* o *morfotáctica sufijal*.

cas; por ejemplo, estudios sobre universales lingüísticos (Greenberg, 1963; Bybee, 1985) han propuesto que algunas categorías aparecen antes que otras en la mayoría de las lenguas. Por otra parte, en lenguas aglutinantes, la secuencialidad de prefijos, bases y sufijos es fundamental para estudiar la formación de palabras.

Así, el objetivo de este trabajo de investigación es desarrollar un procedimiento automatizado que describa parte de la morfología del español. Específicamente, se trata de obtener un método automático que reciba el mínimo de información lingüística *a priori* y descubra, a partir de un corpus, los sufijos y su secuencialidad en dicha lengua.

Contar con un método como éste, podría permitir al lingüista estudiar el sistema morfológico de la lengua con una mirada neutra, permitiendo que emerjan las regularidades del sistema lingüístico. Podría decirse que la idea de este proyecto de investigación es, a la manera de lingüistas como Harris o Hockett, proponer un método que permita descubrir las unidades morfológicas y la morfotáctica de una lengua desconocida.

Expondré en este capítulo los problemas de investigación que intento resolver. Además consigno los objetivos que pretendo alcanzar, las preguntas de investigación que guían mi trabajo y la metodología utilizada para desarrollar el método. También ofrezco un resumen del contenido de cada capítulo de esta tesis.

Planteamiento del problema

En esta sección presento distintos problemas que he observado, por un lado, sobre algunas propuestas prominentes de análisis morfológico automático y, por otro, sobre la poca atención que las mismas han puesto en el descubrimiento automático de la morfotáctica.

Hasta la fecha, la manera común en lingüística computacional[8] para describir la morfología de las lenguas es mediante un conjunto de reglas elaboradas a partir del conocimiento del lingüista y no descubiertas automáticamente a partir de corpus. En general, se dan por sentados los morfemas de la lengua y su ordenamiento, con lo cual se crean reglas de reconocimiento y de generación de palabras.

Aunque este tipo de métodos computacionales basados en reglas tienen tiempo de ser estudiados y engloban conocimiento lingüístico relevante (por lo que no deben ser descartados) corren el riesgo de volver su objeto de estudio el aparato de descripción y no la lengua. Además, ante una lengua desconocida, pueden intentar forzarla al conjunto de reglas prestablecidas.

Actualmente, el método estándar de análisis morfológico computacional está basado en lo que se conoce como fonología de dos niveles (Koskenniemi, 1983; 1984; Antworth, 1990)[9]. Este método permite reconocer y generar palabras de una lengua a partir de una lista de morfemas, sus rasgos sintácticos y un conjunto de reglas fonológicas de transformación, todos estos elementos previamente elaborados por el investigador.

Por otra parte, hoy en día predominan los modelos morfológicos generativistas que suelen presuponer los morfemas con los que se formulan reglas para describir su ordenamiento (Spencer, 1991). Cuando los lingüistas computacionales adoptan estos modelos

[8] Definiré la lingüística computacional como el estudio de las lenguas naturales mediante procedimientos computacionales Para mí, el objeto de estudio de esta disciplina es la lengua; sin embargo, se han dado distintas definiciones que involucran otros aspectos. Por ejemplo, Kay (2003) menciona dos objetivos de la lingüística computacional: avanzar en la teoría lingüística y desarrollar soluciones prácticas. También se ha visto a esta disciplina como el estudio de los sistemas computacionales que permiten interpretar (comprender) y generar lenguaje natural (Grishman, 1991, pág. 15).

[9] Describiré este método más adelante en la sección 3.4.

asumen que existe una morfología única e ideal, y que ésta puede describirse a partir solamente de la reflexión de un hablante.

Por lo anterior, inspirado en el trabajo de los lingüistas prechomskianos, la idea central de este trabajo no es presuponer los morfemas, sino descubrirlos y examinarlos en su entorno para encontrar y describir, mediante una gramática de estados finitos, las regularidades de su ordenamiento.

Para descubrir unidades morfológicas, el distribucionalismo hizo propuestas para identificarlas a partir solamente de la forma, dejando de lado el significado. Por ejemplo, Harris propuso un método para segmentar morfemas basado en la variedad de fonemas anteriores y posteriores a un posible corte morfológico. Entre más variedad de fonemas, es mayor la probabilidad que dicho corte sea una frontera morfológica (Harris, 1955). Esta propuesta se puede considerar el primer trabajo no supervisado de segmentación morfológica.

En términos generales, un método de segmentación no supervisado carece de una entrada de información lingüística explícita, obtenida de la reflexión y el análisis del lingüista. Más bien, esta información se obtiene del análisis automático del corpus. De esta manera, el método de Harris no recibía información explícita de dónde cortar las palabras, sino que a partir del conteo de fonemas se proponían los cortes.

Diversos métodos no supervisados de segmentación morfológica han surgido desde entonces. Uno de ellos fue propuesto a partir de la teoría matemática de la comunicación formulada por Shannon y Weaver (1964). El método consiste en medir la cantidad de información (entropía) asociada a una segmentación. Otro, de la primera mitad de los años

sesenta, se debe a un equipo de investigadores rusos a cargo de N. D. Andreev y está basado en la idea de que los afijos son más frecuentes que las bases[10].

Una propuesta adicional fue dada por De Kock y Bossaert (1978) que toma sus fundamentos del principio de economía de signos o rentabilidad del sistema. Este principio permite suponer que, dados dos segmentos, si el primero pertenece a un conjunto pequeño de segmentos muy frecuentes, mientras el segundo pertenece a un conjunto potencialmente infinito de segmentos de baja frecuencia, se puede proponer un corte morfológico entre esos dos segmentos. Finalmente, también han sido usadas estadística de digramas para determinar valores de independencia o de no asociación entre segmentos (Kageura, 1999).

Recientemente han surgido nuevos métodos no supervisados de segmentación morfológica como los basados en el modelo de longitud de descripción mínima (Goldsmith, 2001; 2006; Creutz y Lagus, 2002; 2004; 2005), optimización mediante algoritmos genéticos (Gelbukh, Alexandrov y Han, 2004) y cálculo de medidas de afijalidad (Medina, 2000; 2003)[11].

Los métodos anteriores descubren unidades morfológicas a partir del procesamiento de corpus y sin reglas o descripciones elaboradas *a priori*. Desafortunadamente, sólo el de Creutz y Lagus (2004; 2005) descubre una morfotáctica y la gran mayoría sólo segmenta las palabras en dos unidades, bases y sufijos, o máximo tres, si considera también los prefijos. En otras palabras, no se descubren las secuencias de sufijos.

Se puede observar entonces que las propuestas de análisis computacional de la morfología de las lenguas no han tomado suficientemente en cuenta el descubrimiento de la

[10] Una descripción de este método puede verse en Medina (2003, págs. 75-80).

[11] Estos métodos serán descritos con detalle en el capítulo 2.

morfotáctica; y las propuestas basadas en reglas, aunque sí lo han hecho, no infieren esa información del corpus y requieren de un aparato de descripción preconcebido.

Dos hechos, al menos, pueden explicar la falta de trabajos sobre descubrimiento automático de la morfotáctica. El primero es que la segmentación morfológica tiene su principal uso en la regularización de las palabras de un documento para su procesamiento automático, por ejemplo, para la recuperación de información o la minería de textos (Hull, 1996; Paik et al., 2011). En este proceso de regularización lo que importa es eliminar los afijos de las palabras y no rescatarlos para su análisis, este proceso se llama truncamiento (*stemming*).

El segundo hecho es la abundancia de experimentos en lenguas de morfología simple, como la del inglés. En esta lengua la combinación morfológica se da con una base y algunos prefijos y sufijos. En lenguas aglutinantes como el alemán o el finlandés, por el contrario, es posible encontrar concatenados afijos y bases en múltiples combinaciones. De hecho, son algunos trabajos para el finlandés los que comienzan a tomar en cuenta el descubrimiento de una morfotáctica (Creutz y Lagus, 2004; 2005).

Estos trabajos no sólo mejoraron sus resultados al involucrar la morfotáctica de las palabras, sino que también lograron determinar qué afijo pertenecía a qué base. De lo anterior se explica que estudiar métodos utilizados en lenguas aglutinantes haya sido importante para esta investigación.

Por otro lado, muchos de estos métodos no supervisados también asumen que existe una morfología ideal por lo que abordan el problema de descubrirla como la búsqueda de un modelo optimizado que no admite variación diatópica ni diacrónica. Lo desafortunado de esto es que los detalles del método para buscar este modelo se vuelven lo más importante y deja de ser importante el estudio del lenguaje.

Un método que no asume una morfología única e ideal, y que aborda el problema de descubrirla mediante conceptos y descripciones lingüísticas es el desarrollado por Medina (2000; 2003). Éste ha permitido obtener catálogos de prefijos y sufijos del español y otras lenguas no emparentadas como el chuj (lengua maya) (Medina y Buenrostro, 2003), la lengua checa (Medina y Hlaváčová, 2005) y el tarahumara (Medina y Alvarado, 2006). Por estas razones éste será el método que utilizaré para desarrollar mi investigación. El problema es que este método no separa las secuencias de afijos, por lo que será necesario hacer las modificaciones pertinentes.

Pensando que es posible obtener de forma automática una descripción morfológica del español (sufijos y su morfotáctica), otro problema es definir la mejor manera de representarla. En lingüística computacional, son los autómatas de estados finitos[12] (equivalentes a una gramática de estados finitos) los que se utilizan clásicamente para representar la morfología incluso de las lenguas más complejas (Sproat, 1992; Jurafsky y Martin, 2009; Goldsmith, 2010).

Emplear un autómata o gramática de estados finitos para describir la morfotáctica sufijal del español no es sólo una cuestión de representación del resultado del método que propondré, también es la suposición de que es posible tratar la morfología de una lengua como un lenguaje regular[13]. De hecho, para la morfología de la lengua española se intuye que una representación así es suficiente. Es pertinente recordar que la gramática de estados finitos será una representación de la descripción morfológica del español que se construirá automáticamente a partir del corpus y no será diseñada manualmente.

[12] Más adelante (capítulo 3) defino en qué consiste un autómata de estados finitos y una gramática de estados finitos, así como su equivalencia.

[13] Una discusión sobre la pertinencia de usar estas gramáticas para la morfología de las lenguas se puede ver en Sproat (1992) y Anderson (1992).

Una ventaja de hacerlo de esta manera es la posibilidad de motivar numerosos trabajos futuros de investigación de variación morfológica. Esto es, los fenómenos de variación, tanto dialectal como diacrónica, podrán hacerse visibles al aplicarles a muestras de varios registros el método que se espera desarrollar. Esto es, para diferentes estados de lengua cabe esperar diferentes conjuntos de morfemas que caractericen a cada estado y, por ende, diferentes gramáticas de estados finitos. Así, el método que se desarrollará podrá facilitar comparaciones dialectales y diacrónicas de una misma lengua en el nivel morfológico.

Una vez generada la gramática de estados finitos, valdrá la pena comparar esta información de carácter morfológico, inherente a los corpus, con la información elaborada por lingüistas mediante la introspección y/o métodos empíricos manuales. Esto es, se pueden descubrir datos interesantes aún para lenguas ya estudiadas porque el conocimiento que aporta el método es empírico, en el sentido de basarse en información dura presente en corpus y no en la intuición educada de un ser humano.

También, ya que el método será no supervisado (sin información lingüística explícita *a priori*) podrá ser utilizado en distintos corpus de distintas lenguas. Además, como ya se mencionó, es posible pensar que servirá para realizar comparaciones en el plano diacrónico y diatópico de una misma lengua.

Preguntas y objetivos de investigación

Planteo en esta sección las preguntas que trataré de responder en mi trabajo de investigación y los objetivos que deberé cumplir para intentar dar solución a algunos problemas planteados en la sección anterior.

Como expuse, existen métodos de análisis morfológico automático que describen la morfotáctica de algunas lenguas naturales, pero crean esta descripción de forma manual (son métodos supervisados) y generalmente a partir de la reflexión de un solo hablante.

Por otro lado, los métodos no supervisados que descubren unidades morfológicas a partir de corpus no toman en cuenta el descubrimiento de su morfotáctica y los que lo hacen sólo han tratado lenguas aglutinantes. Así, se puede decir que no hay un método automático no supervisado que descubra la morfotáctica de lenguas predominantemente afijales.

Ya que los fenómenos morfológicos son vastos y complejos en las distintas lenguas, este proyecto de investigación se centrará sólo en la lengua española y especialmente en su morfología concatenativa afijal (sufijación).

Dado lo anterior, surge la pregunta: ¿es posible la generación automática de un aparato formal de descripción morfológica a partir de corpus, que dé cuenta de los sufijos y sufitáctica del español? Se intuye que sí es posible generarlo y se propone como aparato de descripción una gramática de estados finitos.

Describir la afitáctica de una lengua mediante una gramática de estados finitos conlleva el ver a esta parte de la morfología como un lenguaje regular (de estados finitos). Dado que se ha criticado el uso de estas gramáticas para estudiar la morfología[14], cabe la siguiente pregunta ¿una gramática de estados finitos es suficiente como aparato formal de descripción morfológica de los sufijos y sufitáctica del español? Al respecto, se intuye que sí es suficiente.

Por tanto, el presente trabajo de investigación tiene el objetivo de desarrollar un método automático no supervisado para generar, a partir de corpus y mediante una gramática

[14] Véase por ejemplo Sproat (1992) y la crítica de Anderson (1992, págs. 387-391).

de estados finitos, una descripción morfológica del español, acotada al descubrimiento de sus sufijos y su morfotáctica.

Este objetivo puede descomponerse en los siguientes objetivos específicos:

1. Descubrir, a partir de corpus y mediante un método no supervisado de segmentación morfológica automática, los sufijos y sufitáctica de la lengua española.
2. Generar, a partir de los sufijos y sufitáctica descubiertos, una gramática de estados finitos que describa la morfotáctica del español.

Delimitación

Pongo en esta sección algunas consideraciones a propósito del alcance de mi investigación. En primer lugar, desarrollaré mi propuesta tomando como lengua de estudio el español, aunque cabe recordar que el método será no supervisado, por lo que se espera que funcione en otras lenguas de morfología similar, en especial lenguas flexivas como el italiano, francés, portugués o inglés.

De los numerosos fenómenos morfológicos del español, limitaré mi estudio a la morfología concatenativa y solamente la sufijal. Esto se explica porque mi interés está en descubrir la morfotáctica y considero que limitando la cantidad de fenómenos concatenativos a la sufijación es más factible formular un primer método automático que sirva de base para incorporar a futuro los demás[15].

[15] No tomo en cuenta la composición porque impone el descubrimiento de varias bases en la palabra. Tampoco la prefijación por su peculiar comportamiento en español, que ha llevado a considerarla más como composición que derivación o flexión (Moreno de Alba, 1996; Varela y García, 1999). La parasíntesis es otro fenómeno que queda fuera de mi investigación, ya que las gramáticas de estados finitos no pueden representar fenómenos discontinuos.

En lo que corresponde a la descripción que espero obtener automáticamente, no planteo hacer una distinción entre derivación y flexión, es decir, el método no marcará automáticamente cuáles sufijos son flexivos y cuáles derivativos[16]. Tampoco se propone involucrar la clase de palabra en el análisis automático para separar, por dar un ejemplo, bases y afijos nominales de verbales[17]. Esta distinción sería importante para discriminar afijos formalmente parecidos, piénsese en la –a como marca verbal (*cant-a*) y en la marca de género (*niñ-a*).

Otros aspectos no involucrados en el desarrollo del método son la identificación de alomorfos, en el sentido de que el método no propondrá si un sufijo es o no un alomorfo de otro. Además, aunque la relación entre fonología y morfología es muy estrecha, dejaré de lado el estudio de los fenómenos morfofonológicos y me centraré sólo en aspectos morfológicos. Esto no quiere decir que no sea importante tomar en cuenta la fonología.

Cabe recordar que para obtener la descripción morfológica del español procesaré computacionalmente un corpus[18], lo que implica la ausencia de análisis semántico sobre los textos. A propósito del corpus, no haré ninguna modificación para corregir errores de escritura o transcripción.

Finalmente, no adoptaré una postura teórica o corriente lingüística para definir dónde "deberían" ser segmentadas las palabras. Aunque sea difícil despegarse de la formación

[16] Aunque desde la lingüística se ha estudiado mucho la distinción entre flexión y derivación (Beard, 1998; Stump, 1998; Anderson, 1985), computacionalmente es muy difícil distinguirlas cuando se dan por afijación. En este sentido podría ayudar que se ha observado que la flexión siempre está más lejos de la base, esto es, más pegada al límite de la palabra (Greenberg, 1963).

[17] No se propone esto porque sería necesaria mi intervención para etiquetar las palabras a procesar, lo que cambiaría el carácter no supervisado del método; aunque sería interesante considerar a futuro métodos no supervisados de agrupamiento de palabras.

[18] Describiré el corpus de estudio en la siguiente sección.

que uno ha adquirido, preferiré tener, en la medida de lo posible, la mirada de un lingüista que se enfrenta a la tarea de describir una lengua desconocida[19].

Metodología

En esta sección se puntualizan dos aspectos importantes de mi trabajo. Por un lado, las perspectivas metodológicas adoptadas y, por otro, los pasos que llevé a cabo para desarrollar mi investigación.

Mi trabajo está guiado por la perspectiva metodológica de la lingüística computacional. Esta perspectiva tiene la ventaja de trabajar con instrumentos lógicos llamados programas de computadora que procesan grandes cantidades de datos lingüísticos. En esta disciplina es posible repetir varias veces un experimento mediante la ejecución de estos programas con distintas modificaciones. Por lo anterior, los experimentos realizados fueron planteados de esta manera.

Ya que este método fue desarrollado a partir del procesamiento automático de corpus, es decir, a partir de datos empíricos, fue necesario contar con una muestra representativa de la lengua de estudio. De esta manera se puede esperar que la morfotáctica inferida del corpus sea la morfotáctica del español.

Según Biber (1993), para que un corpus logre el mayor grado de representatividad su diseño debe cumplir con algunas características, siendo las más importantes las siguientes. Primero se partir de una definición lo más precisa posible de la población que se intenta estudiar. Segundo se debe elaborar una estratificación de esta población, esto es, definir los

[19] Una virtud de los métodos automáticos no supervisados es que permiten que el investigador se aleje de la lengua de estudio ya que el método por sí mismo no toma preferencias por uno u otro análisis, sólo sigue mecánicamente los pasos predefinidos por el investigador.

géneros, temas y registros de los documentos que integrarán el corpus. También es recomendable definir el tamaño de muestra textual, es decir, qué cantidad de texto será extraído de cada documento. Finalmente se debe decidir la estrategia de recopilación de las muestras textuales, por ejemplo, mediante una selección aleatoria.

En lugar de construir un nuevo corpus para esta investigación, se decidió utilizar como corpus de estudio el Corpus del Español Mexicano Contemporáneo (CEMC) ya que cumple con las características expuestas. Además, es el único corpus existente diseñado bajo criterios estadísticos como una muestra representativa del léxico del dialecto mexicano del español (Lara y Ham Chande, 1974). Si está representado el léxico, se puede pensar que también está representada gran parte de la morfología del español mexicano, por lo que los resultados que se obtengan de mi trabajo podrán asumirse como regularidades de esta variante dialectal.

Por otra parte, como establecen McEnery y Wilson (1996, pág. 32), un corpus representativo de una lengua suele verse como una referencia estándar de esa lengua, por lo que se espera que esté disponible para diversos estudios. La ventaja de utilizar un corpus representativo ya estudiado en nuevas investigaciones es que los resultados pueden ser comparados. Además, como indica esos autores, si los resultados entre estudios varían, se deberá en menor medida a los datos y más a la metodología utilizada para el análisis.

A propósito de lo anterior, el CEMC ya había sido utilizado como corpus de estudio para desarrollar un método de segmentación morfológica automática que descubre sufijos y prefijos. Por tanto, debido a todas las características mencionadas, se considera justificado el uso del CEMC como corpus de estudio para esta investigación. En seguida describiré brevemente sus características generales.

El CEMC cuenta con aproximadamente dos millones de palabras (ocurrencias) distribuidas en diversos géneros seleccionados para representar tres niveles de lengua: culta, sub-culta y no-estándar. El corpus se forma de párrafos obtenidos aleatoriamente (no consecutivos) de alguna de las 996 obras o transcripciones de grabaciones. Cada texto tiene una longitud de alrededor de dos mil palabras. Con el fin de mostrar la variedad de géneros y el porcentaje de datos por cada uno, replico la base estadística del DEM en la Tabla 0.1, tomada de Lara y Ham Chande (1974, pág. 260).

Tabla 0.1 Base estadística de DEM

Tomada de Lara y Ham Chande (1974, pág. 260)

Total de la muestra: 100%		*Porcentajes por géneros*	
Lengua Culta:	66.80%		100%
		Literatura	22.45
		Periodismo	26.34
		Ciencia	26.94
		Técnica	15.26
		Discurso político	2.69
		Religión	1.79
		Habla de la Ciudad de México	4.49
Lengua Sub-culta:	11.70%		100%
		Literatura popular	53.00
		Conversaciones grabadas	47.00
Lengua No-estandar	21.50%		100%
		Textos regionales	60.46
		Documentos de antropólogos	15.34
		Jergas	13.95
		Conversaciones grabadas	10.25

Adoptar la perspectiva de la lingüística computacional y un corpus representativo del español mexicano fueron dos decisiones metodológicas primordiales de mi investigación. Ahora presento los pasos que desarrollé para llevar a cabo mi trabajo.

El primero paso fue el estudio y caracterización de la morfotáctica en general y de la morfotáctica sufijal del español en particular. Con ello fue posible conocer el fenómeno lingüístico que describí automáticamente. Además resaltó el hecho de que para descubrir la morfotáctica primero era necesario descubrir las unidades morfológicas.

Por lo anterior, el segundo paso fue conocer las características de algunos métodos no supervisados que descubren unidades lingüísticas mediante segmentación morfológica automática. En este paso se revisó a fondo el método seleccionado para segmentar los tipos de datos del corpus de estudio. Éste método (Medina, 2000; 2003) calcula un índice de afijalidad para cada posible corte morfológico de una palabra.

Este índice de afijalidad se obtiene mediante la cuantificación de tres características lingüísticas de los afijos: son segmentos más gramaticales que las bases, no ocurren aislados y se combinan con muchos otros segmentos de baja frecuencia (bases). Las medidas utilizadas para cuantificar dichas características son la entropía asociada a un segmento, la cantidad de cuadros en los que participa el segmento y un índice de economía que mide la capacidad combinatoria del segmento[20].

El trabajo de ese autor permitió obtener un catálogo de afijos del español mediante la segmentación de las palabras una sola vez en el valor máximo de afijalidad, esto es, se dividía la palabra en una base y un sufijo (o en una base y un prefijo). Por tanto, el sufijo resultante podía ser un sufijo individual o varios sufijos concatenados. Entonces, fue necesario indagar cómo modificar este método para obtener todos los sufijos posibles.

[20] Los detalles de este método se presentarán en la sección 2.5.

Véase por ejemplo la Tabla 0.2 con los índices de afijalidad para la palabra DEFINICIONES. Segmentar en el valor más alto da como resultado DEFINI~CIONES, proponiéndose un segmento ~CIONES que deberá dividirse en dos sufijos (~CION~ES).

Tabla 0.2 Índices de afijalidad para la palabra DEFINICIONES

D	E	F	I	N	I	C	I	O	N	E	S
	0	0	0	0.2304	0.3333	**0.8398**	0.1269	0.581	0.1628	0.1087	0.2254

Otro caso es el de la Tabla 0.3, donde se pueden ver los índices de afijalidad de la palabra ALARMANTES. Cortar en el valor máximo separaría el segmento final ALARMANTE~S, dejando adherido a la base un segmento pertinente que también debería ser segmentado: ~ANTE~.

Tabla 0.3 Índices de afijalidad para la palabra ALARMANTES

A	L	A	R	M	A	N	T	E	S
	0	0	0.1738	0.3634	0.5021	0.1061	0.536	0.07867	**0.8298**

Después de la revisión sobre métodos de segmentación morfológica, decidí estudiar las gramáticas y autómatas de estados finitos. Este estudio tuvo como metas entender las bases de estos formalismos, determinar cuál de los dos era mejor estrategia para describir la morfotáctica del español y establecer su equivalencia. Se decidió que era mejor idea crear la gramática como un autómata de estados finitos.

Para el problema de la representación de la morfología de una lengua, se puede ver un autómata de estados finitos como un conjunto de estados, representados por la letra *q*, de los cuales sólo hay un estado inicial (*q0*) y uno o varios estados finales[21]. Además incluye un conjunto de morfemas y un conjunto de transiciones que indican los estados de salida a

[21] Los detalles de la definición de autómata de estados finitos, su equivalencia con una gramática y la manera de representarlo como un grafo se presentarán en el capítulo 3.

partir de un estado de entrada y un morfema (estas transiciones describen la secuencialidad de los morfemas).

La representación gráfica de un autómata de este tipo se realiza mediante un *grafo* donde los estados se representan con nodos (vértices) en forma de círculos y las transiciones con flechas dirigidas llamadas arcos. Los estados finales se representan con doble círculo. Así, una posible salida del método automático propuesto es mostrada en la Figura 0.1.

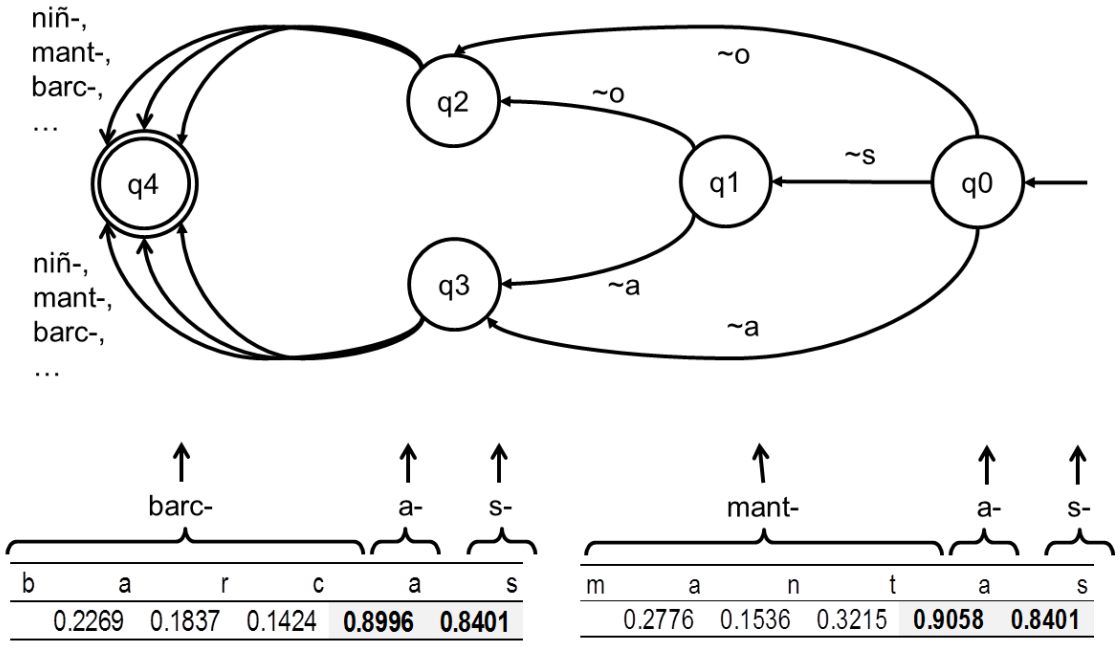


Figura 0.1. Ejemplo de grafo para algunos sufijos flexivos nominales

En el grafo se puede ver la descripción de la secuencialidad de algunos sufijos flexivos. Por ejemplo, la secuencia de estados *q0, q1, q3, q4* describen la secuencialidad de morfemas de la palabra *mant-a-s*. De esta manera, siguiendo la dirección de los arcos se puede avanzar en el grafo y ver las posibilidades de orden y secuencialidad. Se incluyen dos palabras con índices de afijalidad para hacer notar que los segmentos morfológicos que se incorporen al autómata serán propuestos por estos índices.

El siguiente paso en el desarrollo de la investigación fue el descubrimiento de las unidades morfológicas a partir del corpus de estudio. Se modificó el método de segmentación morfológica basado en el cálculo de un índice de afijalidad para obtener todos los sufijos posibles por palabra. Para ello se realizaron diversos experimentos de segmentación probando distintas variantes para calcular los índices de afijalidad. Con el experimento que ofreció mejores resultados se determinó el método de segmentación y se procesó el corpus para producir una lista de tipos de palabras segmentados.

Finalmente, el último paso consistió en la generación automática de la gramática de estados finitos para describir la sufitáctica del español. Un programa de computadora fue creado para tomar los tipos de palabras segmentados y construir el autómata. En este paso también se realizó la evaluación de resultados.

Plan de la tesis

Presento en esta sección de forma resumida el contenido de cada capítulo de la tesis. En la presente introducción, además de este plan, se presenta el planteamiento del problema, que puntualiza la falta de un método automático no supervisado de descubrimiento de la morfotáctica. Luego se exponen las preguntas y los objetivos de investigación de este trabajo. En seguida se hace una delimitación del alcance del método automático que se propondrá. Finalmente se exponen las perspectivas metodológicas adoptadas y los pasos que se llevaron a cabo para realizar este trabajo de investigación.

El capítulo uno, *Morfotáctica*, presenta la caracterización de este fenómeno lingüístico. Para ello se brinda su definición y algunas explicaciones sobre su naturaleza. Después se describe la morfotáctica sufijal del español, consignando los sufijos flexivos y derivati-

vos, tanto nominales como verbales. Al final se presenta un procedimiento para determinar esquemas morfotácticos de las lenguas humanas.

El segundo capítulo está dedicado a los métodos no supervisados de segmentación morfológica. Después de algunas generalidades, se describen cuatro métodos. El primero es el método más referenciado en este tipo de trabajos computacionales (*Linguistica*), el segundo es un método desarrollado para lenguas aglutinantes (*Morfessor*), el tercero es un método basado en optimización mediante algoritmos genéticos, y el cuarto calcula un índice de afijalidad para descubrir unidades morfológicas. Este último fue el método empleado en la tesis. El capítulo cierra con algunas observaciones generales sobre estos métodos.

En el capítulo tres, se presentan en primer lugar algunos conceptos básicos para entender las gramáticas formales y los autómatas de estados finitos. Luego se da la definición de una gramática formal y se describen sus tipos. En seguida se presentan la definición, tipos y representaciones de los autómatas de estados finitos. Después se consigna la equivalencia entre gramáticas regulares y autómatas. Se presentan también dos variantes de autómata: los autómatas probabilísticos y los modelos ocultos de Markov. Finalmente se detalla la manera en cómo la morfología computacional ha representado la morfotáctica de las lenguas.

El capítulo cuatro, *Experimentos de segmentación morfológica automática*, presenta los experimentos realizados para modificar el método de segmentación basado en el cálculo de afijalidad. Primero se describe un experimento que sirvió como primer acercamiento al problema. Luego se explica el proceso para definir un conjunto de experimentos para buscar la estrategia final de segmentación. En seguida se detalla la manera de evaluar estos experimentos, que incluye la constitución de un corpus de evaluación segmentado manual-

mente. Después se discuten los resultados de los experimentos y se brindan algunas conclusiones sobre el proceso de segmentación automática.

El capítulo cinco está dedicado al experimento de generación automática del autómata. Se presenta primero el procedimiento general para construirlo con el fin de discutir algunos aspectos de diseño del mismo. Luego se consigna el algoritmo computacional. En seguida se abordan los detalles de los experimentos realizados. Después se discuten los resultados y se presenta la evaluación del autómata final.

El capítulo de conclusiones ofrece un resumen de experimentos, la descripción final del método propuesto y la revisión de objetivos y preguntas de investigación. Además, se puntualizan los problemas del método propuesto, sus ventajas y el trabajo futuro, antes de presentar las conclusiones finales.

1. Morfotáctica

En este capítulo expondré la definición del concepto de morfotáctica y revisaré diferentes posturas que tratan de explicar su naturaleza. Después, revisaré brevemente la morfología sufijal del español y su morfotáctica, con el fin de sentar las bases para mi trabajo de investigación. Así, el objetivo de este capítulo es caracterizar mi objeto de estudio y mostrar la factibilidad de crear un método que pueda describirlo automáticamente.

1.1. Definición de morfotáctica

Esta sección está dedicada a establecer la definición de morfotáctica que utilizaré en adelante, para ello, partiré de algunas definiciones propuestas por distintos autores.

Los trabajos en morfología de distintas lenguas, especialmente los que adoptan una mirada tipológica, han dejado bien claro que los fenómenos morfológicos son variados y complejos (Sapir, 1954; Bybee, 1985; Anderson, 1992). También han establecido que existe predominio de la morfología concatenativa (Sproat, 1992, pág. 44) y en especial de la sufijal: "Cada idioma posee uno o más métodos formales para indicar la relación de un concepto secundario con respecto al concepto primario del elemento radical. Algunos de estos procedimientos gramaticales, como la sufijación, están extraordinariamente difundidos" (Sapir, 1954, pág. 71).

Al respecto, como se verá en seguida, todo indica que los estudios de morfotáctica se basan en este tipo de morfología, ya que la adición de material fonológico (o escrito)

permite hablar de orden y secuencialidad[22]. Entonces, ya que el español exhibe principalmente una morfología concatenativa y predominantemente sufijal, merece un estudio desde esta perspectiva.

Si se observan las palabras del ejemplo (1.1), cabe preguntar ¿por qué en (1.1a) se intuye un orden correcto de los segmentos de la palabra y en los ejemplos de (1.1b) no?[23] La respuesta es que los fenómenos morfológicos concatenativos están guiados por ciertas pautas de ordenamiento, es decir, las lenguas tienen una morfotáctica.

(1.1) a. cre~ar~la
b. *cre~la~ar, *ar~cre~la

Por ejemplo, Lara llama morfotáctica "a la característica que tienen todas las expresiones verbales de una lengua, de ordenar sus morfemas en una secuencia determinada o en varios esquemas secuenciales" (2006, pág. 65). Una definición similar dice que "the study of the arrangement of morphemes in linear sequence, […], is morphotactics" (Crystal, 2003, pág. 300).

En primer lugar se puede ver que la morfotáctica de una lengua tiene que ver con el orden y secuencialidad de los morfemas en las palabras. Por una parte, el orden se refiere a la posición de los morfemas ligados con respecto a la base de la palabra. Nida (1949, págs. 68-71) llamó a esto *relaciones estructurales y posicionales entre morfemas* y Hockett (1971, pág. 287) las llamó *clases posicionales*.

[22] En el caso de morfologías no concatenativas, es posible pensar que las reglas de transformación sean parte de su morfotáctica; sin embargo, esto no será indagado en mi investigación.

[23] Es posible ver en este ejemplo otras segmentaciones, una de ellas podrías ser la separación de la vocal temática de la raíz verbal (cre~a~r~la), otra quizás separaría la vocal final como marca de género del enclítico (cre~ar~l~a). Sin embargo, lo pertinente en este caso es la intuición que como hablantes tenemos de un orden de elementos.

Es posible clasificar a los morfemas ligados por su orden de aparición en prefijos (***in***-*confesable*), si preceden a la base, y sufijos (*blanc*-***o***), si aparecen después. Se han propuesto también los infijos, cuando el segmento si se insertan al interior de la base. Según Pena (1999), en español hay presencia de infijos en la derivación apreciativa[24], por ejemplo, con el segmento –it– en: *lej*-***it***-*os*, *azuqu*-***ít***-*ar*. Otro tipo de afijos son los circunfijos. Estos son discontinuos y rodean la base, son la combinación de un prefijo y un sufijo dependientes entre sí, por ejemplo: *sombra* > ***en***-*sombr*-***ec***-*er*, *rojo* > ***en***-*roj*-***ec***-*er* Es posible llamar de manera genérica a todos estos segmentos como afijos[25].

La secuencialidad, por otra parte, se refiere al encadenamiento de morfemas. Hockett (1971, pág. 131) se refirió a ésta como *ordenamiento morfemático*. Para Nida (1949, pág. 76) se trata de un análisis distribucional de morfemas, específicamente de estructuras morfológicas complejas. Este autor propone tres estructuras: tema ligado más morfema ligado, tema libre más morfema ligado, y tema libre o ligado más tema ligado o libre.

Las secuencias de morfemas ligados presentan un ordenamiento fijo (*capital*-*iz*-*ar* vs **capital*-*ar*-*iz*). Es generalmente aceptado que dicho ordenamiento está determinado por un mecanismo que impone restricciones. Como mencionan Jurafsky y Martin, la morfotáctica es "the model of morpheme ordering that explains which classes of morphemes can follow other classes of morphemes inside a word" (2009, pág. 53). De manera concreta, como lo dice Sproat, la morfotáctica es "the ordering restrictions on morphemes" (1992, pág. 83).

[24] Si en realidad se trata de infijación es un problema que no atenderé por el momento, aunque hay debate al respecto.

[25] No menciono los interfijos ya que en español presentan ciertas inconsistencias en su definición (Pena, 1999, pág. 4326).

Así, la morfotáctica también es el estudio que describe las restricciones de ordenamiento de morfemas y paradigmas de morfemas. Para lograr describir estas restricciones es necesaria la intervención humana, ya que éstas se explican generalmente con relación a las categorías gramaticales y sintácticas de la palabra cuando se adhiere cada afijo.

Para efectos de mi trabajo de investigación, que busca proponer un método no supervisado, conviene más definir la morfotáctica como la característica de las lenguas de ordenar sus morfemas en secuencias determinadas, dejando de lado por el momento el estudio de sus restricciones de ordenamiento. Con base en esta definición de morfotáctica, para este trabajo la *afitáctica* o *morfotáctica afijal* se entenderá como la descripción de orden y secuencialidad de los afijos de una lengua. De la misma forma, la *sufitáctica* o *morfotáctica sufijal* será el orden de sufijos de una lengua.

1.2. La naturaleza de la morfotáctica

En esta sección se revisarán brevemente algunas posturas teóricas que tratan de explicar las restricciones de ordenamiento de afijos. A pesar de que el método que se propondrá no tomará en cuenta un análisis de este tipo, es pertinente conocer algunos aspectos sobre la naturaleza de la morfotáctica.

Como se verá, no parece haber una postura que explique todos los tipos de ordenamiento en la palabra. En términos generales, en las restricciones de ordenamiento están involucrados los siguientes aspectos. Primero, la tendencia a preferir que algunos afijos estén más cerca de la raíz o base debido a la relevancia que para el hablante tenga la información asociada al morfema. Segundo, utilizar la aparición de los afijos como rastro de algún cambio en la sintaxis. Tercero, en la adjunción de morfemas se involucran las catego-

rías de palabras (adjetivo, sustantivo, etcétera) tanto de la base como de la base más los afijos.

1.2.1. Constituyentes inmediatos

El estructuralismo propuso un análisis de la palabra basado en una jerarquía de constituyentes inmediatos. Un ejemplo de este análisis puede verse en la Figura 1.1, tomado de Anderson (1992, pág. 13).

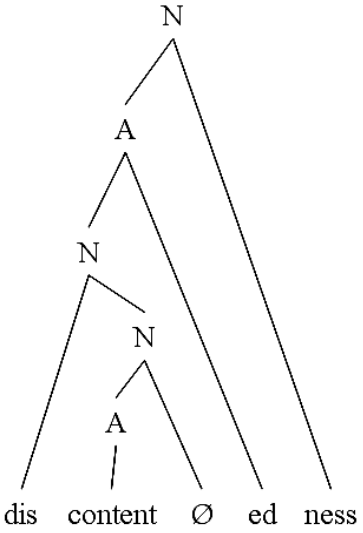


Figura 1.1 Estructura de constituyentes inmediatos de *discontentedness*
Tomada de Anderson (1992, pág. 13)

Para el estructuralismo, la clase de palabra determina los morfemas que pueden o no adherirse en la construcción de una palabra (Sproat, 1992, págs. 83-84). De esta manera, se tendrían afijos que sólo pueden unirse a verbos, adjetivos o sustantivos. Además, cada morfema traería consigo un cambio de categoría para la palabra que se está formando, producto de su unión con otro morfema. Esto último se puede ver en el árbol de constituyentes que muestra la Figura 1.1.

1.2.2. Restricciones sintácticas

Por otro lado, la gramática generativa propuso que la naturaleza de la morfotáctica era totalmente sintáctica. Se decidió que el ordenamiento de morfemas podía explicarse mediante los mismos esquemas que la sintaxis, por tanto, para esta corriente de la lingüística no era necesario el estudio de la morfotáctica (Anderson, 1992, págs. 15-17).

Según esta postura, los componentes de la gramática de una lengua son las reglas de estructura de frase y las transformaciones, que producen la representación superficial de una oración. Por ejemplo, Chomsky (1984) explica las variantes morfológicas de los verbos auxiliares del inglés mediante una regla que selecciona el verbo junto con un afijo. Luego una transformación pone al afijo en el lugar correcto en la estructura superficial.

El hecho de que las reglas y transformaciones tuvieran acceso libre a los morfemas de la lengua no hizo necesario un componente morfológico para los generativistas. De la misma forma, la morfotáctica fue vista como una parte de la sintaxis. Hoy en día la morfología tiene su lugar también en esta postura teórica y se sigue estudiando la relación entre estos dos niveles (Piera y Varela, 1999).

1.2.3. Principio de espejo entre sintaxis y morfología

Es claro que ciertos fenómenos morfológicos tienen una conexión estrecha con la sintaxis. Un ejemplo de éstos son los cambios de valencia de un verbo que se marcan morfológicamente. Me refiero a los procesos de construcción de pasivas, antipasivas, aplicativas y causativas, llamados, en términos de la gramática generativa, reglas o procesos de cambio de función gramatical.

Normalmente, cuando la correspondencia entre la función semántica y la función gramatical se interrumpe por uno de estos procesos, el verbo sufre cambios morfológicos

(Katamba y Stonham, 2006, pág. 287). Se cree que estas marcas morfológicas que sufre el verbo permiten al oyente recuperar la función sintáctica subyacente, aunque esto está en discusión.

Al respecto, lo pertinente para esta investigación es que en muchas lenguas cada cambio de función gramatical corresponde a un cambio morfológico. Esto es conocido como principio de espejo (*Mirror Principle*). El siguiente ejemplo de la lengua luganda (1.2), tomado de Katamba y Stonham (2006, pág. 289), lo muestra claramente[26]:

(1.2)

a. Sujeto Verbo
Abaana basoma
Niños leer
'Los niños leen'

b. Causativa
Sujeto Verbo Objeto
Nnaaki asom-es-a abaana
Nnaaki leer-CAUS-BSV niños
'Nnaaki hace que los niños lean'

c. Causativa pasiva
Sujeto Verbo Objeto
Abaana basom-es-ebw-a Nnaaki
niños leer-CAUS-PASS-BSV Nnaaki
'Los niños son puestos a leer por Nnaaki'

CAUS= causativa, BSV=sufijo verbal básico, PASS= pasiva

Según lo expuesto en (1.2) se da primero el proceso de creación de la causativa y después el de la pasiva. Morfológicamente hay una correspondencia ya que se adhiere primero el sufijo de causativa (~es) y después el de pasiva (~ebw). La idea detrás es que cada adición de rol en la construcción sintáctica se ve reflejada en la morfología del verbo con la adición de un morfema.

[26] También pueden verse más ejemplos en Sproat (1992, pág. 85).

El principio de espejo ha sido cuestionado con evidencia en distintas lenguas[27], por lo que resulta difícil sostener que éste baste para explicar la naturaleza de la morfotáctica. Katamba y Stonham (2006, págs. 291-293) mencionan algunos ejemplos donde se viola el principio de espejo. Uno es del náhuatl clásico, donde la adjunción de morfemas sigue un orden definido sin aparente relación con tal principio. El otro es del bantú, en esta lengua todo sufijo verbal pronunciado con un solo sonido tiende a aparecer después de todos los otros sufijos, lo que habla de una motivación más bien fonológica.

1.2.4. Universales lingüísticos

Greenberg (1963) hace una serie de observaciones de carácter tipológico sobre la morfología de un conjunto de 30 lenguas. Según este autor, existe predominio de la sufijación sobre la prefijación y de éstas sobre fenómenos de morfología afijal discontinua[28]. Observa además que lenguas exclusivamente prefijales son muy raras, mientras que las exclusivamente sufijales no lo son.

Producto de sus observaciones, Greenberg propone una serie de universales lingüísticos de los cuales rescataré aquí sólo los que son pertinentes para mi trabajo de investigación. Según el universal 28, la derivación siempre estará más cerca de la base cuando aparezca con flexión; no importa si se dan por prefijación o sufijación.

El universal 29 establece que si una lengua tiene flexión, siempre tiene derivación. En cuanto a las categorías flexivas verbales, siempre que se presenten en la lengua categorías de persona y número o de género, habrá también categorías de modo y tiempo (universal 30). Sobre el orden de categorías flexivas nominales, se reporta en el universal 39 que si

[27] Para leer una discusión al respecto véase Katamba y Stonham, (2006), sección 12.6.

[28] Incluye aquí la infijación, intercalación (por ejemplo, el cambio vocálico en lenguas semíticas) y ambifijación (llamada tradicionalmente circunfijación o parasíntesis).

aparecen las de número y caso, ambas como prefijos o sufijos, la de número casi siempre está más cerca de la raíz.

Los universales llevan a Greenberg a proponer la idea de jerarquías de proximidad, esto es, que ciertos elementos deben estar más cerca de un elemento central[29]. Éste podría ser la base o raíz de la palabra. Una consecuencia de esta jerarquía sería que elementos más cercanos a la raíz (por ejemplo el número gramatical) aparezca en mayor número de lenguas.

Bybee (1985) también realiza un estudio de diversas lenguas. Coincide con Greenberg en el orden entre derivación y flexión. Además, propone que se trata de un nivel de relevancia de significado lo que determina el hecho de que la derivación aparezca más cercana a la base. De acuerdo con esta relevancia sería posible predecir el orden de morfemas: a mayor relevancia, mayor cercanía (Bybee, 1985, pág. 34).

Bajo esta mirada, Bybee explica el universal 39 de Greenberg: la categoría flexiva de número aparece más cerca de la base o raíz por ser más relevante para el significado del sustantivo en el que aparece (una entidad o varias entidades). El caso gramatical, por su parte, es relevante al interior de la oración y por eso aparece más lejos en comparación con el número.

Bybee analiza 50 lenguas y encuentra pocas excepciones en su predicción del ordenamiento de morfemas. El orden de categorías flexivas que propone sería: aspecto, tiempo, modo y persona. Según esta autora, es bastante estricto el orden de aspecto y tiempo (no encontró contraejemplos), y más flexible el de modo y persona.

[29] No es claro si esta proximidad es de tipo semántico o de relevancia para el hablante.

1.2.5. Morfología léxica

Una idea básica de esta propuesta teórica (llamada LPM por sus siglas en inglés: *Lexical Phonology and Morphology*) es que el elemento principal del análisis morfológico es la palabra y no el morfema. Además, propone como principio fundamental que existe un lazo muy fuerte entre las reglas morfológicas, que construyen la palabra, y las reglas fonológicas, que determinan la manera de pronunciarla. Ambas reglas se distribuyen en niveles jerarquizados (*strata*) dentro del léxico, uno debajo del otro, dando organización al componente morfológico de una lengua (Katamba y Stonham, 2006).

Según esta postura, la naturaleza de la morfotáctica está basada en esto niveles jerarquizados (Kiparsky, 1983; Sproat, 1992). Cada nivel se encarga de cierta parte de la morfología, por ejemplo, de ciertos afijos. Así, dependiendo del nivel al que pertenece el afijo, se da el ordenamiento. Además, cada nivel contiene reglas particulares, de manera que el producto de las reglas de un nivel alimenta al siguiente, de allí el ordenamiento de morfemas.

En esta postura los afijos del inglés se pueden dividir en neutrales, que no cambian fonológicamente la base a la que se adhieren, y no neutrales, que sí afectan fonológicamente algún segmento de la base. Por ejemplo, –less sería un sufijo neutral ya que no cambia de lugar el acento de la base a la que se une *home/home-less*; por otro lado, el sufijo –ic sería no neutral porque sí cambia de lugar el acento de la base *strategy/strategic*.

La división de afijos del párrafo anterior permite ordenarlos en niveles. El nivel I sería el de los afijos no neutrales. El nivel II sería el de afijos neutrales. Su orden de adjunción en la base se explica por el orden de los niveles, esto es, siempre se adhieren primero los afijos de nivel I y luego los de nivel II.

Se puede ver un ejemplo en (1.3), tomado de Kiparsky (1983, pág. 3). El sufijo –ian pertenece al nivel I, por eso aparece primero. Luego se adhiere el sufijo –ism, que pertenece al nivel II.

(1.3) a. Mendel-ian-ism, mongol-ian-ism.
b. *Mendel-ism-ian, *mongol-ism-ian.

En este caso, el sufijo –ian es no neutral (nivel I) porque mueve el acento a la sílaba anterior. Por el contrario, –ism es un sufijo neutral (nivel II), esto es, no mueve el acento de la base a la que se adhiere (Kiparsky, 1983; Katamba y Stonham, 2006).

En el caso de afijos del mismo nivel también existe un orden determinado en la palabra. Sin embargo, este ordenamiento no está predefinido, más bien se presenta de acuerdo con el proceso de formación de la misma. Cada regla morfológica incluye la clase de palabra sobre la que actúa y la clase de palabra que entrega. Así, el orden de afijos está determinado por la clase de palabra de la base. En (1.4) se puede ver un ejemplo tomado de Katamba y Stonham (2006, pág. 115).

(1.4) a. 'home$_N$-less$_A$-ness$_N$
b. *'home$_N$-ness$_N$-less$_A$

Tanto –less como –ness son sufijos neutrales, por lo que ambos pertenecen al nivel II. El orden de estos sufijos está determinado por la clase de palabra de la base de derivación. Así, el sufijo –ness no puede unirse a sustantivos (1.4b), sino a adjetivos, por lo que aparece después del sufijo –less, que sí puede unirse a sustantivos y los convierte en adjetivos (1.4a).

Ahora bien, cuando ocurren en la misma palabra afijos derivativos y flexivos que pertenecen al mismo nivel, los primeros se adhieren a la base antes que los segundos. En

otras palabras, entra en juego otro ordenamiento de reglas: las reglas de derivación suceden antes de las de flexión.

La propuesta de la morfología léxica ha sido debatida principalmente porque el ordenamiento puede ser explicado de otras maneras, dejando de lado los niveles jerarquizados (Sproat, 1992, pág. 88)[30].

1.3. La morfotáctica del español

Las secciones anteriores han permitido conocer el concepto de morfotáctica y algunas posturas que tratan de explicar su naturaleza. Indagaré ahora sobre el objeto de estudio de mi investigación: la morfotáctica del español.

En primer lugar, hay que decir que el español puede concatenar un número considerable de afijos, como se puede ver en el ejemplo (1.5), aunque estas formas no son muy frecuentes. En el ejemplo hay dos prefijos unidos a una base, a la que después se unen tres sufijos.

(1.5) Anti-re-elec-cion-ista-s

Como se puede ver, para estudiar la morfotáctica es necesario conocer los elementos que se concatenan. Por tanto, describiré en seguida, de forma general, la morfología sufijal del español, ya que, como dije en la introducción de esta tesis, para mi investigación sólo tomaré en cuenta los sufijos. Los prefijos y otros fenómenos concatenativos se dejarán para trabajo futuro.

[30] Una crítica extensa sobre esta postura teórica puede verse en el capítulo 7 de Katamba y Stonham (2006).

No intentaré hacer un estudio extenso y detallado de los fenómenos involucrados en la sufijación, ya que mi interés está únicamente en poner las bases para el desarrollo de mi trabajo. Por ello no abundaré ni en posturas teóricas ni en explicaciones semánticas. Trato, más bien, de dar una aproximación de lo que espero descubrir con el método automático, así que pondré especial énfasis en determinar cuáles son los segmentos morfológicos con los que se realizan los fenómenos flexivos y derivativos del español.

1.3.1. Morfología sufijal del español

El español cuenta con diversos fenómenos morfológicos tanto concatenativos como no concatenativos, aunque indudablemente es una lengua predominantemente sufijal. Tanto la flexión como la derivación se dan por medio de la concatenación de segmentos finales. Es más, todas las marcas gramaticales, tanto verbales como nominales (género, número, tiempo, modo, aspecto y persona) se realizan mediante sufijación, con excepción de algunos casos supletivos más o menos comunes (*padre* vs *madre*, *soy* vs *fui*).

1.3.1.1. Flexión

La flexión es el fenómeno morfológico que modifica una palabra para crear paradigmas que permitan expresar variaciones funcionales o gramaticales de la palabra original. También se ha propuesto que la unidad que se flexiona no es la palabra sino el tema (Hockett, 1971; Pena, 1999) o el lexema (Stump, 1998).

Los paradigmas flexivos suelen ser muy regulares en el sentido de que una palabra acepta todas las variantes morfológicas de un paradigma determinado, aunque existen excepciones como los verbos que se conjugan sólo en algunas personas o la presencia de formas supletivas. También, se puede decir que las marcas flexivas de una palabra son obliga-

torias. Además, para hablar de flexión es común poner como condición que la palabra flexionada no cambie de categoría gramatical.

En español existe tanto flexión nominal como verbal y ambas son realizadas mediante la adición de sufijos. En los apartados subsecuentes esbozaré algunas características de estas flexiones.

1.3.1.1.1. Flexión nominal

Las formas nominales en español presentan flexión de género y número. Por un lado, el número se expresa formalmente con la ausencia o presencia de los sufijos *–s* o *–es*, siendo un fenómeno morfológico con bastante regularidad. El género, por otro lado, no es tan regular y está asociado a los sufijos *–a*, *–e*, y *–o* (Ambadiang, 1999).

El género

Según Ambadiang (1999), en la flexión nominal de género es posible encontrar pares regulares (*gato/gata*), palabras que alternan sufijo pero que cambian semánticamente (*manzana/manzano*), unidades léxicas independientes (*padre/madre*), y palabras únicas que no tiene género opuesto (*víctima*).

En el caso de pares semánticamente distintos, pero que alternan segmento final de género (*manzana/manzano*), éstos suelen analizarse como casos de derivación y no de flexión. De esta manera, la terminación *–o* de *manzano* sería un sufijo derivativo que formaría derivados con significado de "árbol que produce".

En el caso de los sustantivos animados, el género corresponde habitualmente con el sexo de los referentes (*niño/niña*). Si no se da esta correspondencia, la tendencia es encontrar una sola forma ya sea femenina o masculina (*víctima*).

Aunque la correlación entre género gramatical y sexo del referente es bastante general, no existe una relación obligatoria. Se pueden encontrar sustantivos que pertenecen a un determinado género sin importar el referente (*foca, piloto*). En estos casos se suele marcar el género mediante una palabra adjunta (*foca macho, la piloto*).

En términos generales, la asignación del género en los sustantivos inanimados se da con base en una organización en clases léxicas (Ambadiang, 1999, pág. 4854). Por ejemplo, los días, meses, años, siglos, idiomas, vinos y puntos cardinales son masculinos; mientras que las carreteras, horas, islas, montañas y letras del alfabeto son sustantivos femeninos.

Claro que hay excepciones en esta organización, como en las estaciones del año: *el invierno* (masculino) y *la primavera* (femenino). En otros casos es posible encontrar palabras que mantienen su significado en ambos géneros (*la azúcar/ el azúcar*), las que pueden tener ambos géneros en singular, pero no en plural (*el mar/la mar/los mares/*las mares*)[31], y las que cambian de significado con el cambio de género (*la capital/el capital*).

Según Ambadiang (1999, pág. 4872), hay dos posturas para el análisis morfológico del género. La más restrictiva sólo considera que hay morfemas de género cuando se dan pares de palabras con alternancia de segmentos (desinencias) y estas conservan un solo significado (*niño/niña*). En los casos de pares con cambio de significado o formas léxicas independientes, los segmentos finales no son considerados morfemas, aunque éstos coincidan formalmente con las marcas prototípicas (*naranja/naranjo, caballo/yegua*).

La otra postura, menos restrictiva, considera que el género puede presentarse mediante diversas marcas, por lo que los segmentos *–o, –e, –a* y *–Ø* son considerados alomorfos del morfema de género. Para efectos de mi investigación, asumiré esta postura y tomaré

[31] Utilizo el asterisco (*) como marca de agramaticalidad.

a los segmentos *–a, –e, –o* como marcas de género. Con lo anterior esperaría que estos segmentos aparezcan en la morfotáctica generada automáticamente.

El número

El número se formula de manera muy regular en español. En el caso del singular no hay marca asociada, mientras que el plural se expresa con las terminaciones *–s* o *–es*. En cuanto a su semántica, el número no sólo significa una oposición cuantitativa (*casa/casas*), sino también una diferencia de intensidad (*agua/aguas*), cambio de matiz de significado (*belleza/bellezas*) o no produce cambio (*pantalón/pantalones*) (Ambadiang, 1999, págs. 4886-4889).

La selección del segmento que marca plural se puede explicar por las siguientes reglas. Si la palabra en singular termina en vocal no acentuada o *–é*, entonces se añade *–s*. Si la palabra en singular acaba en consonante distinta de *–s* o vocal acentuada seguida de *–s* entonces se añade *–es*. Si la palabra en singular termina en vocal no acentuada seguida de *–s* no hay cambio en plural. Finalmente, aunque la norma sea la adición de *–es*, los sustantivos terminados en vocal acentuada *–í, –ú, –á* y *–ó* presentan doble forma de plural, una con *–s* y otra con *–es* (*colibrís, colibríes*).

Hay tres posturas que intentan explicar morfológicamente la presencia del segmento *–es* en los plurales del español (Ambadiang, 1999, pág. 4892). La primera toma este segmento como alomorfo de *–s*, la segunda propone el apócope de *–e–* en la forma singular, y la tercera plantea la epéntesis de *–e–* en el plural.

En lo que respecta a los pronombres, sólo hay adjunción de –s en *ella/ellas, la/las, le/les, lo/los*. Por otro lado, la mayoría de los adjetivos posesivos, demostrativos y artículos

presentan plural regular. Además, la marcas de plural *–es* y *–s* aparecen en pronombres relativos, interrogativos, cuantificadores, conjunciones y números ordinales.

Ya que el plural es muy regular en español, es de esperarse que la representación morfológica generada incluya ambos sufijos.

1.3.1.1.2. Flexión verbal

Los verbos del español se agrupan en tres conjugaciones de acuerdo con la terminación de su infinitivo. La forma de infinitivo se ha utilizado tradicionalmente como representante de una conjugación o grupo de variantes de un verbo. Así, se tiene el grupo de verbos de la primera conjugación, que asocia infinitivos terminados en –ar (*trabajar*); el de la segunda conjugación, que incluye infinitivos terminados en –er (*obtener*); y los verbos de la tercera conjugación con infinitivos terminados en –ir (*vivir*).

Según Alcoba (1999) es posible analizar la estructura del verbo en dos componentes principales. El primero es una parte invariable llamada raíz, que da el significado léxico. El segundo es una parte variable, formada por la vocal temática (VT) específica de cada conjugación (–a–, –e–, –i–) y por la terminación o desinencia, constituida por los morfemas de tiempo-aspecto-modo (TAM) y número-persona (NP). Cabe resaltar que hay otras propuestas sobre la estructura del verbo, también consignadas por Alcoba, que proponen más o menos los mismos constituyentes.

Casos especiales son las llamadas formas no finitas o no personales del verbo: infinitivo, gerundio y participio. En éstas están presentes sólo la raíz verbal, la vocal temática y una marca específica para cada forma: –r para infinitivo, –ndo para gerundio y –do para participio. Según Alcoba (1999, pág. 4923) estas marcas deben considerarse flexivas, aunque hay discusión al respecto.

Las variantes flexivas del verbo (conjugaciones) siguen patrones bastante regulares, de tal manera que pueden asociarse a tres modelos, uno para cada conjugación. A los verbos con este comportamiento se les llama verbos regulares. Existen también otros verbos, algunos de uso muy frecuente, que se comportan de manera irregular, esto es, que no siguen los patrones normales de conjugación. Estos son llamados verbos irregulares.

Más allá de discutir una postura teórica que explique la estructura del verbo en español, mi objetivo es conocer algunas posibilidades de segmentación para compararlas con las que emerjan del corpus de estudio mediante el método automático. Por tanto, en seguida consigno dos propuestas de segmentación para verbos regulares de las tres conjugaciones. La primera es del Diccionario del Español de México (DEM)[32] y la segunda es de Alcoba (1999).

Tabla 1.1 Segmentación de verbos regulares del DEM y de Alcoba

	DEM			ALCOBA		
	1ª conjug.	2ª conjug.	3ª conjug.	1ª conjug.	2ª conjug.	3ª conjug.
	AMAR	COMER	SUBIR	CANTAR	TEMER	PARTIR
	INDICATIVO					
	PRESENTE					
	(am–)	(com–)	(sub–)	(cant–)	(tem–)	(part–)
1s		–o			–o	
2s	–as	–es		–a–s	–e–s	
3s	–a	–e		–a	–e	
1p	–amos	–emos	–imos	–a–mos	–e–mos	–i–mos
2p	–an	–en		–a–n	–e–n	
2p	–áis	–éis	–ís	–á–is	–é–is	–í–s
3p	–an	–en		–a–n	–e–n	

[32] Diccionario del Español de México (DEM) http://dem.colmex.mx, El Colegio de México, A.C., [15 de noviembre de 2012]

Tabla 1.1 Segmentación de verbos regulares del DEM y de Alcoba (continuación)

	DEM			ALCOBA		
	1ª conjug. AMAR	2ª conjug. COMER	3ª conjug. SUBIR	1ª conjug. CANTAR	2ª conjug. TEMER	3ª conjug. PARTIR
	PRETÉRITO					
	(am–)	(com–)	(sub–)	(cant–)	(tem–)	(part–)
1s	–é	–í		–é	–í	
2s	–aste	–iste		–a–ste	–i–ste	
3s	–ó	–ió		–ó	–ió	
1p	–amos	–imos		–a–mos	–i–mos	
2p	–aron	–ieron		–a–ro–n	–ie–ro–n	
2p	–asteis	–isteis		–a–ste–is	–i–ste–is	
3p	–aron	–ieron		–a–ro–n	–ie–ro–n	
	FUTURO					
	(amar–)	(comer–)	(subir–)	(cant–)	(tem–)	(part–)
1s		–é		–a–ré	–e–ré	–i–ré
2s		–ás		–a–rá–s	–e–rá–s	–i–rá–s
3s		–á		–a–rá	–e–rá	–i–rá
1p		–emos		–a–re–mos	–e–re–mos	–i–re–mos
2p		–án		–a–rá–n	–e–rá–n	–i–rá–n
2p	–áis	–éis		–a–ré–is	–e–ré–is	–i–ré–is
3p		–án		–a–rá–n	–e–rá–n	–i–rá–n
	COPRETÉRITO					
	(am–)	(com–)	(sub–)	(cant–)	(tem–)	(part–)
1s	–aba	–ía		–a–ba	–í–a	
2s	–abas	–ías		–a–ba–s	–i–a–s	
3s	–aba	–ía		–a–ba	–í–a	
1p	–ábamos	–íamos		–á–ba–mos	–í–a–mos	
2p	–aban	–ían		–a–ba–n	–í–a–n	
2p	–abais	–íais		–a–ba–is	–í–a–is	
3p	–aban	–ían		–a–ba–n	–í–a–n	
	POSPRETÉRITO					
	(amar–)	(comer–)	(subir–)	(cant–)	(tem–)	(part–)
1s		–ía		–a–ría	–e–ría	–i–ría
2s		–ías		–a–ría–s	–e–ría–s	–i–ría–s
3s		–ía		–a–ría	–e–ría	–i–ría
1p		–íamos		–a–ría–mos	–e–ría–mos	–i–ría–mos
2p		–ían		–a–ría–n	–e–ría–n	–i–ría–n
2p		–íais		–a–ría–is	–e–ría–is	–i–ría–is
3p		–ían		–a–ría–n	–e–ría–n	–i–ría–n

Tabla 1.1 Segmentación de verbos regulares del DEM y de Alcoba (continuación)

	DEM			ALCOBA		
	1ª conjug.	2ª conjug.	3ª conjug.	1ª conjug.	2ª conjug.	3ª conjug.
	AMAR	COMER	SUBIR	CANTAR	TEMER	PARTIR
SUBJUNTIVO						
PRESENTE						
	(am–)	(com–)	(sub–)	(cant–)	(tem–)	(part–)
1s	–e	–a		–e	–a	
2s	–es	–as		–e–s	–a–s	
3s	–e	–a		–e	–a	
1p	–emos	–amos		–e–mos	–a–mos	
2p	–en	–an		–e–n	–a–n	
2p	–éis	–áis		–é–is	–á–is	
3p	–en	–an		–e–n	–a–n	
PRETÉRITO						
	(am–)	(com–)	(sub–)	(cant–)	(tem–)	(part–)
1s	–ara/ase	–iera/iese		–a–ra/–a–se	–ie–ra/ie–se	
2s	–aras/ases	–ieras/ieses		–a–ra–s/–a–se–s	–ie–ra–s/–ie–se–s	
3s	–ara/ase	–iera/iese		–a–ra/–a–se	–ie–ra/ie–se	
1p	–áramos /ásemos	–iéramos/iésemos		–á–ra–mos/ –á–se–mos	–ié–ra–mos/ –ié–se–mos	
2p	–aran/asen	–ieran/iesen		–a–ra–n/–a–se–n	–ie–ra–n/–ie–se–n	
2p	–arais/aseis	–ierais/ieseis		–a–ra–is/–a–se–is	–ie–ra–is/–ie–se–is	
3p	–aran/asen	–ieran/iesen		–a–ra–n/–a–se–n	–ie–ra–n/–ie–se–n	
FUTURO						
	(am–)	(com–)	(sub–)	(cant–)	(tem–)	(part–)
1s	–are	–iere		–a–re	–ie–re	
2s	–ares	–ieres		–a–re–s	–ie–re–s	
3s	–are	–iere		–a–re	–ie–re	
1p	–áremos	–iéremos		–á–re–mos	–ié–re–mos	
2p	–aren	–ieren		–a–re–n	–ie–re–n	
2p	–areis	–iereis		–a–re–is	–ie–re–is	
3p	–aren	–ieren		–a–re–n	–ie–re–n	
IMPERATIVO						
	(am–)	(com–)	(sub–)	(cant–)	(tem–)	(part–)
2s	–a	–e		–a	–e	
2s	–e	–a				
2p	–ad	–ed	–id	–a–d	–e–d	–i–d
2p	–en	–an				

Tabla 1.1 Segmentación de verbos regulares del DEM y de Alcoba (continuación)

DEM			ALCOBA		
1ª conjug.	2ª conjug.	3ª conjug.	1ª conjug.	2ª conjug.	3ª conjug.
AMAR	COMER	SUBIR	CANTAR	TEMER	PARTIR
NO PERSONALES					
(am–)	(com–)	(sub–)	(cant–)	(tem–)	(part–)
–ar	–er	–ir	–a–r	–e–r	–i–r
–ando	–iendo		–a–ndo	–ie–ndo	
–ado	–ido		–a–do	–i–do	

Considero que estas propuestas dan muestra de dos extremos. Por un lado, el DEM propone pocas segmentaciones ya que no descompone los elementos flexivos, lo que da como resultado que siempre haya dos segmentos. Además, en el caso del futuro de indicativo y del pospretérito, propone mantener completa la forma de infinitivo (*amar-é, amar-ía*). Por otro lado, Alcoba, como ya lo había mencionado, propone varios cortes que corresponden a la separación de vocal temática, terminaciones de tiempo-aspecto-modo y terminaciones de número-persona.

1.3.1.2. Derivación

La derivación es el fenómeno morfológico que modifica una palabra para crear una nueva con un significado diferente y por lo general con una nueva categoría gramatical. De acuerdo con Pena (1999), la derivación se encarga de la generación de nuevos temas (unidades que queda después de eliminar los morfemas flexivos).

En las siguientes secciones presento algunas generalidades sobre la derivación nominal y verbal del español. Al igual que en la flexión, mi interés estará puesto en determinar los sufijos y su ordenamiento, ya que esperaría que ambos se reflejaran en la descripción morfológica que generaré automáticamente.

1.3.1.2.1. Derivación nominal

El estudio de la derivación nominal del español enfrenta algunos retos que presentaré de manera breve con el fin de entender la complejidad de este fenómeno morfológico. De acuerdo con Santiago y Bustos (1999, pág. 4507), estos problemas se presentan al tratar de determinar los siguientes aspectos de los sufijos derivativos: a) su inventario y características formales, b) su segmentación, c) sus fenómenos morfofonológicos, d) su semántica, e) sus alternancias y f) su variación dialectal.

Según estos autores, no hay consenso en el inventario de sufijos derivativos nominales del español. Por un lado, hay sufijos que derivan tanto sustantivos como adjetivos, lo que dificulta proponer la separación entre derivación nominal y derivación adjetival. Por otro lado, no resulta sencillo determinar si algunos segmentos son sufijos independientes o alomorfos de un solo sufijo. Al respecto, hay dos posturas generales. La primera considera que hay sufijos distintos cuando se dan cambios formales. La segunda prefiere ver como alomorfos a los sufijos en distribución complementaria y parecido formal.

Otros problemas se agregan cuando se trata de separar el sufijo de la base[33]. Éstos se pueden dar en casos donde los sufijos muestran diferencias formales mínimas con una distribución dudosa o impredecible. También en derivados deverbales que producen indecisiones; por ejemplo, si tomar la vocal temática como parte de la base o del sufijo. Finalmente, en situaciones donde la segmentación produce bases de derivación muy dudosas.

Complicaciones adicionales al estudio de la derivación nominal presentan los fenómenos morfofonológicos que se dan principalmente en la base, algunos de manera inconsistente. Ejemplos de estos son: monoptongaciones (*sentim***ie***nto/sentim***e***ntal*), alternancias

[33] Véase también Moreno de Alba (1986).

vocálicas (*j***o***ven/j***u***ventud*), perdida de vocal final (*van***o***/vanidad*) y alternancias consonánticas (*públi***c***co/publi***c***idad*), entre otros.

Los últimos problemas que mencionan Santiago y Bustos (1999) tienen que ver con la inconsistencia en la asignación de contenido semántico por parte de los sufijos, muchas veces porque asignan contenido muy especializado. Además, la poca sistematicidad en la alternancia de sufijos, principalmente por la presencia de diferentes sufijos que asignan el mismo contenido semántico. Finalmente, la existencia de variaciones dialectales que afectan la productividad, la semántica y la selección de algunos sufijos.

Considero que los problemas anteriores dan muestra, por un lado, de la complejidad de elaborar un método que descubra automáticamente el inventario de sufijos derivativos y, por otro, de que los resultados de un método automático no coincidirán completamente con el análisis humano. Por otro lado, si descubrir la lista de sufijos y su ordenamiento ya es complicado, intentar distinguir entre sufijos flexivos y derivativos lo es más, por lo que no trato este problema en mi investigación.

Tomaré de Moreno de Alba (1986) su inventario de sufijos derivativos del español, ya que el estudio que hizo para obtenerlos se basa en corpus y en especial en uno de español mexicano, lo que coincide con mi metodología de trabajo. La Tabla 7.1 del anexo A muestra la lista de sufijos, sus alomorfos, una brevísima descripción y un ejemplo tomado de su estudio.

Este inventario de sufijos derivativos del español permite ver, al menos, dos problemas con los que tendrá que lidiar el método automático. El primero es la diversidad de alomorfos de ciertos sufijos, por ejemplo –adura, –atura, –idura, –tura y –ura. Lo más probable es que el método busque regularidades y no obtenga tanta variedad de sufijos. El segundo es la coincidencia formal entre ciertos sufijos derivativos y flexivos, piénsese en los

segmentos finales –ía, –aría, –o y –a. Éstos coinciden con marcas flexivas verbales y nominales.

1.3.1.2.2. Derivación verbal

En español se pueden derivar verbos de distintas categorías como: adjetivos (*blanco/blanquear*), sustantivos (*burbuja*/*burbujear*), verbos (*dormir/adormecer*), pronombres (*tú/tutear*) y adverbios (*adelante/adelantar*). La derivación se puede dar por concatenación de sufijo o por adjunción de sufijo y prefijo, lo que se ha llamado parasíntesis (***en***-*roj*-***ecer***). Para mi investigación, como ya había mencionado, no tomaré en cuenta procesos parasintéticos del español.

Tradicionalmente, se reconocen como sufijos derivadores de verbos a los siguientes: –ar, –ear, –ecer, –ificar e –izar (Serrano-Dolader, 1999; Beniers, 2004). Además, hay una serie de segmentos que pueden anteceder a los mencionados sufijos y que se unen a la base de derivación, como –et– en *toqu**et**ear* o –urr– en *cant**urr**ear*. Al respecto, Beniers (2004, pág. 143) los agrupa como alomorfos de un sufijo –VC.

En seguida haré una breve descripción de los sufijos derivativos basándome en Beniers (2004), quien realizó un estudio sobre el mismo corpus que utilizo en mi investigación, el CEMC. Pondré énfasis en el aspecto formal, más que en el semántico, por la naturaleza de mi investigación.

El sufijo –ar genera verbos a partir de sustantivos, adjetivos y adverbios. Algunos de los fenómenos que ocurren en la base son: elisión de vocal final (*adelant**e**/adelant**ar***) o cierre de vocal final (*concept**o**/concept**u**ar*). También se dan casos de producción de un verbo en forma transitiva e intransitiva (*adelante/adelantar/adelantarse*). Como dice Beniers (2004, pág. 72), este tipo de derivación produce vacilación en su direccionalidad, ya que es

igual de válido derivar sustantivos de verbos que verbos de sustantivos (¿*abogar* > *abogado* o *abogado* > *abogar*?).

El sufijo –ear permite derivar verbos de sustantivos (*boicot/boicot**ear***), adjetivos (*redondo/redond**ear***) y verbos (*bailar/bailot**ear***). Entre los fenómenos morfofonológicos que ocurren se puede dar elisión de vocal (*cábul**a**/cabul**ear***) y, al igual que para el sufijo –ar, el acento pasa a la marca de infinitivo (*líder/liderear*).

En el caso del sufijo –ecer, se forman verbos preferentemente de adjetivos (*claro/**es**clar**ecer***) y en menor medida de sustantivos (*flor/flor**ecer***) y verbos (*dormir/**a**dorm**ecer***). El sufijo –ificar forma verbos de adjetivos (*eléctrico/electr**ificar***) y de sustantivos (*código/cod**ificar***); también es común encontrar bases de derivación cultas, como *petrificar*. Finalmente, -izar produce verbos a partir de sustantivos (*mártir/martir**izar***), algunas veces con presencia de sufijo –AL (*norma*/*norm**alizar***); y también de adjetivos, muchos terminados en –al (*vital/vit**alizar***). En estos tres sufijos también se da pérdida de vocal final de la base.

1.4. Procedimiento para determinar esquemas morfotácticos

En Lara (2006) encontré una propuesta para determinar los esquemas morfotácticos de una lengua. Para este autor, estos esquemas son como los esquemas silábicos, es decir, "patrones canónicos" propios de una lengua. La propuesta para determinar el esquema morfotáctico de una palabra es la siguiente (Lara, 2006, pág. 66):

(1.6) a) Segmentar la secuencia [de fonemas o letras] en morfemas.
b) Probar la cohesión que hay entre ellos.
c) Determinar el orden en que aparecen.

Considero pertinente discutir esta propuesta para ver si es posible proponer un método automático que se base en ella. Por tanto, en los siguientes párrafos analizaré cada uno de los pasos.

La segmentación de palabras en morfemas ha sido una de las tareas más importantes de los estudios morfológicos. Por su parte, la morfología computacional cuenta también con propuestas para descubrir morfemas de manera automática. Así, llevar a cabo el primer paso que propone el método de (1.6) es factible, por lo que en el capítulo 2 revisaré algunos métodos automáticos de segmentación morfológica.

Para analizar el segundo paso (1.6b), es necesario entender el término *cohesión*. Al hablar de este término, Lara se refiere a "una especie de pegamento o 'glutinosidad'" (2006, pág. 67) entre morfemas. Esta idea de glutinosidad fue propuesta por Medina (2000; 2003). Para Lara, la cohesión se puede detectar al intentar insertar elementos entre ellos. Ésta puede ser alta, como sería entre la raíz verbal y la vocal temática del español, entre las cuales no es posible intercalar ningún segmento. También puede ser cohesión media, por ejemplo entre bases nominales y flexiones de género, ya que es posible insertar algunos morfemas derivativos (*-it-* o *-uch-*). Finalmente, la cohesión puede ser tan baja como entre palabras, aunque algunas tienden a aparecer muy pegadas con otras, como los clíticos.

Es en el estudio de lenguas aglutinantes donde cobra mayor relevancia el análisis de cohesión entre morfemas y el descubrimiento de esquemas morfotácticos, ya que estas lenguas pueden encadenar grupos de alta cohesión (núcleos morfemáticos) para formar nuevos

significados[34]. Por otro lado, en lenguas como el español la tendencia es que las palabras tengan un solo núcleo morfemático.

Si bien Lara habla de intercalar elementos para medir la cohesión entre morfemas, utilizaré para mi investigación un método automático que propone cuantificar esta glutinosidad mediante el cálculo de ciertas medidas de afijalidad (Medina, 2000; 2003), el cuál revisaré en el capítulo 2. Es este método el que propone la idea de glutinosidad para referirse a la fuerza de adhesión entre unidades lingüísticas.

El último paso (1.6c) consiste en determinar el orden de morfemas en la palabra. Este orden sigue un esquema morfotáctico específico de la lengua, que forma parte de un conjunto de esquemas posibles. En morfologías concatenativas, estos esquemas describirían los morfemas que se prefijan y se sufijan a las bases. Lara (2006, pág. 81) propone los siguientes esquemas morfotácticos más frecuentes del español, véase (1.7).

(1.7) Lexema ligado (raíz verbal) + gramema de persona: *am+o*.

Lexema libre: *ducha*.

Lexema ligado + gramema de género + gramema de número: *niñ+o+s*.

Lexema ligado + gramema derivativo + gramema de género + gramema de número: *niñ+it+o+s*.

Gramema preposicional + lexema ligado + gramema derivativo + gramema de género + gramema de número: *anti+american+ist+a+s*.

Idear un método automático que ofrezca los resultados que propone Lara no es nada sencillo. Se requeriría distinguir entre morfemas libres y ligados, entre derivación y flexión, y entre categorías flexivas (nominales y verbales). Por eso, el método que propondré para inferir la morfotáctica involucrará sólo el análisis de segmentos, con la única distinción entre bases y sufijos.

[34] En el capítulo 2 revisaré un método automático desarrollado para lenguas aglutinantes que precisamente se preocupa por analizar la morfotáctica de cada palabra.

El método seguirá, en términos generales, la propuesta de Lara. Para los pasos uno y dos, utilizaré un método automático de segmentación morfológica que plantea calcular la glutinosidad como medida de adhesión entre segmentos mediante un índice de afijalidad. Este método lo explicaré en el capítulo 2, junto con otros métodos de segmentación.

Luego, para el paso final de describir el ordenamiento, utilizaré un autómata de estados finitos. Este "dispositivo" abstracto, usado regularmente en la morfología computacional, será descrito en el capítulo 3. Finalmente, a partir del autómata, será posible generar una lista de patrones que describan la morfotáctica del español. Los llamaré *patrones morfotácticos* para distinguirlos de los esquemas morfotácticos de la propuesta de Lara.

El siguiente capítulo estará dedicado al estudio de algunos métodos de segmentación morfológica automática, dentro de los cuales se encuentra el método que utilizaré para realizar mi propuesta.

2. Métodos de segmentación morfológica automática

El capítulo anterior me permitió definir mi objeto de estudio. Presenté primero la definición de morfotáctica y algunas posturas que tratan de explicar su naturaleza. Luego consigné datos sobre la morfología sufijal del español con el fin de conocer su morfotáctica y su inventario de sufijos. Al final, analicé un procedimiento para determinar esquemas morfotácticos que coincide con el método automático que propondré.

Como se vio, para describir la morfotáctica del español es necesario identificar los segmentos morfológicos que forman las palabras. Por lo anterior, en este capítulo expondré algunas propuestas automáticas de segmentación morfológica. Pondré especial atención en propuestas no supervisadas, esto es, que requieran el mínimo de información lingüística *a priori*.

Haré primero una revisión general sobre diversos métodos, para luego describir en detalle algunos de ellos. El primero es el método más referenciado en trabajos de segmentación automática no supervisada para lenguas flexivas. El segundo es un método desarrollado para lenguas aglutinantes que toma en cuenta la morfotáctica de la palabra. El tercero es una propuesta de optimización con algoritmos genéticos. Finalmente, el cuarto es un método para cuantificar la fuerza de adhesión entre segmentos (glutinosidad).

2.1. Generalidades sobre los métodos de segmentación

No han sido pocos los acercamientos con los que se ha abordado el problema de descubrir automáticamente unidades morfológicas en corpus, por lo que en esta sección brindo un panorama muy general sobre algunos de ellos y de las características que distinguen unos de otros[35].

La morfología computacional es el tratamiento de los fenómenos morfológicos de las lenguas naturales mediante procedimientos automáticos (simbólicos, estadísticos o una combinación de ambos). Dada la complejidad de la morfología, estos estudios son de diversa naturaleza.

Por ejemplo, están aquellos cuyo fin es la generación de un conjunto de reglas de reconocimiento y generación de palabras. Otros estudios buscan encontrar reglas para manipular los cambios en la morfofonología (morfofonémica) de los segmentos morfológicos. Un ejemplo más es el conjunto de estudios centrados en el descubrimiento de unidades morfológicas principalmente en lenguas de morfología concatenativa, como el inglés o español. Es en este último grupo de estudios que se centra mi trabajo.

[35] Para leer sobre otros métodos de segmentación morfológica automática no considerados en este apartado pueden consultarse las siguientes fuentes. En Medina (2003) hay una revisión de las primeras propuestas de métodos automáticos desde la de Harris. En los artículos de Creutz y Lagus (2002; 2004; 2005) y Creutz, (2003) se revisan, aunque brevemente, distintos métodos contemporáneos. Goldsmith (2010) ofrece una revisión más amplia de diversos métodos en el marco de la segmentación morfológica general. Finalmente, Hammarström y Borin (2011) ofrecen un estudio comparativo de casi 200 métodos de segmentación morfológica no supervisada desde el método de Harris.

En términos generales, el descubrimiento de unidades morfológicas se lleva a cabo mediante un procedimiento de segmentación. Es decir, tomar decisiones sobre dónde cortar una cadena hablada o una cadena de texto. Las dos unidades que se descubren son: palabras y morfos[36]. Por un lado, el descubrimiento de palabras cobra relevancia en leguas escritas que no usan espacios entre palabras o en la segmentación de discurso hablado (cadenas de fonemas o fonos). Por otro lado, los morfos se descubren a partir de las palabras ya identificadas en un texto.

Para visualizar el alcance y las limitantes de los procedimientos de segmentación automática, conviene recordar la amplia diversidad de fenómenos morfológicos en las distintas lenguas humanas. Esto conlleva que no serán las mismas estrategias las que se tomen en lenguas de poca morfología, que en lenguas aglutinantes donde el encadenamiento de unidades puede ser considerable, véase un ejemplo del turco en (2.1) tomado de Sproat (1992, pág. 20).

(2.1)
çöp+lük+ler+imiz+de+ki+ler+den+mi+y+di
gargabe+AFF+PL+1P/PL+LOC+REL+PL+ABL+INT+AUX+PAST
'was it from those that were in our garbage cans?'

Existen métodos computacionales que utilizan recursos lingüísticos prefabricados que incluyen el análisis morfológico de un humano (métodos supervisados). Estos se contraponen a los métodos donde no hay recursos de ese tipo y el mismo método propone cierto análisis a partir del mínimo de información lingüística *a priori* (métodos no supervisados). Para esta tesis, los segundos son fundamentales.

[36] Hockett (1971) propone el término *morfo* como la representación de un morfema. En lingüística computacional, se ha adoptado este término como la realización gráfica (ortográfica) de un morfema (Sproat, 1992, pág. 247).

Ejemplo del primer tipo de métodos es el trabajo de Sproat et al. (1996), quienes propusieron un método para segmentar, en palabras, textos de caracteres chinos. Recuérdese que el chino o el japonés no utilizan espacios para delimitar palabras escritas. Por el contrario, en lenguas como el español es posible utilizar el espacio y los signos de puntuación como marcadores de segmentación de palabras, aunque con esto no todos los problemas estén resueltos[37].

Como anotan estos autores, la segmentación de palabras depende del sistema de escritura de la lengua. En algunas, como el español, se tendrá la posibilidad de hablar de palabras ortográficas (separadas por espacio o signos), en otras no, como el chino. Por tanto, el problema de segmentar palabras en chino es más difícil que en español.

Su método está basado en una lista de palabras y afijos creada manualmente y modelada (representada) como un transductor de estados finitos[38]. La secuencia de símbolos chinos (*hanzis*) a segmentar se modela con un autómata de estados finitos (un aceptador, computacionalmente hablando). Para la segmentación, un procedimiento transforma el autómata aceptador en transductor cuya ruta menos costosa es la segmentación propuesta como correcta.

Otra propuesta que incluye recursos hechos a mano es la de Teahan et al. (2000), también para la segmentación de palabras en chino. Estos autores utilizan un corpus segmentado previamente por un humano. El procedimiento de segmentación está basado en

[37] Piénsese en las abreviaturas, cantidades o cifras, contracciones, etcétera.

[38] El concepto de transductor de estados finitos está basado en el concepto de autómata de estados finitos, que explicaré en el capítulo 3. Por ahora puedo adelantar que un autómata es una representación de una cadena de símbolos (lenguaje). El transductor representa pares de símbolos que se corresponden, algunas veces por reglas lingüísticas que convierten un símbolo en otro.

crear un modelo de compresión de texto que se utiliza para insertar espacios en el corpus a segmentar.

Es cierto que, como dicen estos autores, lo único necesario para contar con un segmentador para otra lengua es otro corpus segmentado, ya que el procedimiento de segmentación sería el mismo, pero estos recursos no abundan ni son fáciles de conseguir o adaptar. Cada día hay más corpus etiquetados de más lenguas, pero las diferencias en los criterios de constitución y etiquetado pueden llegar a dificultar su adopción. Además, existen lenguas de bajos recursos computacionales, como muchas lenguas mexicanas, de las que no se cuenta con corpus electrónicos en particular porque no tienen sistema de escritura.

He mencionado dos métodos para segmentar palabras a partir de un corpus. Pero son de mayor interés para mi investigación los métodos para segmentar palabras en morfos. A propósito de ellos, es importante decir que la gran mayoría se ha desarrollado para lenguas de morfología concatenativa relativamente simple, como el inglés. En este capítulo también tomaré en cuenta métodos desarrollados para lenguas como el español, que tiene una morfología flexiva más compleja que la del inglés, y un método desarrollado para lenguas aglutinantes para llevar esta investigación a un extremo interesante.

Los métodos de segmentación morfológica pueden obtener, en términos generales, dos salidas. La primera es una lista de bases y afijos (comúnmente llamado lexicón de morfos[39]) donde se pierde la relación de qué afijos pertenecen a qué bases. La segunda es una descripción morfológica, que puede incluir paradigmas de bases y afijos, o la morfotáctica

[39] Utilizaré el término lexicón en sentido computacional para referirme a la lista de bases y afijos utilizada para procesamiento automático.

de cada palabra. Cuando la descripción incluye una morfotáctica, ésta puede tener los siguientes niveles de detalle[40]:

(2.2) a) prefijo–base–sufijo
b) (prefijo*–base–sufijo*)+

La primera representación (2.2a) es la que comúnmente obtienen los métodos automáticos que se han desarrollado para lenguas flexivas con poca morfología, como el inglés, o con más morfología como el español o el francés. Lo que me gustaría resaltar es que estos métodos sólo cortan dos veces la palabra, una para determinar un prefijo y otra para determinar un sufijo, por lo que pueden obtener resultados como los siguientes (para una palabra del español y otra del francés), tomados de Goldsmith (2001, págs. 180-181): *acontecimiento-s* y *abolitionniste-s*, donde no hay separación de los sufijos derivativos.

La otra representación (2.2b) se ha obtenido principalmente para lenguas aglutinantes donde se puede encontrar un conjunto de prefijos unidos a una base seguida de varios sufijos y todo esto seguido de otra base con sus respectivos afijos. Esta descripción también sería necesaria para lenguas como el español con la idea de obtener varios sufijos por palabra (*acontec-imiento-s)* o en compuestos morfológicos como *saca-corcho-s*. Este último nivel de detalle es más complejo porque conlleva cierta jerarquía de elementos que indique qué afijo corresponde a qué base, por ejemplo [*saca-[[corcho]-s]]*[41].

Es posible clasificar los distintos métodos de segmentación morfológica por la estrategia principal que utilizan; aunque algunos son realmente una combinación de varias estra-

[40] En esta notación utilizada en la representación de lenguajes regulares, el asterisco '*' significa que el término anterior se repite cero o más veces y el signo '+' que se repite una o más veces (Karttunen, Chanod y Grefenstette, 1996, pág. 308; Hopcroft, Motwani y Ullman, 2001)

[41] Véase Val Álvarado (1999, págs. 4788-4799) para una descripción de este tipo de compuestos.

tegias. Primero están los métodos basados en un conteo de letras anteriores y posteriores a una posible segmentación. Harris (1955) propone este método y se convierte en el primer método no supervisado de segmentación morfológica[42].

El método se basa en contar la variedad de fonemas potenciales anteriores y posteriores a un posible corte morfológico (se prueban todos los posibles cortes de una palabra). Entre más variedad de fonemas potenciales, mayor la probabilidad de una frontera morfológica, ya que esa variedad representa mayor incertidumbre (Harris, 1955). Otros métodos basados en el método de Harris son los de Déjean (1998), Ando y Lee (2000) y parte del de Goldsmith (2001).

También se pueden encontrar métodos basados en similitud semántica (Schone y Jurafsky, 2000; 2001), que utilizan la propuesta llamada *Semántica Latente*, y otros basados en similitud ortográfica (Neuvel y Fulop, 2002; Baroni, Matiasek y Trost, 2002). Otro tipo de métodos engloba un conjunto bastante grande de propuestas para crear modelos probabilísticos. La siguiente lista muestra algunos ejemplos de estos trabajos:

- Longitud de descripción mínima (Deligne y Bimbot, 1997; De Marcken, 1995; Creutz y Lagus, 2002; 2005; Kit y Wilks, 1999; Goldsmith, 2001).
- Entropía del modelo morfológico (Redlich, 1993).
- Probabilidad máxima (*Maximum Likelihood*) (Creutz y Lagus, 2002).
- Modelos bayesiano (Brent, 1999; Creutz, 2003; Creutz y Lagus, 2004).

Finalmente también hay métodos de optimización del modelo morfológico (Gelbukh, Alexandrov y Han, 2004; Gelbukh et al. 2008; Lara Reyes, 2008). En las si-

[42] Harris propone también distintas variantes del método que no describiré, como contar los distintos fonemas una y dos posiciones antes del posible corte.

guientes secciones profundizaré en algunos de estos métodos con el fin de compararlos con el método que utilicé en mi trabajo de investigación.

2.2. *Linguistica*

En este apartado describiré con detalle el algoritmo para segmentación morfológica que propone Goldsmith (2001; 2006; 2010). Este algoritmo fue implementado en un programa de computadora llamado *Linguistica* y hoy en día es comúnmente usado como estándar de comparación (*gold standard*) para trabajos de segmentación morfológica del inglés y otras lenguas.

El trabajo de Goldsmith, a decir de él mismo, tiene dos objetivos. Un objetivo práctico que consiste en contar con un analizador morfológico para varias lenguas que pueda ser usado en tareas como recuperación de documentos o traducción automática. También tiene un objetivo teórico que es conocer cuánta información *a priori* requiere un dispositivo (programa) de inducción de morfología capaz de hacer un análisis sobre la estructura del lenguaje muy cercano al que haría un lingüista.

Para comenzar a entender su método de descubrimiento morfológico, se puede decir que lleva a cabo dos grandes pasos:

1. Utilizar un conjunto de heurísticas para proponer segmentaciones.
2. Evaluar el proceso de segmentación con el modelo de Longitud de Descripción Mínima (*Minimum Description Length*, MDL).

Para el primer paso, utiliza dos tipos de heurísticas. Las primera son heurísticas "de fuerza bruta" (*bootstraping*), que proponen segmentaciones de palabras en bases y afijos. Las segundas son heurísticas que este autor llama incrementales, éstas proponen modifica-

ciones a las primeras segmentaciones. Luego, de acuerdo con una evaluación con base en el modelo MDL, se decide cuáles modificaciones se aceptan y cuáles se rechazan.

El modelo MDL, adaptado al problema de segmentación morfológica, establece una medida de longitud de descripción del corpus (*C*) tomando en cuenta un modelo morfológico probabilístico (*M*). Dicha descripción se obtiene mediante la suma de la longitud del modelo morfológico más la longitud de la compresión del corpus. La manera de calcular la longitud de descripción mínima se puede ver en la ecuación (2.3), tomada de Goldsmith (2006, pág. 355).

$$DescriptionLength(C, M) = length(M) + \log_2 \frac{1}{prob(C|M)} \qquad (2.3)$$

Como puede verse, la longitud de compresión del corpus se calcula mediante el logaritmo base 2 del recíproco de la probabilidad asignada al corpus (C) dado el modelo morfológico (M). Entre más alta sea la probabilidad condicional de que la morfología describa al corpus, *prob(C|M),* mejor será la morfología como modelo, pero ya que se utiliza el recíproco, entonces el número menor será indicador del mejor modelo morfológico.

En resumen, la morfología *M* que minimice la función de (2.3) es la mejor morfología del corpus. El primer término de esa función expresa qué tan compacta es la morfología y el segundo expresa qué tan bien esa morfología describe al corpus en cuestión.

Para determinar la longitud de la morfología, Goldsmith usa tres elementos que se generan automáticamente mediante las heurísticas antes mencionadas:

- La lista de bases.
- La lista de afijos.

- La lista de estructuras combinatorias (*signatures*) que almacenan qué bases pueden aparecen con qué afijos.

En la Figura 2.1 se puede ver un ejemplo de una de estas estructuras combinatorias. Así, parte del tamaño del modelo morfológico está basado en el cálculo de los apuntadores que asocian los afijos y las bases.

$$\begin{Bmatrix} crawl \\ jump \\ walk \end{Bmatrix} \begin{Bmatrix} NULL \\ ed \\ ing \\ s \end{Bmatrix}$$

Figura 2.1. Estructuras combinatorias (*signatures*)
Tomada de Goldsmith (2006, pág. 355)

El método de Goldsmith determina segmentaciones iniciales mediante las siguientes heurísticas:

- Método de Harris usando sólo los sucesores frecuentes.
- Búsqueda de estructuras combinatorias entre los segmentos propuestos por el método de Harris.

La idea de usar los sucesores frecuentes consiste en contar el número de letras distintas que aparecen inmediatamente después de un segmento. Entre más sucesores, mayor es la probabilidad de una segmentación. Por ejemplo, en (2.4a) el número de sucesores frecuentes después del segmento *gover-* es de uno (sólo la letra *n*). En cambio, después del segmento *govern-* hay seis: *e, i, m, o, s, espacio/signo* (2.4b).

(2.4)
a) *gover-n, gover-ned, gover-ning, gover-nment, gover-nor, gover-ns.*
b) *govern, govern-ed, govern-ing, govern-ment, govern-or, govern-s.*

Goldsmith restringe el método de Harris para segmentar sólo cuando se propongan bases de tres o más letras y cuando el valor del sucesor frecuente de la letra anterior y pos-

terior sea exactamente igual a 1. Una vez obtenidos dos segmentos por palabra se forman estructuras combinatorias tomando el primer segmento como base y el segundo como sufijo. En seguida se filtran algunas de estas estructuras de acuerdo con los siguientes criterios:

- Se aceptan las estructuras donde los sufijos aparezcan en por lo menos tres palabras.
- Se aceptan las estructuras con más de 25 bases asociadas.
- Si las estructuras tienen menos de 25 bases, se aceptan aquellas con al menos dos sufijos de al menos dos letras de longitud.

El siguiente paso es revisar las estructuras para evaluar si es posible pasar letras finales de las bases a los sufijos. Para saber cuántas letras pasar, se calcula la entropía de los segmentos finales no mayores de cuatro letras. Cualquier cambio en las segmentaciones es evaluado mediante el modelo MDL.

Después se llevan a cabo más ajustes a las segmentaciones:

- Se revisan las bases para saber si alguna puede segmentarse usando las bases y sufijos ya descubiertos.
- Se revisan bases y sufijos para determinar más estructuras combinatorias.
- Se toman palabras que comiencen con las bases determinadas y se establecen los segmentos finales como sufijos si aparecen en al menos tres palabras.
- Se toman las palabras que terminan con los sufijos descubiertos, los segmentos iniciales se establecen como bases si forman con el sufijo una estructura combinatoria ya existente. Cuando no forman una estructura se evalúa la aceptación del segmento como base si disminuye la longitud de descripción del modelo.

Finalmente, se buscan alomorfos de bases ya descubiertas usando dos procedimientos. El primero es buscar alomorfos que cambien sólo por la pérdida de la letra vocal final

(*lov-/ love-*). El segundo es buscar alomorfos donde uno de ellos sufra cambio en la letra vocal final, por ejemplo *y* > *i*. Sobre la evaluación, el método obtuvo 72% de exactitud (*accuracy*) en las primeras 300,000 palabras del corpus Brown.

Llama la atención de este método el uso de las estructuras combinatorias (*signatures*) para validar las segmentaciones. Esto ayuda a que los segmentos propuestos por el método sean morfológicamente más pertinentes, ya que su pertenencia a una de estas estructuras resalta su carácter combinatorio.

También resulta interesante cómo esta propuesta trata el problema de descubrir la morfología de una lengua como la búsqueda de un conjunto "óptimo" de bases, afijos y estructuras combinatorias; sin embargo, esta estrategia conlleva la presuposición de que existe una morfología única e ideal a la que el método aspira a llegar, lo cual es discutible.

2.3. *Morfessor*

Morfessor es el nombre que Creutz y Lagus (2005) le dieron a un conjunto de métodos que han desarrollado al menos desde al año 2002 para segmentar palabras del finlandés en morfos. En esta sección reviso estos métodos ya que, como lo he mencionado, estudiar la segmentación automática en una lengua aglutinante permite tener una perspectiva más amplia de la complejidad computacional del problema a resolver. Además, algunos de estos métodos toman en cuenta la morfotáctica de la palabra.

Los dos métodos desarrollados inicialmente por estos autores son llamados *morfessor baseline* (Creutz y Lagus, 2002; Creutz, 2003). El primero está basado en la generación de una lista de morfos (lexicón) a partir de segmentaciones aleatorias de las palabras y evaluadas por funciones de costo. El menor costo de una función significa el lexicón más compacto que mejor describe el corpus.

El segundo consiste en generar un lexicón de un corpus de entrada y generar un corpus artificial a partir de este lexicón mediante procedimientos probabilísticos. Cada palabra del corpus artificial es segmentada *n* veces hasta que el corpus generado es exactamente el corpus de entrada. El tamaño del lexicón de morfos, la longitud de cada uno de ellos, los caracteres que los forman, su orden y su frecuencia intervienen en el cálculo de funciones de probabilidad. Sus resultados para el inglés, comparados con *Linguistica*, fueron peores o semejantes. Por otra parte, en finlandés estos métodos superaron al método de Goldsmith.

Según los autores, uno de los problemas de estos métodos es que palabras muy frecuentes quedan sin segmentar y las poco frecuentes muy segmentadas. Esto se debe principalmente a que al poner la palabra más frecuente de manera completa en el lexicón se logra el menor costo. Es importante mencionar que el lexicón es *plano*, en el sentido de que no refleja la estructura interna de las palabras, por ejemplo, qué sufijo pertenece a qué base.

El siguiente método propuesto por Creutz y Lagus (2004) es llamado *morfessor categories-ML*. Usa probabilidad máxima (*Maximun Likelihood* , ML) y asocia los morfos a tres categorías: prefijos, sufijos y bases. En este método se descubre una morfotáctica gracias a las transiciones de una categoría a otra.

Al reflexionar en la complejidad del finlandés, decidieron analizar la palabra como una combinación de bases y afijos alternando libremente, y resolvieron tomar en cuenta dependencias en la secuencia de los morfos (morfotáctica). Se basan en tres supuestos:

a) Los morfos pueden pertenecer a tres categorías: bases, prefijos y sufijos.
b) No se pueden tener ciertas secuencias como sufijo a inicio de palabra, prefijo a final de palabra y prefijo seguido de sufijo sin pasar por una base.
c) Las categorías tienen ciertas características:

i. Lo afijos tienen información sintáctica, son de propósito general. Éstos son usados con un gran número de otros morfos.

ii. Las bases tienen información semántica y forman un conjunto más grande que el de los afijos.

iii. Las bases no son muy cortas con el fin de distinguirse unas de otras.

Para asignar probabilidades a cada posible segmentación de una palabra utilizaron modelos ocultos de Markov[43]. La probabilidad de una segmentación para una palabra *w* en varios morfos $\mu_1\mu_2...\mu_k$ está dada por la fórmula de (2.5), tomada de Creutz y Lagus (2004, pág. 45).

$$p(\mu_1\mu_2 \dots \mu_k|w) = \left[\prod_{i=1}^{k} p(C_i|C_{i-1}) \cdot p(\mu_i|C_i)\right] \cdot p(C_{k+1}|C_k) \qquad (2.5)$$

En la fórmula anterior, $p(C_i|C_{i-1})$ expresa la probabilidad de transición de la categoría de un morfo a la siguiente. Además, la probabilidad de un morfo μ_i dada una categoría C_i para ese morfo está expresada por $p(\mu_i|C_i)$. Finalmente, C_{k+1} representa el final de palabra, por lo que $p(C_{k+1}|C_k)$ expresa la probabilidad de transición de la categoría del último morfo hacia el final de la palabra.

En resumen, la probabilidad de una segmentación está dada por el producto de las probabilidades de transición entre categorías, desde el inicio de palabra, C_0, hasta el final de palabra, C_{k+1}, y la probabilidad de que cierta categoría se asigne a cada morfo.

[43] Un modelo oculto de Markov es un autómata de estados finitos cuyas transiciones tienen asignadas probabilidades (Charniak, 1996, pág. 32).

Una característica interesante del método (y del finlandés) es que permite que un morfo pueda funcionar como base o como afijo dependiendo de la palabra[44]. En seguida pongo los pasos generales del procedimiento de segmentación:

a) Segmentación inicial con el método de Creutz y Lagus (2002).
b) Dada la segmentación anterior se asignan las categoría más probables:
    i) Prefijo, si es difícil predecir el siguiente morfo usando la medida de perplejidad[45].
    ii) Sufijo, si es difícil predecir el anterior morfo usando la medida de perplejidad.
    iii) Base, mediante una función basada en la longitud en letras.
    iv) Se agrega la categoría ruido (*noise*) para morfos que no caen en ninguna de las anteriores.
c) Se segmentan palabras formadas por otros morfos ya descubiertos, excepto:
    i) Si se segmenta en morfos ruido.
    ii) Si es una secuencia de categorías no permitida.
    iii) Si los segmentos tienen baja probabilidad.
d) Se eliminan los morfos ruido. Se unen al morfo adyacente, prefiriendo:
    i) Morfos pequeños.
    ii) Morfos ruido.
    iii) Bases.

[44] Según estos autores *pää* es prefijo en la palabra 'pää+aihe+e+sta' y base en 'pää+hän' (Creutz y Lagus, 2004, pág. 49).

[45] La perplejidad es una medida basada en la entropía cruzada (*cross entropy*). En palabras de Manning y Schütze "a perplexity of *k* means that you are as surprised on average as you would have been if you had had to guess between *k* equiprobable choices at each step" (1999, pág. 78).

Para evaluar el método, compararon sus resultados con los de un corpus segmentado con un analizador morfológico hecho manualmente. Para el finlandés este método resultó mejor que el método del 2003 y que *Linguistica* de Goldsmith. Sus resultados fueron del 79% de precisión para 16 millones de palabras en finlandés.

Según los autores, el método disminuyó errores gracias a la concatenación de morfos ruido, a la resegmentación de palabras con morfos ya descubiertos y al uso de categorías. Sin embargo, generalizó y segmentó sufijos donde no había. Un ejemplo de segmentación y de asignación de categorías para una palabra del finlandés puede verse en (2.6), basado en Creutz y Lagus (2004, pág. 49).

(2.6) bahama – saar – et
BASE – BASE – SUFIJO
Bahama – isla – PL.
'Islas Bahamas'

El último método propuesto por estos autores, *morfessor categories-MAP* (Creutz y Lagus, 2005), usa el enfoque de *Máximo a Posteriori* (MAP), equivalente al modelo de Longitud de Descripción Mínima descrito en el apartado (2.2). Algunas características importantes del método son que genera un lexicón jerárquico, donde cada palabra está formada por cadenas de caracteres o por morfos, que a su vez pueden estar formados recursivamente por otros morfos.

Nuevamente se toma en cuenta una representación de la morfotáctica de las palabras, ya que cada una se representa como un modelo oculto de Markov que incluye cuatro categorías de morfos: prefijo (PRE), sufijo (SUF), base (STM) y no-morfema (NON).

Este método propone encontrar, mediante funciones de probabilidad, el mejor lexicón de morfos, esto es, el conjunto más compacto de morfos que describa lo mejor posible

al corpus. El problema entonces se puede ver como la búsqueda de un lexicón que maximice la probabilidad condicional $p(lexicon|corpus)$, esto es, la probabilidad de un lexicón dado en corpus. La ecuación de (2.7), tomada de Creutz y Lagus (2005), expresa el cálculo de esta probabilidad donde $arg \max_{lexicon}$ indica que se buscan los valores que maximicen las probabilidades.

$$arg \max_{lexicon} p(lexicon|corpus) = arg \max_{lexicon} p(corpus|lexicon) \cdot p(lexicon) \qquad (2.7)$$

Por una parte, la probabilidad de que el corpus sea descrito por el lexicón, $p(corpus|lexicon)$, se obtiene de los modelos ocultos de Markov que representan la morfotáctica de las palabras del corpus de entrada. Por otra parte, la probabilidad del lexicón de morfos, $p(lexicon)$, se determina a su vez por dos probabilidades que, a grandes rasgos, intentan describir la forma y "significado" de los morfos.

La forma se describe mediante la probabilidad de que el morfo esté formado de letras o de que esté formado por submorfos. El "significado" se describe mediante las probabilidades de la frecuencia del morfo, su tamaño en letras y la perplejidad a su izquierda y derecha (qué tanto se puede predecir el morfo anterior o siguiente).

Como reportan los autores, este método superó bastante los resultados de *Linguistica* para el finlandés y logró rebasar sus métodos anteriores (*morfessor baseline* y *morfessor categories-ML*). Un ejemplo de la segmentación para la palabra finlandesa "oppositiokansanedustaja" (miembro del parlamento de la oposición) se puede ver en la Figura 2.2.

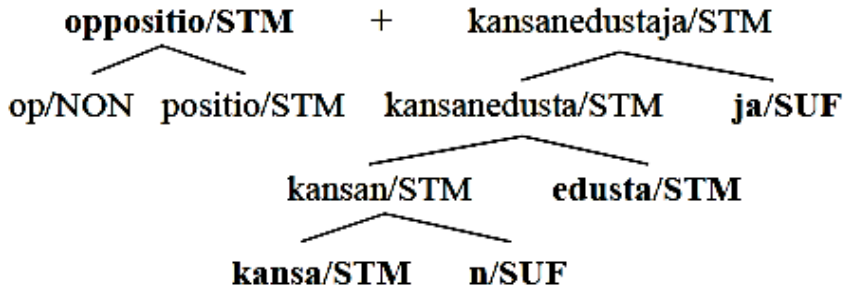


Figura 2.2 Ejemplo de segmentación del método *morfessor categories-MAP*
Tomado de Creutz y Lagus (2005, pág. 108)

Una diferencia interesante entre las primeras propuestas de estos autores (*morfessor baseline*) y las últimas (*morfessor categories*) es la incorporación de características lingüísticas. Las primeras trataron el problema del descubrimiento morfológico solamente como un proceso de encontrar un lexicón compacto de segmentos aleatorios, lo que no dio buenos resultados.

Después, en las últimas propuestas, a los segmentos aleatorios se les ve como parte de una morfotáctica, lo que mejora los resultados. También se formulan probabilísticamente algunas características lingüísticas de los morfos, como su tamaño, su contenido de información (semántica o gramatical) y su nivel combinatorio; aspectos que también benefician al método. Es interesante que los resultados finales reflejen la complejidad morfológica de una lengua aglutinante.

2.4. Optimización mediante algoritmos genéticos

En esta sección describo, a grandes rasgos, un método que asume el problema de segmentación morfológica como un problema de optimización para el cual utiliza un algoritmo genético.

Gelbukh, Alexandrov y Han (2004) proponen este método para determinar segmentos flexivos de lenguas sufijales como el español o el inglés. Se basan en la idea de obtener un modelo morfológico óptimo, cuya principal característica es la de contar el mínimo número de bases y sufijos capaces de describir el corpus de análisis (prefiriendo un conjunto menor de sufijos que de bases). Toman dos hipótesis de partida: (i) las palabras del corpus se forman de dos segmentos, una base y un sufijo, y no se toma en cuenta ningún otro tipo de fenómeno morfológico; (2) para el aprendizaje de una lengua se requiere del mínimo esfuerzo, por lo que se requiere del mínimo número de bases y sufijos.

Definen tres conjuntos, *V* para el conjunto de palabras, *S* para el conjunto de bases y *E* para el conjunto de segmentos finales. La idea es entonces buscar el mínimo tamaño de *S*+*E*, prefiriendo los casos donde *E* es menor.

Utilizan la estrategia de algoritmos genéticos[46] para encontrar el mejor grupo de bases y sufijos. En términos generales, forman cadenas binarias (cromosomas) que indican la ausencia (0) o presencia (1) de una base o sufijo. El tamaño del cromosoma es igual al tamaño de *S*+*E*. La función a minimizar en cada mutación se puede ver en (2.8), tomada de Gelbukh, Alexandrov y Han (2004, pág. 436).

[46] Como método computacional, los algoritmos genéticos simulan el proceso de evolución de los organismos vivos con la idea de resolver problemas de diversa índole. Una de sus características principales es la búsqueda de una solución a partir de la exploración de una enorme cantidad de posibles soluciones (Holland, 1992). Éstas se representan como cadenas de ceros y unos, así el objetivo se vuelve la búsqueda de una cadena particular. Estas cadenas pueden verse como cromosomas formados de secuencias de genes que representan individuos de una población. Para explorar diversas soluciones (crear nuevas generaciones de individuos), las cadenas son modificadas repetidamente mediante operaciones de mutación, cruza y reproducción. Una función matemática de aptitud o idoneidad (*fitness*) selecciona sólo algunas cadenas para producir la siguiente generación. El procedimiento se detiene cuando se ha llegado a un estado deseable, por ejemplo, que la función ya no se pueda minimizar más.

$$|S| + 0.000001|E| + |V \setminus (S + E)| \rightarrow \min \quad (2.8)$$

En la fórmula anterior, $|S|$ representa el número de bases, $|E|$ representa el número de segmentos finales y la expresión $|V \setminus (S + E)|$ representa el número de palabras de V que no fueron segmentadas dados los elementos de S y E (donde el símbolo $\setminus$ indica diferencia de conjunto). Para dar preferencia a un conjunto más pequeño de elementos de E, se utiliza el producto $0.000001|E|$. Los resultados de este método fueron prometedores.

Posteriormente se hicieron modificaciones a este método y se realizaron nuevos experimentos (Lara Reyes, 2008). Uno de ellos fue con palabras desprovistas de marcas flexivas, con el objeto de encontrar sufijos derivativos (Gelbukh et al., 2008). En este nuevo experimento, la principal modificación al método fue la búsqueda de paradigmas de sufijos.

En el primer paso del nuevo método se obtiene una lista de bases y sufijos. Además, se determinan todos los paradigmas posibles, esto es, grupos de sufijos que acompañan a varias bases. Luego, se eliminan de la lista inicial los sufijos que no sean parte de algún paradigma de al menos dos elementos o cuando pertenezcan a un paradigma de baja frecuencia. En cuanto al uso del algoritmo genético, este método incluyó básicamente los mismo pasos que el anterior.

En una comparación manual de cien tipos de palabras del español segmentadas con este método y con el método de Goldmith, se obtuvieron resultados muy similares: 84% de precisión para este método contra 87% de precisión de *Linguistica*.

Uno de los aspectos más interesantes de la propuesta presentada en esta sección es cómo se adoptó un procedimiento computacional que simula la evolución de los seres vivos al problema del análisis morfológico de una lengua. Desafortunadamente, ya que un algo-

ritmo genético explora una enorme cantidad de posibles soluciones, llegar a la solución final suele tomar bastante tiempo.

También se puede resaltar que el primer acercamiento de este método intentó descubrir la morfología sólo a partir del tamaño de los conjuntos de bases y sufijos. Lo anterior conlleva ver a la morfología sólo como un conjunto mínimamente redundante de morfemas. Esta idea no es el mejor, ya que la morfología de una lengua natural, además de la economía, también involucra aspectos combinatorios. Esto se reflejó en los últimos experimentos realizados por los autores, en los cuales se utilizaron combinaciones de bases y sufijos para validar segmentaciones.

2.5. Índice de afijalidad

Esta sección está dedicada al método que utilizaré para realizar mi investigación. Es un método no supervisado que obtiene una lista de bases y afijos (prefijos y sufijos) a partir de un corpus. Considero que es una buena opción para el desarrollo de mi propuesta ya que cuenta con las siguientes características: no es supervisado, está basado en conceptos y caracterizaciones lingüísticas, fue probado en español, y obtuvo buenos resultados incluso en otras lenguas no emparentadas al español.

Medina (2000; 2003) propone este método no supervisado de segmentación morfológica basado en la cuantificación de qué tan gramatical es una unidad lingüística. Una unidad es más gramatical si aporta más estructura que significado a la lengua. Al interior de la palabra, las unidades más gramaticales son los afijos, al exterior son principalmente las conjunciones, preposiciones, pronombres (especialmente los clíticos) y artículos.

Para medir que tan gramatical es un segmento al interior de la palabra, ese autor propone cuantificar su afijalidad. Para el exterior, propone calcular la cliticidad de las pala-

bras. Según su propuesta, ambas medidas pueden entenderse como caracterizadoras de una misma fuerza de adhesión entre unidades lingüísticas: la glutinosidad. Ya que mi investigación involucra sólo la segmentación de palabra en unidades más pequeñas, tomaré de su propuesta la parte que corresponde a la cuantificación de la afijalidad.

Para obtener una medida de afijalidad, Medina propone cuantificar tres características de un afijo. Así, se espera que los afijos de una lengua tengan mayor afijalidad que las bases. Ya que la glutinosidad indica qué tanto se adhieren dos segmentos, habría mucha glutinosidad entre una base y un sufijo. Al interior de las bases cabe esperar la máxima glutinosidad posible.

Las características cuantificadas de los afijos son: (i) no ocurren aislados, sino como parte de las palabras, (ii) ocurren en contextos similares y se combinan con bases de relativa baja frecuencia, y (iii) tienen contenido más gramatical. Para cuantificar la característica (i) se utiliza una medida de cuadros, para la (ii) se usa una medida de economía, y para la (iii) se calcula la entropía. En seguida describo cada una de estas medidas.

2.5.1. Medida de cuadros

Esta subsección explica el concepto de *cuadro* propuesto por Greenberg (1967) como primera medida de afijalidad. Un cuadro se puede entender como una estructura combinatoria en donde participan cuatro segmentos, dos iniciales (*cas~*, *sill~*) y dos finales (*~a*, *~ita*) que al combinarse forman cuatro palabras del corpus (*cas~a, cas~ita, sill~a , sill~ita*). En un cuadro es posible tener un segmento nulo (ø) como en *in~cauto, in~feliz, ø~cauto, ø~feliz*.

La medida de cuadros de un segmento se determina por el número de cuadros en el que participa. Para contarlos se toma en cuenta sólo aquellos donde el segmento fijo perte-

nece a un conjunto más pequeño de segmentos más frecuentes que el conjunto de segmentos alternantes (Medina, 2003, pág. 105).

Dada una segmentación $a_{i,j}$::$b_{i,j}$, donde *i* es el índice de la palabra examinada y *j* el índice de la posición en la palabra donde se hace la segmentación, entonces el número de cuadros del segmento *j* de una palabra *i* se denota por $c_{i,j}$.

2.5.2. Medida de entropía

La segunda medida de afijalidad involucrada es la medida de entropía. La entropía (Shannon y Weaver, 1964) es usada como una medida de la cantidad de información que contienen todos los segmentos con que se combina el afijo. La idea es que los segmentos más gramaticales (afijos) contienen menos cantidad de información que los segmentos de contenido (bases).

Por tanto, dado que una base contiene más información comparada con la de un afijo, es posible encontrar picos de entropía al interior de una palabra que indiquen fronteras entre bases y afijos. Esta propuesta sigue la intuición de Harris de contar la cantidad de fonemas anteriores y posteriores a un corte morfológico.

Así, dada una segmentación $a_{i,j}$::$b_{i,j}$, se calcula la entropía a la izquierda $H(i,j)^{izq}$ de una segmentación *j* de la palabra *i* de la siguiente manera, véase (2.9) tomada de Medina (2003, pág. 108), que es una adaptación de la fórmula de Shannon.

$$H(i,j)^{izq} = -\sum_{x=1}^{f(a_{i,j})} p(b_{x,j}|\, a_{i,j}) \times log_n(p(b_{x,j}|a_{i,j})) \qquad (2.9)$$

En la fórmula anterior, $p(b_{x,j}|a_{i,j})$ se refiere a la probabilidad asociada a cada segmento que alterna con $a_{i,j}$, esto es, la probabilidad de seleccionar ese segmento del conjunto

de posibles segmentos alternantes. Una descripción de esta probabilidad se puede ver en la fórmula de (2.10), tomada de Medina (2003, pág. 107), donde $B_{i,j}$ es el conjunto de los segmentos a la derecha de $a_{i,j}$ con posibilidad de ser seleccionados y | $B_{i,j}$ | el tamaño de este conjunto.

$$0 \leq p\left(b_{k,j}\middle|a_{i,j}\right) = \frac{f\left(b_{k,j}\right)}{f\left(a_{i,j}\right)} \leq 1, \qquad b_{k,j} \in B_{i,j}, \qquad k = 1, 2, 3, \ldots |B_{i,j}| \qquad (2.10)$$

En la fórmula anterior, $f\left(b_{k,j}\right)$ es la frecuencia de cada segmento del conjunto $B_{i,j}$ y $f\left(a_{i,j}\right)$ es la frecuencia asociada al segmento $a_{i,j}$. Para entender mejor el cálculo de esta probabilidad daré un ejemplo tomado del mismo autor. Para la segmentación *p::reviamente* se tuvieron asociados 7,206 tipos de palabras que comenzaban con *p*– y 2,184 tipos que comenzaban con *pr*–. Entonces,

$f\left(a_{i,j}\right) = 7206,$

$f\left(b_{k,j}\right) = 2184,$

$p(b_{k,j}|a_{i,j}) = 2184 \div 7206 = 0.303081.$

Luego la entropía asociada a la segmentación *p*::*r* es:

$H(i,j)^{izq} = -p \times \ln p = -0.303081 \times -1.193755 = 0.361804.$

Finalmente la entropía total de la segmentación $p::B_{i,1}$ se calcula a partir de todas las entropías de las segmentaciones que comienzan con *p*–. Según los experimentos de ese autor, los picos de entropía calculados de derecha a izquierda son mejores indicadores del final de una base y por tanto del inicio del sufijo.

2.5.3. Medida de economía

Explicaré en esta subsección la medida de economía. Para calcular esta medida se retoma el principio de economía de signos de De Kock y Bossaert, utilizado en su propuesta de segmentación morfológica: "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" (Medina, 2003, pág. 111).

Este principio coincide con el hecho de que los afijos tengan alta capacidad combinatoria, lo que permite caracterizarlos como unidades que aportan economía al sistema lingüístico. Así, para medir esta capacidad combinatoria de los afijos es necesario cuantificar la cantidad de signos con los que se combinan. Entonces, dada una segmentación $a_i::b_i$, si a_i pertenece a un conjunto potencialmente infinito de segmentos poco frecuentes, mientras que b_i pertenece a un conjunto pequeño de segmentos muy frecuentes, entonces a_i sería una base y b_i un afijo .

Las fórmulas de (2.11) permiten calcular los índices de economía de las segmentaciones para prefijos y sufijos. En esas expresiones, $A_{i,j}$ es el conjunto de segmentos alternantes a la izquierda del segmento $b_{i,j}$ y $B_{i,j}$ es el conjunto de alternantes de $a_{i,j}$, donde $a_{i,j} \in A_{i,j}$ y $b_{i,j} \in B_{i,j}$. Además, $|A_{i,j}|$ es el tamaño de $A_{i,j}$ y $|B_{i,j}|$ es el tamaño de $B_{i,j}$.

Hay algunas restricciones que deben cumplir los elementos de los conjuntos mencionados. La primera impide contar segmentos con menor frecuencia que la de su acompañante, la idea detrás es eliminar posibles bases de los dos lados, ya que las bases deben ser menos frecuentes que los afijos. Por tanto, $A^p_{i,j}$ sería el conjunto de supuestos prefijos y $B^s_{i,j}$ el de supuestos sufijos, donde $A^p_{i,j} \in A_{i,j}$ y $B^s_{i,j} \in B_{i,j}$, además $\left|A^p_{i,j}\right|$ es el número de elementos de $A^p_{i,j}$ y $\left|B^s_{i,j}\right|$ el número de elementos en $B^s_{i,j}$.

a) Prefijos: $k^{p}_{i,j} = \frac{\left|B_{i,j}\right| - |B^{s}_{i,j}|}{|A^{p}_{i,j}|}$

b) Sufijos: $k^{s}_{i,j} = \frac{\left|A_{i,j}\right| - |A^{p}_{i,j}|}{|B^{s}_{i,j}|}$ (2.11)

La segunda restricción elimina del conteo de segmentos alternantes aquellos que comparten el mismo fonema adyacente a la segmentación. Según la propuesta de De Kock y Bossaert, si en los supuestos afijos hay alguno que aparece con varias bases que coincidían en el mismo fonema (o letra) adyacente al afijo, se puede suponer que ese fonema es del supuesto afijo y no de la base. Igualmente, si en las supuestas bases hay una que aparecía con varios afijos que tengan el mismo fonema adyacente a la base, se puede suponer que el fonema pertenece a la base.

La tercera restricción requiere que los elementos de los conjuntos de supuestos afijos y supuestas bases participen en por lo menos un cuadro. Así, la economía de una segmentación se obtiene al comparar los tamaños de los conjuntos después de haberles aplicado las restricciones. Si a la izquierda de una segmentación hay mayor cantidad de supuestas bases ($\left|A_{i,j}\right| - \left|A^{p}_{i,j}\right|$) que de supuestos afijos a la derecha ($\left|B^{s}_{i,j}\right|$), entonces el segmento $b_{i,j}$ es una sufijo. También, si hay mayor cantidad de supuestas bases a la derecha ($\left|B_{i,j}\right| - \left|B^{s}_{i,j}\right|$) que de supuestos afijos de la izquierda ($\left|A^{p}_{i,j}\right|$), el segmento $a_{i,j}$ es un prefijo.

2.5.4. Combinación de medidas

Ya revisadas las tres medidas de afijalidad, en esta subsección expongo la forma en que son combinadas para obtener el índice de afijalidad que permite la segmentación morfológica de palabras. Medina pone a competir estas medidas de segmentación entre ellas, y con otras basadas en estadísticas de digramas, con la idea de conocer su eficiencia en la tarea de

segmentación. La evaluación en 836 tipos de palabras del CEMC propuso como mejores medidas la de cuadros, entropía y economía. De hecho, la combinación de economía y entropía resultó aún mejor (95% de aciertos).

Ese autor atribuye los resultados a que las medidas caracterizan verdaderas propiedades lingüísticas de los afijos. Esto permite ver a un afijo como una unidad que se ha desgastado fonológica y semánticamente, de tal manera que aparece adherido a otras unidades, es muy frecuente y participa en gran número de estructuras combinatorias. Además, contiene poca cantidad de información (en sentido técnico) y se adhiere a muchas unidades para darles estructura.

En consecuencia, se propuso un índice de afijalidad que se obtiene con el promedio normalizado de las tres medidas, véase la fórmula de (2.12) tomada de Medina (2003, pág. 130). Este índice fue calculado a partir de los tipos de palabras del CEMC sin modificaciones y modificando algunos caracteres para lograr una representación fonológica del mismo. Luego se segmentaron todos los tipos de palabra y los resultados fueron del 90.41% de tipos bien segmentados en una muestra de 836 tipos.

$$AF^{n}(s_x) = \frac{\frac{c_x}{max c_i} + \frac{h_x}{max h_i} + \frac{k_x}{max k_i}}{3} \qquad (2.12)$$

Con la idea de ejemplificar el resultado de cuantificar la afijalidad de las posibles segmentaciones de una palabra, en seguida pongo algunos ejemplos[47]. En la Tabla 2.1 se puede ver el ejemplo del cálculo de medidas de derecha a izquierda para la palabra *casa*,

[47] Estos ejemplos los obtuve con medidas calculadas de derecha a izquierda y con la representación fonológica propuesta por Medina (2003, pág. 358). Para distinguir esta representación de la representación ortográfica utilizaré para la primera las diagonales, véanse más detalles en la sección 5.2 y Tabla 5.1.

representada fonológicamente como /KASA/. Dado que es una representación fonológica, también incluye la forma verbal *caza.*

Tabla 2.1 Medidas de afijalidad para la palabra /KASA/

	K	A	S	A
Entropía		2.459	1.726	**2.628**
Cuadros		267	0	**14669**
Economía		**0.9963**	0	0.9466
Afijalidad		0.6512	0.2189	**0.9834**

Como se puede ver, la medida de entropía calculada es más alta al final del segmento /KAS~/, debido a que esta palabra alterna con otras similares como /KASITA/ o /KASOTAS/. La medida de cuadros también es más alta en este segmento, de hecho es muy alta, lo que refleja la capacidad combinatoria del segmento /~A/ en posición final. En cambio, la medida de economía es más alta en la segmentación /K~ASA/, aunque no rebasa por mucho a la segunda segmentación /KAS~A/. Finalmente, el índice de afijalidad que combina las tres medidas propone la segmentación esperada y separa la marca de flexión que puede ser nominal (*cas-a*) o verbal (*caz-a*).

Otro ejemplo es el de la Tabla 2.2. En él se puede observar cómo la medida de entropía es más alta en la segmentación /PAST~ELES/, que se explica por la relación de esta palabra con otras similares, como /PAST~A/ y /PAST~O/. Después del segmento /PAST~/, la variedad de signos es más alta y por tanto es más difícil predecir el siguiente. De hecho, esta segmentación coincide con la separación del sufijo derivativo –el concatenado con la marca de flexión de plural –es*:* /PAST~ELES/.

Tabla 2.2 Medidas de afijalidad para la palabra /PASTELES/

	P	A	S	T	E	L	E	S
Entropía	0	0	1.673	**2.22**	1.62	2.039	1.362	
Cuadros	0	0	0	72	0	**929**	160	
Economía	0	0	0	0	0	**0.9903**	0.9438	
Afijalidad	0	0	0.2512	0.3333	0.2432	**0.9729**	0.5795	

En este ejemplo, son las medidas de cuadros, de economía y el índice de afijalidad las que coinciden en el valor más alto, proponiendo la segmentación /PASTEL~ES/, que separa de manera pertinente la flexión de número[48].

Pongo un ejemplo final para mostrar que el índice de afijalidad también presenta problemas en algunas segmentaciones. En la Tabla 2.3 se pueden ver las medidas de afijalidad calculadas para la palabra /ENSEÑANSA/.

Tabla 2.3 Medidas de afijalidad de la palabra /ENSEÑANSA/

	E	N	S	E	Ñ	A	N	S	A
Entropía	0	0	0	0	2.44	1.01	1.726	**2.628**	
Cuadros	0	0	0	0	68	26	**243**	0	
Economía	0	0	0	0	0.2794	0	**0.9835**	0	
Afijalidad	0	0	0	0	0.4974	0.1281	**0.8855**	0.3333	

Esta vez, ninguno de los valores más altos de las tres medidas, ni el del índice de afijalidad, proponen la segmentación en el sufijo derivativo esperado: –anza. En su lugar la medida de entropía propone un segmento /~A/ asociado a una marca flexiva inexistente en esta palabra. Las medidas de cuadros y economía proponen la segmentación /ENSEÑAN~SA/, que si bien no es la esperada, seguramente responde a la economía del segmento final /~SA/, que debe aparecer en muchas otras palabras, y a la aparición de la forma libre /ENSEÑAN/.

Resulta pertinente recordar que el resultado de aplicar este método al corpus de estudio fue un catálogo de afijos. Dentro de éste había tanto afijos individuales como afijos concatenados, es decir, sólo se segmentaba una sola vez cada tipo de palabra. Ya que éste

[48] Si se usara el índice de afijalidad para proponer más segmentaciones en la palabra, tomando sus valores más altos, el segundo valor más alto propondría la segmentación /PASTEL~E~S/. Esto es lógico ya que la economía asociada al segmento /~S/ es muy alta porque corresponde a una marca flexiva de plural, aunque no en esta palabra. Los problemas asociados a utilizar este índice para obtener varias segmentaciones los expondré con mayor detalle en la sección 4.2.3.

será el método que utilizaré en mi investigación, deberé hacer las modificaciones y experimentos pertinentes para determinar cómo obtener todos los sufijos posibles de cada tipo de palabra, esto se describe en el capítulo 4.

En cuanto a la aplicación de este método a otras lenguas, sólo comentaré que se ha utilizado para distintas lenguas no emparentadas y para determinar afijos (sufijos y prefijos) tanto flexivos como derivativos (Medina, 2007). Por ejemplo, se ha empleado para obtener un catálogo de prefijos derivativos en lengua checa (Medina y Hlaváčová, 2005); afijos de flexión verbal en chuj, lengua maya (Medina y Buenrostro, 2003); y sufijos derivativos en tarahumara, lengua de la familia yuto-azteca (Medina y Alvarado, 2006).

2.5.5. Aspectos computacionales

Este apartado describe brevemente la estructura de almacenamiento de los tipos de palabras del corpus, que sirve para medir la afijalidad entre segmentos. El programa que crea y utiliza esta estructura está desarrollado con el lenguaje de programación C++.

Para el cálculo de las medidas, los tipos de palabra del corpus son representados en dos estructuras arbóreas donde cada nodo corresponde a una letra. Una de las estructuras está organizada a partir de la primera letra y hasta la última, la otra está organizada en sentido contrario, de la última a la primera. El nodo raíz de la primera estructura lleva a los nodos que corresponden a la primera columna de letras de todos los tipos de palabras, y de éstos se puede ir a los nodos de la segunda columna, y así sucesivamente.

En cada nodo se almacenan distintas frecuencias, además de cada una de las medidas de afijalidad y el índice que combina las tres. Haciendo uso de estas estructuras arbóreas, se proponen las segmentaciones en los valores más altos del índice de afijalidad de los tipos de palabras del corpus. Además, un procedimiento del programa permite incorporar

nuevas palabras a las estructuras arbóreas, calcular sus índices y segmentarlas. Finalmente, otro procedimiento obtiene el catálogo de afijos del corpus.

2.6. Observaciones sobre los métodos de segmentación

Expongo en esta sección mis observaciones sobre los métodos de segmentación que describí en las secciones anteriores. Comparo sus características generales y las perspectivas con las que abordan el problema de segmentación.

En cuanto a la perspectiva para resolver el problema, observo una clara preferencia por intentar construir lo que computacionalmente se llama *modelo morfológico*[49]. Luego, mediante diversas estrategias se busca "el mejor modelo" que describa el corpus. *Linguistica, Morfessor* y el método de algoritmos genéticos trabajan de esta manera. Partir de la idea de crear un modelo de este tipo y buscar optimizarlo conlleva la suposición de que existe una morfología única, óptima o ideal[50].

Por otra parte, el método del índice de afijalidad no propone de inicio un modelo morfológico que va mejorando, sino que descubre las unidades morfológicas mediante la cuantificación de características lingüísticas.

[49] Aquí modelo morfológico se refiere el conjunto de unidades que se descubren y no al método utilizado para descubrirlas. Como dice Goldsmith "Thus morphological models offer a level of segmentation that is typically larger than the individual letter and smaller than the word" (2010, pág. 7).

[50] Si bien éste no es el lugar para desarrollar una discusión sobre la construcción de la gramática en la mente de los hablantes, sí quiero dejar sentada esta pregunta lingüística ¿existe una morfología ideal, un modelo único y optimizado, o será más bien que cada hablante construye su propia morfología? No son pocos los cuestionamientos implicados en esta pregunta, por lo que evito entrar en mayor detalle al respecto.

En términos generales, los tres métodos que proponen y buscan un modelo morfológico ideal tratan de que sea el más compacto, pero que al mismo tiempo contenga segmentos pertinentes. *Linguistica* y *Morfessor* lo hacen a través de métodos probabilísticos. El método del algoritmo genético no usa probabilidades sino una función que debe ser minimizada.

Esta idea del modelo más compacto coincide con el comportamiento económico del sistema lingüístico y en especial con la economía que se da en la morfología. El método del índice de afijalidad también contempla esta característica de la morfología al calcular la medida de economía de cada segmento.

Es interesante que el método de algoritmos genéticos requiera el mínimo de información lingüística y el mínimo de elaboración de su modelo; sin embargo, es el que prueba la mayor cantidad de posibles segmentaciones en las que muchas pueden ser poco pertinentes, porque así funcionan los algoritmos genéticos. Inclusive, en la versión que utiliza estructuras combinatorias para filtrar segmentos, la cantidad de posibles segmentaciones que prueba el algoritmo es bastante grande.

En este sentido, *Linguistica* utiliza muchas heurísticas y estructuras combinatorias (*signatures*) para disminuir la cantidad de segmentaciones posibles a evaluar. Por su parte, *Morfessor* (*Morfessor-MAP*) usa un elaborado conjunto de probabilidades para guiar las segmentaciones e incorpora una morfotáctica que le permite mejorar sus resultados, especialmente para el finlandés. Esto es lógico, ya que el inglés necesita considerar poco o nada una morfotáctica debido a su morfología flexiva simple.

El método del índice de afijalidad no elabora un modelo probabilístico o función de optimización para evaluar las segmentaciones. Si bien utiliza algunas heurísticas para condicionar ciertos cálculos, básicamente obtiene medidas para todo corte dentro de la palabra.

Este método también utiliza estructuras combinatorias que son los llamados cuadros, aunque en la salida del método no se presenten como lo hace *Linguistica*.

A pesar del complejo entramado de probabilidades condicionales que incluye el método de *Morfessor*, es interesante la manera en como incorpora conceptos lingüísticos. Este método modela probabilísticamente características formales y "semánticas" de los segmentos. También llama la atención cómo elabora probabilidades para la morfotáctica de las palabras, tanto para la secuencia de categorías, como para la secuencia de segmentos.

Todos los métodos dejan clara la importancia de utilizar estructuras combinatorias para obtener segmentos morfológicos más pertinentes. Además, el uso de la morfotáctica de la palabra en el caso de *Morfessor* brindó la posibilidad de describir una jerarquía de morfos de las palabras. Lo anterior apunta a que utilizar características lingüísticas mejora el resultado de los métodos morfológicos automáticos no supervisados. Piénsese en los experimentos dedicados sólo a minimizar la redundancia de la lista de morfos, sus resultados no fueron los mejores porque el descubrimiento de la morfología no es solamente un proceso de compresión de datos.

Sobre el resultado que ofrecen estos métodos, sin contar las últimas propuestas de Creuz y Lagus, los demás métodos sólo separan la base de un prefijo y/o de un sufijo, pero no dan cuenta del encadenamiento de afijos. Esto es, sólo se separa la palabra en dos elementos, base y un sufijo o base y un prefijo. El interés de mi investigación es dar un paso hacia adelante y descubrir la morfotáctica, esto es, la secuencia de afijos[51].

[51] A futuro sería muy interesante buscar un método como el Creutz y Lagus utilizable en español para generar una jerarquía de bases y afijos. Esto ayudaría al estudio de la composición. Por ahora, dado que sólo contemplo a los sufijos, considero que es posible prescindir de esta jerarquía.

Por tanto, considero que para la perspectiva de mi trabajo, que no presupone una morfología ideal, el método que calcula un índice de afijalidad es la opción más pertinente. Este método parte de conceptos y caracterizaciones lingüísticas. Además, toma en cuenta estructuras combinatorias. Fue probado para español y otras lenguas con buenos resultados. Otra característica es que se basa en la cuantificación de la adhesión entre unidades lingüísticas (glutinosidad), que fundamenta parte de la propuesta para la determinación de esquemas morfotácticos de Lara (véase la sección 1.4).

Finalmente, hablando del método que propondré para la descripción morfotáctica del español, resta analizar la manera de representar la secuencialidad de sufijos. Como se mencionó, ha sido tradicional en la morfología computacional el uso de gramáticas formales en forma de autómatas de estados finitos. Por eso, el siguiente capítulo estará dedicado a revisar estos formalismos computacionales.

3. Gramáticas formales y autómatas de estados finitos

En la introducción de mi trabajo de investigación mencionaba que parte de la morfología concatenativa de las lenguas naturales puede verse como un lenguaje regular. Esto conlleva que dicho lenguaje deba describirse mediante una gramática del mismo tipo. Por esta razón, en este capítulo describiré diversos aspectos relacionados con las gramáticas formales.

Además, en el capítulo 1 consigné la necesidad de contar con una manera de representar la secuencialidad de morfemas una vez que éstos han sido identificados en las palabras de una lengua. Luego, esta descripción permitiría conocer los patrones morfotácticos de la lengua de estudio. Ya que los mecanismos estándar en morfología computacional para representar la morfotáctica han sido los autómatas de estados finitos, en este capítulo también presento los fundamentos para conocer estos mecanismos.

Creo conveniente presentar ambos formalismos en el mismo capítulo ya que la teoría de las gramáticas formales y la teoría de autómatas tienen muchos aspectos en común.

3.1. Conceptos básicos

En esta sección presento algunos conceptos básicos que me permitirán exponer de manera más clara los conceptos de gramáticas formales y autómatas. Las definiciones que expondré se enmarcan en las áreas de la lingüística matemática y computacional, por lo que no coincidirán necesariamente con las definiciones que provengan de otras perspectivas lingüísticas.

El primer concepto es el de *alfabeto*. Desde la lingüística matemática, éste se puede definir como un conjunto de símbolos, finito y no vacío, que se representa con la letra griega $\sum$. En (3.1) se pueden ver dos ejemplos de alfabetos. El primero (3.1a) es un alfabeto binario que incluye sólo dos elementos: 0 y 1. El segundo (3.1b) es un alfabeto que corresponde al conjunto de todas las letras minúsculas.

(3.1) a) Un alfabeto binario: $\sum = \{0, 1\}$.
b) El conjunto de todas la letras minúsculas: $\sum = \{a, b, ..., z\}$.

Dado el concepto de alfabeto, es posible definir el concepto de *cadena* como una secuencia finita de símbolos seleccionados de un alfabeto. Por ejemplo, la secuencia 01101 es una cadena obtenida del alfabeto binario presentado en (3.1a). Un tipo de cadena especial es la cadena vacía, que contiene cero símbolos y es representada comúnmente con los símbolos *e* o ϵ.

El conjunto de todas las cadenas obtenidas de un alfabeto se representa como $\sum^*$. Por ejemplo, en (3.2) se da una expresión que representa todas las cadenas posibles del alfabeto binario ejemplificado arriba.

(3.2) $\{0, 1\}^* = \{e, 0, 1, 00, 01, 10, 11, 000, ...\}$

Una operación común entre cadenas es la concatenación, que permite obtener una cadena formada por la yuxtaposición de las cadenas originales. Por ejemplo, si se tiene la cadena x = 01101 y la cadena y = 110, su concatenación sería xy = 01101110.

El tercer concepto importante es el de *lenguaje formal*. Se puede definir como un conjunto de cadenas que pertenecen al conjunto de todas las cadenas posibles generadas de un alfabeto (lo que arriba se definió como $\sum^*$). El concepto de lenguaje formal es importante para la lingüística porque, como dicen Hopcroft, Motwani y Ullman (2001, pág. 30), de

una manera simple se puede ver a toda lengua natural como un conjunto de cadenas. Un lenguaje formal puede formarse de un conjunto infinito de cadenas, pero siempre estará restringido a un conjunto finito de símbolos de un alfabeto.

Con estos tres conceptos es posible exponer ahora dos formalismos muy importantes para la lingüística computacional, en primer lugar las gramáticas formales y en segundo los autómatas de estados finitos.

3.2. Gramáticas formales

Si se adopta la perspectiva matemática de considerar un lenguaje como un conjunto de cadenas construidas a partir de un alfabeto finito de símbolos, se debe tener una gramática que describa de manera precisa dicho conjunto de cadenas. Este tipo de gramática se conoce como gramática formal y en esta sección expondré sus antecedentes matemáticos, su definición y sus características principales.

3.2.1. Antecedentes

Esta subsección dará cuenta de los antecedentes matemáticos que dieron fundamento al concepto de gramática formal, impulsado por Chomsky durante los años 50.

La teoría de las gramáticas formales está basada en los llamados sistemas *semi-Thue* (Wall, 1972, pág. 207). Por tanto, con el fin de conocer los fundamentos de estos sistemas, revisaré en seguida algunos sistemas que sirven de antecedente.

El primero es el *sistema axiomático*, definido como un conjunto de tres elementos (*A, S, P*) donde:

1. *A* es un conjunto finito de símbolos llamado alfabeto.
2. *S* es un conjunto de cadenas formadas a partir de *A* llamadas *axiomas*.

3. *P* es un conjunto de relaciones que se establece entre todas las cadenas posibles de *A* (*A**). Estas relaciones, llamadas *producciones* o *reglas*, están formadas por dos elementos, como se verá en seguida.

Las reglas de un sistema axiomático pueden describirse como dos secuencias de cadenas (x_1, x_2,…, x_{n-1}, x_n), donde se dice que x_n se deduce de (x_1, x_2,…, x_{n-1}). Una notación alternativa es x_1, x_2,…, $x_{n-1} \rightarrow x_n$. Una secuencia ordenada de cadenas, como y_1, y_2,…, y_m, se llama *derivación* de y_m si y sólo si cada cadena en la secuencia es un axioma o se deriva de una regla de *P*. Además, si una cadena es obtenida a partir de una derivación, se dice que es un *teorema* del sistema.

Un ejemplo de un sistema axiomático tomado de Wall (1972, pág. 198) mostrará los conceptos anteriores, véase (3.3).

(3.3) $A = \{a, b\}$
$S = \{aa, bb\}$
$P = \{(x, y)\ \epsilon\ A^* \times A^* | \ y = axa \text{ v } y = bxb\}$

La regla de *P* se explicaría como la inserción del primer elemento de la regla (*x*) en el segundo elemento de la misma (*y*), lo que obliga a que se haga ya sea entre dos *a* (*axa*) o entre dos *b* (*bxb*), de acuerdo con el segundo elemento de la regla. Este sistema axiomático produciría el conjunto infinito de pares ordenados: {(*e, aa*), (*e, bb*), (*a, aaa*), (*a, bab*), (*b, aba*), (*b, bbb*), (*aa, aaaa*),...}, con una notación alternativa: {$e \rightarrow aa$, $e \rightarrow bb$, $a \rightarrow aaa$, $a \rightarrow bab$, $b \rightarrow aba$, $b \rightarrow bbb$, $aa \rightarrow aaaa$,...}. Estos pares son también el conjunto de reglas posibles del sistema.

Decía que este tipo de sistemas puede producir derivaciones y teoremas. En seguida ejemplifico esto. Una derivación del sistema expuesto en (3.3) sería la secuencia *bb, abba, aabbaa*, ya que la cadena *aabbaa* es deducida de la regla $abba \rightarrow aabbaa$, la cadena *abba*

es deducida de la regla *bb* → *abba*, y la cadena inicial *bb* es un axioma. Además, ya que *aabbaa* fue obtenida a partir de una derivación, se puede decir que sería un teorema del sistema.

Por el contrario, la secuencia *bb*, *baab* no es una derivación, ya que la cadena *baab* no proviene de ninguna regla en *P*. Tampoco la secuencia *ab*, *aaba*, *baabab* es una derivación, debido a que la cadena inicial *ab* no es un axioma.

A partir de estos sistemas surgen los sistemas axiomáticos extendidos, con la diferencia de que existen dos tipos de símbolos en el alfabeto. De esta manera se tendrían dos alfabetos, el básico y el auxiliar, los cuales no comparten ningún símbolo. Otra diferencia con los sistemas axiomáticos es que en los sistemas extendidos los teoremas sólo contienen símbolos del alfabeto básico.

Formalizando, se puede definir un sistema axiomático extendido como un conjunto de cuatro elementos (*A, B, S, P*) donde:

1. *A* es un conjunto finito de símbolos llamado alfabeto auxiliar.
2. *B* es un conjunto finito de símbolos llamada alfabeto básico, que no comparte símbolos con *A* (ni *A* con *B*).
3. *S* es el conjunto de todas las cadenas posibles formadas por la unión de los dos alfabetos, llamadas axiomas. Éstos podrían ser expresados por un conjunto finito de axiomas esquemáticos.
4. *P* es un conjunto de relaciones que se establece entre todas las cadenas posibles de la unión de *A* y *B*, (*A* U *B*)*. Estas relaciones, llamadas producciones o reglas, podrían ser expresadas por un conjunto finito de reglas esquemáticas.

En este tipo de sistemas, no todas las derivaciones terminan en un teorema. Sólo son teoremas aquellas cadenas finales formadas por elementos de *B*. Cuando la derivación termina en un teorema se llama *prueba*.

Se toma nuevamente un ejemplo de Wall (1972, pág. 200) para ejemplificar un sistema axiomático extendido, véase (3.4).

(3.4) $A = \{M\}$
$B = \{a, b\}$
$S = \{M\}$
$P = \{ \alpha M\beta \rightarrow \alpha aMa\beta$
$\alpha M\beta \rightarrow \alpha bMb\beta$
$\alpha M\beta \rightarrow \alpha aa\beta$
$\alpha M\beta \rightarrow \alpha bb\beta\}$

Dado el sistema axiomático extendido anterior, donde α y β son cualquier cadena formada por la unión de símbolos de *A* y *B*, (*A* U *B*)*, se pueden revisar algunas secuencias de cadena para ejemplificar una derivación, un teorema y una prueba.

La secuencia de cadenas *M, aMa, aaMaa, aabMbaa* puede considerarse una derivación, ya que todas las cadenas se producen por una regla del sistema o son un axioma; sin embargo, no se puede considerar una prueba, ya que la cadena final *aabMbaa* contiene un símbolo del alfabeto auxiliar (*M*). Por otro lado, la secuencia de cadenas *M, aMa, aaMaa, aabbaa,* además de ser una derivación, sí termina en un teorema (cadena formada sólo por elementos del alfabeto básico, *B*) y por tanto es una prueba.

Ya definidos los sistemas axiomáticos extendidos, puedo explicar los sistemas *semi-Thue*. Se llamaron así porque fueron estudiados por primera vez por Axel Thue. Éstos son

sistemas axiomáticos extendidos, es decir, formados por los mismos cuatro elementos (*A, B, S, P*), donde las reglas son binarias y de la forma[52]:

$$\alpha x \beta \rightarrow \alpha y \beta$$

En estas reglas *x* e *y* son cadenas fijas formadas a partir de la unión de los dos alfabetos (A U B)*, mientras que α y β son cadenas variables tomadas de la misma unión (A U B)*. Esta diferencia con los sistemas axiomáticos extendidos se traduce en que las reglas están restringidas, esto es, que reemplazan una cadena fija por otra cadena fija.

Otra diferencia es que los sistemas axiomáticos anteriores tenían la posibilidad de contar con reglas del tipo $x_1, x_2, \ldots, x_{n-1} \rightarrow x_n$. En cambio, los sistemas *semi-Thue* sólo tienen una cadena a la derecha y una a la izquierda. Esta situación modifica el concepto de derivación. En un sistema *semi-Thue*, una derivación es una secuencia ordenada de cadenas $y_1, y_2, \ldots, y_m$ donde obligatoriamente y_m es un axioma y cada cadena, excepto y_1, es deducida de la cadena inmediata anterior por una regla de *P*. Véase el siguiente ejemplo (3.5).

(3.5) $A = \{C, D, E, F, G, H\}$
$B = \{a\}$
$S = \{HFGa\}$
$P = \{ FG \rightarrow DGaa$
$FD \rightarrow DF$
$HD \rightarrow HC$
$CD \rightarrow FC$
$CG \rightarrow FFGa$
$HF \rightarrow E$
$EF \rightarrow E$
$EG \rightarrow E$
$Ea \rightarrow a\}$

[52] La diferencia entre un sistema *Thue* y un *semi-Thue* es que el primero incluye el esquema inverso $\alpha y \beta \rightarrow \alpha x \beta$.

Dado el sistema *semi-Thue* anterior, tomado de Wall (1972, págs. 203-204), una derivación de *a* sería la que muestro en (3.6). A un lado de cada cadena se muestra la regla que la produjo. Véase como la derivación parte de un axioma y termina en un símbolo del alfabeto básico.

(3.6) *HFGa (axioma)*
EGa (por HF → E)
Ea (por EG → E)
a (por Ea → a)

3.2.2. Definición

Una vez expuestos los antecedentes de la sección anterior, en esta sección expondré la definición de gramática formal y algunas de sus características.

La teoría de gramáticas formales parte de los sistemas *semi-Thue* con algunas modificaciones (Wall, 1972, pág. 207). El alfabeto básico es llamado vocabulario terminal (V_T) y el alfabeto auxiliar es llamado vocabulario no terminal (V_N). Estos dos vocabularios no comparten elementos entre sí y a la unión de ambos se le llama vocabulario (V).

Las producciones de la gramática son llamadas reglas gramaticales y consisten en un conjunto finito de reglas esquemáticas de la forma abreviada $\varphi \rightarrow \psi$, las cuales se leen como ‘φ es reescrito como ψ‘. Otra diferencia es que una gramática cuenta con un sólo axioma, que es el símbolo *S* de *sentence*. Así, se pude definir formalmente una gramática *G* como un conjunto de cuatro elementos (V_N, V_T, P, S) (Hopcroft y Ullman, 1969, pág. 11).

Chomsky propuso el uso de algunas convenciones para representar los elementos de una gramática formal. En términos generales, los símbolos son representados con las prime-

ras letras del alfabeto y las cadenas de símbolos con las letras finales. La Tabla 3.1 muestra estas convenciones en forma detallada.

Tabla 3.1: Convenciones para elementos de una gramática formal

	Símbolos individuales (primeras letras del alfabeto)	Cadenas (últimas letras del alfabeto)
No terminal	*A, B, C,...*	*..., X, Y, Z*
Terminal	*a, b, c,...*	*..., x, y, z*
No especificado	*α, β, γ,...*	*..., χ, ψ, ω*

En una gramática formal, una derivación sucede cuando todas las cadenas son obtenidas de alguna regla, sin ser requisito que la primera cadena sea un axioma. Dado lo anterior, se dice que la primera cadena domina a la última. Una derivación puede ser una *derivación terminada* si y solo si la última cadena no puede ser reescrita por ninguna regla de la gramática. Este concepto permite definir una *cadena terminal* (*x*), ésta es generada por una derivación terminal que comienza con el axioma *S* y debe estar formada por símbolos terminales ($x \in V_T^*$).

El término de cadena terminal es la base para definir el lenguaje asociado a una gramática *L*(*G*). Éste es el conjunto de cadenas terminales generadas por la gramática. De manera formal se puede establecer la siguiente expresión (Hopcroft y Ullman, 1969, pág. 11):

$$L(G) = \{w | w \text{ está en } V_T^* \text{ y } S \overset{*}{\Rightarrow}_G w\} \tag{3.7}$$

Si dos gramáticas G_1 y G_2generan el mismo lenguaje, se puede decir que son equivalentes $L(G_1) = L(G_2)$. Tomaré de Wall (1972, pág. 209) la siguiente gramática a manera de ejemplo para mostrar varios de los conceptos anteriores, véase (3.8).

(3.8) $G = (V_N , V_T , \{S\}, P)$

$V_N = \{S, A, B, C\}$

$V_T = \{a, b, c\}$

$P = \{1.\ S \rightarrow ABC$

$2.\ A \rightarrow aA$

$3.\ A \rightarrow a$

$4.\ B \rightarrow Bb$

$5.\ B \rightarrow b$

$6.\ BC \rightarrow Bcc$

$7.\ ab \rightarrow ba\}$

Esta gramática cuenta con un solo axioma y dos conjuntos distintos de símbolos V_N y V_T, además, todas las reglas son producciones de un sistema *semi-Thue*. Entonces, la secuencia *BCA, BccA, BbccA, BbccaA* puede considerarse una derivación de *G*, aunque no comience con el axioma; sin embargo, no es una derivación terminada ya que la última cadena, *BbccaA*, podría ser reescrita.

Un ejemplo de derivación terminada sería la secuencia de cadenas *BCA, BccA, BbccA, BbccaA, Bbccaa, bbccaa*, ya que la última cadena incluye sólo símbolos terminales y no puede ser reescrita. La cadena *bbccaa*; sin embargo, no es una cadena terminal porque la derivación no comienza con el axioma.

La secuencia de cadenas *S, ABC, aBC, aBcc, abcc, bacc* también muestra una derivación terminada. Además, en este caso la cadena final *bacc* sí es una cadena terminal debido a que está compuesta únicamente por símbolos terminales, no puede ser reescrita y la derivación inició con el axioma *S*.

3.2.3. Tipos de gramáticas y lenguajes

Ya establecida la definición de gramática formal, revisaré en esta subsección la clasificación de gramáticas formales y la relación que existe entre ellas y los lenguajes que describen. Se han estudiado varios tipos de gramáticas formales de acuerdo con el nivel de restricción de sus reglas. Los principales tipos de gramáticas son (Hopcroft y Ullman, 1969, págs. 13-15; Wall, 1972, págs. 211-212):

Gramáticas tipo 0. Llamadas sistemas de reescritura no restringida ya que no tienen ninguna restricción en sus reglas; son básicamente los sistemas *semi-Thue*.

Gramáticas tipo 1. Llamadas gramáticas sensibles al contexto. Cada regla es de la forma $\varphi A\psi \rightarrow \varphi\omega\psi$, donde φ y ψ podrían ser nulas. La cadena que no puede ser nula es ω, por lo que en cada regla un solo símbolo no terminal (A) es reescrito como una cadena no nula. Se llama gramática sensible al contexto porque sus reglas indican que A se reescribe como ω cuando aparece con la cadena φ a su izquierda y la cadena ψ a su derecha. En lingüística, este tipo de reglas se expresan como $A \rightarrow \omega/\varphi\_\_\_\psi$.

Gramáticas tipo 2. Llamadas gramáticas libres de contexto. Cada regla es de la forma $A \rightarrow \omega$, donde ω no puede ser nula ($\omega \neq e$). De esta manera, en este tipo de gramáticas las cadenas φ y ψ son nulas. Por tanto, el símbolo A puede ser reescrito como ω sin importar el contexto en el que aparece; de allí el nombre del tipo de gramática.

Gramáticas tipo 3. Llamadas gramáticas regulares o de estados finitos. Cada regla es de la forma $A \rightarrow xB$ o de la forma $A \rightarrow x$, donde x no puede ser nula ($x \neq e$). Este tipo de gramática agrega la restricción de que el lado derecho de cada regla debe ser una cadena formada de símbolos terminales seguida de, a lo mucho, un símbolo no terminal.

Ya que las gramáticas de tipo 3 son las pertinentes para mi trabajo, pongo en seguida un ejemplo tomado de Wall (1972, pág. 212) para dar cuenta de la forma de las reglas.

(3.9) $G = (V_N , V_T , \{S\}, P)$

$V_N = \{S, A, B, C\}$

$V_T = \{a, b, c\}$

$P = \{1.\ S \rightarrow bA$

$2.\ A \rightarrow bA$

$3.\ A \rightarrow aB$

$4.\ B \rightarrow ab$

$5.\ B \rightarrow cC$

$6.\ C \rightarrow c\}$

Se puede observar en (3.9) que a la izquierda de cada regla sólo hay un símbolo no terminal, excepto en la regla 1 que tiene al axioma ($S \rightarrow bA$). Además, a la derecha de las reglas sólo aparece un símbolo terminal seguido de, a lo mucho, uno no terminal ($B \rightarrow cC$).

Los lenguajes producidos por cada tipo de gramática reciben su nombre de acuerdo con el tipo que los genera. Entre clasificar una gramática o un lenguaje, es más fácil clasificar la gramática, ya que basta observar las formas de las reglas y detectar las restricciones que imponen (Wall, 1972, pág. 213). Así, existen lenguajes:

Tipo 0, llamados conjuntos recursivamente enumerables.

Tipo 1, llamados lenguajes sensibles al contexto.

Tipo 2, llamados lenguajes libres de contexto.

Tipo 3, llamados lenguajes regulares, conjuntos regulares o lenguajes de estados finitos.

Cuando propongo que es posible describir la morfotáctica del español, en especial la sufijal, mediante una gramática de estados finitos, estoy caracterizando esta morfología

como un lenguaje regular. Esto no es nuevo, ya que tradicionalmente la morfología computacional ha visto la morfología concatenativa de esta manera.

De hecho, se propuso la llamada morfotáctica de estados finitos, que abordaré más adelante. Lo que ahora me importa resaltar es que las gramáticas de tipo 3 son equivalentes a los llamados autómatas de estados finitos. El lenguaje que genera una gramática de este tipo es exactamente el conjunto de cadenas que acepta un autómata (Hopcroft y Ullman, 1969, pág. 15). Por tanto, es importante que revise en la siguiente sección el concepto de autómata de estados finitos.

3.3. Autómatas de estados finitos

La teoría de las gramáticas formales, revisada en el apartado anterior, y la teoría de autómatas tienen correspondencias en muchos de sus aspectos más importantes. Por esta razón y en especial por la equivalencia entre una gramática tipo 3 y un autómata de estados finitos, revisaré en esta sección los aspectos más importantes a propósito de estos "dispositivos abstractos" llamados autómatas.

3.3.1. Definición

En esta subsección presento la definición de autómata, un concepto que se volverá de suma importancia en mi investigación ya que podría servir para describir la morfotáctica inferida del corpus. En términos muy generales, un autómata es un dispositivo o máquina abstracta que recibe una entrada con la cual realiza algunas operaciones de acuerdo con un conjunto de instrucciones predefinidas (Wall, 1972, pág. 254).

Un autómata puede verse como un sistema que siempre está en un estado, el cual le permite recordar una parte de su historia. La historia completa del sistema está definida por

un conjunto finito de estados, por lo que se recuerda lo importante y se olvida lo demás. La ventaja de tener un número finito de estados es que el sistema se puede implementar con un número definido de recursos (Hopcroft, Motwani, y Ullman, 2001, págs. 2-3). El estudio de la teoría de autómatas comenzó en los años 50 y, desde la perspectiva computacional, es relevante porque permite la creación de modelos para dispositivos de hardware y software.

En particular, los autómatas de estados finitos pueden funcionar de dos maneras, como aceptadores o como generadores. Los autómatas aceptadores reciben como entrada una cadena de símbolos de un lenguaje, ejecutan un número finito de pasos y se detienen en un estado que permite saber si la cadena fue aceptada o rechazada (Wall, 1972; Jurafsky y Martin, 2009).

Es común asociar el funcionamiento de un autómata de estados finitos aceptador como la operación de una lectora que recibe una cinta de entrada. Esta cinta estaría dividida en cuadros que contendrían cada uno de los símbolos de una cadena de entrada, escritos de izquierda a derecha. Los cuadros que no tuvieran símbolo de entrada serían marcados con un símbolo nulo, por ejemplo #.

El autómata (la lectora) comenzaría a leer los símbolos de la cinta (cadena de entrada). Si la lectura se detiene antes de llegar al final de la cinta, se dice que el autómata está bloqueado y la cinta de entrada es rechazada. Por otro lado, si el autómata lee todos los símbolos de la cinta hasta detenerse en un símbolo nulo (fin de los símbolos de la cinta), se revisa el estado del autómata. Si el autómata se quedó en un estado marcado como estado final, o estado de aceptación, entonces la cinta de entrada es aceptada, de otra manera la cinta es rechazada.

Formalmente, un autómata de estados finitos se puede definir como un conjunto de cinco elementos (Jurafsky y Martin, 2009, pág. 28):

(3.10) $Q = \{q_0, q_1, q_2 \ldots q_{N-1}\}$ Un conjunto finito de N estados. Se puede representar como K

$\sum$ Un conjunto finito de símbolos de un alfabeto de entrada

q_0 Un estado inicial

F Un conjunto de estados finales, $F \subseteq Q$

$\delta(q, i)$ Funciones de transición que, dado un estado $q \in Q$ y un símbolo de entrada $i \in \sum$, regresan un nuevo estado $q' \in Q$

El procesamiento de una cadena de entrada comenzaría en el estado inicial del autómata (q_0). Este procesamiento sería dirigido por el conjunto finito de transiciones, que podrían verse también como tripletas de la forma (a_i, q_i, q_k), donde q_i y q_k son estados y a_i es un símbolo de entrada. Si el autómata está en q_i y lee el símbolo a_i de la cadena entrada, el autómata cambia al estado q_k (la cabeza de la lectora se mueve un cuadro a la derecha).

Como se mencionó, un autómata de estados finitos también puede verse como un dispositivo de generación de cadenas de un lenguaje, autómata generador. En este caso, y pensando en el dispositivo de la lectora de cinta, se pensaría en una cinta donde se imprimirían los símbolos de una cadena cada vez que la lectora avanza un cuadro a la derecha según determinadas instrucciones. Pensando en el autómata como conjunto de estados, cada tripleta (a_i, q_j, q_k) indicaría a qué estado cambiar y qué símbolo generar.

3.3.2. Tipos

Una vez definido un autómata de estados finitos y explicado su funcionamiento, en esta subsección presento los dos tipos principales de autómatas. Éstos pueden clasificarse en *determinísticos*, si en cada instrucción el autómata pasa a un solo estado, y *no determinísticos*, si pasa a un conjunto de estados. Dicho en otras palabras, un autómata determinístico

sólo está en un estado a la vez, mientras que el no determinístico puede estar en varios estados al mismo tiempo.

Para todo autómata no determinístico existe su equivalente determinístico, que acepta el mismo conjunto de cadenas, es decir, el mismo lenguaje. Por lo general, el autómata no determinístico es más fácil de diseñar y su autómata determinístico equivalente tendría aproximadamente el mismo número de estado o, en el peor de los casos, tendría 2^n estados. En Hopcroft, Motwani y Ullman (2001, págs. 61-64) se encuentra una discusión de cómo hacer equivalentes dos autómatas de distinto tipo.

Un autómata determinístico puede definirse formalmente como el siguiente conjunto de elementos (K, $\sum$, δ, $q0$, F), donde:

K es un conjunto finito, no vacío, de estados. También se representa como Q.

$\sum$ es un conjunto finito, no vacío, de símbolos, llamado alfabeto de entrada.

δ es una función de transición que recibe dos argumentos: una cadena y un estado, y regresa un estado.

$q0$ es un elemento de K que representa el estado inicial.

F es un subconjunto de K, estos estados representan el conjunto de estados finales o estados de aceptación.

Un autómata de estados finitos tiene asociado un lenguaje, que consiste en el conjunto de todas las cadenas que el autómata acepta. Se comienza con un estado y un símbolo, luego se revisa si existe una transición que involucre ambos. En caso de existir, la transición regresa el siguiente estado. Con este nuevo estado y el siguiente símbolo, se busca nuevamente una transición, que regresará un siguiente estado. Se continúa así con todos los símbolos hasta llegar al último estado, si este estado es uno de los estados finales o de acep-

tación (*F*), entonces la cadena es aceptada, de lo contrario es rechazada y no forma parte del lenguaje del autómata.

En un autómata no determinístico todos los elementos son iguales a los de un autómata determinístico excepto las funciones de transición. Éstas regresan cero, uno o varios estados (los autómatas determinísticos sólo regresan uno). Esta situación provoca que se evalúen varios caminos para la misma cadena de entrada, por lo que ésta será aceptada si al menos uno de ellos comienza con el estado inicial y termina en un estado de aceptación. Visto de otro modo, dada una transición que regrese varios estados, el autómata no determinístico avanzará a todos esos estados, después ejecutará la siguiente transición y tal vez alguno de los caminos anteriores termine y otros avancen.

La definición formal de un autómata no determinístico es la misma que la de uno determinístico ($K, \sum, \delta, q0, F$). La diferencia está en el valor que regresa la función de transición, ésta regresa un subconjunto de estados de *K*.

3.3.3. Representaciones

Presentaré en esta subsección dos maneras de representar un autómata de estados finitos. La primera es con un diagrama de estados o diagrama de transiciones y la segunda con una tabla de transiciones.

El diagrama de estados es un grafo dirigido formado de nodos (vértices) y arcos. Los nodos, dibujados mediante círculos, representan los estados del autómata. Los arcos son líneas con flechas que representan las transiciones, estos arcos tienen asociado un símbolo y parten de un estado para llegar a otro. Los estados finales son representados con círculos dobles, el estado inicial puede tener una flecha apuntando hacia él si no es el estado *q0*.

Para generar un diagrama de transiciones a partir de un autómata ya construido se pueden seguir estos pasos:

1. Para cada estado de *Q* se crea un nodo, es decir, se dibuja un círculo etiquetado con su nombre.

2. Para cada función de transición, por ejemplo $\delta(q, a) = p$, se dibuja un arco del nodo *q* al nodo *p*, etiquetado como *a*. Si varios símbolos provocan una transición de *q* a *p*, se debe poner un solo arco etiquetado con la lista de símbolos.

3. El estado inicial *q0* tiene una flecha opcional dirigida hacia él. La flecha no se origina en ningún nodo.

4. Los nodos de aceptación o nodos finales (*F*) tienen doble círculo.

En la Figura 3.1 se puede ver el diagrama de estados que representa un autómata de estados finitos para una pequeña parte de la morfología del español. En éste se pueden ver los elementos mencionados anteriormente. El estado inicial *q0* tiene la flecha que apunta hacia él. Los estados de aceptación tienen doble círculo y los arcos llevan el símbolo que permite pasar de un estado a otro. Incluyo un símbolo *base–* para representar un conjunto de posibles bases nominales.

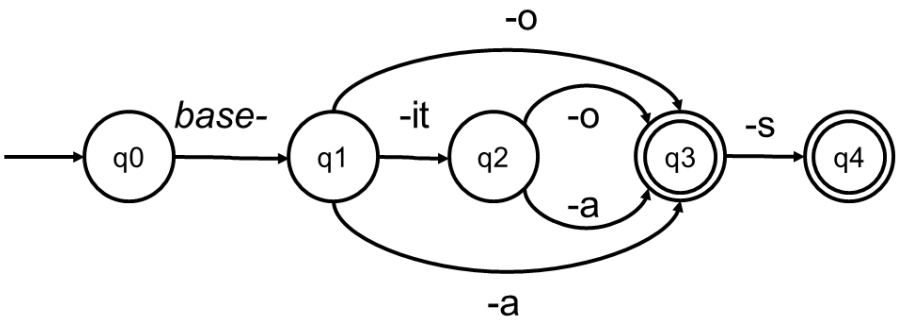


Figura 3.1 Ejemplo de un diagrama de estados

Las transiciones del autómata podrían describirse de la siguiente manera. La primera tendría asociado como símbolo cualquier cadena con categoría nominal y se definiría como $\delta(q0, base–) = q1$, esto es, si el autómata se encuentra en el estado inicial y lee como símbolo una cadena de categoría nominal, pasa al estado *q1*[53]. La segunda transición podría ser $\delta(q1, –o) = q3$, esto es, si el autómata se encuentra en el estado *q1* y lee como símbolo el segmento *–o*, pasa al estado *q3*. Obsérvese que éste es un estado de aceptación (doble círculo).

Estableciendo este autómata como aceptador, si se da como entrada al autómata la cadena *puebl-o*, como una secuencia de símbolos *puebl-* y *-o*, con base en las dos transiciones explicitadas en el párrafo anterior, se puede decir que el autómata sí acepta esta cadena de símbolos. El primer símbolo *puebl-*, es un nominal, por lo que el autómata pasa del estado *q0* al estado *q1*. El segundo símbolo es *–o*, y dado que sí existe una transición asociada a este símbolo, el autómata pasa del estado *q1* al estado *q3*. Como no hay más símbolos en la cadena de entrada, la revisión termina y el autómata acepta la cadena ya que se ha quedado en un estado de aceptación.

Con el fin de ejemplificar que la construcción de un autómata que intente reflejar la morfología de una lengua no es una tarea simple, incluso si se restringe solamente a la morfología sufijal del español, desarrollaré algunos ejemplos ayudándome del autómata anterior.

Supóngase al autómata como aceptador y la cadena de entrada *testig-o*. Si bien esta palabra no es de flexión regular y, como mencioné en su momento, es dudoso que el seg-

[53] Otra representación de esta transición sería (*base–, q0, q1)*

mento –o pueda ser considerado como flexión de número, el autómata acepta la cadena mediante las transiciones: *q0, q1 q3*.

Supóngase ahora el autómata como generador. Si se asume *testig–* como una base nominal, como se hizo en el párrafo anterior, el autómata generaría las siguientes cadenas *testig–o*, **testig–a* y **testig–it–o*, entre otras. Lo anterior plantea algunos cuestionamientos: ¿debería el autómata separar las bases nominales que tienen flexión regular de las que tienen flexión irregular en el caso del género? ¿Cómo puede saber un método automático si una palabra se flexiona en todo el paradigma?

Creo que estos ejemplos dejan ver algunas de las dificultades de idear un autómata que intente represente la morfotáctica del español y de la necesidad de considerar si el autómata se asume como aceptador o generador.

Otra manera de representar un autómata es mediante una tabla de transiciones. Ésta expresa las funciones de transición, $\delta(q, a) = p$, de la siguiente manera: en los renglones de la tabla se ponen los estados y en las columnas se ponen los símbolos de entrada; en la intersección de un estado (q) y un símbolo (a) se escribe el estado que resulta de la transición (p). El estado inicial se marca con una flecha y los estados resultantes o de aceptación con un asterisco (*) o con dos puntos (:). En caso de que no exista una transición definida para una determinada combinación de estado y símbolo, el espacio en la tabla se llena con un símbolo nulo Ø.

La siguiente tabla de transiciones (Tabla 3.2) sería equivalente al diagrama de estados de la Figura 3.1.

Tabla 3.2 Ejemplo de una tabla de transiciones

	Símbolos de entrada				
Estados	*base*	*-o*	*-it*	*-s*	*-a*
→0	1	Ø	Ø	Ø	Ø
1	Ø	3	2	Ø	3
2	Ø	3	Ø	Ø	3
*3	Ø	Ø	Ø	4	Ø
*4	Ø	Ø	Ø	Ø	Ø

Obsérvese como, dado el estado inicial (*q0*) y una *base* nominal, la intersección de renglón y columna marca como salida el estado *q1*, lo que es equivalente a la función de transición δ(*q0*, *base*-) = *q1*.

En la tabla de transiciones de un autómata no determinístico, las intersecciones entre columnas y renglones tendrían como valor un conjunto de estados. Se debe recordar que un autómata no determinístico puede estar en varios estados a la vez. En el caso de un diagrama de estados sólo basta agregar los arcos necesarios para representar un autómata no determinístico.

Como se pudo ver en esta subsección, tanto el diagrama de estados como la tabla de transiciones parecen buenas opciones para representar la morfología sufijal del español, a pesar de que sólo se mostró una pequeña parte (el sufijo –it de diminutivo y las marcas de flexión nominal). Esto hace pensar en la conveniencia de construir un autómata de estados finitos en lugar de una gramática. Además, como ya mencionaba, los autómatas han sido la manera tradicional de representar la morfotáctica. Será pertinente entonces revisar la equivalencia entre un autómata y una gramática en la siguiente subsección.

3.3.4. Equivalencia entre gramática y autómata

En esta subsección abordo el tema de la equivalencia entre una gramática tipo 3 y un autómata de estados finitos. Esta equivalencia se da porque el lenguaje que acepta un autómata de estados finitos es un lenguaje regular, esto es, el mismo lenguaje que describe una gramática tipo 3.

Por otro lado, la teoría de autómatas y lenguajes formales ha desarrollado procedimientos para generar una gramática a partir de un autómata y viceversa (Wall, 1972). En seguida describo estos procedimientos.

Dada en una gramática formal $G = (V_N , V_T , \{S\}, P)$ de tipo 3 con reglas de la forma $A \rightarrow aB$ o $A \rightarrow a$, es posible construir un autómata M que acepte el mismo lenguaje L mediante los siguientes pasos:

1. El vocabulario terminal V_T es el alfabeto de entrada de M.
2. Los miembros del vocabulario no terminal V_N son los estados de M, más un nuevo estado qF.
3. Para cada regla $A \rightarrow aB$ en G, se crea la transición (a, A, B) en M, y para cada regla $A \rightarrow a$ en G, se crea la transición (a, A, qF) en M.
4. S es el estado inicial de M.
5. qF es el único estado final de M.

De manera inversa es posible construir una gramática a partir de un autómata. Dado un autómata no determinístico $M = (K, \sum, \delta, q_0, F)$ se puede construir una gramática G con el siguiente procedimiento:

1. El alfabeto de entrada $\sum$ se convierte en el vocabulario terminal V_T.
2. El conjunto de estados K se convierte en el vocabulario no terminal V_N.

3. Cada transición (a, q_i, q_j) en M se convierte en una regla en G de la forma $q_i \rightarrow aq_j$. Si q_j es un estado final, se agrega la regla $q_i \rightarrow a$.
4. El símbolo inicial de G es q_0.

A manera de ejemplo, muestro en la Figura 3.2 la gramática equivalente al autómata presentado anteriormente, el cual repito para facilitar su comparación.

GRAMÁTICA

$G = (V_N , V_T , \{q0\}, P)$

$V_N = \{q0, q1, q2, q3, q4\}$

$V_T = \{base-, -o, -it, -a, -s\}$

$P = \{$

R1. $q0 \rightarrow base\text{-}q1$,
R2. $q1 \rightarrow -oq3$,
R3. $q1 \rightarrow -itq2$,
R4. $q1 \rightarrow -aq3$,
R5. $q2 \rightarrow -oq3$,
R6. $q2 \rightarrow -aq3$,
R7. $q3 \rightarrow -s$
R8. $q1 \rightarrow -o$
R9. $q1 \rightarrow -a$
R10. $q2 \rightarrow -o$
R11. $q2 \rightarrow -a\}$

AUTÓMATA

$M = (K, \sum, \delta, q_0, F)$

$K = \{q0, q1, q2, q3, q4\}$

$\sum = \{base-, -o, -it, -a, -s\}$

$F = \{q3, q4\}$

$\delta = \{(base-, q0, q1),$
$(-o, q1, q3),$
$(-it, q1, q2),$
$(-a, q1, q3),$
$(-o, q2, q3),$
$(-a, q2, q3),$
$(-s, q3, q4)\}$

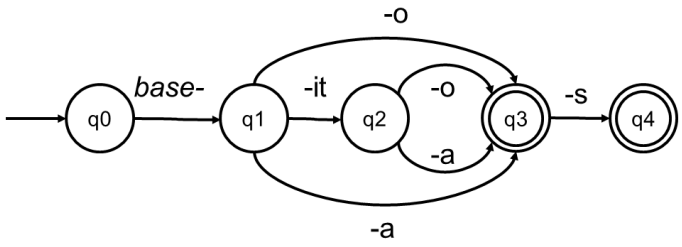


Figura 3.2 Ejemplo de gramática y autómata equivalentes

Daré el ejemplo de una derivación terminada para mostrar la equivalencia entre la gramática y el autómata de la figura anterior, aunque sería necesaria una demostración exhaustiva para confirmarlo. Tanto en Wall (1972, pág. 263) como en Hopcroft y Ullman

(1969, págs. 33-35) se puede ver una comprobación de la equivalencia entre un autómata y una gramática.

Con base en la gramática anterior, dada la base nominal *niñ–*, es posible aplicar la siguiente secuencia de reglas para derivar la cadena terminal *niñ-o-s*. Esta cadena sería generada también por la secuencia de transiciones de los estados *q0, q1, q3, q4*.

R1. *q0* → *base-q1*: *niñ-q1*

R2. *q1* → *-oq3* : *niñ-oq3*

R7. *q3* → -s : *niñ-o-s*

En conclusión, si un autómata acepta exactamente las mismas cadenas terminales generadas por una gramática, se puede hablar de que son equivalentes. Esta equivalencia se vuelve fundamental para el desarrollo de mi investigación. Por tanto, ya que una gramática de estados finitos es equivalente a un autómata de estados finitos y éste ha sido el mecanismo tradicional para representar la morfotáctica en morfología computacional, construiré un autómata, y no una gramática, para representar la morfotáctica de mi corpus de estudio. Además los autómatas cuentan con una representación gráfica en forma de diagrama de transiciones, que me parece mejor representación para su análisis.

3.3.5. Autómatas probabilísticos y modelos ocultos de Markov

Con el fin de mostrar algunas de las variantes de autómata de estados finitos que la teoría ha propuesto, en esta sección describiré dos tipos de autómatas.

Es posible agregar a las transiciones de un autómata la probabilidad de generar un nuevo estado a partir del estado inicial de la transición. Cuando se han agregado probabilidades a todas las transiciones se habla de un autómata de estados finitos probabilístico, también conocido como “cadena de Markov” (Charniak, 1996, pág. 32).

En este tipo de autómatas, la suma de las probabilidades de los arcos que salen de un estado debe sumar uno. Además, la probabilidad de generar una cadena se obtiene del producto de probabilidades de los arcos utilizados para generarla. Estos autómatas probabilísticos imponen la condición de que a partir de un estado y dado un símbolo de salida, solamente hay un estado siguiente.

Una variante de autómata de estados finitos probabilístico, ampliamente usada en la lingüística computacional, es el modelo oculto de Markov. Según Charniak (1996, pág. 43), es posible definirlo de manera formal como un conjunto de cuatro elementos (*q0, K, W, E*), donde *K* es un conjunto de estados, *q0* es el estado inicial ($q0 \in K$), *W* es un conjunto de símbolos de salida y *E* es un conjunto de transiciones.

En estos modelos, una transición es un conjunto de cuatro elementos (q^i, q^j, w^k, p), donde q^i es el estado en donde comienza la transición ($q^i \in K$), q^j es el estado en donde termina la transición ($q^i \in K$), w^k es un símbolo de salida ($w^k \in W$) y p es la probabilidad de tomar esa transición.

En estos modelos es posible tener un estado del que partan varias transiciones con el mismo símbolo de salida, pero que lleven a diferentes estados. Esta situación hace imposible saber el estado actual del modelo a partir únicamente del símbolo de salida, porque pueden ser varios estados. Por consiguiente, no se conoce la secuencia de estados que toma el modelo con sólo ver la cadena, lo que le da el carácter de modelo "oculto". Claro que se han desarrollado procedimientos para tomar la secuencia de transiciones más apropiada.

La probabilidad de una transición se define a partir de dos probabilidades: la probabilidad del símbolo de salida (w^k) y la probabilidad de pasar al siguiente estado (q^j) a partir de estado actual (q^i). Tomar sólo la información del estado anterior para obtener la probabilidad de la transición al estado siguiente es una suposición de estos modelos que permite

simplificar los cálculos necesarios para representar una cadena de símbolos; aunque en los fenómenos representados por el modelo esto no sea cierto.

El autómata de estados finitos que será generado automáticamente en esta tesis podría servir de base para generar a futuro un autómata probabilístico o un modelo oculto de Markov. Claro que primero se deberá indagar la ventaja de representar la morfotáctica mediante estos modelos.

3.4. Representación computacional de la morfotáctica

En esta sección abordo algunas propuestas de representación y tratamiento automático de la morfología de lenguas naturales. Pondré especial interés en los trabajos que representan computacionalmente la morfotáctica, lo que me permitirá establecer cómo se ha tratado este problema anteriormente.

Los mecanismos estándar en morfología computacional para representar la morfotáctica de las lenguas son los autómatas de estados finitos (Sproat, 1992, pág. 124). Los trabajos basados en este tipo de mecanismos se pueden englobar en la llamada *morfotáctica de estados finitos*. Ésta asume que la aparición de un morfema depende sólo del morfema que le antecede. Con esta idea, los autómatas resultan buenos[54] para representar fenómenos de morfología concatenativa y no así otros tipos de fenómenos morfológicos.

La Figura 3.3 muestra el ejemplo de un autómata construido para una parte de la morfología del inglés, tomado de Sproat (1992, pág. 127). Este autómata estaría asociado a cadenas como *nominalization, hospital, hospitalization* y *moralize*.

[54] Según Sproat resultan "sufficiently powerful" (1992, pág. 127).

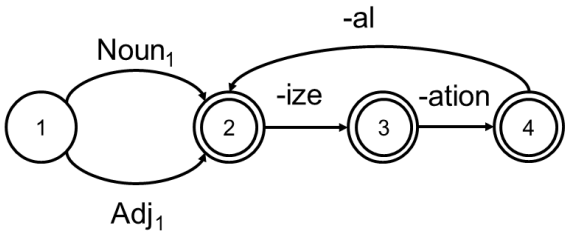


Figura 3.3 Ejemplo de autómata para una parte de la morfología del inglés
Tomado de Sproat (1992, pág. 127)

En realidad son varios los fenómenos que estos mecanismos no pueden representar, esto debido a la complejidad morfológica de las lenguas. Por ejemplo, tienen dificultad para representar fenómenos morfológicos discontinuos donde la aparición de un morfema depende de otro que no se encuentra adyacente (Sproat, 1992). Un ejemplo de estos fenómenos es la parasíntesis del español.

Según Anderson (1992, págs. 387-391), para estos mecanismos también son problemáticos los fenómenos de infijación y reduplicación, así como los cambios vocálicos, la metátesis y el truncamiento de material fonológico o escrito.

Otra característica de los autómatas como el que se mostró en la Figura 3.3 es que su cadena de entrada debe estar segmentada, lo que implica que se conoce su segmentación *a priori*. Para evitar esta situación, es posible construir otro tipo de representaciones como una red de discriminación (*discrimination network*) o *trie*.

Esta es una red de nodos donde cada uno representa una letra. El nodo raíz apunta a una letra inicial de todas las palabras que se intentan representar. Además, es posible agregar a los nodos información de la categoría a la que pertenece la palabra (sustantivo, adjetivo). La Figura 3.4 muestra un ejemplo reelaborado de Sproat (1992, pág. 129).

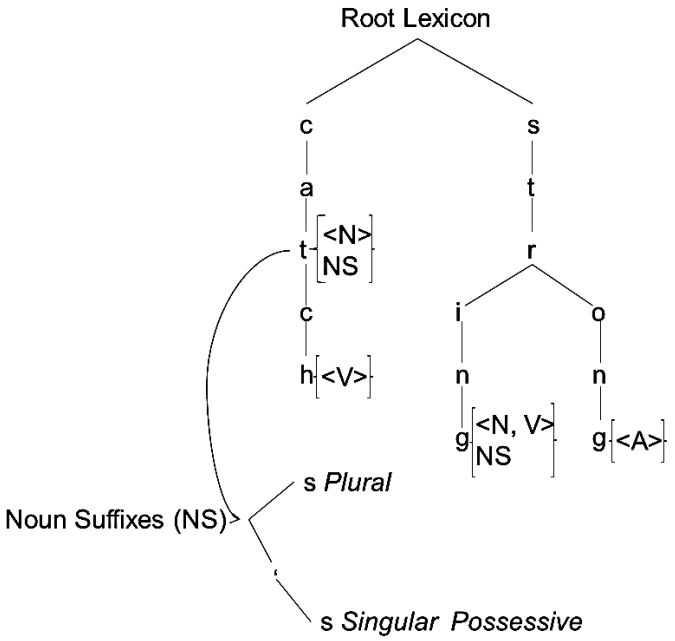


Figura 3.4 Ejemplo de red de discriminación o *trie*
Reelaborado de Sproat (1992, pág. 129)

En la figura anterior se pueden ver las categorías asociadas a cada palabra en la letra final correspondiente. Además, hay una marca que indica que esa palabra cuenta con un sufijo nominal. El conjunto de sufijos nominales está representado en otra red separada.

Tanto los autómatas de estados finitos como las redes *trie* no contemplan cambios en la forma fonológica (o escrita) de la lista de palabras que representan (morfofonología). Por lo anterior, los morfólogos computacionales buscaron una representación que contemplara estos cambios. El resultado fue la adopción de la fonología de dos niveles, también llamada fonología de estados finitos (Koskenniemi, 1983; Antworth, 1990).

La fonología de dos niveles está basada en transductores de estados finitos (Sproat, 1992; Jurafsky y Martin, 2009). Éstos son autómatas con un alfabeto formado de pares de símbolos. Utilizando la idea de un autómata como un lector de cinta, en el caso de los transductores habría dos cintas en lugar de una. La primera se llama cinta superior o cinta

léxica y la segunda se llama cinta inferior o cinta de superficie. El transductor avanzaría leyendo la cinta superior e imprimiendo (o reconociendo) el símbolo de la cinta inferior. La Figura 3.5 presenta un ejemplo tomado de Sproat (1992, pág. 133).

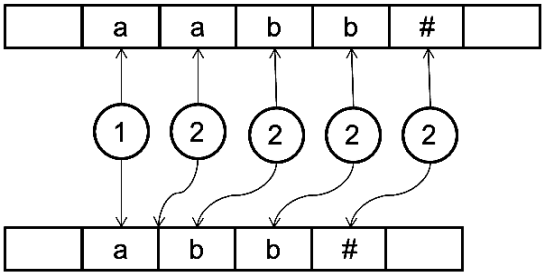


Figura 3.5 Ejemplo de transductor de estados finitos
Tomado de Sproat (1992, pág. 133)

El transductor de la figura anterior muestra el proceso de aceptación de las cadenas (*aabb, abb*). El símbolo # representa el fin de cadena. Los círculos con números se refieren los estados del transductor ejemplificado por este par de cintas. Véase cómo el segundo símbolo de la cinta léxica no corresponde a ningún símbolo en la cinta de superficie, lo que permite transformar la cadena *aabb* en *abb*.

Estos transductores sirvieron de manera afortunada para representar reglas de transformación fonológica propuestas por la lingüística, lo que motivó su amplia aceptación y estudio. Cuando la fonología comenzó a trabajar con conjuntos de reglas ordenadas, se propuso utilizar transductores en cascada. En éstos, la salida de un transductor era la entrada del siguiente. Había tantos transductores como reglas ordenadas.

Si bien la idea de estos transductores en cascada era teóricamente posible, su implementación se volvió difícil. Por lo anterior se buscaron soluciones computacionales para reducir la cantidad de estados y de procesamiento. Así, surgió la propuesta de Koskenniemi

(1983), conocida como morfología de dos niveles. En ésta, las reglas fonológicas, y por tanto los transductores, funcionan en paralelo y ya no en cascada.

La morfología de dos niveles permitió tratar buena parte de los fenómenos morfofonológicos de distintas lenguas, lo que a su vez provocó el desarrollo de diversos programas de computadora de análisis morfológico automático (Antworth, 1990; Sproat, 1992; Jurafsky y Martin, 2009, págs. 365-367). Incluso hubo propuestas para morfología no concatenativa (Kay, 1987).

Las propuestas anteriores son una muestra de las posibilidades de representación y procesamiento morfológico automático. Lo que quisiera resaltar de ellas es que las representaciones (autómatas y transductores) son elaboradas manualmente. Esto puede dar pie a que el objeto de estudio del investigador se vuelva el artefacto abstracto de descripción y no la lengua.

Revisar las propuestas anteriores me permitió resaltar algunas de las virtudes y defectos de los autómatas de estados finitos como representaciones de la morfotáctica de las lenguas. A pesar de sus limitaciones, considero que para las características morfológicas del español, los autómatas de estados finitos pueden representar gran parte de su morfotáctica afijal. Por tanto, gracias a lo revisado en este capítulo, establezco que la manera más pertinente para representar la morfotáctica sufijal del español será mediante la construcción de un autómata, el cual puede convertirse mediante el procedimiento descrito arriba en una gramática de estados finitos.

4. Experimentos de segmentación morfológica automática

En este capítulo describo el procedimiento de descubrimiento de bases[55] y sufijos del español mediante la segmentación morfológica automática de un conjunto de tipos de palabras obtenidas de un corpus.

Como ya mencioné, utilizaré el método basado en el cálculo de un índice de afijalidad como estrategia de segmentación morfológica automática para luego descubrir los patrones morfotácticos del español de México. Antes de exponer cómo utilizaré este método, considero necesario hacer algunas observaciones preliminares.

El índice de afijalidad ha sido utilizado para inferir un catálogo de sufijos del español basado en un solo corte al interior de cada palabra, como quedó establecido en la sección 2.5. Obviamente, el hecho de tener un corte por palabra conlleva que los segmentos resultantes puedan ser sufijos individuales o grupos de sufijos concatenados, además de las bases.

Para mi investigación, es necesario identificar a cada uno de los sufijos de una palabra, ya que de esta manera podré establecer sus patrones de secuencialidad. Un primer acercamiento a la resolución de este problema podría consistir en descomponer los grupos sufijales concatenados en sufijos individuales a partir del catálogo inferido; sin embargo, hay dos problemas. Primero, en el catálogo se ha perdido la información que relaciona a cada sufijo o grupo de sufijos con sus bases. Segundo, tendría que distinguirse entre sufijos

[55] Se debe recordar que utilizaré *base* como el segmento sobre el que operan los fenómenos tanto de flexión como de derivación, véase Pena (1999, pág. 4318).

individuales y grupos de sufijos para segmentar sólo los segundos, problema que no es trivial.

Por lo anterior, opté mejor por segmentar cada palabra del CEMC mediante una estrategia formulada a partir de las medidas involucradas en el cálculo de la afijalidad. Para determinar la mejor estrategia a seguir, efectué un experimento de segmentación como primer acercamiento. Los detalles al respecto son expuestos en la sección subsecuente.

Como resultado de ese primer experimento, constaté la necesidad de hacer un análisis más profundo para establecer una mejor estrategia de segmentación. Por consiguiente, definí un conjunto de intuiciones de segmentación que me permitieron diseñar un grupo de experimentos donde distintas estrategias fueron probadas. En secciones posteriores expongo las intuiciones, experimentos, su evaluación y la selección de la estrategia final que será utilizada para la segmentación que permitirá inferir la morfotáctica del español de México.

4.1. Primer acercamiento a la segmentación automática

De manera intuitiva y como primer acercamiento a la resolución de mi problema de segmentación, realicé un experimento que consistió en segmentar las palabras en los picos del índice de afijalidad. Esta afijalidad fue calculada utilizando un promedio normalizado de las tres medidas: entropía, economía y cuadros. Esta manera de calcular el índice ya había sido propuesta en investigaciones anteriores (Medina, 2003).

Dado un conjunto de índices de afijalidad af_i^k calculados para una palabra, hay un pico de afijalidad cuando el valor de ese índice es mayor al valor anterior y posterior, esto es, cuando $af_{i-1}^k < af_i^k > af_{i+1}^k$, donde *k* es la longitud de la palabra más 1 (el final de la

palabra). El procedimiento de segmentación comienza de derecha a izquierda asumiendo que el final de palabra tiene valor cero de afijalidad, lo que permite sufijos de una letra. Adicionalmente se prohibieron segmentaciones en los tres primeros caracteres de la palabra para evitar segmentaciones al interior de las bases, aunque el español cuenta con bases cortas, por ejemplo *am-ar*.

Obsérvese el ejemplo de la Tabla 4.1, en él se encuentran los índices de afijalidad calculados para la palabra UTILIZADOS[56]. Siguiendo el procedimiento expuesto, el primer pico de afijalidad corresponde al valor 0.8269, que daría como resultado la segmentación UTILIZADO~S. Los siguientes picos de afijalidad son 0.9421 y 0.4984, dejando la segmentación final como UTIL~IZ~ADO~S.

Tabla 4.1 Índices de afijalidad de la palabra UTILIZADOS

U	T	I	L	I	Z	A	D	O	S
	0.1634	0.1854	0.1123	**0.4984**	0.1818	**0.9421**	0.3585	0.8103	**0.8269**

La principal ventaja de esta estrategia es que no se pone ningún umbral para decidir dónde segmentar; sin embargo, valores bajos de afijalidad pueden ser tomados en cuenta.

Se realizaron las modificaciones al programa de computadora para segmentar las palabras de un corpus. Después fue necesario encontrar una manera de evaluar si el procedimiento de segmentación propuesto trabajaba aceptablemente. Al respecto, en sistemas de procesamiento de lenguaje natural hay dos criterios de evaluación: intrínseco y extrínseco (Spärck-Jones y Galliers, 1996).

El primero (intrínseco) se lleva a cabo mediante una evaluación que compara el análisis automático con el análisis de un especialista. Para el problema que me atañe, esto se traduce en comparar la segmentación automática con la segmentación manual de un con-

[56] Este primer acercamiento lo realicé con la representación ortográfica del corpus.

junto de palabras. Desafortunadamente no pude encontrar un corpus de español segmentado morfológicamente para esta evaluación, por lo que decidí realizar una evaluación extrínseca.

La evaluación extrínseca determina la efectividad de un método de acuerdo con su utilidad para realizar otra tarea de procesamiento de lenguaje natural. Por tanto, se decidió evaluar la segmentación automática mediante una tarea de resumen automático de documentos (Méndez-Cruz et al. 2013). En términos generales, esta tarea consiste en extraer los enunciados más relevantes de un documento (Torres-Moreno, 2011).

Que un programa de computadora determine cuáles enunciados son más relevantes es complejo, pero se han propuesto numerosos acercamientos para resolver este problema. Uno de ellos es el que proponen Torres-Moreno et al. (2009) y que se ha implementado en el sistema CORTEX. Este sistema genera resúmenes a partir de documentos usando un modelo de espacio vectorial.

CORTEX utiliza un algoritmo de decisión que combina diversas métricas para asignarle a cada enunciado del documento un valor de relevancia. El resumen producido es una concatenación de los enunciados con los valores más altos, según una tasa de compresión expresada en número de enunciados, de palabras o porcentaje del documento.

Para realizar los resúmenes, CORTEX preprocesa los documentos y después los representa como una matriz γ (véase Figura 4.1), donde cada elemento γ_i^{μ} representa el número de ocurrencias de la palabra *i* en el enunciado μ ($1 \leq i \leq M$ palabras, $1 \leq \mu \leq P$ enunciados). También se crea otra matriz ξ para representar la presencia y ausencia de una palabra en un enunciado.

$$\gamma = \begin{bmatrix} \gamma_1^1 & \gamma_2^1 \dots & \gamma_i^1 \dots & \gamma_M^1 \\ \gamma_1^2 & \gamma_2^2 \dots & \gamma_i^2 \dots & \gamma_M^2 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \gamma_1^P & \gamma_2^P \dots & \gamma_i^P \dots & \gamma_M^P \end{bmatrix}, \quad \gamma_i^\mu \in \{0,1,2 \dots\}$$

Figura 4.1 Matriz γ de ocurrencias de palabras por enunciado en CORTEX

A partir de las matrices mencionadas, CORTEX calcula distinta información matemática y estadística para obtener métricas de frecuencias, entropía, medidas de Hamming y otras. No describo aquí los cálculos de estas métricas, pero éstos pueden encontrarse en Torres-Moreno et al. (2009) y Torres-Moreno (2011). El algoritmo de decisión combina todas las métricas y calcula un valor de relevancia para cada enunciado. Luego, como se había dicho, los enunciados con los valores más altos conforman el resumen del documento.

Parte del preprocesamiento de documentos que realiza CORTEX consiste en quitar algunas palabras y regularizar las restantes mediante un procedimiento de lematización (*lemmatization*). Este procedimiento tiene el objetivo de reducir la cantidad de variantes ortográficas de una palabra, lo que permite reducir el espacio de representación y mejorar los cálculos matemáticos.

La lematización consiste en asignar una palabra como representante de un conjunto de variantes morfológicas. Por ejemplo, las flexiones verbales son representadas por la forma en infinitivo y los sustantivos por el singular masculino. CORTEX realiza esta labor mediante un diccionario donde busca la palabra del documento y obtiene su representante.

Una estrategia alternativa utilizada por CORTEX para disminuir las variantes de una palabra es eliminar sus sufijos flexivos. Esto regulariza las palabras a una forma trunca. Este procedimiento se llama truncamiento (*stemming*).

El procedimiento de truncamiento que utiliza este sistema es el algoritmo de Porter (1980), que está basado en reglas elaboradas manualmente donde los sufijos de la palabra son eliminados o sustituidos por otros más cortos. Este procedimiento fue desarrollado originalmente para el inglés, pero existen versiones para otras lenguas como el español o el francés. Opcionalmente, CORTEX también puede realizar ambos procesos, lematización y luego truncamiento, como otra estrategia de regularización de palabras.

Para la evaluación extrínseca se decidió desarrollar un truncador de palabras a partir de la segmentación morfológica. Este truncador sería utilizado por CORTEX para generar los resúmenes automáticos. Afortunadamente, CORTEX cuenta con una arquitectura modular que permitió ensamblar el truncador desarrollado.

La idea detrás de esto fue que si los resúmenes mejoraban, entonces había indicio de que la segmentación morfológica era eficiente desde una perspectiva extrínseca. Si el método de segmentación trabajaba de manera regular, entonces el truncamiento sería también regular y el sistema de resumen automático aprovecharía esta regularidad para sus objetivos.

Se consideró interesante probar tres estrategias de truncamiento: (i) truncar en el primer pico de afijalidad a la derecha, (ii) truncar en el primer pico de afijalidad a la izquierda, y (iii) truncar en el valor máximo de afijalidad. Por ejemplo, para la palabra segmentada UTIL~IZ~ADO~S (véase arriba Tabla 4.1) la estrategia (i) daría como resultado la palabra truncada UTIL~IZ~ADO~, la estrategia (ii) propondría la palabra truncada UTIL~, y la estrategia (iii) plantearía la palabra truncada UTILIZ~.

Un método automático no supervisado tiene la ventaja de ser relativamente independiente de la lengua, por lo que se realizaron experimentos en español, francés e inglés. Además, los resultados de un método de procesamiento estadístico, como el que calcula el

índice de afijalidad, pueden ser sensibles al tamaño de datos con los que es alimentado. Entonces, se pensó en evaluar si había relación entre la cantidad de palabras para generar las medidas (entropía, cuadros y economía) y las segmentaciones resultantes. Por lo anterior se utilizaron tres tamaños de corpus de entrenamiento, esto es, tres cantidades de palabras para generar las medidas de afijalidad.

Lo anterior dio como resultado la siguiente configuración de experimentos de truncamiento para cada lengua (véase Tabla 4.2).

Tabla 4.2. Configuración de experimentos de truncamiento

	Truncamiento en:		
Tamaño del corpus de entrenamiento	Valor máximo (vM)	Pico más a la derecha (R)	Pico más a la izquierda (L)
100 mil palabras	vM100	R100	L100
200 mil palabras	vM200	R200	L200
500 mil palabras	vM500	R500	L500

Para evaluar si las estrategias de truncamiento mejoraban los resultados del resumidor, se pusieron a competir con las estrategias utilizadas por éste: lematización con diccionario (lemm), truncamiento con el algoritmo de Porter (stem) y ambos procedimientos (lems). Fueron agregadas dos pruebas más: truncar arbitrariamente a seis caracteres (fixed) y no modificar la palabra (raw). La Tabla 4.3 muestra todas las estrategias probadas en los experimentos de resumen automático para la evaluación extrínseca.

Tabla 4.3 Conjunto de experimentos para la evaluación extrínseca

Estrategia	Descripción
vM100	Valor máximo de afijalidad y corpus de entrenamiento de 100 mil palabras
vM200	Valor máximo de afijalidad y corpus de entrenamiento de 200 mil palabras
vM500	Valor máximo de afijalidad y corpus de entrenamiento de 500 mil palabras
R100	Pico de afijalidad más a la derecha y corpus de entrenamiento de 100 mil palabras
R200	Pico de afijalidad más a la derecha y corpus de entrenamiento de 200 mil palabras
R500	Pico de afijalidad más a la derecha y corpus de entrenamiento de 500 mil palabras
L100	Pico de afijalidad más a la izquierda y corpus de entrenamiento de 100 mil palabras
L200	Pico de afijalidad más a la izquierda y corpus de entrenamiento de 200 mil palabras
L500	Pico de afijalidad más a la izquierda y corpus de entrenamiento de 500 mil palabras
lemm	Lematización con diccionario
stem	Truncamiento con algortimo de Porter
lems	Lematización y luego truncamiento
fixed	Truncamiento a seis caracteres
Raw	Ninguna modificación a la palabra

Se utilizaron diversos textos para conformar el corpus de evaluación del resumidor. Para el inglés se tomaron 50 grupos de textos provenientes de la tarea dos de la competencia internacional *DUC 2004*[57]. Para el español se utilizaron ocho artículos del área biomédica obtenidos de la revista especializada *Medicina Clínica*[58]. Con respecto al francés se utilizó el corpus *Canadien French Sociological Articles* (Torres-Moreno et al., 2010) de la revista electrónica especializada *Perspectives interdisciplinaires sur le travail et la santé* (PISTES)[59].

En lo que atañe al corpus de entrenamiento se experimentó, como había dicho, con tres distintos tamaños: 100, 200 y 500 mil ocurrencias de palabras. En el caso del español,

[57] http://duc.nist.gov/duc2004.

[58] http://zl.elsevier.es/es/revista/medicina-clinica-2.

[59] http://www.elsevier.es/revistas/ctl_servlet?_f=7032&revistaid=2

utilicé extractos del CEMC. Para el inglés utilicé 24 documentos extraídos de la tarea siete de la competencia *INEX 2012* (*Tweet Contextualization Track*)[60]. Para el francés, conformé el corpus de diversas fuentes[61].

Se generaron los resúmenes de cada documento del corpus de evaluación con el sistema CORTEX utilizando cada una de las variantes de normalización de palabras expuestas arriba (véase Tabla 4.3). Luego fue necesario evaluar la calidad de estos resúmenes para lo cual se utilizaron los resúmenes elaborados por humanos. Esta evaluación se realizó con el sistema ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004). Este sistema realiza evaluaciones semiautomáticas que miden la similitud entre el resumen automático y el resumen manual. Se utilizaron medidas de similitud basadas en pares de palabras contiguas (bigramas) y no contiguas (bigramas con huecos). Los primeros son llamados ROUGE-2 y los segundos ROUGE-SU4.

Los resultados de los experimentos para español pueden verse en la Figura 4.2. Puede observarse que los mejores resúmenes se obtuvieron con la estrategia de cortar las palabras en el pico de afijalidad más a la izquierda y con un corpus de entrenamiento de 500 mil palabras (L500). De hecho, la segunda mejor estrategia es la misma pero con un corpus de 200 mil palabras.

[60] https://inex.mmci.uni-saarland.de/tracks/qa/2012/.

[61] Los detalles de este corpus y del corpus de evaluación, así como de los resultados obtenidos, se pueden en Méndez-Cruz et al. (2013).

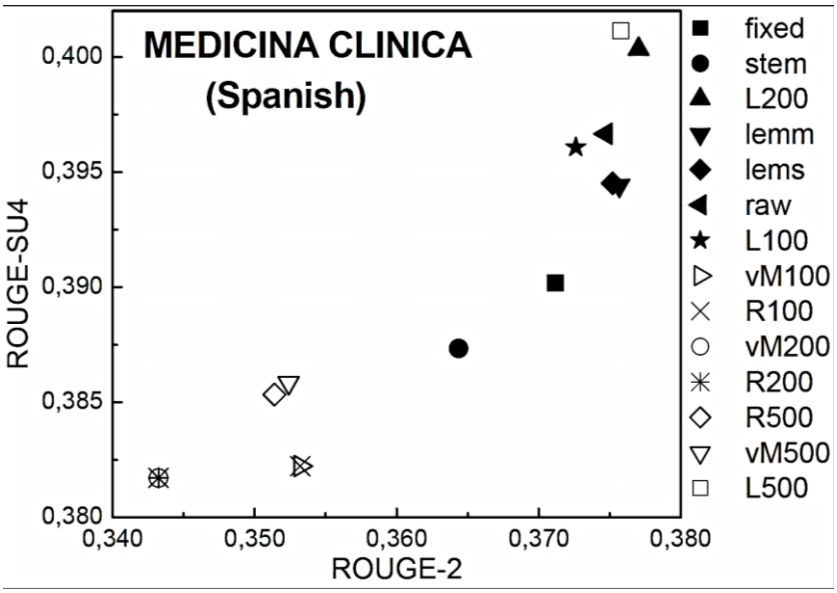


Figura 4.2 Resultados de la evaluación extrínseca para español

Los peores resultados para español se dan con resúmenes que usaron la estrategia ya sea de segmentar en el pico más a la derecha o en el valor más alto de afijalidad. En el caso del francés (véase Figura 4.3), el mejor método fue también el método L500.

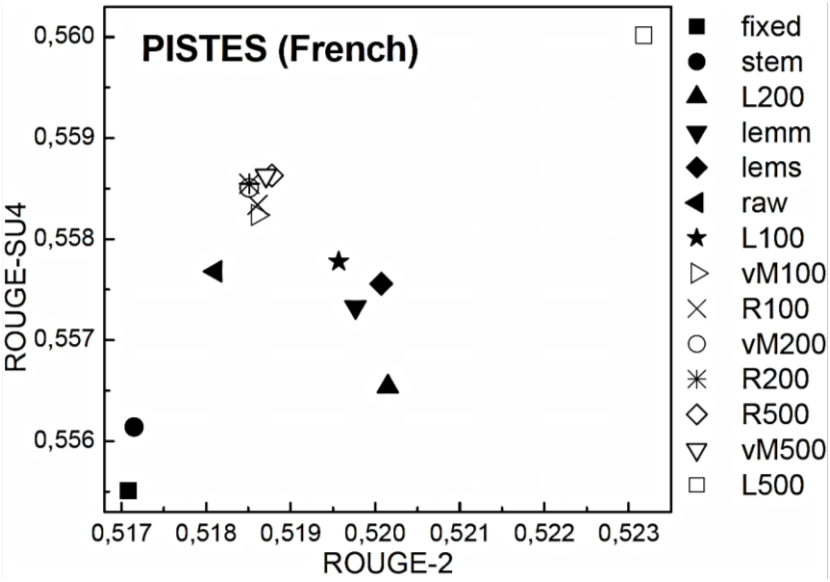


Figura 4.3 Resultados de la evaluación extrínseca para francés

Para esta lengua las peores estrategias fueron el truncamiento con el algoritmo de Porter y la falta de regularización de las palabras. En general, los resultados obtenidos llevan a pensar que para español y francés CORTEX mejora cuando hay mayor truncamiento. Esto es, cortar en el pico de afijalidad más a la izquierda implica que las palabras son desprovistas de todo sufijo flexivo o derivativo.

En contraste, los resúmenes para inglés no mejoraron con el método de corte en el pico de afijalidad más a la izquierda, por el contrario, éste fue el método con peores resultados como puede verse en la Figura 4.4. En esta lengua, fueron mejores resúmenes los que utilizan lematización con diccionario (lemm) y los que cortan en el pico más a la derecha o en el valor más alto de afijalidad (vM100 y R100). Además, se logran buenos resultados con corpus de entrenamiento pequeño, esto es, de 100 mil palabras.

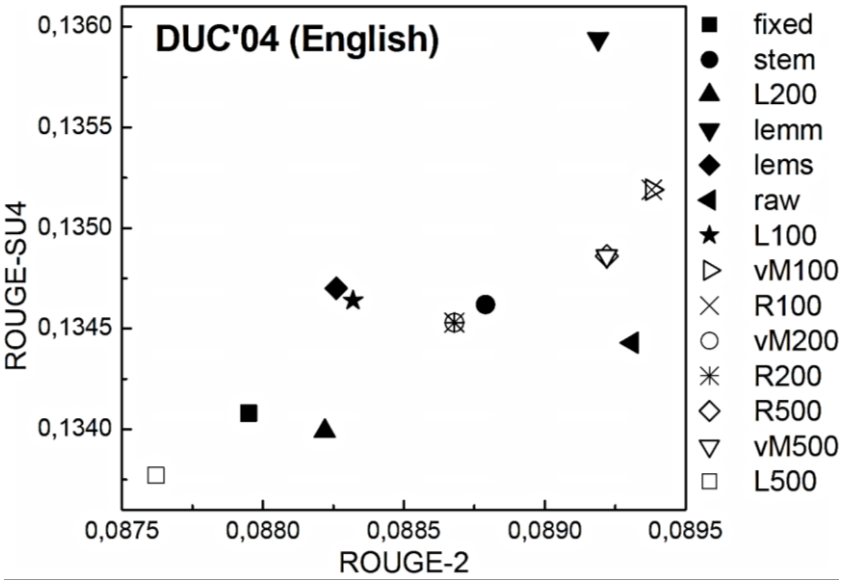


Figura 4.4 Resultados de la evaluación extrínseca para inglés

Los experimentos de evaluación demostraron que la estrategia de segmentación morfológica funcionó bien para realizar la tarea de resumen automático de documentos, aunque fue mejor para español y francés que para inglés. Una desventaja de esta evaluación

extrínseca es el desconocimiento de qué tanto el funcionamiento interno de CORTEX ayudó a que se dieran esos resultados.

Con el truncador desarrollado para los experimentos anteriores se participó en la competencia *INEX 2012* (Méndez-Cruz, Soriano-Morales y Medina-Urrea, 2011), específicamente en la tarea de contextualización de *tweets* en inglés (*Tweet Contextualization Track*). Esta tarea consistió en obtener textos de la Wikipedia en inglés relacionados con el tema de cada *tweet*. Los contextos finales obtenidos de estos textos no debían superar las 500 palabras, por lo que se utilizó CORTEX con el truncador desarrollado para realizar los resúmenes de los textos extraídos de la Wikipedia.

Nuevamente se probaron tres tamaños de corpus de entrenamiento para calcular las medidas de afijalidad: 100, 200 y 500 mil palabras. Los resúmenes de 1,133 *tweets* contextualizados fueron enviados a la competencia. La evaluación de la contextualización estuvo a cargo de los organizadores (SanJuan et al. 2011).

La mejor posición obtenida para el resumidor con el truncador fue el lugar 9 de 27 lugares. Esta posición se logró con un corpus de 500 mil palabras. Las otras variantes lograron los lugares 15 (200 mil) y 17 (100 mil), lo que da muestra de que el truncamiento con mayor corpus de entrenamiento ayudó a CORTEX a obtener mejores resultados en esta competencia.

Finalmente, se decidió que era necesario experimentar con otras estrategias de segmentación, por lo que me di a la tarea de proponer varios experimentos que revelaran una mejor manera de segmentar las palabras. También se vio la necesidad de realizar una evaluación intrínseca para evaluar qué estrategia de segmentación convenía adoptar.

En los siguientes apartados expongo cómo se llegó a la formulación de las intuiciones de segmentación, los experimentos y su evaluación.

4.2. Definición del conjunto de experimentos

Esta sección está dedicada a exponer el diseño de los experimentos de segmentación morfológica automática. Por un lado, se realizaron algunos experimentos basados en estrategias de segmentación que habían sido propuestas en trabajos anteriores de investigación, pero que no se habían llevado a cabo (Medina, 2000; 2003); estas son puntualizadas en el apartado 4.2.1.

Además, hice otros experimentos que fueron diseñados a partir de intuiciones formuladas en este trabajo de investigación. Para elaborar estas intuiciones, primero recuperé de manera puntual varios antecedentes establecidos en las investigaciones anteriores a propósito del cálculo de la afijalidad. Estos antecedentes se pueden ver en el apartado 4.2.2.

Después combiné estos antecedentes con mis propias observaciones sobre el cálculo de las medidas de afijalidad de un conjunto pequeño de palabras; observaciones que consigno en el apartado 4.2.3. Con toda esta información, elaboré las intuiciones que guiaron los experimentos de segmentación realizados, éstos pueden verse en las secciones 4.2.4 y 4.2.5.

4.2.1. Estrategias de segmentación propuestas anteriormente

Las siguientes estrategias de segmentación morfológicas ya habían sido propuestas en investigaciones previas sobre el índice de afijalidad (Medina, 2000, pág. 108; 2003, pág. 133), aunque no se habían probado. Todas se basan en un índice de afijalidad obtenido a partir del cálculo del promedio de las medidas normalizadas, ya sea de dos de ellas (entropía y economía) o de las tres (entropía, economía y cuadros).

(a) Segmentar cuando el valor de afijalidad sea mayor a cero.

(b) Segmentar en el valor más alto de afijalidad.

(c) Segmentar cuando el valor de afijalidad sea mayor a un valor umbral (0.5).

(d) Segmentar recursivamente hacia la izquierda en el valor más alto de afijalidad.

(e) Segmentar recursivamente hacia la derecha en el valor más alto de afijalidad.

En seguida discuto cuáles de estas estrategias utilizaré en mi trabajo de investigación. La estrategia propuesta en (a), como ya fue dicho en trabajos anteriores, produciría una gran cantidad de afijos por palabra, mucho de ellos falsos, ya que la gran mayoría de posibles segmentos llevan cierto grado de afijalidad y muy pocos están desprovistos de ella.

Por ejemplo, véanse los índices de afijalidad de la palabra ALARMANTES en la Tabla 4.4. La estrategia de segmentar cuando el valor de afijalidad sea mayor a cero daría como resultado la segmentación ALA~R~M~A~N~T~E~S. Como puede verse, se esperarían muchos afijos falsos (sobresegmentación) y por consiguiente no tomaré en cuenta esta estrategia.

Tabla 4.4 Índices de afijalidad de la palabra ALARMANTES

A	L	A	R	M	A	N	T	E	S
	0	0	0.1738	0.3634	0.5021	0.1061	0.536	0.07867	0.8298

Descarto utilizar también la estrategia de (b) porque propone segmentar la palabra sólo una vez. Esta fue la estrategia seguida para obtener el catálogo de afijos en las investigaciones anteriores.

Las estrategias (c), (d) y (e) sí son tomadas en cuenta en mi investigación ya que por medio de ellas es posible obtener varios sufijos por palabra. En los siguientes párrafos hago algunas observaciones adicionales sobre estas estrategias.

La estrategia (c) despierta la curiosidad por saber qué tan pertinente es el uso de un umbral en el proceso de segmentación, lo que lleva a la cuestión de qué valor conviene po-

ner a ese umbral. Dado que el índice de afijalidad está normalizado (va de cero a uno), tomar valores arriba de la mitad (0.5) parece en principio buena idea. Lo que no debe olvidarse es que entre más alto sea el umbral, mayor será el nivel de afijalidad que se exija para que un segmento sea considerado como un afijo.

En lo que toca a las estrategias (d) y (e), éstas ponen de manifiesto la importancia de considerar la direccionalidad de la segmentación, es decir, hacia dónde se realizan los cortes en la palabra (izquierda o derecha). Además, proponen usar el valor más alto de afijalidad en lugar de un umbral. Esta idea resalta el hecho de que cada segmento de un palabra conlleva cierto nivel de afijalidad y que son los más altos los que estarían asociados al fenómeno de afijación.

Con relación a la direccionalidad y recordando los resultados obtenidos en los experimentos reportados en investigaciones anteriores, donde el catálogo de afijos estaba encabezado por los sufijos con mayor afijalidad y que estos coincidían con los sufijos más flexivos, ¿se puede esperar que una estrategia de corte hacia la derecha proporcione buenos resultados, cuando se sabe que en español los sufijos flexivos están más a la derecha de un palabra? Para resolver esta cuestión se llevaron a cabo experimentos en ambas direcciones.

4.2.2. Antecedentes sobre el cálculo de la afijalidad

Consigno en esta subsección el conjunto de antecedentes que, a propósito de la segmentación morfológica, fueron expuestos en trabajos previos de investigación sobre el cálculo de la afijalidad. Estos antecedentes me permitieron inferir algunas ideas sobre la segmentación morfológica que a la larga me ayudaron a generar algunas intuiciones. Varios de estos antecedentes ya fueron mencionados en la sección 2.5, pero los repito aquí de manera puntual

con el fin de hacer más clara la discusión sobre el diseño de los experimentos de segmentación.

Antecedentes a propósito de las medidas de afijalidad:

A.1 Los cuadros son una medida de la validez de una segmentación.

A.2 El valor más alto de entropía, calculada de derecha a izquierda de una palabra, indica dónde termina una base y comienza un sufijo o cadena de sufijos.

A.3 En el caso de los sufijos, la economía asociada a un corte será más alta si el segmento del lado derecho es muy frecuente y pertenece a un conjunto relativamente pequeño de segmentos.

Pongo en seguida algunos breves comentarios sobre estos antecedentes. Del primer antecedente (A.1) se puede inferir que es posible esperar cortes más precisos si se combina la medida de cuadros con cualquier otra medida de afijalidad. Para el caso de A.2 es posible pensar que la medida de entropía es buena candidata para descubrir la base de un palabra. Finalmente, de A.3 se intuye que la medida de economía es buena candidata para descubrir sufijos flexivos, ya que son más frecuentes y aportan más economía al sistema lingüístico.

Las investigaciones anteriores ya habían hecho clara la necesidad de combinar las medidas de afijalidad en lo que se ha llamado un índice de afijalidad. En lo que respecta a la combinación de medidas recupero los siguientes antecedentes:

A.4 "La cualidad que tiene [un segmento] de ser 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" (Medina, 2003, pág. 128).

A.5 La combinación del índice de economía y de entropía mejoró considerablemente los resultados, en comparación con la combinación de los tres índices.

El antecedente A.4 muestra que es factible multiplicar los valores de las medidas en lugar de promediarlos para obtener el índice de afijalidad. Del antecedente A.5 se puede espera que combinar la medida de entropía y la medida de economía sea una buena estrategia para obtener cortes en las palabras, aunque habrá que averiguar si funciona igual de bien para varios cortes que para uno solo.

4.2.3. Reflexiones sobre las medidas de afijalidad

Revisé de manera general aproximadamente 500 palabras y con detalle un subconjunto de 60 de ellas con la idea de observar el comportamiento de las medidas de entropía, economía y cuadros. El resultado de estas observaciones se muestra a continuación. Éste, combinado con los antecedentes expuestos en el apartado previo, me permitió elaborar algunas intuiciones.

Ya el antecedente A.2 me permitía suponer que la entropía es buena candidata para descubrir la base de una palabra. Sin embargo, observé que su efectividad es variable y obtiene mejores resultados en palabras con sufijos derivativos.

Véase por ejemplo el inciso (a) de la Tabla 4.5, donde el valor máximo de entropía calculado para la palabra NIÑO propone la segmentación N~IÑO. Esto se debe a la relación entre este palabra y palabras como CARIÑO, GUIÑO, LAMPIÑO o PATIÑO; sin embargo, ésta no es una segmentación correcta (~IÑO no es un sufijo derivativo en el palabra NIÑO). En cambio, para el palabra DEFINICIÓN, véase (b) de la misma tabla, el valor máximo de entropía sí separa correctamente el sufijo derivativo de la base.

Tabla 4.5 Medidas de entropía para las palabras NIÑO y DEFINICIÓN

N	I	Ñ	O
	1.957	1.281	0.8643

(a)

D	E	F	I	N	I	C	I	Ó	N
	0	0	0	1.735	**2.061**	1.05	0.6541	1.336	1.468

(b)

Pongo en seguida otro ejemplo. A pesar de que en (a) de la Tabla 4.6 la entropía propone la segmentación equivocada C~ANCIÓN, en (b) es capaz de separar la raíz de la flexión verbal (CANT~AREMOS) y en (c) la base del sufijo derivativo (VENG~ANZA).

Tabla 4.6 Entropías para CANCIÓN, CANTAREMOS y VENGANZA

C	A	N	C	I	Ó	N
	1.561	1.132	1.05	0.6541	1.336	1.468

(a)

C	A	N	T	A	R	E	M	O	S
	1.895	0.9992	2.179	**2.713**	1.414	1.52	0.8699	1.216	1.301

(b)

V	E	N	G	A	N	Z	A
	0	0	1.099	**2.468**	0.878	1.496	1.061

(c)

Ahora bien, si el antecedente A.1 proponía la medida de cuadros como validadora de una segmentación, entonces ¿será posible evitar los cortes equivocados que propone la entropía si se combina con la medida de cuadros? Como se puede ver en (a) de la Tabla 4.7, el valor máximo de entropía coincide con un valor muy pequeño de cuadros y es el valor máximo de la medida de cuadros el que propone la segmentación correcta NIÑ~O.

Tabla 4.7 Medidas de afijalidad para NIÑO y CANCIÓN

	N	I	Ñ	O
Entropía		**1.957**	1.281	0.8643
Cuadros		21	0	**242312**
Economía		0.8095	0	**0.9992**

(a)

	C	A	N	C	I	Ó	N
Entropía		**1.561**	1.132	1.05	0.6541	1.336	1.468
Cuadros		0	0	0	0	**655**	0
Economía		0	0	0	0	**0.9252**	0

(b)

En el caso de (b) de la Tabla 4.7, el valor máximo de entropía coincide con un valor cero de cuadros. Si combinara estas medidas con un producto, se cancelaría el corte propuesto por la entropía. Hasta el momento las principales intuiciones que se obtiene son que tal vez sea posible descubrir las bases de las palabras con una combinación de entropía y cuadros, y que tal vez conviene multiplicar las medidas.

La Tabla 4.7 también mostró que los valores máximos de las medidas de cuadros y economía coinciden. Esto no sucede siempre, como expongo a continuación, pero el comportamiento de estas medidas deja entrever una posibilidad interesante. Pude observar que cuando los valores máximos de las dos medidas coinciden se trata generalmente de sufijos flexivos o sufijos de verboides (que son muy productivos), véanse los ejemplos de la Tabla 4.8.

Tabla 4.8 Medidas de afijalidad para ELIMINAR y NIÑOS

	E	L	I	M	I	N	A	R
Entropía	0.6365	0.6931	2.02	2.462	1.647	**2.72**	1.068	
Cuadros	0	0	40	0	6	**291188**	11244	
Economía	0	0	0	0	0.8333	**0.9463**	0.8972	

(a)

	N	I	Ñ	O	S
Entropía	**1.643**	1.32	1.216	1.301	
Cuadros	0	0	230703	**253968**	
Economía	0	0	0.9992	**1**	

(b)

En el caso de la derivación, vi de manera recurrente que los valores máximos de estas dos medidas no coincidían, por ejemplo, esto sucedió sistemáticamente con sufijos derivativos como –ACIÓN, –ANTE y –ANZA como lo muestro en los casos de la Tabla 4.9.

Tabla 4.9 Medidas de afijalidad para ELIMINACIÓN, DIBUJANTE y CONFIANZA

	E	L	I	M	I	N	A	C	I	Ó	N
Entropía	0	0.6365	1.777	1.855	1.55	**2.453**	1.05	0.6541	1.336	1.468	
Cuadros	0	0	12	0	2	**71548**	1094	0	388	0	
Economía	0	0	0	0	0.5	0.7816	0	0	**0.9974**	0	

(a)

	D	I	B	U	J	A	N	T	E
Entropía	0	0	1.099	1.004	**2.652**	0.5838	1.781	1.254	
Cuadros	0	0	0	0	**36238**	2233	6499	0	
Economía	0	0	0	0	0.5688	0	**0.9994**	0	

(b)

	C	O	N	F	I	A	N	Z	A
Entropía	0	0	1.055	1.149	**2.468**	0.878	1.496	1.061	
Cuadros	0	0	0	0	**2993**	0	22	0	
Economía	0	0	0	0	0	0	**0.9545**	0	

(c)

Como puede observarse, es la medida de cuadros la que propone el corte en los sufijos derivativos. En cambio, la medida de economía propone un tipo de sufijo más generalizador, un segmento más frecuente y regular. Piénsese, por ejemplo, en que el segmento

~ÓN es una parte constante de los segmentos ~ACIÓN, ~CIÓN, ~ICIÓN, ~IÓN y ~UCIÓN. Ya el antecedente A.3 describía esto cuando indicaba que en el caso de los sufijos, la economía asociada a un corte será más alta si el segmento del lado derecho es muy frecuente y pertenece a un conjunto relativamente pequeño de segmentos.

Sin embargo, la falta de coincidencia de corte entre las medidas de cuadros y economía también se da en casos de flexión verbal, véase por ejemplo la Tabla 4.10, lo que no permite generalizar una manera de distinguir automáticamente flexión de derivación. La relación entre estas dos medidas y la capacidad de usarlas para distinguir entre tipos de sufijos podrían ser evaluadas en futuras investigaciones.

Tabla 4.10 Medidas de afijalidad para la palabra CANTAREMOS

	C	A	N	T	A	R	E	M	O	S
Entropía		1.895	0.9992	2.179	**2.713**	1.414	1.52	0.8699	1.216	1.301
Cuadros		3	0	303	**274560**	9016	12750	25991	0	0
Economía		0	0	0.1551	0.9296	0.8136	0.9385	**0.9994**	0	0

Otra de las intuiciones obtenidas de esta exploración es que diferentes combinaciones de las medidas de afijalidad pueden dar buenos resultados. Así, tal vez es posible identificar una base combinando, por ejemplo, entropía y cuadros, y descubrir los afijos con entropía y economía.

4.2.4. Intuiciones sobre la segmentación morfológica

En esta sección plasmo las intuiciones que sirvieron de guía para la realización de los experimentos de segmentación. En ellas están involucrados todos los aspectos discutidos en los apartados anteriores. Esto es, tomo en cuenta las estrategias de segmentación y los antecedentes sobre el cálculo de la afijalidad; además, involucro las observaciones que surgieron de la exploración de medidas calculadas para algunas palabras.

Es importante recordar que en mi trabajo de investigación el descubrimiento de unidades morfológicas está restringido a bases y sufijos, dejando el fenómeno de prefijación para trabajo futuro. Por lo anterior, asumo de manera simplista que la base está al comienzo de la palabra.

Agrupo las intuiciones de acuerdo con tres ideas generales que muestro a continuación.

1. Utilizar una combinación de medidas para descubrir bases (por ejemplo, la combinación de entropía y cuadros) y otra distinta para descubrir sufijos (por ejemplo, entropía y economía).
2. Utilizar la misma combinación de medidas (por ejemplo, sólo la combinación de entropía y cuadros) para descubrir tanto bases como sufijos.
3. Volver a calcular las medidas de afijalidad de la palabra después de descubrir cada segmento, por ejemplo, cortar todas las palabras en dos segmentos y tomar ya sea el de la derecha o el de la izquierda como la entrada de una nueva medición de afijalidades.

La primera idea propone el uso de ciertas medidas para descubrir la base y de otras para descubrir los sufijos. Responde a la curiosidad por saber si es posible caracterizar de manera distinta a estas unidades morfológicas utilizando las medidas de afijalidad.

Primeramente, de los antecedentes se sabe que la medida de entropía es buena candidata para descubrir bases (antecedente A.2); aunque, como se vio en la exploración de medidas, su efectividad es variable. Entonces, se puede pensar que la combinación de ésta con la medida de cuadros, que sirve como medida de validez de una segmentación (antecedente A.1), permitirá descubrir bases de manera más precisa.

Ahora bien, asumiendo que se descubre la base, es necesario determinar los cortes para los sufijos. Si la medida de economía está más asociada a unidades más afijales (antecedente A.3) y su combinación con la medida de entropía dio buenos resultados para el descubrimiento de sufijos en experimentos anteriores de un solo corte por palabra (antecedente A.5), entonces se puede proponer el uso de esta combinación de medidas para determinar los cortes para cada sufijo. De esta manera surge la siguiente intuición, representada gráficamente por la Figura 4.5.

I1. Es posible descubrir la base de una palabra mediante la combinación de las medidas de entropía y cuadros, y después descubrir los sufijos mediante la combinación de las medidas de economía y entropía.

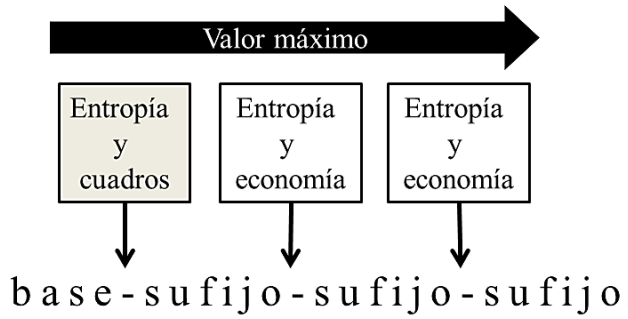


Figura 4.5 Utilizar entropía y cuadros para descubrir la base

Al reflexionar sobre la intuición anterior, es posible preguntarse si la estrategia contraria puede ser plausible. Es decir, descubrir primero el sufijo más a la derecha (más afijal) y luego proponer cortes para obtener los restantes sufijos hasta llegar al corte correspondiente a la base. Esta idea involucra la combinación de las mismas medidas pero en orden inverso. Surge entonces la siguiente intuición, representada gráficamente en la Figura 4.6.

I2. Es posible descubrir el sufijo más a la derecha mediante la combinación de las medidas de entropía y economía, y después descubrir los sufijos restantes hasta encontrar la base mediante la combinación de las medidas de entropía y cuadros.

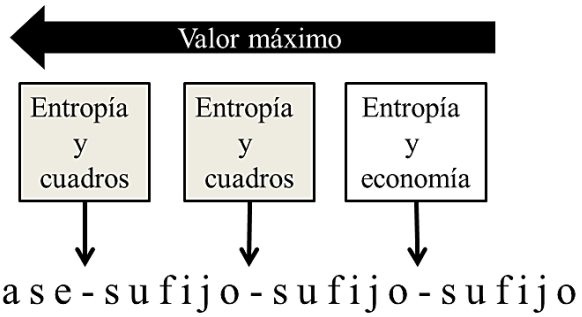


Figura 4.6 Utilizar entropía y economía para descubrir último sufijo

Además, sería pertinente experimentar con un cambio en la manera de combinar las medidas, esto es, pienso que sería bueno multiplicarlas ya que un promedio entre ellas siempre dará como resultado cierta cantidad de afijalidad aunque haya valores de cero para la medida de cuadros. Así, al multiplicarlas si la cantidad de cuadros es cero, el corte será anulado. Entonces para los experimentos usé tanto el valor normalizado del producto de las medidas, como el promedio de valores normalizados.

La segunda idea es que una sola combinación de medidas de afijalidad (por ejemplo, entropía y economía) basta para segmentar una palabra en todas las unidades posibles (bases y sufijos); en contraste con utilizar una combinación de medidas para determinar la base y otra para determinar los sufijos.

El antecedente de que la afijalidad más alta describe cuantitativamente a las unidades más afijales (antecedente A.4) lleva a pensar que el valor máximo de afijalidad de una palabra correspondería al último sufijo: el sufijo más flexivo, y por tanto más afijal. En una estrategia de segmentación basada en esta idea, se harían cortes sucesivos hacia la izquierda de la palabra, donde el último corte descubriría la base, lo que significa descubrirla únicamente por su posición sin distinguirla de los sufijos usando sus características cuantitativas.

Sin embargo, también se asumió que el valor máximo de afijalidad no descubre el último sufijo de la palabra, sino que permite separar la base del resto de los sufijos aglutinados. Luego, con cortes sucesivos hacia la derecha, sería posible separar estos sufijos.

Se utilizó también una estrategia de segmentación basada en todos los valores de afijalidad mayores a 0.5. En ésta no es pertinente la dirección hacia donde se hace la segmentación y sería de esperarse que hubiera más segmentos por palabra que en las otras estrategias. Con el fin de ejemplificar las tres estrategias anteriores véase la Tabla 4.11 Medidas de afijalidad para la palabra CANTEN, si la palabra fuera segmentada en valores máximos (sólo mayores a 0.5) hacia la izquierda a partir del primer valor máximo, se obtendría la segmentación CAN~T~EN. Si el procedimiento fuera hacia la derecha, la segmentación sería CANT~E~N. Luego, si se toman todos los valores mayores a 0.5 la segmentación sería CAN~T~E~N dando como resultado cuatro segmentos y no tres.

Tabla 4.11 Medidas de afijalidad para la palabra CANTEN

C	A	N	T	E	N
	0.2861	0.1399	0.5291	**0.9185**	0.6077

Las intuiciones resultantes de lo expuesto en los párrafos anteriores se expresan a continuación, cada una incluye una representación gráfica.

I3. El valor máximo de afijalidad de una palabra permite descubrir el último sufijo y luego los sufijos restantes hasta encontrar la base.

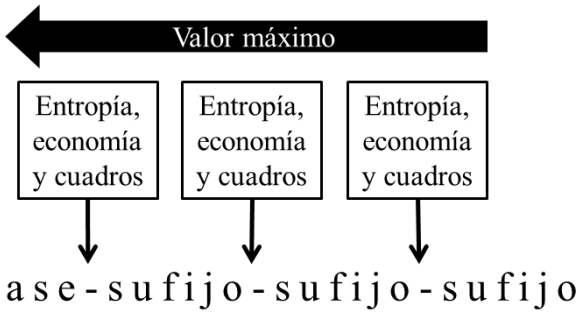


Figura 4.7 Valor máximo de afijalidad para descubrir sufijos y luego base

I4. El valor máximo de afijalidad de una palabra permite descubrir la base y luego los sufijos restantes hasta encontrar el último sufijo.

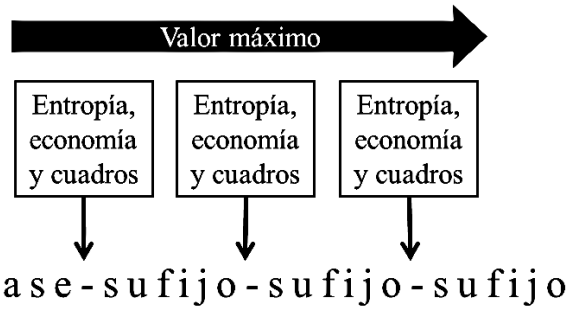


Figura 4.8 Valor máximo de afijalidad para descubrir base y luego sufijos

I5. Los valores de afijalidad de una palabra mayores a 0.5 permiten descubrir la base y los sufijos.

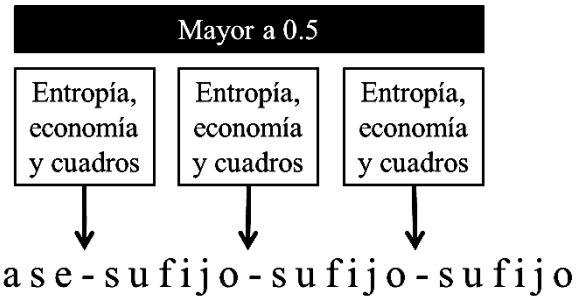


Figura 4.9 Afijalidad mayor a 0.5 para descubrir base y sufijos

Para los experimentos que ayudarán a comprobar estas intuiciones, es necesario probar con dos maneras de calcular la afijalidad. Debe recordarse que en las investigaciones

anteriores se había calculado este índice con la combinación de las tres medidas (entropía, economía y cuadros) y con la combinación sólo de la medida de economía y de entropía (antecedentes A.4 y A.5). Además, los experimentos contemplan dos formas de combinar las medidas de afijalidad: mediante un promedio y mediante un producto.

Otra idea diferente sería pensar en cortar la palabra en el valor máximo de afijalidad y volver a calcular las medidas de afijalidad de los segmentos resultantes, ya sea el de la derecha, el de la izquierda o ambos. Si se decide volver a calcular las medidas de afijalidad sólo para el segmento del lado derecho, se estaría asumiendo que el primer corte descubrió la base. Si se decide volver a calcular las medidas de afijalidad sólo para el segmento de la izquierda, se asumiría que el primer corte descubrió el sufijo más externo de la palabra.

Se trata de una estrategia basada también en los valores máximos de afijalidad, pero con la diferencia de que se estaría tomando siempre el primer valor máximo para proponer un corte, en comparación con las estrategias anteriores, en las que se toma el primer valor máximo, luego el segundo valor máximo y así sucesivamente.

La intuición derivada de las ideas anteriores se expone a continuación. Debajo está una representación gráfica.

I6. El procedimiento recursivo de segmentar en el valor máximo de afijalidad y volver a calcular las medidas de afijalidad para los segmentos resultantes permite descubrir la base y los sufijos de una palabra.

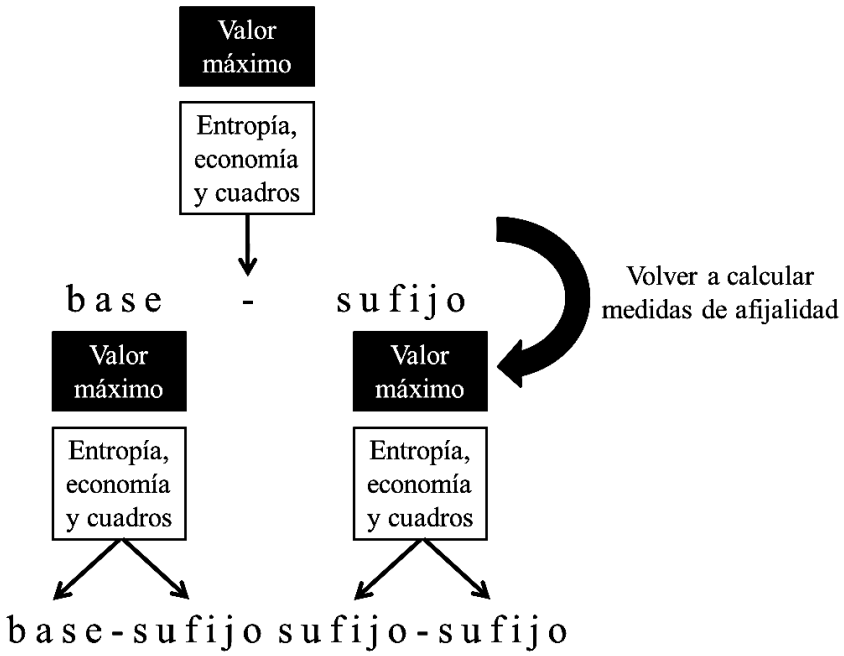


Figura 4.10 Procedimiento recursivo para descubrir bases y sufijos

Se dejará para trabajo futuro la comprobación de esta intuición, por lo cual no habrá experimentos relacionados con ella. En el apartado siguiente se describen los experimentos que realicé en lo que a la segmentación morfológica se refiere.

4.2.5. Experimentos

Como parte de las ideas expuestas en los apartados anteriores, se pueden identificar ciertas condiciones involucradas en una posible estrategia de segmentación morfológica. En seguida expongo estas condiciones y después los experimentos.

Dado que la afijalidad puede ser calculada de distintas maneras, es posible determinar ciertas condiciones implicadas en dicho cálculo. Por un lado se encuentran las medidas combinadas (entropía, economía y cuadros) o, mejor dicho, qué combinación hacer de ellas. Por ejemplo, combinar entropía con cuadros, o combinar entropía con economía, o combinar las tres.

Por otro lado, es posible cambiar la manera de combinarlas, ya que se puede utilizar un promedio, un producto, etcétera[62]. En el caso de un producto de medidas, si una de ellas tiene valor cero, todo el cálculo será cero y el posible corte será anulado sin importar que las otras medidas sean muy altas. Por el contrario, si se utiliza un promedio, siempre se obtendrá un valor aunque alguna de las medidas tenga valor de cero.

Tómense como ejemplo los cálculos mostrados en la Tabla 4.12. En este caso el índice de afijalidad muestra el promedio de valores normalizados. Como se trata de un promedio de las tres medidas, siempre hay un valor de afijalidad, a pesar de que hay muchos valores de cero para las medidas de cuadros y economía. Si la manera de combinar las medidas fuera con un producto, la medida de afijalidad sería cero para todos los cortes, excepto el corte en ~ÓN.

Tabla 4.12 Medidas de afijalidad para la palabra CANCIÓN

	C	A	N	C	I	Ó	N
Entropía		**1.561**	1.132	1.05	0.6541	1.336	1.468
Cuadros		0	0	0	0	**655**	0
Economía		0	0	0	0	**0.9252**	0
Afijalidad		0.3333	0.2418	0.2243	0.1397	**0.9519**	0.3136

Una vez decidida la manera de calcular la afijalidad para una palabra, es necesario decidir también dónde hacer los cortes para la segmentación. De manera muy general, pueden observarse dos posibilidades: (i) segmentar cuando los valores superen un umbral, y (ii) segmentar en los valores más altos (valores máximos). Como se había mencionado, conviene tomar en cuenta los valores más altos ya que la afijalidad permite calcular qué tanto un segmento es un afijo, por lo que a mayor afijalidad, mayor cualidad de afijo; aunque puede haber otras posibilidades no contempladas aquí.

[62] Una discusión sobre la pertinencia de combinar estas medidas y de cómo hacerlo se puede encontrar en Medina (2003).

Si se decide segmentar el palabra en los valores más altos, es necesario poner un umbral que limite lo que se considera "alto", de lo contrario habrá sobresegmentación ya que un valor muy pequeño puede ser el más alto entre valores aún más pequeños. Ejemplificaré a qué me refiero tomando la idea de segmentar una palabra con cortes sucesivos hacia la izquierda en los valores máximos de afijalidad.

En la Tabla 4.13 se observa que el primer valor máximo de afijalidad es 0.8298, lo que propondría un corte en ALARMANTE~S. Luego, de los valores restantes, el máximo es 0.536, con un segundo corte en ALARMAN~TE~S. Después, el siguiente valor máximo es 0.5021, con corte en ALARM~AN~TE~S. Si no hay un umbral para los valores más altos, el procedimiento sigue y los nuevos valores máximos serían 0.3634 (ALAR~M~AN~TE~S) y finalmente 0.1738 (ALA~R~M~AN~TE~S).

Tabla 4.13 Medidas de afijalidad para la palabra ALARMANTES

A	L	A	R	M	A	N	T	E	S
	0	0	0.1738	0.3634	0.5021	0.1061	0.536	0.07867	0.8298

Como se pudo observar, sin un umbral, el procedimiento de cortes sucesivos sobresegmentaría las palabras. Es importante entonces combinar la estrategia de valores más altos con la imposición de un umbral que determine qué valores máximos serán seleccionados. Para todos los experimentos que realicé, puse un umbral de 0.5.

Como ya se dijo, existen también distintas maneras de usar los valores máximos de una palabra. Una de ellas estaría basada en la direccionalidad, esto es, tomar los valores máximos hacia la derecha o hacia la izquierda. El procedimiento consistiría en cortar la palabra en el primer valor más alto de afijalidad y luego cortar en los valores más altos en una dirección: izquierda o derecha. Las diferencias de segmentación entre una dirección y otra se pueden ver en el siguiente ejemplo.

Para la segmentación de la palabra NIÑOS (Tabla 4.14), primero se toma el valor máximo de afijalidad (0.9305), que da como resultado la segmentación NIÑO~S. Si el procedimiento estuviera basado en una direccionalidad hacia la izquierda, el segundo corte sería en NIÑ~O~S. En cambio, si el procedimiento estuviera basado en una direccionalidad hacia la derecha, ya no habría otro corte.

Tabla 4.14 Índice de afijalidad para la palabra NIÑOS

N	I	Ñ	O	S
	0.3333	0.2678	0.8824	0.9305

No está por demás decir que hay muchas otras posibles qué considerar en un procedimiento de segmentación basado en las medidas de afijalidad. Por ejemplo, se puede pensar en usar otras medidas o distintas maneras de combinarlas, pero ya no se tomarán en cuenta. A manera de resumen, la Tabla 4.15 contiene las condiciones involucradas en una posible estrategia de segmentación y las realizaciones de estas condiciones.

Tomé en cuenta todas las posibles combinaciones de esas condiciones, lo que arrojó un conjunto de dieciséis experimentos; algunos con menor o mayor probabilidad de ser exitosos. Estos experimentos se resumen en la Tabla 4.16.

No está de más recordar que las investigaciones anteriores habían desarrollado un programa de computadora que calcula las medidas de afijalidad de las palabras de un corpus para generar un catálogo de afijos. Modifiqué este programa para realizar los experimentos y obtener un conjunto de tipos de palabras segmentados.

Tabla 4.15. Condiciones involucradas en la segmentación

Condición	Explicación	Realización
Medidas combinadas	Involucra una selección de qué medidas serán combinados para calcular la afijalidad.	1) Entropía y economía. 2) Entropía y cuadros. 3) Entropía, economía y cuadros.
Manera de combinación	Involucra las operaciones matemáticas que combinarán las medidas para calcular la afijalidad.	1) Promedio de índices normalizados. 2) Multiplicación de medidas y normalización del producto.
Direccionalidad de la segmentación	Involucra la dirección hacia donde se toman los valores máximos una vez que se ha determinado el primer corte.	1) Derecha. 2) Izquierda.
Valor en el que se segmenta	Involucra la condición que debe cumplir el valor de afijalidad para decidir si se segmenta la palabra.	1) Mayor a 0.5. 2) En el valor máximo

Fue necesaria una evaluación de las segmentaciones generadas por cada experimento. Esto me llevó a optar por una estrategia de evaluación que detallo en la sección subsecuente.

Tabla 4.16. Experimentos de segmentación realizados

#	Combinación de medidas	Manera de combinarlas	Direccionalidad	Segmentar en	Descripción
1	Entropía-cuadros y entropía-economía	Promedio de valores normalizados	Derecha	Valor máximo	Segmentar en el valor máximo del promedio de valores normalizados de entropía y cuadros. Luego segmentar hacia la derecha en los valores máximos del promedio de valores normalizados de entropía y economía.
2	Entropía-cuadros y entropía-economía	Producto normalizado	Derecha	Valor máximo	Segmentar en el valor máximo del producto normalizado de entropía y cuadros. Luego segmentar hacia la derecha en los valores máximos del producto normalizado de entropía y economía.
3	Entropía-economía y entropía-cuadros	Promedio de valores normalizados	Izquierda	Valor máximo	Segmentar en el valor máximo del promedio de valores normalizados de entropía y economía. Luego segmentar hacia la izquierda en los valores máximos del promedio de valores normalizados de entropía y cuadros.
4	Entropía-economía y entropía-cuadros	Producto normalizado	Izquierda	Valor máximo	Segmentar en el valor máximo del producto normalizado de entropía y economía. Luego segmentar hacia la izquierda en los valores máximos del producto normalizado de entropía y cuadros.

Tabla 4.16. Experimentos de segmentación realizados (continuación)

#	Combinación de medidas	Manera de combinarlas	Direccionalidad	Segmentar en	Descripción
5	Entropía, economía y cuadros	Promedio de valores normalizados	Izquierda	Valor máximo	Segmentar en el valor máximo del promedio de valores normalizados de entropía, economía y cuadros. Luego segmentar hacia la izquierda en los valores máximos del mismo promedio de valores.
6	Entropía, economía y cuadros	Promedio de valores normalizados	Derecha	Valor máximo	Segmentar en el valor máximo del promedio de valores normalizados de entropía, economía y cuadros. Luego segmentar hacia la derecha en los valores máximos del mismo promedio de valores.
7	Entropía, economía y cuadros	Producto normalizado	Izquierda	Valor máximo	Segmentar en el valor máximo del producto normalizado de entropía, economía y cuadros. Luego segmentar hacia la izquierda en los valores máximos del mismo producto normalizado.
8	Entropía, economía y cuadros	Producto normalizado	Derecha	Valor máximo	Segmentar en el valor máximo del producto normalizado de entropía, economía y cuadros. Luego segmentar hacia la derecha en los valores máximos del mismo producto normalizado.

Tabla 4.16. Experimentos de segmentación realizados (continuación)

#	Combinación de medidas	Manera de combinarlas	Direccio-nalidad	Segmentar en	Descripción
9	Entropía y economía	Promedio de valores nor-malizados	Izquierda	Valor máximo	Segmentar en el valor máximo del promedio de valores normalizados de entropía y economía. Luego segmentar hacia la izquierda en los valo-res máximos del mismo promedio de valores.
10	Entropía y economía	Promedio de valores nor-malizados	Derecha	Valor máximo	Segmentar en el valor máximo del promedio de valores normalizados de entropía y economía. Luego segmentar hacia la derecha en los valo-res máximos del mismo promedio de valores.
11	Entropía y economía	Producto normalizado	Izquierda	Valor máximo	Segmentar en el valor máximo del producto normalizado de entropía y economía. Luego segmentar hacia la izquierda en los valores máximos del mismo producto normalizado.
12	Entropía y economía	Producto normalizado	Derecha	Valor máximo	Segmentar en el valor máximo del producto normalizado de entropía y economía. Luego segmentar hacia la derecha en los valores má-ximos del mismo producto normalizado.

Tabla 4.16. Experimentos de segmentación realizados (continuación)

#	Combinación de medidas	Manera de combinarlas	Direccio-nalidad	Segmentar en	Descripción
13	Entropía, economía y cuadros	Promedio de valores nor-malizados		Mayor 0.5	Segmentar cuando el promedio de valores nor-malizados de entropía, economía y cuadros sea mayor a 0.5.
14	Entropía, economía y cuadros	Producto normalizado		Mayor 0.5	Segmentar cuando el producto normalizado de entropía, economía y cuadros sea mayor a 0.5.
15	Entropía y economía	Promedio de valores nor-malizados		Mayor 0.5	Segmentar cuando el promedio de valores nor-malizados de entropía y economía sea mayor a 0.5.
16	Entropía y economía	Producto normalizado		Mayor 0.5	Segmentar cuando el producto normalizado de entropía y economía sea mayor a 0.5.

4.3. Evaluación de la segmentación automática

En esta sección describo la estrategia que utilicé para evaluar los experimentos de segmentación morfológica y los resultados obtenidos. Gracias a la revisión que hice de un conjunto de fuentes documentales sobre morfología del español (consignada en la sección 1.3), recopilé 1,600 palabras segmentadas de acuerdo con cada fuente. Este conjunto de palabras constituyó el corpus de evaluación (evaluación intrínseca)[63]. Luego segmenté automáticamente las mismas palabras de acuerdo con cada experimento y comparé la segmentación obtenida contra la segmentación propuesta por las fuentes documentales mediante una evaluación estricta.

Para calcular las medidas de afijalidad de las palabras del corpus de evaluación utilicé un corpus con 965,565 tipos de palabras formado por una lista de palabras proporcionada por el Laboratorio de Lenguaje Natural y Procesamiento de Texto del IPN (Gelbukh y Sidorov, 2003) utilizada en un sistema de análisis morfológico automático supervisado, los vocablos del Diccionario del Español de México (2010) y los tipos de palabras del CEMC.

El resultado de la evaluación de cada experimento fue el número de palabras segmentadas automáticamente que coincidieron con la segmentación propuesta por las fuentes documentales. En términos generales, el mejor experimento fue el que obtuvo mayor número de coincidencias. La constitución del corpus se describe en el apartado 4.3.1, y los resultados y conclusiones de la evaluación en el 4.3.2.

[63] El corpus está disponible en http://www.corpus.unam.mx/morfotactica/corpusEvalSeg.csv.

4.3.1. Constitución del corpus de evaluación

Como se mencionó, el corpus de evaluación consta de 1,600 palabras tomadas de distintas fuentes. Fueron seleccionadas palabras tanto del fenómeno de flexión como de derivación, únicamente nominal y verbal. Los porcentajes de cada fenómeno del corpus se muestran en la Tabla 4.17. Luego, las fuentes usadas para obtener los ejemplos de cada fenómeno se enlistan en la Tabla 4.18.

Tabla 4.17 Porcentajes de cada fenómeno en el corpus de evaluación

Fenómeno	Palabras	Porcentaje
Flexión nominal	76	5%
Derivación nominal	855	53%
Flexión verbal	490	31%
Derivación verbal	180	11%

Tabla 4.18 Fuentes utilizadas para el corpus de evaluación

Fenómeno	Fuente
Flexión nominal	Ambadiang (1999)
Flexión verbal	Alcoba (1999) y DEM[64]
Derivación nominal	Moreno de Alba (1986)
Derivación verbal	Beniers (2004)

Para la constitución del corpus, escogí mayoritariamente palabras cuyos fenómenos de flexión y derivación fueran regulares (83% del total de palabras). Es decir, aquellas palabras cuyas bases no sufrían modificaciones de forma como producto del fenómeno morfo-

[64] Diccionario del Español de México (DEM) http://dem.colmex.mx, El Colegio de México, A.C., [15 de noviembre de 2012].

lógico, por ejemplo ***ELIMINA***R/***ELIMINA***CIÓN. En el caso particular de la flexión verbal, tomé como regulares los ejemplos de las conjugaciones regulares.

Ya que el método de segmentación automática está basado en coincidencias de segmentos, esto es, busca regularidades en la forma de las palabras, es de esperarse que trabaje mejor para fenómenos de flexión y derivación que no afecten la forma de la base.

Sin embargo, a manera de experimento, incluí también ejemplos de fenómenos de flexión y derivación irregulares, esto es, donde la base o sufijo sufrían algún cambio de forma (A***G***UA/A***C***UÁTICO) o los verbos pertenecían a los modelos de conjugación irregular. En los apartados subsecuentes abundo en la constitución del corpus.

4.3.1.1. Flexión nominal

Para la flexión de género incorporé pares de palabras muy regulares en su marcación, ya sea por la alternancia de –o/–a o de –e/–a (GAT***O***/GAT***A***, JEF***E***/JEF***A***). También incluí pares que no son equivalentes semánticamente, pero que presentan alternancia de sufijo, como es el caso de MANZAN***O***/MANZAN***A***; tomé como regulares estos casos, ya que en la forma lo son.

El grupo de irregulares lo constituyeron palabras que cambian toda su forma cuando alternan género, como PADRE/MADRE. También los que Ambadiang llama formas únicas o unidades léxicas individuales, que no tienen alternante de género, aunque sí cuentan con sufijo –o/–a/–e (VÁSTAG***O***, VÍCTIM***A***, LUMBR***E***). Además, los sustantivos cuyo género está marcado en alguna palabra que los acompaña, como TESTIGO (***EL*** TESTIGO/***LA*** TESTIGO). Finalmente, incluí algunas palabras que no cuentan con el sufijo de género y en su lugar hay terminación vocálica (TRIB***U***), o consonántica (PARE***D***, VIRU***S***, CLIMA***X***).

En lo referente a la flexión de número, agregué los plurales de algunas palabras, tanto regulares como irregulares, seleccionadas para representar el género (GATOS, GATAS, PADRES, MADRES, PAREDES). Para darle variedad al corpus, fueron seleccionadas palabras plurales que se forman a partir de un singular que termina en vocal acentuada –é (BEB***É***S), y aquellas cuyo singular termina con consonantes como –l, –n, –d y –s (PASTE***L***ES, ALGODO***N***ES, VERDA***D***ES, COMPA***S***ES).

A pesar de que los plurales fueron formados a partir de palabras marcadas como irregulares para el género (PARED), no los marqué como irregulares para el número, ya que se forman con la adición de –(e)s, que es la manera regular de formar el plural (PARED***ES***). Las que sí quedaron marcadas como irregulares fueron palabras cuyo segmento final coincidía con las marcas de número, pero en las cuales no debería haber segmentación (ANÁLISIS, MARTES).

Tanto para el género como para el número, se incorporaron palabras formadas a partir de derivaciones nominales (AGRUPACIONES, JOVENCITOS, DEFECTUOSOS, DEFECTUOSAS, HERMOSÍSIMOS). Finalmente comento que agregué cinco ejemplos de palabras en plural que funcionan como adjetivos o pronombres (ESTAS, ESTOS, NUESTROS, NUESTRAS, NOSOTROS) y el plural de una conjunción (PEROS).

4.3.1.2. Flexión verbal

En el caso de la flexión verbal, tomé dos fuentes para poner ejemplos en el corpus de evaluación (véase arriba Tabla 4.18). Ya que cada fuente propone distinta segmentación, y ambas propuestas me parecen válidas, tomé en cuenta las dos. De esta manera, si el método automático coincidía con alguna de las posibles segmentaciones, entonces se consideraba como un acierto. En seguida comento las distinciones entre las dos fuentes.

La primera fuente (Alcoba) propone separar la vocal temática de la raíz verbal, por ejemplo, TEM~IE~NDO, CANT~A~MOS; además, propone segmentar los morfemas de tiempo-aspecto-modo y de número-persona, por ejemplo, CANT~Á~BA~MOS. Por su parte, el DEM propone dejar la vocal temática unida al sufijo flexivo, por ejemplo, COM~IENDO, AM~AMOS; y no separa los morfemas finales de los verbos, por ejemplo, AM~ÁBAMOS.

Para el caso del futuro de indicativo, el DEM formula una segmentación del tipo AMAR~ÁS, que da muestra del fenómeno histórico de formación de este tiempo. En cambio, la primera fuente propone AM~A~RÁ~S, que corresponde a una segmentación de la vocal temática y los morfemas de tiempo-aspecto-modo y número-persona. Lo mismo sucede para el pospretérito, donde el DEM plantea una segmentación como CANTAR~ÍAMOS y Alcoba una como CANT~A~RÍA~MOS.

En otros aspectos, como ya lo decía, los verbos concernientes a los modelos de conjugación regular fueron marcados como regulares en el corpus de evaluación, mientras que los marcados como irregulares fueron los que pertenecen a los modelos irregulares. Tomé seis verbos regulares, tres de cada fuente, de Alcoba: TEMER, PARTIR y CANTAR; y del DEM: COMER, SUBIR y AMAR. Incorporé las formas no personales y todas las conjugaciones de estos verbos, con excepción de las formas compuestas.

Para los verbos irregulares me basé en los modelos de conjugación irregular del DEM. De ellos elegí los siguientes verbos, indico entre paréntesis el modelo de conjugación: AGRADECER (1a), CAER (1d), DESPERTAR (2a), ADQUIRIR (2b), SOÑAR (2c), JUGAR (2d), MEDIR (3a), CONSTRUIR (4), ANDAR (5), PRODUCIR (7a), CABER (10a), QUERER (11a), TENER (12a), VENIR (12b), DECIR (13), IR (19).

Por cada uno de los verbos irregulares seleccioné ejemplos de conjugación sólo de un tiempo, aquel donde aparecía alguna irregularidad, por ejemplo, de AGRADECER tomé AGRADEZCA, AGRADEZCAS, AGRADEZCAMOS, AGRADEZCAN, AGRADEZCÁIS. Descarté los verbos marcados por el DEM como poco usados y cuando la raíz de las conjugaciones resultaba menor de tres letras (CA~Í).

4.3.1.3. Derivación nominal

Basé la selección de ejemplos de derivación nominal en la obra de Moreno de Alba (1986). Ésta incluye tanto sustantivos como adjetivos derivados. Fueron incorporados un total de 885 ejemplos correspondientes a 188 sufijos derivativos[65]. En la Tabla 7.1 del anexo A presento la lista completa de sufijos derivativos usados para obtener los ejemplos

Para darle variedad al corpus, por cada alomorfo tomé ejemplos de los distintos tipos de derivación que consigna Moreno de Alba: palabras derivadas y palabras relacionadas (1986, págs. 15-16). A pesar de que mi estudio no involucra el carácter semántico de los sufijos, lo cual me impide descubrir el tipo de relación entre la palabra derivada o relacionada con su palabra base de derivación (primitiva), me pareció buena idea tomar en cuenta estos dos tipos.

En los casos de derivación a partir de verbos (ABURRIR/ABURRICIÓN) intenté, en la medida de lo posible, tomar ejemplos de las tres conjugaciones verbales. Además, en el caso de sufijos que forman derivados a partir de distintas clases de palabras, hice lo posible por incluir alguna palabra de cada una (IGUAL/IGUALDAD, VECINO/VECINDAD).

[65] Utilicé el capítulo V "Sufijos ordenados por su forma (alomorfos)" para tomar los ejemplos, aprovechando que este autor los presenta agrupados en alomorfos. Después revisé el capítulo II "Inventario de sufijos y voces derivadas" para incluir los sufijos faltantes que no estaban contemplados en el capítulo V.

Las palabras etiquetadas como irregulares en esta sección del corpus fueron aquellas en donde ocurría algún tipo de modificación en la base. Algunos ejemplos de éstas son: cambios en consonantes y vocales, muchas veces por que el derivado conserva la forma latina, A***G***UA/A***C***UÁTICO, DICTA***D***OR/DICTA***T***ORIAL, S***E***GUIR/S***I***GUIENTE, J***O***-VEN/J***U***VENTUD; eliminación de consonante INTERCEP**T**AR/INTERCEPCIÓN, AN-TECE**D**ER/ANTECESOR; diptongación PR***O***BAR/PR***UE***BA; cambio de acento F***A***BRI-CAR/F***Á***BRICA; monoptongación, algunas veces por influencia de la raíz latina T***IE***M-PO/T***E***MPORADA, SENTIM***IE***NTO/SENTIM***E***NTAL; cambio en el sufijo derivativo –iente, CONSTITUIR/CONSTITU***YENTE***; y otros MES/MENSUAL.

Por otro lado quiero resaltar que Moreno de Alba (1986) incluye un sufijo derivativo –V que coincide con las marcas de género –o/–a/–e (REPART~O, SIEMBR~A, INTÉRPRET~E). Estos ejemplos los etiqueté como derivación y no como flexión. Finalmente, como es de esperarse, en muchos casos se encontraron morfemas de género acompañando a los morfemas derivativos (CHIQUITITO); en esta situación marqué los ejemplos como de derivación y flexión.

4.3.1.4. Derivación verbal

De los ejemplos de Beniers (2004), tomé sólo de los sufijos derivativos –ear, –ecer, –ificar e –izar. No incluí derivados a partir del sufijo –ar ya que no hay distinción con la marca de infinitivo. En el caso de los sufijos sí considerados, esperaba una segmentación del tipo ~IFIC~AR.

La mayoría de las palabras fueron incluidas en infinitivo, pero seleccioné una de cada sufijo para conjugarla en los tiempos del indicativo: ARPONEAR, FLORECER, EJEMPLIFICAR y HORRORIZAR. Con la idea de darle variedad al corpus, tomé tanto ejemplos

de derivaciones que Beniers llama postsustantivas (HORROR/HORRORIZAR) como postadjetivas (ACTUAL/ACTUALIZAR).

Incorporé escasas palabras irregulares cuando aparecía una modificación en el derivado, como cambios consonánticos CHI***C***O/CHI***QU***EAR y cambios de acento HORROR/HORRORIC***É***. Ya que limité mi investigación únicamente al fenómeno de sufijación, evité cualquier palabra con parasíntesis, fenómeno común en este tipo de derivación (ATARDECER).

Para terminar este apartado comento que estos verbos derivados fueron segmentados de acuerdo con la propuesta de Alcoba (CHISMOS~E~A~R) y con la del DEM (CHISMOS~E~AR). Lo mismo hice para los verbos que fueron conjugados.

4.3.1.5. Enclíticos

Agregué al corpus de evaluación verbos con enclíticos para indagar qué sucedía en los experimentos. En seguida describo la estrategia que seguí para recolectar los ejemplos. Primero busqué en el CEMC los seis verbos regulares ya incluidos en el corpus de evaluación y tomé las formas con enclíticos. Obviamente no aparecieron todas las combinaciones posibles de clíticos con esos verbos, por lo que después busqué, también en el CEMC, las combinaciones faltantes de clíticos con cualquier otro verbo. Cuando existían, tomé ejemplos de las tres conjugaciones.

Intenté contar con al menos tres ejemplos por cada grupo de enclíticos, así que si no aparecían en el CEMC[66] utilicé el corpus de Mark Davies[67] para completarlos, evitando

[66] Diccionario del Español de México. *Corpus del Español Mexicano Contemporáneo* (CEMC). <http://www.corpus.unam.mx/cemc>, software AMATE ver. 1.0, [13/02/2013].

[67] Davies, Mark. (2002-) *Corpus del Español: 100 million words, 1200s-1900s*. Disponible en http://www.corpusdelespanol.org.

siempre verbos irregulares. Finalmente no aparecieron ejemplos de verbos con combinaciones de enclíticos: ~mele, ~meles, ~melas, ~tele, ~teles y ~nosles; por tanto no los tomé en cuenta.

Estas palabras con enclíticos fueron segmentadas de acuerdo con la propuesta de Alcoba y con la del DEM. Además, para cada una de ellas, agregué una segmentación separando las marcas de género y número de *les, los, las, le, la, lo* (Ambadiang, 1999). Lo anterior dejó cuatro posibles combinaciones de segmentaciones para infinitivos con enclíticos:

1. COM~É~R~NOS~LOS
2. COM~ÉR~NOS~LOS
3. COM~É~R~NOS~L~O~S
4. COM~ÉR~NOS~L~O~S.

En el apartado subsecuente describo los resultados obtenidos en los experimentos de segmentación.

4.3.2. Resultados de la evaluación

En este apartado se señalan los resultados de la evaluación de los experimentos de segmentación. En lo que respecta al total de palabras marcadas como regulares, en la Tabla 4.19 muestro las medidas de precisión alcanzadas por cada experimento. El método que obtuvo mejores resultados hace cortes sucesivos hacia la izquierda en el valor máximo del promedio de las tres medidas de afijalidad. Este método obtuvo un 33.8% de precisión.

Tabla 4.19 Medidas de precisión para palabras regulares

Combinación de medidas	Manera de combinarlas	Direccionalidad	Segmentar en	Precisión
Entropía, economía y cuadros	Promedio de valores normalizados	Izquierda	Valor máximo	33.8%
Entropía-economía y entropía-cuadros	Promedio de valores normalizados	Izquierda	Valor máximo	29.2%
Entropía y economía	Producto normalizado	Izquierda	Valor máximo	28.8%
Entropía, economía y cuadros	Promedio de valores normalizados		Mayor 0.5	28.8%
Entropía, economía y cuadros	Producto normalizado		Mayor 0.5	26.9%
Entropía y economía	Promedio de valores normalizados	Izquierda	Valor máximo	26.7%
Entropía, economía y cuadros	Producto normalizado	Izquierda	Valor máximo	26.7%
Entropía y economía	Producto normalizado		Mayor 0.5	25.9%
Entropía y economía	Promedio de valores normalizados		Mayor 0.5	25.7%
Entropía-economía y entropía-cuadros	Producto normalizado	Izquierda	Valor máximo	25.7%
Entropía, economía y cuadros	Producto normalizado	Derecha	Valor máximo	23.9%
Entropía y economía	Promedio de valores normalizados	Derecha	Valor máximo	20.7%
Entropía, economía y cuadros	Promedio de valores normalizados	Derecha	Valor máximo	20.7%
Entropía-cuadros y entropía-economía	Producto normalizado	Derecha	Valor máximo	20.5%
Entropía y economía	Producto normalizado	Derecha	Valor máximo	20.3%
Entropía-cuadros y entropía-economía	Promedio de valores normalizados	Derecha	Valor máximo	19.5%

El segundo lugar en precisión (29.2%) fue el experimento que propone determinar primero el sufijo más externo, con el promedio de las medidas de entropía y economía, y luego los demás sufijos hasta llegar a la base con el promedio de entropía y cuadros. Este experimento tuvo una diferencia mínima de precisión con los dos siguientes experimentos, que obtuvieron el 28.8%. Uno de ellos segmenta con cortes hacia la izquierda en el valor máximo del producto de entropía por economía, y el otro corta las palabras cuando el promedio de las tres medidas supera el valor de 0.5.

El experimento con resultados más bajos (19.5%) fue el que segmenta primero en el valor máximo del promedio de entropía y cuadros, y después corta sucesivamente hacia la derecha en el promedio de entropía y economía. De hecho, la variante de este mismo experimento que utiliza el producto de las medidas en lugar del promedio también logró bajos resultados: fue el antepenúltimo lugar.

Se puede ver claramente en la Tabla 4.19 que los métodos que cortan hacia la izquierda obtienen los mejores resultados, mientras que los que cortan hacia la derecha alcanzan los peores. Puede distinguirse también cierto predominio del uso de valores máximos sobre el uso de un umbral de 0.5, ya que los tres mejores experimentos segmentan en el valor máximo.

Sobre qué medidas combinar y la manera de hacerlo (producto o promedio) no veo una tendencia clara. Por el resultado de los dos primeros experimentos, parece plausible combinar las tres medidas mediante un promedio en lugar de un producto. Sin embargo, observando de forma global la tabla de resultados, hay experimentos que logran el mismo nivel de precisión usando dos o tres medidas y usando producto o promedio.

Los resultados me llevan a considerar pertinente la tercera intuición de segmentación, que repito en seguida:

I3. El valor máximo de afijalidad de una palabra permite descubrir el último sufijo y luego los sufijos restantes hasta encontrar la base.

Esta intuición conlleva la idea de que hacer cortes sucesivos hacia la izquierda en el valor máximo de afijalidad es buena estrategia para segmentar morfológicamente una palabra. Además, encierra el hecho de que el primer corte tiende a separar el sufijo más externo, de otra manera los cortes hacia la izquierda no darían buenos resultados. Por tanto, esta intuición confirma que la afijalidad puede revelar los sufijos de una palabra, lo que me lleva a considerar plausible el descubrimiento de los patrones morfotácticos haciendo uso de este método de segmentación.

El hecho de que el segundo mejor experimento fuera el que hace un primer corte en el valor máximo del promedio de entropía y economía, y luego realiza cortes sucesivos hacia la izquierda, confirma que la afijalidad tiende a descubrir primero el sufijo más externo. Por lo anterior, creo que la segunda intuición tampoco puede ser totalmente rechazada. Repito esta intuición con fines explicativos.

I2. Es posible descubrir el sufijo más a la derecha mediante la combinación de las medidas de entropía y economía, y después descubrir los sufijos restantes hasta encontrar la base mediante la combinación de las medidas de entropía y cuadros.

En términos generales, las intuiciones I1 y I4 proponen que el valor máximo de afijalidad permite segmentar la base de una palabra y los cortes sucesivos hacia la derecha permiten determinar los sufijos. De acuerdo con los resultados obtenidos por los experimentos no es posible aceptar estas intuiciones.

Hubo un aspecto que llamó mi atención y que comento en seguida. La mayoría de errores de todos los experimentos fueron por subsegmentación de palabras, esto es, la seg-

mentación automática propuso menos cortes en comparación con los del corpus de evaluación.

Al respecto, los experimentos que menos subsegmentaron fueron los que se basaron en cortes cuando la afijalidad superó el 0.5 de afijalidad (no en valores máximos); sin embargo, también fueron los que más sobresegmentaron y por eso no resultaron ser los mejores. Por otro lado, el experimento con mejores resultados de precisión subsegmentó mayoritariamente, pero sobresegmentó muy poco y en consecuencia tuvo más ciertos que los demás experimentos.

Otra cuestión interesante fue el comportamiento en ciertos grupos de palabras del experimento con mejores resultados. En el caso del fenómeno de flexión, este método obtuvo el segundo lugar con un 49.5% de precisión y para el fenómeno de derivación fue el primer lugar, pero con un 31.6%. También para las palabras nominales (incluidas flexión y derivación) resultó ser el mejor con un 29.8% de precisión, pero fue segundo lugar para los palabras verbales (flexión y derivación) con un 39.1%.

De lo anterior puedo concluir que este experimento funcionó mejor para la flexión y en especial para la verbal ya que, aunque para ésta no obtuvo el primer lugar, sí obtuvo mejores niveles de precisión. Una explicación para este resultado sería que los ejemplos verbales incluidos en el corpus de evaluación presentan mayor regularidad que los nominales; también porque en la derivación nominal se incluyó una gran variedad de sufijos derivativos y sus alomorfos, lo que hizo más difícil la tarea para el método automático.

Una situación especial con este experimento se dio en formas verbales con enclíticos. Su desempeño bajó considerablemente hasta llegar a un 14.29% de precisión. En este grupo de palabras, el método que combina las tres medidas mediante un promedio y seg-

menta cuando la afijalidad es mayor a 0.5 fue el que alcanzó mejores resultados con un 46.67% de precisión.

De hecho, los experimentos con mejores resultados para palabras con enclíticos fueron los que cortaron arriba del umbral de 0.5, seguidos por los que cortaron en el valor máximo y luego hacia la derecha. Lo anterior habla de una naturaleza distinta entre los enclíticos y los sufijos, puesta en evidencia por la tendencia desigual de los resultados de los experimentos.

Con el fin de discutir con un poco más de detalle las segmentaciones obtenidas, pongo en seguida cinco tablas con cincuenta palabras cada una. Son ejemplos tomados aleatoriamente, segmentados de manera automática con el experimento que obtuvo mejores resultados. Las tablas incluyen las segmentaciones según el corpus de evaluación. Se presentan primero las segmentaciones que no coinciden y después, separadas por una línea gruesa, las que sí coinciden. En seguida de cada tabla agrego una breve discusión de los resultados.

Tabla 4.20 Ejemplos de segmentaciones para flexión nominal

Manual	Automática	Comentario
EST~O~S	EST~OS	No separa marcas de género y número
LIBR~O~S	LIBR~OS	
LIBR~A~S	LIBR~AS	
EST~A~S	EST~AS	
PERO~S	PER~O~S	Sobresegmentaciones en algunos casos por separación de marcas equivocadas de género, número y enclíticos
MUJER~ES	MUJ~ER~ES	
NUESTR~O~S	NUEST~R~O~S	
NUESTR~A~S	NUES~T~R~A~S	
ALGODON~ES	ALGO~D~ON~ES	
PASTEL~ES	PAS~TE~LES	
CABALLO~S	CABAL~LO~S	
VERDAD~ES	VERDA~D~E~S	
COMPAS~ES	COM~P~ASES	
MONJ~E	MONJ~E	Segmentación correcta
MONJ~A	MONJ~A	
JEF~E	JEF~E	

Tabla 4.20 Ejemplos de segmentaciones para flexión nominal (continuación)

Manual	Automática	Comentario
JEF~A	JEF~A	
LOB~O	LOB~O	
LOB~A	LOB~A	
LOB~O~S	LOB~O~S	
LOB~A~S	LOB~A~S	
GAT~O	GAT~O	
GAT~A	GAT~A	
GAT~O~S	GAT~O~S	
GAT~A~S	GAT~A~S	
BARC~O	BARC~O	
NIÑ~O	NIÑ~O	
NIÑ~A	NIÑ~A	
NIÑ~O~S	NIÑ~O~S	
NIÑ~A~S	NIÑ~A~S	
MANZAN~O	MANZAN~O	
MANZAN~A~S	MANZAN~A~S	
MANZAN~A	MANZAN~A	Segmentación correcta
MANZAN~O~S	MANZAN~O~S	
SUEL~O	SUEL~O	
SUEL~A	SUEL~A	
SUEL~A~S	SUEL~A~S	
BARC~A	BARC~A	
BARC~A~S	BARC~A~S	
LIBR~A	LIBR~A	
LEÑ~A	LEÑ~A	
LEÑ~O	LEÑ~O	
BARC~O~S	BARC~O~S	
BEBÉ~S	BEBÉ~S	
MADRE~S	MADRE~S	
PADRE~S	PADRE~S	
YEGUA~S	YEGUA~S	
JÓVEN~ES	JÓVEN~ES	
DIOS~ES	DIOS~ES	
PARED~ES	PARED~ES	

Nótese en los ejemplos de flexión nominal de la Tabla 4.20 el buen desempeño del experimento de segmentación. También obsérvese que los casos en que no coinciden los cortes es porque hay sobresegmentación, como VERDA~D~E~S, MUJ~ER~ES o NUES~T~R~A~S. Entre estas sobresegmentaciones hay algunas que separan segmentos que coinciden con enclíticos, por ejemplo, PAS~TE~LES o CABAL~LO~S.

Este tipo de segmentaciones son esperadas en un método como el que utilizo, ya que éste funciona basado en la comparación de segmentos sin utilizar otro tipo de información, así que llega a generalizar segmentaciones muy económicas.

Tabla 4.21 Ejemplos de segmentaciones para derivación nominal

Manual	Automática	Comentario
CHIL~EN~O JAL~ÓN COMBIN~AD~O ERR~ÓNE~O SERV~IDOR~A EGO~ÍSMO~S	CHILE~N~O JAL~Ó~N COMB~IN~ADO ERRÓ~NE~O SERVIDO~R~A EGOÍS~MO~S	Segmentaciones cuestionables
ACOMOD~AD~O	ACOMOD~ADO	Segmentación válida
INSTITU~TO CAR~IÑO TEND~ENCIA~S HOMBR~ECITO~S PLANET~ARIO~S CIENT~ÍFIC~A~S ERR~ÓNE~O~S COMPRENS~IV~A~S REPET~ITIV~A CONDUC~TA~S HERMOS~ÍSIM~A TEJ~ID~O	INSTITUT~O CARIÑ~O TENDENCIA~S HOMBRECITO~S PLANETARI~O~S CIENTÍFIC~A~S ERRÓNE~O~S COMPRENSIV~A~S REPETITIV~A CONDUCT~A~S HERMOSÍSIM~A TEJID~O	Tendencia a separar marcas flexivas sin separar sufijo derivativo
ARRIB~OTA HELAD~OT~E ESCOND~ITE TRANSPAR~ENCIA FLOR~ECITA TEMBL~OR PAST~EL	ARRIBO~TA HELADO~TE ESCONDI~TE TRANSPARENC~IA FLOREC~ITA TEMBLO~R PASTE~L	Tendencia a segmentar en supuesto sufijo más corto y económico
GRAND~OT~E~S FRANC~ES~ES BIBLIO~TECA~S ITALI~AN~O~S	GRANDO~T~E~S FRAN~CES~ES BIBLI~OTECA~S ITAL~IAN~O~S	Segmentaciones muy cercanas a la esperada
DOBL~AJE	DOBLAJE	No hubo segmentación
TUR~ISTA	TUR~IS~T~A	Sobresegmentación
HOSPITAL~ARI~A~S ROJ~IZ~O~S ARROY~UELO GOLP~IZA SOMBR~ERO COMPAÑ~ER~O~S	HOSPITAL~ARI~A~S ROJ~IZ~O~S ARROY~UELO GOLPI~ZA SOMBR~ERO COMPAÑ~ER~O~S	Segmentación correcta

Tabla 4.21 Ejemplos de segmentaciones para derivación nominal (continuación)

Manual	Automática	Comentario
FLOJ~ERA	FLOJ~ERA	Segmentación correcta
MANZAN~ILLA	MANZAN~ILLA	
DESQUICI~AMIENTO	DESQUICI~AMIENTO	
NARIZ~ÓN	NARIZ~ÓN	
ELIMIN~ACIÓN	ELIMIN~ACIÓN	
ACEPT~ACIÓN	ACEPT~ACIÓN	
MAGN~ITUD	MAGN~ITUD	
GRAB~ADOR~A	GRAB~ADOR~A	
MUNICIP~AL	MUNICIP~AL	
HERMOS~URA	HERMOS~URA	
EUROP~E~A	EUROP~E~A	
OBJETIV~ISMO	OBJETIV~ISMO	

A diferencia de los resultados de la flexión nominal, la derivación nominal se muestra a primera vista inconsistente. Aquí, la mayoría de las faltas de coincidencia con el corpus de evaluación se dan por subsegmentación, como TEND~ENCIA~S vs TENDENCIA~S o CIENT~ÍFIC~A~S vs CIENTÍFIC~A~S, que cuentan con segmentación flexiva, pero carecen de segmentación derivativa.

También hay una buena cantidad de palabras donde coincide el número de segmentos, pero no el lugar donde se hace el corte, como EGO~ÍSMO~S vs EGOÍS~MO~S o BIBLIO~TECA~S vs BIBLI~OTECA~S. Además hay casos de sobresegmentación, aunque parecen ser pocos ya que en la tabla sólo está TUR~ISTA vs TUR~IS~T~A.

Los casos de coincidencia de segmentación se dan en diversos sufijos, situación afortunada para el método porque significa que logra segmentar distintos fenómenos de derivación, tómese por ejemplo ROJ~IZ~O~S, ARROY~UELO, GOLPI~ZA, SOMBR~ERO, MANZAN~ILLA, DESQUICI~AMIENTO, NARIZ~ÓN, ELIMIN~ACIÓN, MAGN~ITUD y COMPAÑ~ER~O~S.

Una tendencia del método es que éste propone un segmento más corto y por eso más económico (se combina con más segmentos). Véanse por ejemplo los casos de

ARRIB~OTA vs ARRIBO~TA, ESCOND~ITE vs ESCONDI~TE o TRANSPAR~ENCIA vs TRANSPARENC~IA.

Tabla 4.22 Ejemplos de segmentaciones para flexión verbal

Manual		Automática	Comentario
DEM	Alcoba (1999)		
CANT~AMOS	CANT~A~MOS	CAN~T~AMOS	Segmentaciones cuestionables
PART~AMOS	PART~A~MOS	PAR~T~AMOS	
SUB~IEREN	SUB~IE~RE~N	SUBIE~R~E~N	
COM~IDO	COM~I~DO	COM~ID~O	
AM~ÉIS	AM~É~IS	AMÉ~IS	
COM~IERA	COM~IE~RA	COM~IER~A	
PART~IERA	PART~IE~RA	PART~IER~A	
SUB~IERON	SUB~IE~RO~N	SUBIE~RON	Tendencia a juntar VT diptongada a base. Algunas son-pertinentes
SUB~IESE	SUB~IE~SE	SUBIE~SE	
SUB~IESEIS	SUB~IE~SE~IS	SUBIE~SE~IS	
SUB~IÉSEMOS	SUB~IÉ~SE~MOS	SUBIÉ~SEMOS	
AMAR~ÍA	AM~A~RÍA	AM~ARÍA	Tendencia a juntar vocal temática a marcas flexivas. Muchas podrían considerarse pertinentes
AMAR~ÍAN	AM~A~RÍA~N	AM~ARÍAN	
SUBIR~ÍA	SUB~I~RÍA	SUB~IRÍA	
SUBIR~ÁN	SUB~I~RÁ~N	SUB~IRÁ~N	
PARTIR~ÍAMOS	PART~I~RÍA~MOS	PART~IRÍAMOS	
PARTIR~ÍAN	PART~I~RÍA~N	PART~IRÍA~N	
PARTIR~ÁS	PART~I~RÁ~S	PART~IRÁ~S	
PARTIR~ÁN	PART~I~RÁ~N	PART~IRÁ~N	
TEMER~ÍAN	TEM~E~RÍA~N	TEM~ERÍA~N	
CANTAR~ÍAMOS	CANT~A~RÍA~MOS	CANT~ARÍAMOS	
CANTAR~Á	CANT~A~RÁ	CANT~ARÁ	
COMER~ÍAMOS	COM~E~RÍA~MOS	COM~ERÍAMOS	
COM~IERES	COM~IE~RE~S	COM~IERE~S	
PARTIR~ÉIS	PART~I~RÉ~IS	PART~I~RÉIS	Tendencia a separar VT. Podrían considerarse pertinentes
PART~ISTEIS	PART~I~STE~IS	PART~I~STEIS	
PART~IÉRAMOS	PART~IÉ~RA~MOS	PART~IÉ~RAMOS	
CANT~Ó	CANT~Ó	CANT~Ó	Coincide con DEM
CANT~ASTEIS	CANT~A~STE~IS	CANT~ASTEIS	
CANT~ARON	CANT~A~RO~N	CANT~ARON	
CANT~ADO	CANT~A~DO	CANT~ADO	
CANT~ES	CANT~E~S	CANT~ES	
PART~ES	PART~E~S	PART~ES	
PART~ÍAIS	PART~Í~A~IS	PART~ÍAIS	
PART~ÍA	PART~Í~A	PART~ÍA	
SUB~A	SUB~A	SUB~A	
SUB~AN	SUB~A~N	SUB~AN	
AM~ARAN	AM~A~RA~N	AM~ARAN	
AM~ASES	AM~A~SE~S	AM~ASES	

Tabla 4.22 Ejemplos de segmentaciones para flexión verbal (continuación)

Manual		Automática	Comentario
DEM	Alcoba (1999)		
AM~ASEN	AM~A~SE~N	AM~ASEN	Coincide con DEM
AM~ARES	AM~A~RE~S	AM~ARES	
AM~ADO	AM~A~DO	AM~ADO	
COM~IESE	COM~IE~SE	COM~IESE	
COM~IERAIS	COM~IE~RA~IS	COM~IERAIS	
COM~IESEN	COM~IE~SE~N	COM~IESEN	
COM~EMOS	COM~E~MOS	COM~EMOS	
COM~ISTE	COM~I~STE	COM~ISTE	
COM~ER	COM~E~R	COM~ER	
TEM~ISTEIS	TEM~I~STE~IS	TEM~ISTEIS	

La observación más clara al respecto de la flexión verbal es que una considerable cantidad de palabras segmentadas automáticamente tienen un solo corte entre la raíz y la flexión. Además, esta última no cuenta con separación ni de la vocal temática ni de los morfemas de tiempo-aspecto-modo y número-persona (COM~ISTE, PART~IRÍAMOS, CANT~ARON). Por tanto, la segmentación automática coincidió mayoritariamente con la segmentación propuesta por el DEM y no con la de Alcoba.

Para este grupo de palabras, el experimento con mejores resultados globales obtuvo el 49.7% de precisión; sin embargo, hubo otro experimento con resultados de hasta el 54.0% (sólo en este grupo de palabras). Fue el experimento basado en la estrategia que primero corta en el valor máximo del producto de entropía y economía, y luego segmenta hacia la izquierda en el producto de entropía y cuadros. Es importante resaltar que este experimento también segmentó sistemáticamente una sola vez por palabra, separando la flexión de la raíz. El parecer, cuantitativamente no se puede hablar de una separación clara entre la vocal temática y los sufijos de tiempo-aspecto-modo y número-persona (ésta tiende a pegarse a ellos).

Tabla 4.23 Ejemplos de segmentaciones para derivación verbal

Manual		Automática	Comentario
DEM	Alcoba		
FLOR~EC~IERON	FLOR~EC~IE~RO~N	FLOREC~IE~RON	Segmentaciones cuestionables
FLOR~EC~ÍAS	FLOR~EC~Í~A~S	FLORECÍA~S	
FLOR~EC~ÍAIS	FLOR~EC~Í~A~IS	FLOREC~ÍA~IS	
FLOR~EC~ER~EMOS	FLOR~EC~E~RE~MOS	FLORECE~REMOS	
FLOR~EC~ER~ÉIS	FLOR~EC~E~RÉ~IS	FLORECE~RÉIS	
EJEMPL~IFIC~AR~ÁIS	EJEMPL~IFIC~A~RÁ~IS	EJEMPLIFICAR~Á~IS	
COQUET~E~AR	COQUET~E~A~R	COQUE~T~E~AR	
ARPON~E~AR~ÍAIS	ARPON~E~A~RÍA~IS	ARPONE~ARÍA~IS	
ARPON~E~AR~Á	ARPON~E~A~RÁ	ARPONE~ARÁ	
ARPON~E~ABAS	ARPON~E~A~BA~S	ARPONE~ABA~S	
HORROR~IZ~AR~É	HORROR~IZ~A~RÉ	HORRORIZ~ARÉ	
MONOPOL~IZ~AR	MONOPOL~IZ~A~R	MONOPOLIZ~AR	No separa sufijo derivativo, pero separa marcas flexivas
HORROR~IZ~ABAIS	HORROR~IZ~A~BA~IS	HORRORIZ~ABAIS	
HORROR~IZ~ÁBAMOS	HORROR~IZ~Á~BA~MOS	HORRORIZ~ÁBAMOS	
HORROR~IZ~ABA	HORROR~IZ~A~BA	HORRORIZ~ABA	
HORROR~IZ~AR~ÁN	HORROR~IZ~A~RÁ~N	HORRORIZ~ARÁN	
HORROR~IZ~AR~ÍAIS	HORROR~IZ~A~RÍA~IS	HORRORIZ~ARÍAIS	
EJEMPL~IFIC~ABAIS	EJEMPL~IFIC~A~BA~IS	EJEMPLIFIC~ABAIS	
EJEMPL~IFIC~ABA	EJEMPL~IFIC~A~BA	EJEMPLIFIC~ABA	
EJEMPL~IFIC~ÁBAMOS	EJEMPL~IFIC~Á~BA~MOS	EJEMPLIFIC~ÁBAMOS	
EJEMPL~IFIC~AR~É	EJEMPL~IFIC~A~RÉ	EJEMPLIFIC~ARÉ	
EJEMPL~IFIC~AR~ÍAS	EJEMPL~IFIC~A~RÍA~S	EJEMPLIFIC~ARÍAS	
PALID~EC~ER	PALID~EC~E~R	PALIDEC~E~R	
PERMAN~EC~ER	PERMAN~EC~E~R	PERMANEC~E~R	
FLOR~EC~EN	FLOR~EC~E~N	FLOREC~E~N	
FLOR~EC~ER	FLOR~EC~E~R	FLOREC~E~R	
FLOR~EC~E	FLOR~EC~E	FLOREC~E	
FLOR~EC~ER~ÍAN	FLOR~EC~E~RÍA~N	FLORECE~RÍA~N	
ARPON~E~ÁBAMOS	ARPON~E~Á~BA~MOS	ARPONE~ÁBAMOS	
ARPON~E~ASTEIS	ARPON~E~A~STE~IS	ARPONE~ASTEIS	
ARPON~E~AR~ÍAMOS	ARPON~E~A~RÍA~MOS	ARPONE~ARÍAMOS	
ARPON~E~AMOS	ARPON~E~A~MOS	ARPONE~A~MOS	
NOT~IFIC~AR	NOT~IFIC~A~R	NOTIFIC~AR	
RAM~IFIC~AR~SE	RAM~IFIC~A~R~SE	RAMIFIC~ARSE	No separa sufijo derivativo ni enclítico
NACIONAL~IZ~AR	NACIONAL~IZ~A~R	NACION~ALIZ~AR	Segmenta un sufijo derivativo pero no el sufijo derivativo verbal

Tabla 4.23 Ejemplos de segmentaciones para derivación verbal (continuación)

Manual		Automática	Comentario
DEM	Alcoba		
MOM~IFIC~AR~SE MATERIAL~IZ~AR~SE	MOM~IFIC~A~R~SE MATERIAL~IZ~A~R~SE	MOM~IFIC~ARSE MATERIAL~IZ~ARSE	Separa sufijo derivativo, pero no separa enclítico. Podrían considerarse pertinentes
EJEMPL~IFIC~AMOS EJEMPL~IFIC~AS INTENS~IFIC~AR PUR~IFIC~AR CAPITAL~IZ~AR MODERN~IZ~AR INTERIOR~IZ~AR	EJEMPL~IFIC~A~MOS EJEMPL~IFIC~A~S INTENS~IFIC~A~R PUR~IFIC~A~R CAPITAL~IZ~A~R MODERN~IZ~A~R INTERIOR~IZ~A~R	EJEMPL~IFIC~AMOS EJEMPL~IFIC~AS INTENS~IFIC~AR PUR~IFIC~AR CAPITAL~IZ~AR MODERN~IZ~AR INTERIOR~IZ~AR	Coincide con DEM
HORROR~IZ~AS ARPON~E~AR ARPON~E~AN	HORROR~IZ~A~S ARPON~E~A~R ARPON~E~A~N	HORROR~IZ~A~S ARPON~E~A~R ARPON~E~A~N	Coincide con Alcoba
ABUEL~E~Ó ARPON~E~Ó ARPON~E~A	ABUEL~E~Ó ARPON~E~Ó ARPON~E~A	ABUEL~E~Ó ARPON~E~Ó ARPON~E~A	Coincide con ambos

Según los resultados obtenidos, aunque de manera inconsistente, el experimento segmentó mejor los sufijos derivativos –izar (CAPITAL~IZ~AR), –ificar (EJEMPL~IFIC~AS) y –ear (ARPON~E~A~R). Por otro lado, no fue tan bueno para el sufijo –ecer (FLOREC~IE~RON, FLORECÍA~S).

Como en el caso de la flexión verbal, en la derivación verbal el segmento final tiende a mantenerse como una unidad, es decir, sin separación entre vocal temática y morfemas de modo-tiempo-aspecto y número-persona, lo que coincide con la propuesta del DEM. De hecho se ve una tendencia a separar las marcas flexivas, pero sin separar el sufijo derivativo verbal (HORRORIZ~ARÍAIS, EJEMPLIFIC~ABAIS). Dos casos que podrían considerarse pertinentes son MOM~IFIC~ARSE y MATERIAL~IZ~ARSE, ya que separan el sufijo derivativo verbal, aunque no separan el enclítico.

Tabla 4.24 Ejemplos de segmentaciones para enclíticos

Manual				Automática	Comentario
DEM	Alcoba	DEM (género y número)	Alcoba (género y número)		
CRE~É~ME~LO ALIMÉNT~E~LOS PROMET~ÉR~NOS~LAS	CRE~É~ME~LO ALIMÉNT~E~LOS PROMET~É~R~NOS~LAS	CRE~É~ME~L~O ALIMÉNT~E~L~O~S PROMET~ÉR~NOS~L~A~S	CRE~É~ME~L~O ALIMÉNT~E~L~O~S PROMET~É~R~NOS~L~A~S	CREÉM~ELO ALIM~ÉNT~E~LO~S PROMETÉ~R~NOSLAS	Segmentaciones cuestionables
AM~AR~LA AM~AR~LOS AM~ÉMO~NOS CONT~ÁR~NOS~LO QUEM~ÁR~NOS~LA PAS~ÁR~NOS~LAS LANZ~ÁR~NOS~LE QUIT~ÁR~ME~LOS CUID~ÁR~TE~LO CANT~AR~LES CANT~AR~NOS CANT~ÁNDO~ME LLEV~ÁNDO~SE~LOS TOM~ÁR~SE~LO TOST~ÁR~SE~LAS	AM~A~R~LA AM~A~R~LOS AM~É~MO~NOS CONT~Á~R~NOS~LO QUEM~Á~R~NOS~LA PAS~Á~R~NOS~LAS LANZ~Á~R~NOS~LE QUIT~Á~R~ME~LOS CUID~Á~R~TE~LO CANT~A~R~LES CANT~A~R~NOS CANT~Á~NDO~ME LLEV~Á~NDO~SE~LOS TOM~Á~R~SE~LO TOST~Á~R~SE~LAS	AM~AR~L~A AM~AR~L~O~S CONT~ÁR~NOS~L~O QUEM~ÁR~NOS~L~A PAS~ÁR~NOS~L~A~S LANZ~ÁR~NOS~L~E QUIT~ÁR~ME~L~O~S CUID~ÁR~TE~L~O CANT~AR~L~E~S LLEV~ÁNDO~SE~L~O~S TOM~ÁR~SE~L~O TOST~ÁR~SE~L~A~S	AM~A~R~L~A AM~A~R~L~O~S CONT~Á~R~NOS~L~O QUEM~Á~R~NOS~L~A PAS~Á~R~NOS~L~A~S LANZ~Á~R~NOS~L~E QUIT~Á~R~ME~L~O~S CUID~Á~R~TE~L~O CANT~A~R~L~E~S LLEV~Á~NDO~SE~L~O~S TOM~Á~R~SE~L~O TOST~Á~R~SE~L~A~S	AM~ARLA AM~ARLOS AM~ÉMONOS CONT~ÁRNOSLO QUEM~ÁRNOSLA PAS~ÁRNOSLAS LANZ~ÁRNOSLE QUIT~ÁRMELOS CUID~ÁRTELO CANT~ARLE CANT~ARNOS CANT~ÁNDOME LLEV~ÁNDOSELOS TOM~ÁRSELO TOST~ÁRSELAS	Separa marcas verbales pegadas a los enclíticos
EMBOLS~ÁNDO~SE~LA	EMBOLS~Á~NDO~SE~LA	EMBOLS~ÁNDO~SE~L~A	EMBOLS~Á~NDO~SE~L~A	EMBOLS~ÁNDOSE~LA	Separa marcas verbales pegadas a enclíticos. Separa último enclítico
ÉCH~A~TE~LAS COM~ER~LOS	ÉCH~A~TE~LAS COM~E~R~LOS	ÉCH~A~TE~L~A~S COM~ER~L~O~S	ÉCH~A~TE~L~A~S COM~E~R~L~O~S	ÉCH~ATELA~S COM~ERLO~S	Separa marca verbal pegada a los enclíticos. Separa marca de plural
DEC~ÍR~ME~LO HAC~ÉR~ME~LOS PRÉST~A~ME~LA AMÁRR~A~TE~LA PON~ER~TE~LO	DEC~Í~R~ME~LO HAC~É~R~ME~LOS PRÉST~A~ME~LA AMÁRR~A~TE~LA PON~E~R~TE~LO	DEC~ÍR~ME~L~O HAC~ÉR~ME~L~O~S PRÉST~A~ME~L~A AMÁRR~A~TE~L~A PON~ÉR~TE~L~O	DEC~Í~R~ME~L~O HAC~É~R~ME~L~O~S PRÉST~A~ME~L~A AMÁRR~A~TE~L~A PON~É~R~TE~L~O	DEC~ÍR~MELO HAC~ÉR~MELOS PRÉST~A~MELA AMÁRR~A~TELA PON~ÉR~TELO	Separa marcas verbales y enclíticos, pero no segmenta enclíticos

Tabla 4.24 Ejemplos de segmentaciones para enclíticos (continuación)

Manual				Automática	Comentario
DEM	Alcoba	DEM (género y número)	Alcoba (género y número)		
SUPON~ÉR~SE~LE COM~ÉR~SE~LA COM~ER~LO HAC~ÉR~NOS~LOS IMPED~ÍR~NOS~LO PERDÓN~E~SE~NOS OCURR~IÉNDO~SE~LE OPRIM~IÉNDO~SE~LAS COM~IÉNDO~LO PED~ÍR~SE~LO	SUPON~É~R~SE~LE COM~É~R~SE~LA COM~E~R~LO HAC~É~R~NOS~LOS IMPED~Í~R~NOS~LO PERDÓN~E~SE~NOS OCURR~IÉ~NDO~SE~LE OPRIM~IÉ~NDO~SE~LAS COM~IÉ~NDO~LO PED~Í~R~SE~LO	SUPON~ÉR~SE~L~E COM~ÉR~SE~L~A COM~ER~L~O HAC~ÉR~NOS~L~O~S IMPED~ÍR~NOS~L~O OCURR~IÉNDO~SE~L~E OPRIM~IÉNDO~SE~L~A~S COM~IÉNDO~L~O PED~ÍR~SE~L~O	SUPON~É~R~SE~L~E COM~É~R~SE~L~A COM~E~R~L~O HAC~É~R~NOS~L~O~S IMPED~Í~R~NOS~L~O OCURR~IÉ~NDO~SE~L~E OPRIM~IÉ~NDO~SE~L~A~S COM~IÉ~NDO~L~O PED~Í~R~SE~L~O	SUPON~ÉR~SELE COM~ÉR~SELA COM~ER~LO HAC~ÉR~NOSLOS IMPED~ÍR~NOSLO PERDÓN~E~SENOS OCURR~IÉNDO~SELE OPRIM~IÉNDO~SELAS COM~IÉNDO~LO PED~Í~R~SELO	Separa marcas verbales y enclíticos, pero no segmenta enclíticos
PON~ÉR~TE~LAS SERV~ÍR~TE~LAS CÓM~E~LAS RECÓRT~E~LOS AGUÁNT~E~LOS CUMPL~ÍR~SE~LOS PON~ÉR~SE~LES ATRIBU~ÍR~SE~LES	PON~É~R~TE~LAS SERV~Í~R~TE~LAS CÓM~E~LAS RECÓRT~E~LOS AGUÁNT~E~LOS CUMPL~Í~R~SE~LOS PON~É~R~SE~LES ATRIBU~Í~R~SE~LES	PON~ÉR~TE~L~A~S SERV~ÍR~TE~L~A~S CÓM~E~L~A~S RECÓRT~E~L~O~S AGUÁNT~E~L~O~S CUMPL~ÍR~SE~L~O~S PON~ÉR~SE~L~E~S ATRIBU~ÍR~SE~L~E~S	PON~É~R~TE~L~A~S SERV~Í~R~TE~L~A~S CÓM~E~L~A~S RECÓRT~E~L~O~S AGUÁNT~E~L~O~S CUMPL~Í~R~SE~L~O~S PON~É~R~SE~L~E~S ATRIBU~Í~R~SE~L~E~S	PON~ÉR~TELA~S SERV~ÍR~TELA~S CÓM~E~LA~S RECÓRT~E~LO~S AGUÁNT~E~LO~S CUMPL~ÍR~SELO~S PON~ÉR~SELE~S ATRIBU~ÍR~SELE~S	Separa marcas verbales y enclíticos, pero no segmenta enclíticos. Separa marca de plural
PART~IÉNDO~LOS COM~ER~LA SUB~IR~LE	PART~IÉ~NDO~LOS COM~E~R~LA SUB~I~R~LE	PART~IÉNDO~L~O~S COM~ER~L~A SUB~IR~L~E	PART~IÉ~NDO~L~O~S COM~E~R~L~A SUB~I~R~L~E	PART~IÉNDO~LOS COM~ER~LA SUB~IR~LE	Coincide con DEM
PART~IR~NOS COM~IÉNDO~SE COM~ER~ME	PART~I~R~NOS COM~IÉ~NDO~SE COM~E~R~ME			PART~IR~NOS COM~IÉNDO~SE COM~ER~ME	Coincide con ambos

Los bajos resultados obtenidos en las palabras con enclíticos se deben a que, de manera casi generalizada, no hubo segmentación al interior de los grupos de enclíticos, por ejemplo, CONT~ÁRNOSLO y QUEM~ÁRNOSLA. Además, hubo bastantes casos de separación de las marcas verbales, pero los enclíticos se mantuvieron concatenados, como COM~ÉR~SELA, IMPED~ÍR~NOSLO.

Esto da una clara muestra de la naturaleza distinta de los enclíticos y los afijos, de tal manera que el índice de afijalidad no segmenta los primeros. En este sentido las segmentaciones del párrafo anterior serían pertinentes y debieran haberse contado como aciertos. Algo similar ocurre con los casos donde se separan marcas verbales, enclíticos y la marca de plural, como CUMPL~ÍR~SELO~S y PON~ÉR~SELE~S, que resultan bastante pertinentes.

Las segmentaciones que coincidieron se dan cuando sólo hay un enclítico, como en COM~IÉNDO~LO, SUB~IR~LE, PART~IR~NOS, COM~IÉNDO~SE, aunque no sucede en todos los casos de un solo enclítico, por ejemplo, AM~ARLOS o CANT~ARNOS. Para trabajo futuro sería buena idea usar otra medida para separar los enclíticos.

Hasta aquí dejo el análisis de ejemplos de palabras segmentadas. Ahora, otro aspecto que quiero mencionar es el que tiene que ver con las palabras marcadas como irregulares en el corpus de evaluación. Fue una sorpresa que las medidas de precisión fueran mayores en comparación con las obtenidas para las formas regulares. En este caso, el experimento con mejor resultado fue el mismo que para las palabras regulares: cortes sucesivos hacia la izquierda en el valor máximo del promedio de las tres medidas de afijalidad. Sin embargo, su nivel de precisión subió al 41% en comparación con el 33.8% obtenido en palabras regulares. El incremento fue considerable.

Una explicación de esto es el bajo número de palabras irregulares incluidas en el corpus de evaluación y por tanto la menor variedad de sufijos a segmentar (y de alomorfos). Se debe recordar además que para los verbos sólo se incluyeron ejemplos de un solo tiempo y modo, esto es, no se incluyeron todos los tiempos de ambos modos como en el caso de las palabras regulares. Por lo anterior, no se puede afirmar que el experimento resultara mejor para palabras irregulares, sino solamente que es relativamente igual de bueno para este tipo de palabras.

4.4. Observaciones finales

Aunque el nivel de precisión de 33.8% alcanzado por el mejor experimento es bajo, no lo considero desafortunado por las siguientes razones. El corpus de evaluación fue construido con una gran variedad de sufijos derivativos y, sobre todo, de alomorfos de estos sufijos (188 alomorfos). Sirvan de ejemplo los seis alomorfos del sufijo (V)(C)ión propuestos por Moreno de Alba (1986): –ación, –ción, –ición, –ión, –sión y –ución. La determinación de estos alomorfos fue producto de la reflexión humana y difícilmente un procesamiento computacional coincidirá con esa profundidad de análisis.

Además, el método automático busca segmentaciones económicas que a pesar de no coincidir con esa reflexión humana, no dejan de ser válidas ya que la economía es parte de todo fenómeno lingüístico, en especial de carácter morfológico. Véase por ejemplo la Tabla 4.25, donde pongo ejemplos de segmentaciones para el sufijo (V)(C)ión.

Como puede verse, el método propone sólo tres segmentos: ~ACIÓN, ~ÓN y ~CIÓN, algunas veces separando este último en ~CI~ÓN. El experimento segmenta de manera regular, pero las segmentaciones no coinciden con el análisis humano. De hecho, el

segmento –ÓN, no considerado por Moreno de Alba como sufijo, es muy económico ya que se combina con muchas bases.

Tabla 4.25 Ejemplos de segmentaciones para alomorfos del sufijo (V)(C)ión

Alomorfo	Segmentación manual	Segmentación automática	Sufijo propuesto
-ACIÓN	ELIMIN~ACIÓN	ELIMIN~ACIÓN	~ACIÓN
-ACIÓN	ACEPT~ACIÓN	ACEPT~ACIÓN	
-ACIÓN	ACLAR~ACIÓN	ACLAR~ACIÓN	
-ACIÓN	FUNDAMENT~ACIÓN	FUNDAMENT~ACIÓN	
-ACIÓN	AGRUP~ACIÓN	AGRUP~ACIÓN	
-CIÓN	SATISFAC~CIÓN	SATISFACCI~ÓN	~ÓN
-CIÓN	INTERVEN~CIÓN	INTERVENCI~ÓN	
-CIÓN	INDISCRE~CIÓN	INDISCRE~CI~ÓN	
-CIÓN	CONSTITU~CIÓN	CONSTITU~CIÓN	
-CIÓN	PRODUC~CIÓN	PRODUCCI~ÓN	
-ICIÓN	DEFIN~ICIÓN	DEFINI~CIÓN	~CIÓN
-ICIÓN	ABURR~ICIÓN	ABURRI~CIÓN	
-IÓN	DISPERS~IÓN	DISPERSI~ÓN	~ÓN
-IÓN	REUN~IÓN	REUNIÓ~N	
-UCIÓN	EVOL~UCIÓN	EVOLU~CI~ÓN	~CI~ÓN
-UCIÓN	SOL~UCIÓN	SOLU~CI~ÓN	

Lo que quiero resaltar es lo difícil que es encontrar una manera de evaluar un análisis lingüístico automático. ¿La idea es llegar al mismo nivel de análisis de un experto o encontrar un método que logre describir las regularidades de un corpus? Por supuesto que una computadora no puede suplantar la reflexión humana, pero sí puede ayudar a generar inquietudes lingüísticas a partir de los resultados que arroja.

En otros aspectos, sería posible utilizar distintas estrategias de segmentación para distintas clases de palabra y distintos fenómenos morfológicos. Esto se debe a que algunos métodos fueron un poco mejores para nominales y otros para verbos, de la misma forma, algunos fueron un poco mejores para flexión y otros para derivación. Sin embargo, esto conlleva conocer antes de segmentar la clase o el fenómeno morfológico de la palabra.

Esta cuestión no es fácil de resolver. Se tiene, por un lado, la opción de marcar *a priori* las palabras, pero esto dista de mi interés por un método con escasa intervención humana. Por otro lado, se puede implementar algún procedimiento de descubrimiento de clases de palabras o del tipo de fenómeno presente, flexión o derivación, pero ambas son dos tareas suficientemente complejas como para pensar en otro proyecto de investigación.

Ya que mi perspectiva de trabajo es proponer un método automático que describa las regularidades de un corpus sin dar por sentadas las unidades morfológicas, no veo mayor inconveniente en utilizar la estrategia de segmentación que tuvo mejores resultados para incorporarla a mi método de descubrimiento de patrones morfotácticos.

Este capítulo describió el primer paso en el descubrimiento de la morfotáctica del español: el descubrimiento de unidades morfológicas, específicamente bases y sufijos. Mediante un estudio de las medidas de afijalidad (entropía, cuadros y economía) se formularon dieciséis estrategias de segmentación. Se evaluaron y se obtuvo la mejor: hacer cortes sucesivos hacia la izquierda en el valor máximo del promedio de las tres medidas de afijalidad. Así, el siguiente capítulo describe cómo se usó esta estrategia para segmentar los tipos de palabras del CEMC y construir el autómata de estados finitos.

5. Generación automática del autómata de estados finitos

Gracias a los experimentos realizados en el capítulo anterior he podido seleccionar una estrategia de segmentación para descubrir las bases y sufijos del español. Ahora, es necesario describir su ordenamiento mediante un aparato de descripción. En este capítulo describo los aspectos relacionados con la generación del autómata de estados finitos y su evaluación.

En primer lugar expongo el procedimiento para construir automáticamente este aparato de descripción, luego presento los experimentos llevados a cabo. Después ofrezco una evaluación del autómata generado y termino el capítulo con la presentación del método propuesto para descubrir la morfotáctica del español.

5.1. Procedimiento para la generación del autómata

Como se mencionó, una vez que se han descubierto las unidades morfológicas de la lengua de estudio, en este caso particular bases y sufijos, es necesario describir su orden y secuencialidad. Para ello se pensó en utilizar una gramática de estados finitos, ya que es posible ver a la morfología sufijal como un lenguaje regular. Luego, en una revisión de la manera en como la morfología computacional ha descrito la morfotáctica de diversas lenguas, se decidió generar automáticamente un autómata de estados finitos.

La manera de describir la morfotáctica no es asunto trivial. Ya mostraba en la sección 3.4 que existen distintas posibilidades, aunque éstas suelan ser hechas manualmente: autómatas de estados finitos, redes de discriminación (*tries*), autómatas en cascada y trans-

ductores en paralelo. Se descartaron las dos últimas por el hecho de que también representan reglas de transformación fonológica, que no están involucradas en esta investigación.

Construir una red de discriminación (*trie*) era factible, especialmente porque el método del cálculo del índice de afijalidad utiliza una red de letras muy similar (véase sección 2.5.5). Además no describe cambios morfofonológicos, al igual que el método que estoy proponiendo; sin embargo, se decidió utilizar un autómata de estados finitos por las siguientes razones.

Primero, han sido utilizados en la morfología computacional para distintas lenguas dejando de lado el tratamiento de los fenómenos morfofonológicos, lo que coincide con mi enfoque de trabajo. Segundo, el autómata generado podría convertirse a futuro y con relativamente poco trabajo en un autómata probabilístico, inclusive en un modelo oculto de Markov, que pueden ayudar a describir de mejor manera la morfología de una lengua.

Tercero, y más importante, son equivalentes a una gramática de estados finitos como se consignó en el apartado 3.3.4. Ya que el objetivo que se propuso al inicio de este trabajo fue la construcción automática de una gramática, generar el autómata permite cumplir el objetivo dada su equivalencia y porque existen procedimientos mediante los cuales un autómata se puede convertir en una gramática, y viceversa.

En los siguientes apartados explico el procedimiento general de construcción del autómata y luego el algoritmo computacional para construirlo.

5.1.1. Planteamiento general para construir el autómata

En este apartado discuto mi propuesta de construcción del autómata de estados finitos que describe la morfotáctica del español (bases y sufijos). Mi intención es presentar los proble-

mas generales de construir el autómata y exponer la estrategia que decidí utilizar para lidiar con ellos.

Considero que el principal problema de construir automáticamente un autómata morfológico es que el autómata construido genere palabras inexistentes en la lengua. Por tanto, un autómata que represente la morfología de un corpus debe evitar que la secuencia de sus transiciones produzca una palabra inexistente en el corpus, ya que podría ser una palabra inexistente en la lengua de estudio.

En seguida ejemplifico esta situación con cuatro palabras segmentadas que hipotéticamente formarían un corpus: NIÑ~A, NIÑ~O, GAT~A, GAT~O y MONJ~A. Además, simulo un procedimiento automático para crear un autómata a partir de estas palabras, tomando sus segmentos uno por uno y leyendo la palabra de derecha a izquierda, es decir, del último segmento al primero.

Para construir el autómata es necesaria la creación del estado inicial (*q0*), de donde partirán las transiciones. Luego, a medida que se van leyendo los segmentos de cada palabra, éstos se van incluyendo al alfabeto de entrada ($\sum$). Además, se van creando las transiciones (a_i, q_i, q_k), donde q_i y q_k son estados y a_i es un símbolo de entrada. Por ejemplo, para el último segmento (~A) de la primera palabra del corpus (NIÑ~A) se obtendría la transición (~A, *q0*, *q1*). Esto incluye que debe ser creado el estado *q1* y que debe ser agregado al conjunto de estados *K*.

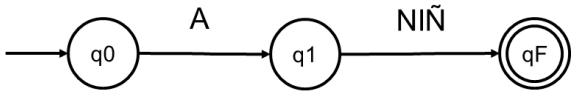
El siguiente segmento de la primera palabra (NIÑ~) se incluye como nuevo símbolo de $\sum$ y se crea la transición (NIÑ~, *q1*, *qF*). Es necesaria la creación del estado final *qF* porque se ha llegado al inicio de la palabra. De esta manera, la secuencia de estados *q0*, *q1*, *qF* asociados a las transiciones que se mencionaron generan los símbolos ~A, NIÑ~, que corresponden a la primera palabra del corpus.

Es posible seguir explicando de esta manera la generación del autómata para el corpus hipotético, pero es más fácil explicarla mediante su representación en forma de grafo. En esta, como se explicó en la sección 3.3.3, cada estado se representa como un nodo (círculo) y las transiciones como arcos (flechas) en donde se ponen sus símbolos asociados.

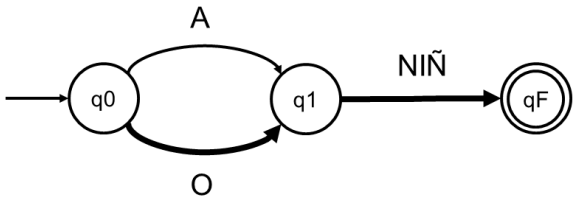
Entonces, procesar la primera palabra (NIÑ~A) daría como resultado el autómata (a) de la figura Figura 5.1. Se pueden ver los estados *q0, q1* y *qF* representados por círculos y los segmentos asociados a las flechas. Recuérdese que el orden de lectura de segmentos es de derecha a izquierda. Se decidió hacer de esta manera porque al final de las palabras del español hay menos variabilidad de segmentos, ya que se trata de los afijos, que forman un conjunto más pequeño de segmentos en relación con el conjunto de bases.

Al procesar la segunda palabra, se pueden unir los arcos de los sufijos ~O y ~A al estado *q1*, logrando aprovechar el arco final asociado a la base NIÑ~, ya que es la misma para los dos palabras, véase autómata (b). La tercer palabra puede quedar representado en el autómata ya creado con sólo agregar al arco final la base GAT~, como se muestra en (c).

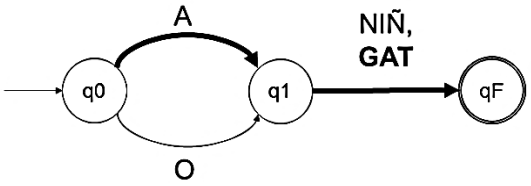
El autómata construido hasta el momento representa también, de manera afortunada, la siguiente palabra del corpus GAT~O sin que sea necesaria ninguna modificación, ya que existe un arco del estado *q0* al estado *q1* con el segmento ~O, véase autómata (d). Finalmente, la palabra MONJ~A podría ser representada, como se hizo antes, aprovechando los arcos ya existentes y sólo agregando la base MONJ~ al arco final como se puede ver en el autómata (e).



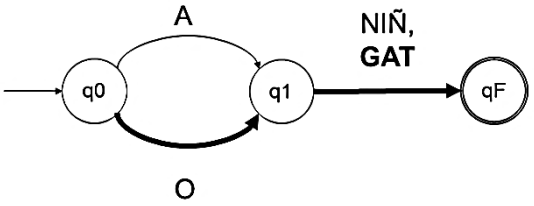
(a)



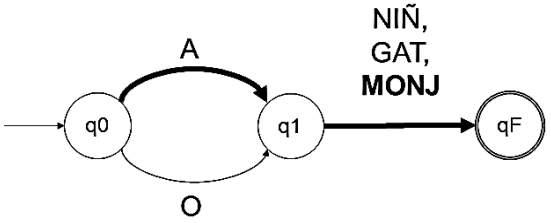
(b)



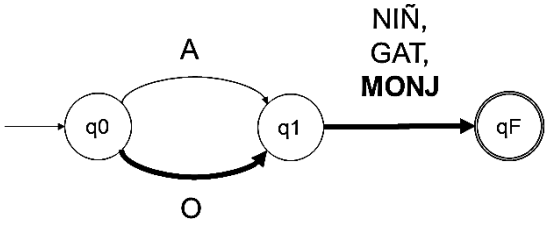
(c)



(d)



(e)



(f)

Figura 5.1. Autómata que produce una palabra inexistente

Con el procedimiento presentado se ha construido un autómata que representa al pequeño corpus hipotético. Cabe resaltar que tres nodos y tres arcos bastaron para tal representación. Desafortunadamente el autómata creado también da cabida a la palabra inexistente *MONJO debido a que la secuencias de estados *q0, q1* generan el segmento ~O, que se asocia al segmento MONJ~ de la siguiente secuencia de estados del autómata *q1*, *qF*, véase el autómata (f). Esta situación impide que se acepte este autómata como representación válida del corpus.

Cuando este tipo de autómatas es hecho manualmente, el investigador es cuidadoso de evitar esta situación. En mi caso, tuve que incorporar al procedimiento automático los mecanismos para cuidar que esto no sucediera. Puedo decir que la estrategia que tomé fue muy conservadora en el sentido de que se evitó al máximo la unión entre nodos, generando siempre nuevas transiciones (arcos) por cada nueva secuencia de sufijos.

Simularé la generación del autómata mediante un procedimiento que toma como base esta estrategia conservadora, utilizando nuevamente el pequeñísimo corpus hipotético presentado antes (NIÑ~A, NIÑ~O, GAT~A, GAT~O y MONJ~A), y haré comparaciones con el procedimiento presentado antes.

El procedimiento conservador produce el mismo autómata para el palabra NIÑ~A que el procedimiento anterior, éste puede verse en (a) de la Figura 5.2. Luego, para procesar la segunda palabra (NIÑ~O) el procedimiento conservador no utiliza los arcos ya construidos. Dado que se trata de una secuencia de sufijos distinta a las secuencias existentes en el autómata actual, se crea un nuevo estado y las transiciones necesarios para representar la nueva palabra. En otras palabras, dado que no hay un camino en el autómata que describa la secuencia de sufijos de la palabra en turno, se crea un nuevo camino para representarla. Esto puede verse en (b) de la Figura 5.2.

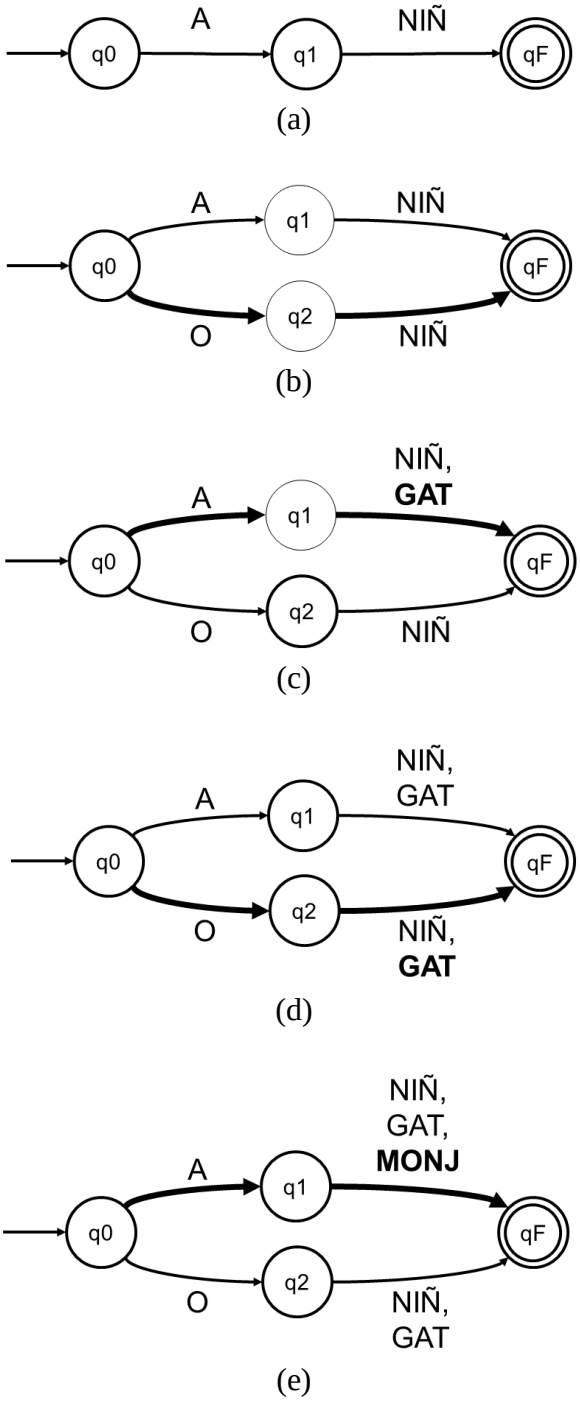


Figura 5.2. Autómata construido con la estrategia conservadora

Para las palabras siguientes (GAT~A y GAT~O) sí se aprovechan las transiciones existentes en el autómata ya que las secuencias de sufijos es la misma, aunque la base es diferente, como se puede ver en (c) y (d), respectivamente. Al final, la última palabra que-

daría representada mediante el camino ya existente en el autómata *q0, q1, qF*, que puede verse en (e) de la misma figura.

El procedimiento conservador generó un grafo con un arco para el segmento final ~A y otro para el segmento final ~O del corpus hipotético. Esto conlleva más estados y más transiciones que el autómata generado con el primer procedimiento. Además, las bases que comparten los sufijos ~A y ~O estarían duplicadas, como se puede ver en los arcos que llegan al estado final.

Después de intentar algunos procedimientos adicionales, decidí tomar esta estrategia conservadora por dos razones. La primera fue tratar de evitar que a medida que se construía el autómata se revisara si las nuevas bases generaban palabras inexistentes en el corpus. La segunda fue que una vez construido el autómata con el procedimiento conservador sería posible pensar en "simplificarlo" o "compactarlo", esto es, buscar modificar las transiciones del autómata para unir estados, lo que podría llevar a eliminar estados y transiciones redundantes.

Por tanto, la generación del autómata se hizo con la estrategia conservadora. Además, con la idea de simplificar la construcción y representación del autómata, decidí crear grupos de bases. De esta manera, en las transiciones que llevan al estado final *qF*, se utiliza el símbolo "base" seguido de un número para representar un conjunto de bases. Esto se puede ver gráficamente en la Figura 5.3, donde base1= {NIÑ, GAT, MONJ} y base2= {NIÑ, GAT}. Este autómata sería equivalente al autómata (e) de la Figura 5.2.

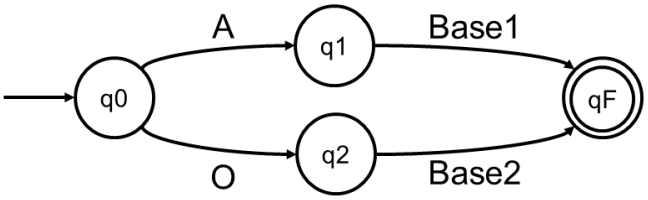


Figura 5.3. Ejemplo de autómata con grupos de bases

Si bien esta estrategia tiene redundancia y cuenta con mayor número de estados y transiciones, brinda la seguridad de no generar ninguna palabra inexistente en el corpus. Además, una vez que el autómata ha sido generado por completo, es posible realizar una tarea de simplificación del mismo para obtener una representación más compacta.

Este procedimiento se vuelve más elaborado a medida que se contemplan más palabras con mayor número de segmentos, pero en términos generales se sigue la misma idea presentada. Ahora expongo en el siguiente apartado el algoritmo computacional que permite generar el autómata de estados finitos.

5.1.2. Algoritmo para construir el autómata

En esta subsección consigno el algoritmo que utilicé para construir el autómata de estados finitos[68]. Éste sigue las consideraciones que expuse en el apartado anterior. En especial, debe recordarse que se procesaron los segmentos de los tipos de palabra de derecha a izquierda, es decir, del último al primer segmento (de los sufijos a la base).

[68] El programa de computadora que implementa este algoritmo fue hecho en Lenguaje C++ con una estrategia de programación orientada a objetos. Por esto, se crearon clases para representar al autómata, sus transiciones, sus estados y sus símbolos. Dos clases adicionales fueron creadas, una para manipular el corpus y otra para construir el autómata. Es en esta última donde se encuentra programado el algoritmo que presento aquí.

La entrada del algoritmo es un archivo con la lista de tipos de palabras segmentados. Se pensó de esta manera para mantener separado el proceso de segmentación morfológica de la construcción del autómata, lo que brinda la posibilidad de usar cualquier estrategia de segmentación y tomar su resultado para este algoritmo.

La salida del algoritmo es un autómata construido como un conjunto de estados y transiciones. Para su análisis y visualización decidí usar un diagrama de estados en forma de un grafo. Ya que resultaba poco práctico construir un solo grafo de todo el autómata dadas sus dimensiones (aproximadamente seis mil estados y siete mil transiciones), decidí fragmentarlo por segmento final de palabra. Esto es, se almacenaron tantos grafos como segmentos finales distintos.

Algoritmo para construir el autómata de estados finitos

```
Inicializar autómata A
Crear estado inicial q0
Crear estado final qF
M = 0   /*Para numerar nuevos estado*/
B = 0    /*Para numerar grupos de bases*/
qN       /* Para almacenar último estado creado en el autómata*/
Para cada palabra pal (ciclo 1):
  Para cada segmento seg de pal, comenzando por el último (ciclo 2):
      Si seg es último segmento de pal
            Si pal tiene un solo segmento (no está segmentada)
                  Si existe transición (baseB, q0, qF)
                        Asignar segmento a grupo de bases baseB
                        Aumentar frecuencia de la transición
```

Algoritmo para construir el autómata de estados finitos (continuación)

Si no existe transición (*baseB, q0, qF*)

B = B+1

Asignar segmento a grupo de bases *baseB*

Crear transición (*baseB, q0, qF*)

Si *pal* no tiene un sólo segmento

Si no existen transiciones en el autómata (autómata vacío)

M = M+1

Crear nuevo estado *qM*

Crear transición (*seg, q0, qM*)

qN = qM;

Si existen transiciones en el autómata (autómata no vació)

Buscar transición que comience en *q0* con símbolo *seg* y que no termine en *qF*

Si existe

Si *A* acepta siguientes segmentos de *pal*

Aumenta frecuencia de transiciones para cada segmento de *pal*

Si *A* no acepta siguientes segmentos de *pal*

M = M+1

Crear nuevo estado *qM*

Crear transición (*seg, q0, qM*)

qN = qM

Si no existe

M = M+1

Crear nuevo estado *qM*

Crear transición (*seg, q0, qM*)

qN = qM

Algoritmo para construir el autómata de estados finitos (continuación)

```
Si seg no es último segmento de pal
        Si seg es primer segmento de pal
                Buscar transición que comience en qN con símbolo baseB y que
                  termine en qF
                Si existe
                        Asignar seg a grupo de bases baseB
                        Aumentar frecuencia de la transición
                Si no existe
                        B = B+1
                        Asignar seg a grupo de bases baseB
                        Crear transición (baseB, qN, qF)
        Si seg no es primer segmento pal
                Busca transición que comience en qN con símbolo seg
                Si no existe
                        M = M+1
                        Crear nuevo estado qM
                        Crear transición (seg, qN, qM)
 Fin ciclo 2
Fin ciclo 1
Fin
```

La Figura 5.4 muestra el ejemplo de un grafo para el segmento final ~MOS. Como mencioné en la subsección anterior, agrupé las bases y numeré estos grupos, esto explica los símbolos base102, base107, base120 y base264. Este autómata describe la morfotáctica del segmento final ~MOS, que se puede ver como la marca flexiva verbal de número-persona. En el autómata se puede observar también la separación de la vocal temática. Re-

sultó afortunado que se mostraran las vocales temáticas de las tres conjugaciones ~A~, ~E~ y ~I~.

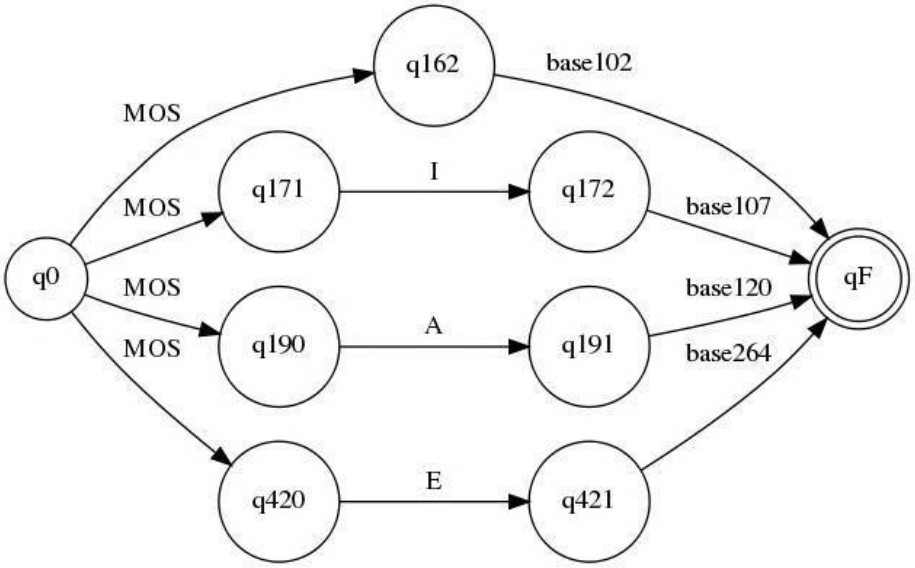


Figura 5.4 Autómata generado para el segmento ~MOS

Se espera que el autómata generado automáticamente describa la morfotáctica del español. Por tanto, sería posible extraer de él los patrones morfotácticos de esta lengua. Estos serían las secuencias de transiciones del autómata que van desde el estado inicial *q0* hasta el estado final *qF*. En el autómata de arriba (Figura 5.4) los patrones morfotácticos serían: Base~MOS, Base~I~MOS, Base~A~MOS y Base~E~MOS.

Como parte de la construcción del autómata se guardó la frecuencia de estos patrones, que por el diseño del autómata resultó equivalente al total de sus bases asociadas. Con el fin de obtener los patrones morfotácticos más pertinentes y una mejor descripción de la morfotáctica del español, decidí eliminar del autómata los patrones (secuencias de transiciones) que no tuvieran una frecuencia mayor o igual al promedio de frecuencias de todos los patrones.

Ya establecido el procedimiento general y el algoritmo para construir automáticamente el autómata, en la siguiente sección consigno los experimentos realizados.

5.2. Experimentos de generación del autómata

En esta sección presento los experimentos que llevé a cabo para generar el autómata que describe la morfotáctica de mi corpus de estudio. En términos generales, realicé dos experimentos. El primero consistió en la generación de un autómata basado en una representación fonológica del corpus y el segundo basado en los caracteres ortográficos sin ninguna modificación, lo que llamaré representación ortográfica.

Las modificaciones utilizadas para obtener la representación fonológica fueron las mismas utilizadas en trabajos anteriores sobre el descubrimiento de unidades morfológicas en el CEMC. Estas consistieron en dejar únicamente el acento de la última vocal y cambiar algunos caracteres por otros como se puede ver en la Tabla 5.1, que es una reelaboración de la Tabla A.7 presentada en Medina (2003, pág. 358).

La primera columna de la Tabla 5.1 muestra los caracteres ortográficos que fueron modificados y los caracteres que los sustituyeron. La segunda columna presenta el fonema representado. Finalmente, la tercera columna describe brevemente el contexto en que se hace la modificación.

Tabla 5.1 Modificaciones a caracteres para representación fonológica

Reelaboración de Medina (2003, pág. 358)

Modificación	Fonema	Contexto
‘v’ → ‘b’	[b]	Todos
‘z’, ‘c’ → ‘s’	[s]	toda ‘z’; ‘ce’, ‘ci’
‘c’, ‘qu’ → ‘k’	[k]	‘ca’, ‘que’, ‘qui’, ‘co’, ‘cu’
‘ch’ → ‘¢’	[č]	todos
‘g’ → ‘g’	[ɣ]	‘ga’, ‘go’, ‘gu’
‘gu’ → ‘g’	[ɣ]	‘gue’, ‘gui’
‘g’ → ‘j’	[h]	‘ge’, ‘gi’
‘h’ → ξ	-	todos
‘y’ → ‘i’	[i]	fin de sílaba, después de vocal (‘ay’, ‘ey’, ...)
‘y’, ‘ll’ → ‘y’	[y]	principio de sílaba, antes de vocal
‘rr’ → ‘»’	[r̄]	todos
‘r’ → ‘»’	[r̄]	principio de palabra; o después de sílaba que termina en ‘n’, ‘l’, ‘s’ o ‘b’.
‘r’ → ‘r’	[r]	entre vocales.

Considero interesante generar un autómata a partir de la representación fonológica ya que en ella varios caracteres se convertirían en uno solo, como en ‘qu’ → ‘k’. Además, se perderían distinciones ortográficas entre caracteres que representan el mismo fonema, como en ‘z’, ‘c’ → ‘s’. Esto tiene impacto en la morfología porque desaparece la distinción ortográfica entre alomorfos de sufijos, piénsese por ejemplos en –ción y –sión. Es más, puede producir similitud entre sufijos, piénsese en los sufijos –azo (*bal-azo*) y –so (*suspen-so*), con la representación fonológica ambos compartirían el segmento final –so.

A pesar de lo anterior, se intuye que la descripción de la morfotáctica del español será mejor con un autómata generado a partir de la representación fonológica del CEMC. Así, utilizando el algoritmo presentado anteriormente, se generaron los dos autómatas a partir de los tipos de palabras segmentados. En el siguiente apartado detallo los resultados obtenidos de estos dos experimentos y brindo una evaluación de los mismos.

5.3. Resultados y evaluación de los autómatas

En este apartado presentaré los resultados de los experimentos de generación de los autómatas a partir de los tipos de palabras del CEMC. Además, discutiré las dificultades que involucra la evaluación de estos resultados y mencionaré la estrategia que utilicé para evaluarlos. Después presentaré la evaluación del autómata que, a mi consideración, representa mejor la morfotáctica del corpus de estudio.

Comenzaré con la presentación de algunas características generales sobre los autómatas obtenidos (véase Tabla 5.2).

Tabla 5.2 Características generales de los autómatas obtenidos

	Representación	
	Fonológica	Ortográfica
Tipos de palabras procesadas	76,679	78,249
Estados	7,174	6,417
Transiciones	8,547	7,797
Patrones morfotácticos descubiertos	422	363
Tiempo de generación del autómata (minutos)	22	19

Como puede observarse, el autómata generado con base en la representación ortográfica es más compacto (contiene menos estados y transiciones), además, incluyó menos patrones morfotácticos en comparación con el autómata generado a partir de la representación fonológica. Se debe recordar que llamo patrones morfotácticos a las secuencias de transiciones que van desde el estado inicial hasta el estado final del autómata.

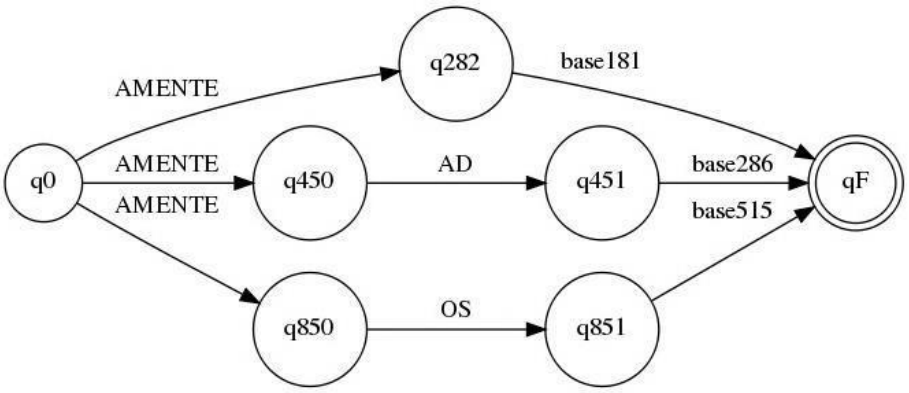
En una comparación basada sólo en los patrones morfotácticos de uno y otro autómata, resultó que ambos autómatas compartieron 350 patrones. Además, observé que el autómata basado en la representación ortográfica no incluyó algunos patrones que a primera

vista me parecen pertinentes y que sí aparecieron en el otro autómata. Listo en seguida estos patrones

/Base~ANDO~LA/
/Base~ANDO~LE/
/Base~ANDO~LO/
/Base~ANDO~LOS/
/Base~ANDO~ME/
/Base~ANDO~SE/
/Base~AR~ES/
/Base~AR~LO/
/Base~AR~SE/
/Base~ARI~A/
/Base~EDAD/
/Base~ENTE/
/Base~ETA/
/Base~I~AMOS/
/Base~IK~AMENTE/
/Base~IK~AS/
/Base~IK~OS/
/Base~IS~AR/
/Base~OTE/
/Base~TIK~A/
/Base~TIK~O/

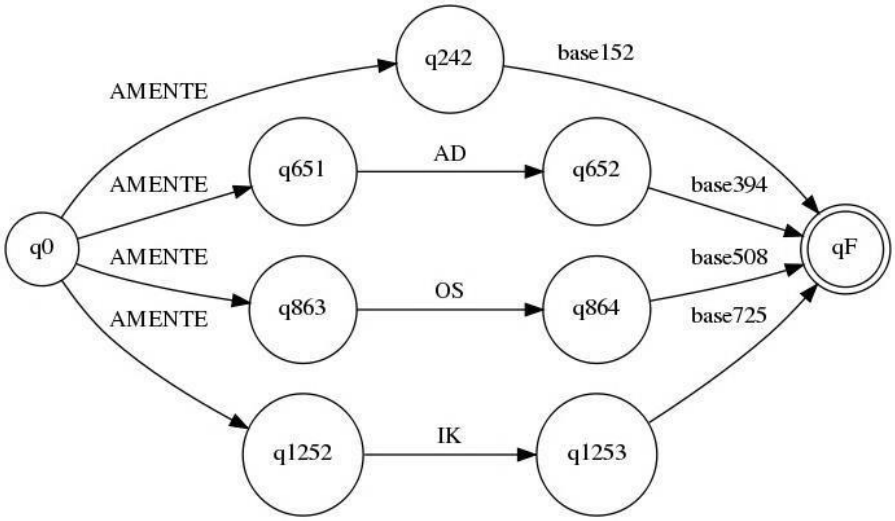
Además, comparando algunos autómatas de segmentos finales, pude observar que el autómata generado con la representación fonológica incluye patrones morfotácticos (secuencias de transiciones) pertinentes que no aparecen en el otro autómata.

Compárense, por ejemplo, los autómatas de la Figura 5.5. El autómata (b) incluye la segmentación del sufijo derivativo formador de adjetivos /~IK~/, y por tanto el patrón /Base~IK~AMENTE/, que no incluye el autómata (a). Del lado derecho de los grafos, se pueden ver algunas bases asociadas a cada camino del autómata.



base181
ABRUPT~
ACCESORI~
ACTIV~
base286
ADECU~
AIR~
AISL~
base515
AFECTU~
AMIST~
ANGUSTI~

(a) representación ortográfica

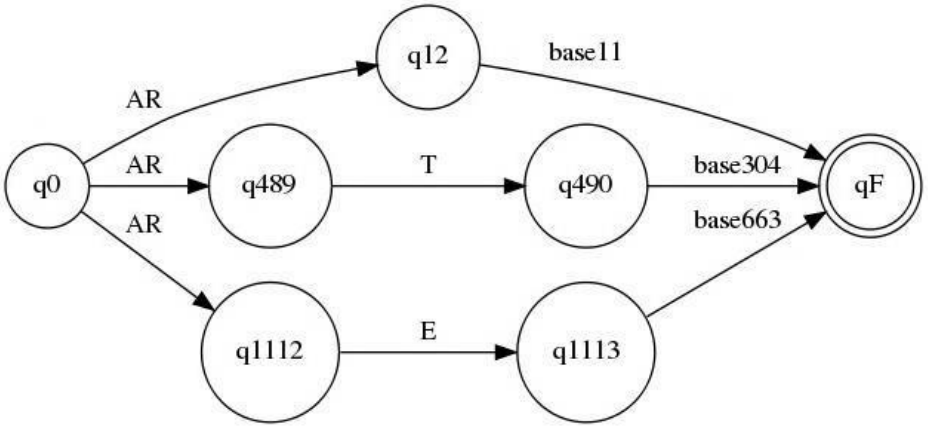


base152
ABID~
ABRUPT~
ACSESORI~
base394
ADEKU~
AIR~
AISL~
base508
AFECTU~
AMIST~
ANGUSTI~
base725
ALFABET~
ANARK~
ANATOM~

(b) representación fonológica

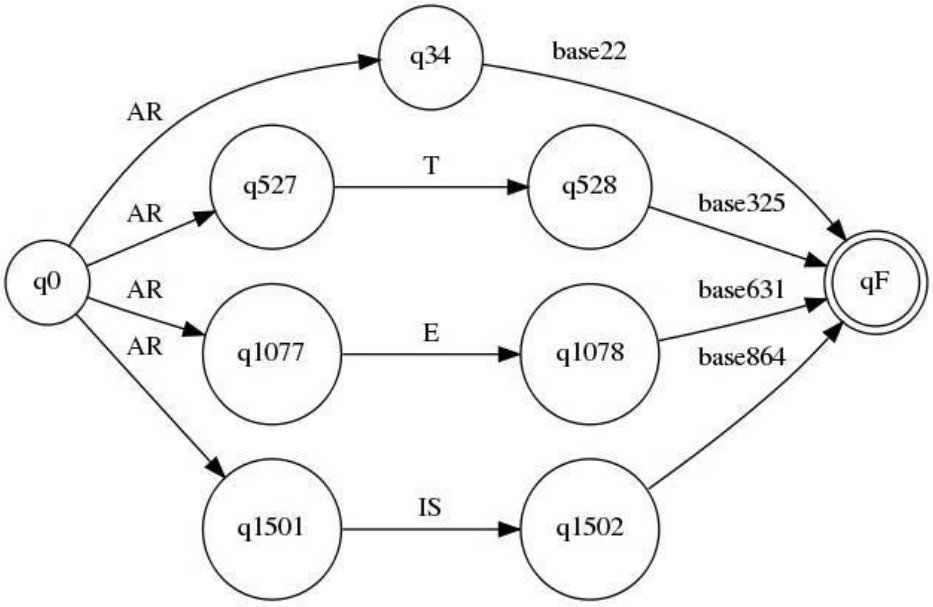
Figura 5.5 Autómatas generados para el segmento /~AMENTE/

Otro ejemplo es el de los autómatas de la Figura 5.6, generados para el segmento final /~AR/. En ellos se puede notar que el autómata generado a partir de la representación fonológica (b) incluye el sufijo /~IS~/, que deriva verbos de sustantivos o adjetivos, y que forma el patrón morfotáctico /Base~IS~AR/. Este patrón no apareció en el otro autómata (a). También se muestran algunas bases asociadas.



base11
ABAL~
ABANDON~
ABLAND~
base304
ACEP~
ADOP~
AFEC~
base663
ALET~
ARR~
BANDE-RILL~

(a) representación ortográfica



base22
ABAL~
ABANDON~
ABANS~
base325
ABUL~
ADOP~
AFEC~
base631
AKA»~
ALET~
BANDERIY~
base864
ANAL~
BITAL~
DEMO-KRAT~

(b) representación fonológica

Figura 5.6 Autómatas generados para el segmento /~AR/

Por lo anterior y dejando para trabajo futuro una comparación exhaustiva, considero para efectos de mi trabajo que el autómata basado en la representación fonológica es mejor representación de la morfotáctica de mi corpus. Por tanto, realizaré la evaluación tomando como base este autómata.

5.3.1. Evaluación

Uno de los aspectos más difíciles de resolver de mi trabajo de investigación fue la evaluación del autómata de estados finitos. Diversas interrogantes son causa de esta situación y en seguida discutiré algunas de ellas.

Algo que incidió en la dificultad de evaluación fue la perspectiva metodológica de mi trabajo. Como ya lo he mencionado, mi interés está en encontrar las regularidades morfológicas del sistema lingüístico del español mediante un método automático. Esto conlleva que la obtención de los resultados esté desprovista de una reflexión humana. La reflexión se hace, en un principio, para definir el método automático y, después, para analizar los resultados. Así, éstos son producto de una sucesión de pasos que se ejecutan mecánicamente sobre todos los casos que cumplan las condiciones previstas en el método.

De lo anterior se desprende que los resultados obtenidos automáticamente no coincidirán por completo con las propuestas consignadas en las gramáticas del español. Aunque sí deberán coincidir en buena medida para considerar al método como pertinente. Surgen entonces los primeros problemas de evaluación: ¿qué tanta coincidencia debe existir para considerar al método como acertado? ¿Contra qué gramática comparar?

Ahora bien, adoptando la forma de evaluación tradicional de la lingüística computacional, que consiste precisamente en comparar los resultados contra una propuesta humana considerada como modelo ideal (*gold standard*), quedaría por resolver la siguiente interrogante: ¿existe un autómata de estados finitos del español hecho manualmente con el que

pueda comparar el autómata generado automáticamente? Desafortunadamente no pude obtener uno[69].

En el supuesto caso de haber contado con un autómata hecho manualmente, hubieran quedado abiertas otras interrogantes. Por ejemplo, nada aseguraba que la forma de construir ese autómata fuera tan cercana a la que yo utilicé, como para dar cabida a una posible comparación de estados y transiciones. Esto es, hay distintas maneras de representar la morfología de una lengua mediante autómatas.

Por ejemplo, ya decía que una forma muy utilizada para esta representación es la fonología de dos niveles. Ésta utiliza las letras (fonemas) como símbolos en lugar de segmentos. Además, usa símbolos de entrada que se reescriben mediante reglas para producir símbolos de salida. El autómata que generé no está diseñado de esta manera.

Por todo lo anterior, decidí hacer una evaluación cualitativa del autómata generado, con especial énfasis en los patrones morfotácticos inmersos en él y basándome en mis conocimientos como lingüista y en la información que recabé sobre la morfología sufijal del español[70]. Es importante señalar que será necesario buscar en un futuro alguna manera cuantitativa de evaluar este autómata. En seguida pongo mi evaluación.

[69] Diversas pueden ser las causas que hicieron difícil encontrar un autómata del español. Por un lado, el autómata podría ser un recurso incorporado en un software comercial y por tanto no ser público. Por otro lado, podría ser que los autómatas de uso público ya no cuenten con una versión utilizable. Por ejemplo, contacté por correo electrónico a la profesora Evelyne Tzoukermann, autora del artículo "A Finite-State Morphological Processor For Spanish" (1990), pero ya no cuenta con la implementación de su trabajo.

[70] Al respecto, cabe resaltar que en las gramáticas y estudios morfológicos no hay apartados dedicados a describir la morfotáctica del español desde una visión general. Lo más común es encontrar artículos especializados para diversos fenómenos, como por ejemplo flexión y derivación, tanto nominal como verbal, y dentro de ellos secciones que mencionan los sufijos o prefijos representativos de cada fenómeno.

En términos generales, el autómata obtenido describe de manera afortunada muchos patrones morfotácticos pertinentes, esto es, que coinciden bien con lo esperado y muestran regularidades morfológicas[71]. También cuenta con patrones no pertinentes, esto es, aquellos que no dan cuenta de la morfotáctica o que no reflejan ninguna regularidad lingüística. Además contiene patrones que a primera vista no parecen pertinentes, pero que analizados con mayor detalle muestran regularidades y tendencias interesantes.

Entonces decidí evaluar con detalle algunos ejemplos de patrones que me parecieron pertinentes y otros evidentemente errados. Con esto prescindí de hacer una evaluación exhaustiva de los 442 patrones morfotácticos descubiertos y de las 55,870 bases involucradas en ellos, que me hubiera llevado mucho tiempo. Presentaré primero los autómatas que me parecen menos pertinentes (errados o muy cuestionables).

Un patrón erróneo es el asociado al segmento final /~D/ (/Base~D/) descrito por el autómata de la Figura 5.7. Se podría pensar que se refiere a la marca verbal de imperativo (*canta-d*); sin embargo, de las bases asociadas sólo una corresponde (/SABE~D/). Del resto de bases hay muchas que con terminación en –d que no debería ser segmentada, como /BERDA~D/, /BONDA~D/ y /PARIDA~D/. Otras como /USTE~D/ y /ALU~D/ tampoco deberían segmentarse, pero se explican porque en el corpus aparece /USTE/ y palabras que comienzan con /ALU~/ (/ALUMBRAR/ o /ALUSIBAS/) lo que produce la propuesta del falso sufijo.

[71] Es necesario hacer una aclaración sobre lo que considero una *regularidad*. Tanto puede referirse a un fenómeno que se presente con mucha frecuencia, como a uno que se presente con relativa baja frecuencia, pero que sea sistemático. Por ejemplo, la presencia de –aba en copretérito para verbos de la primera conjugación e –ía para los de la segunda y tercera es una regularidad, pero también los cambios que se presentan en verbos irregulares, como la aparición de una consonante en *poner/pon**g**o* o *tener/ten**g**o*, es otra regularidad. La primera es muy frecuente, la segunda es menos frecuente pero sistemática para cierto conjunto de verbos.

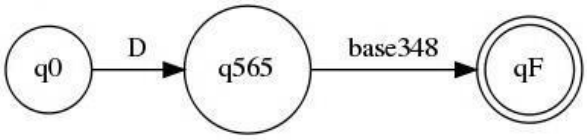


Figura 5.7 Autómata generado para el segmento /~D/

Un autómata cuestionable es el de la Figura 5.8, que representa al patrón morfotáctico /Base~GO/. Está asociado a 31 bases, entre las que destacan derivados como /ARTAS~GO/, /AYAS~GO/, /KASIKAS~GO/ y /NOBIAS~GO/, que se hubiera esperado que generaran el patrón morfotáctico /Base~ASGO/, asociado el sufijo derivativo –azgo consignado en las gramáticas.

De hecho ese patrón nunca se generó y todos los tipos de palabra del corpus con este sufijo derivativo están asociados al autómata del patrón /Base~GO/. Incluso hubo palabras en el corpus que pudieron ayudar a obtener las segmentaciones /KASIK~ASGO/ y /NOBI~ASGO/, como /KASIK~E/ o /NOBI~ESITO/, pero no fue así.

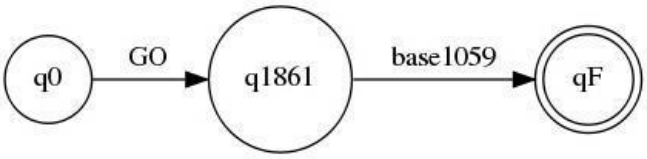


Figura 5.8 Autómata generado para el segmento /~GO/

El resto de tipos de palabras asociados a este autómata tienen una segmentación cuestionable, aunque explicable por su relación con otras palabras, como /BIKIN~GO/-/BIKIN~I/, /MUERDA~GO/-/MUERD~O/, /ESOFA~GO/-/ESOFA~JIKA/[72]. El segmento

[72] Véase cómo la segmentación de la palabra /ESOFA~JIKA/ es cuestionable porque no separa el sufijo esperado –ica; sin embargo, esto se debe al cambio fonológico en la base /g/ > /j/ que provoca que el carácter final de la base se tome como parte del sufijo. Esta es una tendencia del método y explica muchas de las segmentaciones a veces cuestionables.

/~GO/ no es el esperado, pero sí es un segmento con economía y por eso aparece en el autómata.

Otro caso parecido es el del autómata de la Figura 5.9, asociado al patrón /Base~GA/. Este autómata tiene asociadas 23 bases de la cuales ocho son verbales. En ellas se encuentran verbos regulares de la primera conjugación, que no sufren cambio en la base cuando se conjugan y por tanto su segmentación es errónea (/DELE~GA/, /NABE~GA/, /PUR~GA/), ya que se esperaría que la /G/ pertenezca a la base y no al supuesto sufijo. En otras palabras, estas bases deberían estar asociadas al patrón /Base~A/.

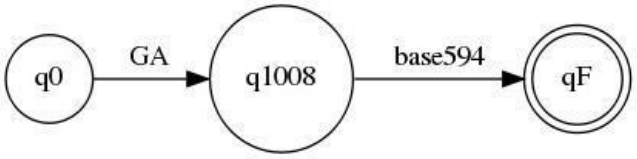


Figura 5.9 Autómata generado para el segmento /~GA/

Las otras bases pertenecen a verbos irregulares de las otras dos conjugaciones /DISTRAI~GA/, /INTERPON~GA/, /PRESUPON~GA/, /SUPERPON~GA/ y /SOBREBEN~GA/. Es muy interesante que si bien /~GA/ no es un sufijo esperado, éste segmento esté rindiendo cuenta de una regularidad morfológica (fenómeno morfofonológico) que modifica la raíz verbal aumentándole un segmento consonántico /g/[73]. Cuando el método compara estas palabras con otras parecidas, propone que el segmento /~G~/ sea parte del sufijo y no de la base, lo cual no es lo esperado, pero está dando cuenta de una regularidad del sistema[74].

[73] Véase por ejemplo Alcoba (1999, pág. 4952).

[74] Es en estos casos donde cabe la pregunta ¿es el patrón encontrado pertinente o no? Considero que sí lo es y será un morfólogo el que decidirá el mejor análisis a partir de este patrón descubierto.

Las demás bases asociadas a este autómata son nominales y presentan segmentaciones erróneas como /KOLE~GA/ o /MAN¢E~GA/, aunque algunas son explicables por sus alternantes como /ANTROPOFA~GA/-/ANTROPOFA~JIA/ o /JUER~GA/-/JUER~SA/[75].

Otro autómata que me pareció cuestionable es el de la Figura 5.10, patrón morfotáctico /Base~SO/, a pesar de que la morfología del español reconoce al sufijo –so como derivador de formas nominales a partir de verbos. Lo consideré así porque entre sus 22 bases asociadas hay derivados de diversa naturaleza junto con segmentaciones erróneas. Como patrón morfotáctico es válido, pero un análisis más detallado revela ciertas inconsistencias.

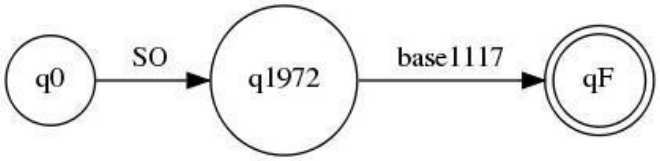


Figura 5.10 Autómata generado para el segmento /~SO/

Hay tres derivados deverbales asociados al sufijo –so. Dos de ellos muestran una regularidad morfofonológica: la pérdida de consonante cuando se adhiere el sufijo /ASEN***D***ER/-/ ASEN~SO/, /PERMI***T***IR/-/PERMI~SO/[76]. El tercero es /DISKAN~SO/ (de *descanso*).

También hay tres derivados con –oso /MASO~SO/, /SELENIO~SO/ y /NITRO~SO/. El caso de /NITRO~SO/ se explica porque alterna con /NITRO***J***ENO/, que comparte la base que propone el método (/NITRO~/) y que es la base que se usa en compuestos que aparecieron en el corpus como /NITROBENSENO/ o /NITRODERIBADO/, por lo que la segmentación no es equivocada. Los dos primeros, aunque cuestionables, con-

[75] Nuevamente el cambio consonántico en la base produce que el segmento que sufre el cambio se pase al supuesto sufijo, como pasa con /ESOFA~GO/-/ESOFA~JIKA/ (véase nota 72).

[76] Véase Moreno de Alba (1986).

firman la tendencia a asociar la parte regular a la base, en este caso el segmento/ ~O~/ (/SELENIO/-/SELENIO~SO/).

Algo similar sucede con /JENERALA~SO/ y /SALIBA~SO/, que se esperarían asociados al patrón del sufijo –azo, pero que aparecen aquí porque dejan la parte regular /~A~/ adherida a la base y no al sufijo, por ejemplo /SALIBA/-/SALIBA~ZO/[77]. El resto son bases con segmentaciones cuestionables ya que ~SO no es sufijo en la palabra (/ILE~SO/, /LAP~SO/, /OKA~SO/, /KUAR~SO/, /MAR~SO/).

Esta situación en la que un patrón morfotáctico pertinente emerge gracias a un grupo de palabras bien segmentadas, pero se extiende a otras de manera cuestionable, se explica porque la segmentación automática involucra dos aspectos en constante pugna. Por un lado, se propone como base el segmento más regular entre palabras semejantes y, por otro lado, los segmentos finales resultantes deben tener gran posibilidad combinatoria, es decir, deben ser económicos. Por lo anterior es comprensible que palabras con terminaciones semejantes tengan sufijos equivocados[78].

Otro autómata cuestionable es el de la Figura 5.11, patrón morfotáctico Base~L, no sólo porque no coincide con los estudios morfológicos, sino también porque tiene asociadas muchas segmentaciones erróneas (45%), como /FRIJO~L/, /PIE~L/, /UTI~L/ o /MIE~L/.

[77] Si se observan las bases asociadas al patrón descubierto /Base~AZO/, se confirma esta situación. Se encuentran bases cuya parte regular no termina en el segmento /~A~/ o éste cambia por otro segmento, por ejemplo /KUARTEL/-/KUARTEL~AZO/, /SILBAT***O***/-/SILBAT~AZO/ y /ESPALDA/-/ESPALDAR~AZO/.

[78] Tómese como ejemplo el caso de los patrones morfotácticos /Base~ETA/ y /Base~ETE/. Ambos resultan ser patrones muy pertinentes, no sólo porque coinciden bien con las propuestas morfológicas de diversos autores, sino también porque la mayoría de sus segmentaciones asociadas son buenas (/ESKOB~ETA/, /LENGU~ETA/, /KUN~ETA/, /ESKUD~ETE/, /SOMBRER~ETE/, /KASK~ETE/); sin embargo, lo anterior no evitó que tuvieran asociados algunos errores (/AGRI~ETA/, /DIESISI~ETE/).

De las segmentaciones restantes, el 37% incluye los sufijos derivativos –al, –ual e –il; sin embargo, resultó más económica la segmentación en ~L.

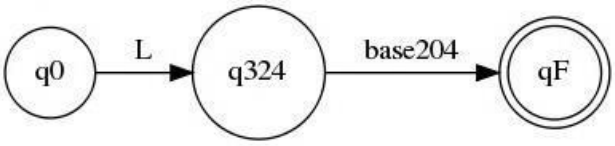


Figura 5.11 Autómata generado para el segmento /~L/

A primera vista el patrón /Base~L/ parece un error, pero si se analizan las bases asociadas es posible descubrir una tendencia. Este patrón tiene asociadas bases que conservan su forma cuando se adhiere el sufijo derivativo, como /TRIBUNA/-/TRIBUNA~L/, /NEURONA/-/NEURONA~L/ o /DOCTRINA/-/DOCTRINA~L/. En cambio, cuando se presenta pérdida de la vocal final, las bases se asocian a otro autómata con el patrón morfotáctico /Base~AL/, también descubierto por el método automático, por ejemplo /AMBIENT***E***/-/AMBIENT~AL/, /BRUT***O***/-/BRUT~AL/ o /TRIUNF***O***/-/TRIUNF~AL/[79].

Esta tendencia de segmentación en /~L/ y /~AL/ se mantuvo en los plurales, dando como resultado los patrones morfotácticos /Base~L~ES/ y /Base~AL~ES/. Creo que lo anterior demuestra que lo que emerge es un conjunto de regularidades, aunque algunas parezcan a primera vista equivocadas.

Otro caso discutible tiene que ver con la presencia de enclíticos. Ya desde el análisis de los experimentos de segmentación había detectado que los enclíticos no se separaban unos de otros, es decir, se mantenían concatenados y la tendencia era a segmentar sólo el enclítico final.

[79] Esto sigue confirmado la tendencia a dejar en la base la parte del sufijo que es constante para muchos tipos de palabras semejantes.

En los autómatas generados se puede observar la misma tendencia a separar sólo el clítico final. Además, en verbos de la primera conjugación, se separa también la marca de gerundio, pero se concatena la marca de infinitivo con el clítico. Por ejemplo, véanse los autómatas asociados al clítico *la* (Figura 5.12), que representan los patrones /Base~LA/, /Base~ANDO~LA/ y /Base~ARLA/.

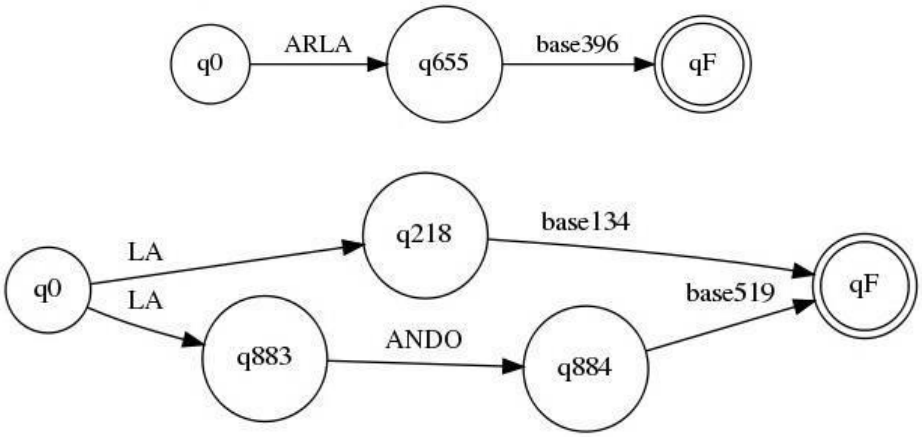


Figura 5.12 Autómatas generados para el segmento /~LA/

Para otros enclíticos ocurre una situación similar, como se puede ver en los siguientes patrones de los enclíticos *las, lo, los, le* y *les*.

/Base~ARLAS/	/Base~E~LO/	/Base~ANDO~LE/
/Base~LAS/	/Base~LO/	/Base~ARLE/
/Base~ANDO~LO/	/Base~ANDO~LOS/	/Base~LE/
/Base~AR~LO/	/Base~ARLOS/	/Base~LES/
/Base~ARLO/	/Base~LOS/	

Lo que sí descubrió el método fueron marcas de flexión de género y número en enclíticos como los siguientes:

/Base~L~A/	/Base~L~ES/	/Base~L~OS/
/Base~L~AS/	/Base~LE~S/	/Base~LO~S/
/Base~L~E/	/Base~L~O/	
/Base~L~E~S/	/Base~L~O~S/	

Otra situación relacionada con los enclíticos es que el método generó patrones morfotácticos donde un sufijo quedó dividido por la separación de un supuesto enclítico. Véase por ejemplo el último camino del autómata de la Figura 5.13, generado para el segmento final ~LE. Este camino se refiere al patrón /Base~AB~LE/, que en realidad corresponde al sufijo derivativo–able, pero por la similitud con el enclítico, el método separó el segmento /~LE/[80].

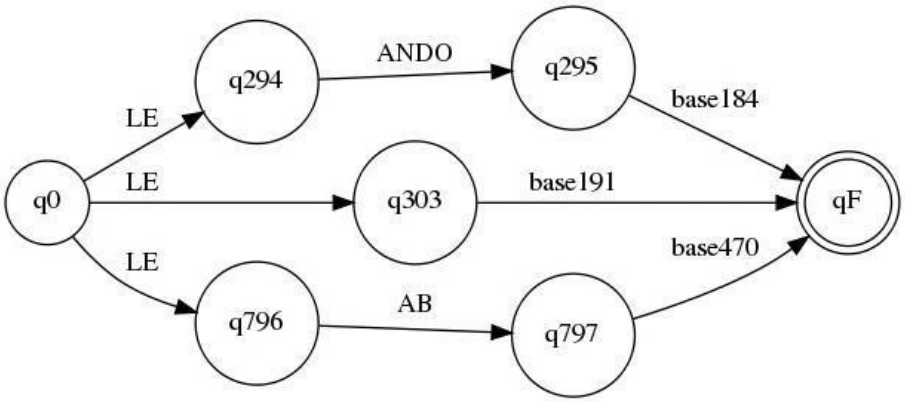


Figura 5.13 Autómatas generados para el segmento /~LE/

Ahora discutiré patrones que me parecen pertinentes (menos cuestionables).Tomo como primer ejemplo el autómata asociado al segmento final /~Ó/, que muestro en la Figura 5.14. Éste incluye algunos patrones morfotácticos que discutiré en seguida.

[80] Cabe mencionar que el método sí generó patrones para este sufijo –able: /Base~ABLE/, /Base~ABLE~S/, /Base~ABLES/ y /Base~BLE/.

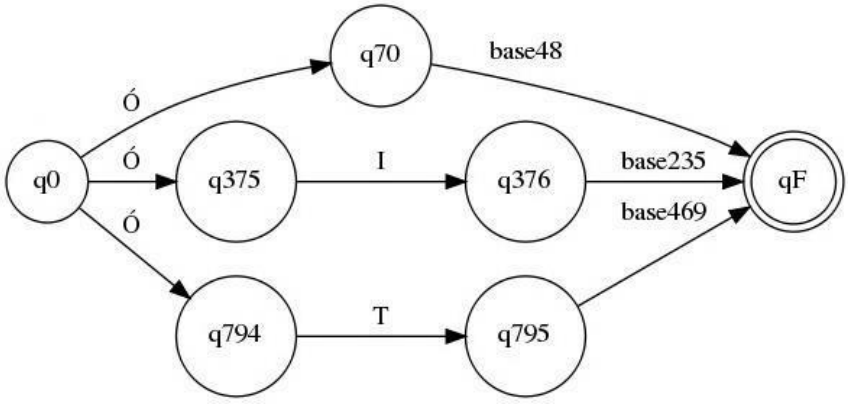


Figura 5.14 Autómata generado para el segmento /~Ó/

El primer patrón morfotáctico del autómata, /Base~Ó/, corresponde a una marca de flexión verbal de pretérito de indicativo. Prueba de ello es que las bases asociadas a este patrón son verbales y casi en su totalidad de la primera conjugación. Las cantidad de bases (1,094) hacen que éste sea el patrón más frecuente para este segmento final, algunos ejemplos son /POSTUL~Ó/, /ESTAF~Ó/, /SUJIRI~Ó/, /OBSEKI~Ó/, /»AY~Ó/ y /ESTRUCTUR~Ó/.

La segunda y tercera secuencias de transiciones, /Base~I~Ó/ y /Base~T~Ó/, corresponden a la misma flexión verbal, pero en ambos casos se separa un segmento antes del sufijo final. Las bases asociadas a estos patrones son también verbales, aunque se asocian considerablemente menos bases (59 y 53 respectivamente).

Una primera explicación de la separación de estos segmentos (/~I~/, /~T~/) se encontraría en la estrategia de segmentación que utilicé. Esta estrategia realiza cortes en los valores más altos de afijalidad de derecha a izquierda de la palabra. Entonces, el primer corte se hace en el sufijo con mayor afijalidad (en este caso /~Ó/), luego se hace otro corte en el siguiente valor más alto de afijalidad, que en el caso de estos tipos de palabra fue para

separar una base hipotética, dejando un segmento intermedio y dando como resultado las segmentaciones finales /~I~Ó/ y /~T~Ó/.

Más allá de explicar el resultado por la mecánica de segmentación, sería pertinente encontrar si estos patrones son reflejo de alguna regularidad morfológica. El patrón morfotáctico /Base~I~Ó/ describe la marca de flexión de pretérito de indicativo, pero para verbos de la segunda y tercera conjugación como /PROMET~I~Ó/ y /DIFUND~I~Ó/. Dar cuenta de esta regularidad me permite considerar este patrón morfotáctico como válido, aunque no sea totalmente acorde con las propuestas teóricas. Cabe aclarar que lo anterior no implica que esta segmentación se dé en todos los casos ya que algunos tipos de palabra no fueron segmentados de esa manera, como /SUJIRI~Ó/ y /OBSEKI~Ó/.

La tercera secuencia de transiciones, patrón /Base~T~Ó/, se explica por la relación que se establece entre formas verbales asociadas a este patrón y los derivados de estas formas que sufren cambio consonántico de /T/ por /S/ en la base, como en /ADOP***T***AR/-/ADOP***S***IÓN/, /AFEC***T***AR/-/AFEC***S***IÓN/ o /INBEN***T***AR/-/INBEN***S***IÓN. Esto se comprueba porque gran parte de los tipos de palabras asociados a este patrón son verbos que sufren este cambio.

Por ejemplo, en las palabras /ADOPTAR/, /ADOPTÓ/ y /ADOPSIÓN/ el corte más regular es en /ADOP~/. Luego, como ya dije, el segmento final /~Ó/ es muy económico, por lo que el método propone, acertadamente, el patrón /Base~T~Ó/. Nuevamente el método está dando cuenta de una regularidad del sistema[81]. También es necesario mencionar que hubo casos donde no hay cambio consonántico en la base y se separa el segmento /~T~/, como /ALIMEN~T~Ó/.

[81] Fue una tendencia en el autómata generado la separación de la consonante /T/, por lo que se generaron un buen número de patrones morfotácticos que la incluyen (/BIOLEN~T~ABA/, /BIOLEN~T~AR/).

Es pertinente mencionar que el patrón morfotáctico /Base~IÓ/ también fue descubierto y representado con el autómata de la Figura 5.15; sus bases asociadas son todas verbos de estas la segunda y tercera conjugación (/ENTEND~IÓ/, /SUSKRIB~IÓ/).

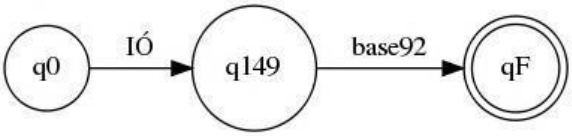


Figura 5.15 Autómata generado para el segmento /~IÓ/

Otros autómatas pertinentes son los relacionados al sufijo derivativo –(V)(C)ión. La Figura 5.16 muestra el autómata generado para el segmento final ~ASIÓN. La primera secuencia de transiciones, patrón morfotáctico /Base~ASIÓN/, es el que cuenta con más bases asociadas (436) de las dos secuencias y todas ellas verbales de la primera conjugación, como /DEKLAR~ASIÓN/, /INAUGUR~ASIÓN/ y /SELEBR~ASIÓN/, por lo que resulta ser un patrón muy pertinente[82].

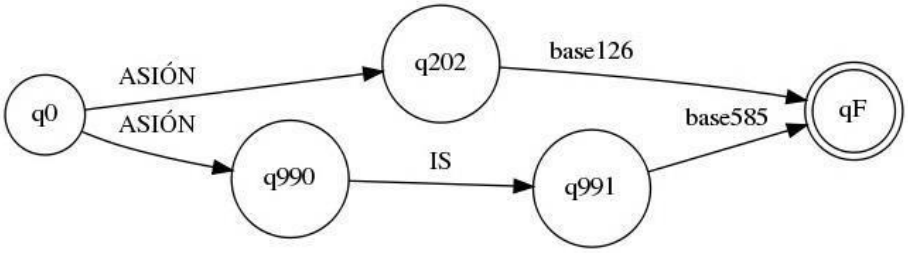


Figura 5.16 Autómata generado para el segmento /~ASIÓN/

El método propone además otra secuencia, patrón morfotáctico /Base~IS~ASIÓN/. Esta secuencia también es pertinente ya que separa el sufijo derivativo –izar que forma, a partir de sustantivos y adjetivos, verbos de la primera conjugación. Entonces el patrón

[82] Resulta interesante en este patrón que la vocal /~A~/ se queda pegada al sufijo y no a la base, como parecía que era la tendencia.

coincide bien con la formación de sustantivos a partir de estos verbos derivados, por ejemplo /ESPESIAL~IS~ASIÓN/, /KAPITAL~IS~ASIÓN/ y /DEMOKRAT~IS~ASIÓN/.

Otro autómata generado fue el de la Figura 5.17. La primera secuencia de transiciones, patrón /Base~SIÓN/, estuvo asociado a 364 bases. Este patrón es pertinente porque incluye bases que no aparecen en el patrón /Base~ASIÓN/. Se trata de aquellas terminadas en consonante, como /C/ o /P/, por ejemplo /INFE***C***~SIÓN/, /ADO***P***~SIÓN/, /DESTRU***C***~SIÓN/, /»EPRODU***C***~SIÓN/ y /DESKRI***P***~SIÓN/.

Este es un patrón que da cuenta de otros fenómenos recurrentes que cambian la base de derivación mediante pérdida de consonante (/INFEC***T***AR/-/INFEC~SIÓN/, /ADOP***T***AR/-/ADOP~SIÓN/), adición de consonante (/DESTRUIR/-/DESTRU***C***~SIÓN/) y cambio de consonante (/«EPRODU***S***IR/-/»EPRODU***C***~SIÓN/, /DESKRI***B***IR/-/DESKRI***P***~SIÓN/).

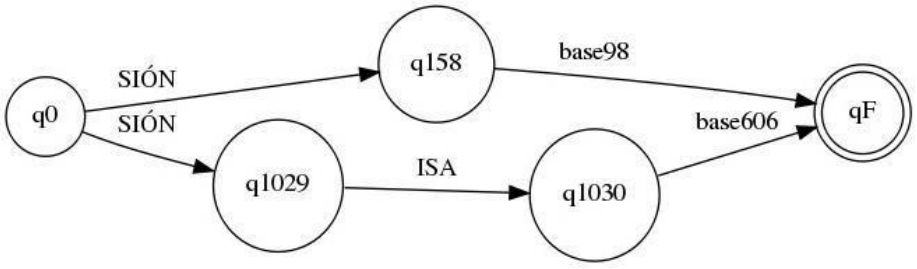


Figura 5.17 Autómata generado para el segmento /~SIÓN/

Hubo tipos de palabras, como /SALIBA~SIÓN/, que se asociaron a este patrón (/Base~SIÓN/), en lugar del patrón de arriba /Base~ASIÓN/. Lo que sucede es que en el corpus sólo aparecieron los siguientes tipos de palabras semejantes:

/SALI~BA/
/SALIBA~LES/
/SALIBA~R/
/SALIBA~SIÓN/
/SALIBA~SO/

Por tanto, la segmentación de sus bases fue más regular al incluir el segmento /~A~/. Esta situación combinada con el hecho de que /~SIÓN/ es un segmento económico da como resultado la segmentación /SALIBA~SIÓN/. Si hubieran aparecido en el corpus otros tipos de palabra como /SALIBÉ/, /SALIBO/ o /SALIBÓ/, seguramente la segmentación hubiera cambiado. Véase por ejemplo el grupo de palabras semejantes a /INAUGUR~ASIÓN/.

/INAUGUR~A/	/INAUGUR~AMOS/	/INAUGUR~ARON/
/INAUGUR~ABA/	/INAUGUR~AN/	/INAUGUR~ARSE/
/INAUGUR~AD~A/	/INAUGUR~ANDO/	/INAUGUR~ASIÓN/
/INAUGUR~AD~A~S/	/INAUGUR~AR/	**/INAUGUR~E/**
/INAUGUR~AD~O/	/INAUGUR~ARÁ/	**/INAUGUR~Ó/**
/INAUGUR~AD~O~S/	/INAUGUR~ARL~A/	
/INAUGUR~AL/	/INAUGUR~ARL~O/	

Se puede notar que el segmento más regular es /INAUGUR~/, gracias a tipos de palabra como /INAUGUR~E/ e /INAUGUR~Ó/. Luego, /~ASIÓN/, que compite en economía con /~SIÓN/, resulta mejor opción porque es económico y permite un conjunto de bases regulares. El resultado es entonces la segmentación /INAUGUR~ASIÓN/.

La otra secuencia de transiciones del autómata de la Figura 5.17, patrón morfotáctico /Base~ISA~SIÓN/, es propuesta de manera afortunada por el método ya que separa dos sufijos derivativos. Los tipos de palabra asociados a este patrón son sustantivos derivados de verbos que a su vez son derivados de sustantivos o adjetivos, por ejemplo /AJIL~ISA~SIÓN/, /BIGOR~ISA~SIÓN/, /MODERN~ISA~SIÓN/, /POLIMER~ISA ~SIÓN/ y /»ASIONAL~ISA~SIÓN/.

El método propuso además una secuencia de transiciones que separa el sufijo /~SIÓN/ en dos segmentos, ésta se puede ver en el autómata de la Figura 5.18[83]. Este pa-

[83] En esa figura hay otra secuencia (/Base~ÓN/) que se refiere al sufijo derivativo que forma aumentativos y apreciativos como /TABL~ÓN/, /SOLTER~ÓN/ o /AMOLAD~ÓN/. Como patrón es pertinente

trón morfotáctico, /Base~SI~ÓN/, resultó asociado a bases verbales de las tres conjugaciones.

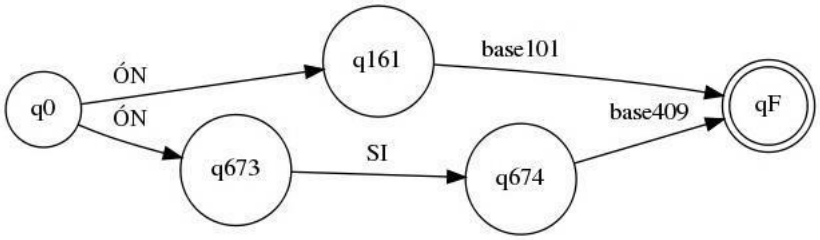


Figura 5.18 Autómata generado para el segmento /~ÓN/

El hecho de que el método propusiera separar el sufijo en dos segmentos se explica, como en casos anteriores, por dos aspectos. Por un lado, en qué tanto se parecen los tipos de palabra entre ellos y, por otro lado, en qué tan económica resulta una determinada segmentación, especialmente de los segmentos finales. El patrón /Base~SI~ÓN/ se produce por el cambio de acento entre pares como los siguientes:

/«EPERKU~SI~ÓN/	/SOBREBALUA~SI~ÓN/	/TRIPULA~SI~ÓN/
/«EPERKU~SI~ONES/	/SOBREBALUA~SI~ONES/	/TRIPULA~SI~ONES/

Se puede ver que el corte en las bases es pertinente, luego, los segmentos /~SIÓN/ y /~SIONES/ comparten el segmento /~SI~/ gracias a la distinción que causa el acento gráfico en /~Ó~/. Lo anterior se combina con el hecho de que el segmento restante /~ÓN/ es de alto nivel combinatorio, prueba de ello es que sí es un sufijo independiente (/ALAMBR~ÓN/, /ALMOAD~ÓN/). Así, resulta lógica la propuesta de segmentación /«EPERKU~SI~ÓN/. Se puede decir que este patrón da cuenta de una regularidad: el cambio de acento gráfico con la adhesión de la marca de plural.

porque da cuenta de estos derivados, pero es justo decir que también incluye una buena cantidad de derivados con los sufijos –ción y –ación.

Se puede ver que algunas segmentaciones, que parecen cuestionables a primera vista, son pertinentes porque muestran regularidades al interior del corpus (y de la lengua). Nuevamente el patrón es pertinente, pero morfológicamente inesperado. Como ya he dicho, mi análisis automático debe complementarse después con un análisis humano que decidirá la mejor descripción morfológica de la lengua de estudio.

Otro autómata asociado al sufijo –(V)(C)ión es el de la Figura 5.19. Este patrón /Base~ISASIÓN/ en lugar de separar, junta dos sufijos. Todas las segmentaciones asociadas a este patrón son sustantivos o adjetivos como /KARBON~ISASIÓN/, /SEMAFOR~ISASIÓN/ y /EXTERIOR~ISASIÓN/. Considero este patrón menos pertinente porque en la búsqueda de la morfotáctica del español, la separación en dos sufijos (/~ISA/ y /~SIÓN/) era más esperada.

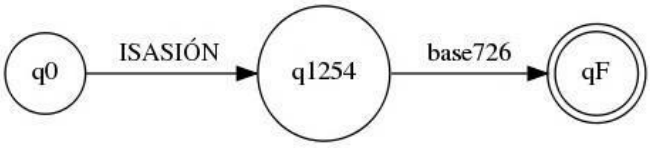


Figura 5.19 Autómata generado para el segmento /~ISASIÓN/

El último autómata propuesto por el método de la familia de sufijos –(V)(C)ión fue el de la Figura 5.20, patrón morfotáctico /Base~IÓN/. El análisis de las bases asociadas a este autómata indica segmentaciones cuestionables como /AB~IÓN/ y /KAM~IÓN/, pero es un patrón pertinente porque incluye bases verbales acordes al sufijo derivativo (/INDIJEST~IÓN/, /DESUN~IÓN/, /AUTOJEST~IÓN/).

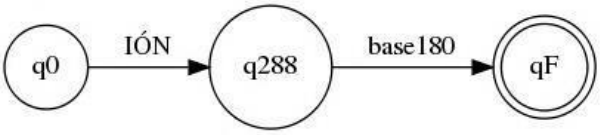


Figura 5.20 Autómata generado para el segmento /~IÓN/

Otros autómatas muy pertinentes fueron los relacionados con los segmentos finales /~AMENTE/ y /~MENTE/. Sus patrones morfotácticos asociados fueron muy regulares e incluyeron todas las formas adverbiales correspondientes. Esto es interesante porque existieron otro patrones parecidos como /Base~ENTE/, /Base~NTE/ y /Base~N~TE/, pero en ninguno se asociaron adverbios[84]. Sólo encontré dos tipos de palabra que fueron mal asociados, /SIMULTANEAME***N*~*N***TE/ e /INKUESTIONABLE***M*~*N***TE/, que son errores de escritura.

El autómata asociado al segmento final /~AMENTE/ se presenta en la Figura 5.21. La primera secuencia de transiciones, patrón morfotáctico /Base~AMENTE/, es la que tuvo más bases asociadas, todas adjetivas (/TONT~AMENTE/, /SOBERBI~AMENTE/, /»EPENTIN~AMENTE/, /EXTRAÑ~AMENTE/). La segunda secuencia, patrón /Base~AD~AMENTE/, da cuenta de manera afortunada de adjetivos con forma de participio (/ORGANIS~AD~AMENTE/, /ESTRUCTUR~AD~AMENTE/, /DESORDEN~AD ~AMENTE/, /ANTISIP~AD~AMENTE/).

[84] De hecho, los patrones /Base~ENTE/ y /Base~NTE/ también fueron pertinentes, relacionados al sufijo derivativo –(V)Vnte.

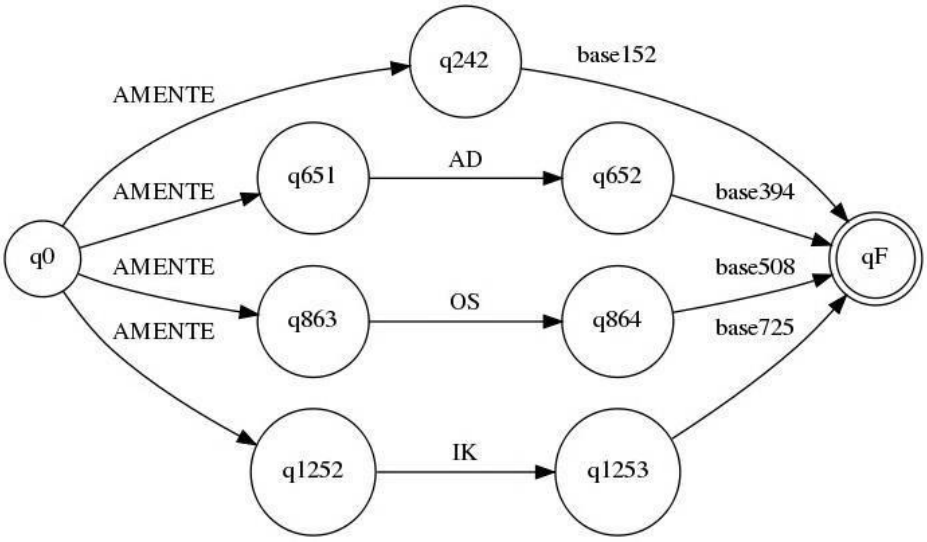


Figura 5.21 Autómata generado para el segmento /~AMENTE/

La tercera secuencia, patrón /Base~OS~AMENTE/, está relacionado con adjetivos derivados mediante sufijo –(u)os(o), como /AFECTU~OS~AMENTE/, /ESPLENDOR~OS~AMENTE/, /MARABIY~OS~AMENTE/ y /PRIMOR~OS ~AMENTE/. Finalmente la cuarta secuencia, patrón /Base~IK~AMENTE/, engloba adjetivos que son derivados con el sufijo –ico, por ejemplo /DEMOKRAT~IK~AMENTE/, /ESTADIST~IK~AMENTE/, /KATEGOR~IK~AMENTE/ y /TELEFON~IK~AMENTE/.

En resumen, el autómata del segmento final /~AMENTE/ (Figura 5.21) muestra de manera afortunada la morfotáctica involucrada en la derivación de este tipo de adverbios mediante cuatro patrones morfotácticos, tres de los cuales dan cuenta de la derivación adjetiva que da paso a la derivación adverbial.

Para el sufijo –mente se generó otro autómata, que se puede ver en la Figura 5.22. Las dos secuencias de transiciones incluyen bases de derivación que no terminan en vocal A (/ALEGR***E***~MENTE/, /DULS***E***~MENTE/, /MAYO***R***~MENTE/), lo que explica el surgimiento afortunado de este otro autómata.

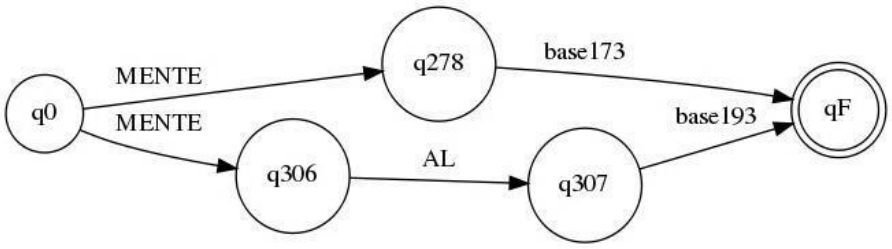


Figura 5.22 Autómata generado para el segmento /~MENTE/

La primera secuencia de transiciones, patrón /Base~MENTE/, está asociado en su gran mayoría a adjetivos terminados en /~E/, /~L/ y /~R/, aunque hubo algunos terminados en /~A/, como /SEMANTIK***A***~MENTE/, /PRESUNTUOS***A***~MENTE/ e /IMPENSAD***A*** ~MENTE/, que se esperaría que hubieran aparecido en los patrones del segmento final /~AMENTE/, discutidos arriba (Figura 5.21).

Si bien la segunda secuencia de transiciones, patrón /Base~AL~MENTE/, da cuenta de manera afortunada de adverbios a partir de adjetivos derivados (/ESPIRITU~AL~MENTE/), es justo decir que también se hubieran esperado otros patrones. Esto se debe a que hay numerosos adjetivos derivados asociados al primer patrón (/Base~MENTE/), como /IMPERSEPTIBLE~MENTE/, /INAGOTABLE~MENTE/, /INDEPENDIENTE~MENTE/ y /PREDOMINANTE~MENTE/, que incluyen otros sufijos derivativos (–ible, –able, –iente y –ante).

Un autómata que también me pareció pertinente fue el del segmento final /~AR/ (véase Figura 5.23). Es un autómata que representa la morfotáctica de la marca de infinitivo de verbos de la primera conjugación. Cuenta con cuatro secuencias de transiciones, que proponen los patrones morfotácticos /Base~AR/, /Base~T~AR/, /Base~E~AR/ y /Base~IS~AR/.

Me parece afortunado porque da cuenta de dos sufijos derivativos que producen verbos a partir de adjetivos y sustantivos, –isar y –ear. Además, el patrón /Base~AR/ es tan regular que tiene asociadas una gran cantidad de bases verbales (1,276). La aparición del segmento /~T~/ (/BIOLEN~T~AR/) se explica por las mismas razones que expuse arriba (véase autómata de la Figura 5.14).

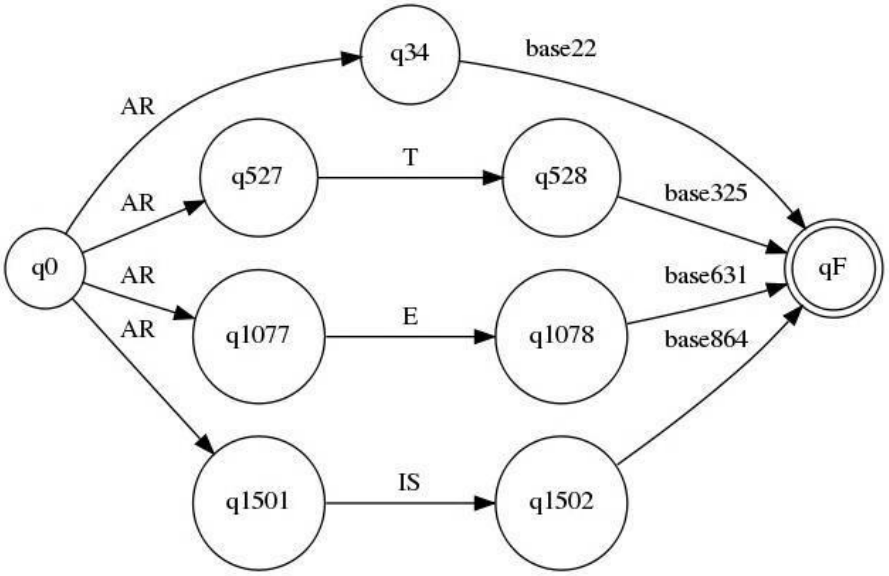


Figura 5.23 Autómata generado para el segmento final /~AR/

Otro grupo de patrones morfotácticos que me parecen pertinentes fueron los relacionados con los sufijos –ativo, –itivo e –ivo. Éstos forman adjetivos a partir de sustantivos, adjetivos y verbos. Resalto estos patrones porque dan muestra de un paradigma morfológico completo y bien organizado; sin embargo, los autómatas que incluyen estos patrones son los autómatas de las marcas de género y número, por lo que no podré mostrarlos ya que son demasiado grandes. El paradigma de patrones morfotácticos se muestra a continuación.

/Base~ATIB~A/
/Base~ATIB~O/
/Base~TIB~A/
/Base~TIB~AS/
/Base~TIB~O/
/Base~TIB~OS/

Los dos primeros patrones surgen porque todas sus bases son verbos de la primera conjugación, por tanto, antes del sufijo derivativo aparece la vocal /~A~/ (/AFIRM~ATIB~O/, /OPER~ATIB~O/, /ESPEKUL~ATIB~O/). Los cuatro patrones siguientes cubren el conjunto de tipos de palabras donde la base de derivación no termina en dicha vocal, ya sea verbal o nominal (/DESKRIP~TIB~O/, /IMPOSI~TIB~O/, /»ESTRIC~TIB~O/, /»ESOLU~TIB~O/).

La formación del paradigma completo de género y número de los cuatro patrones morfotácticos finales se debió a que aparecieron en el corpus ejemplos para cada elemento del paradigma, como se puede ver a continuación. Obsérvese que en estos patrones no hay separación de los sufijos de género y número.

/ATRAC~TIB~A/	/EJEKU~TIB~A/	/INTUI~TIB~A/	/PRODUC~TIB~A/
/ATRAC~TIB~AS/	/EJEKU~TIB~AS/	/INTUI~TIB~AS/	/PRODUC~TIB~AS/
/ATRAC~TIB~O/	/EJEKU~TIB~O/	/INTUI~TIB~O/	/PRODUC~TIB~O/
/ATRAC~TIB~OS/	/EJEKU~TIB~OS/	/INTUI~TIB~OS/	/PRODUC~TIB~OS/

El último autómata que me gustaría discutir es el del segmento final /~MA/, de la Figura 5.20. Éste tuvo sólo 25 tipos de palabras asociados, entre los cuales hay sustantivos femeninos con una segmentación equivocada, como /JIKA~MA/ o /TARI~MA/; sin embargo, también tuvo asociados tipos de palabras que comparten etimología y que le dan sentido al autómata. Éstas se forman con la terminación /~MA/ que viene del sufijo griego –μα que significa resultado de un proceso o acción, por ejemplo /DOG~MA/, /GLAUKO~MA/, /GRANULO~MA/, /EPATO~MA/ y /ENE~MA/, entre otras.

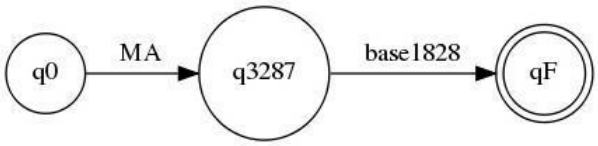


Figura 5.24 Autómata generado para el segmento /~MA/

5.3.2. Tendencias observadas

Para terminar la evaluación, y a manera de resumen, quiero resaltar las tendencias generales que pude observar en el autómata:

i) Hay diferentes secuencias de transiciones asociadas a un mismo fenómeno flexivo o derivativo. Algunas de ellas separan todos los sufijos posibles y otras los unen. Lo afortunado del método es que descubre distintas secuencias asociadas a un mismo fenómeno que hace de ellas un paradigma morfotáctico consistente. Véanse por ejemplo las distintas secuencias obtenidas para el sufijo formador de diminutivos.

/Base~ITA/	/Base~ITAS/	/Base~IT~OS/
/Base~ITO/	/Base~ITOS/	/Base~IT~AS/
/Base~IT~A/	/Base~ITO~S/	/Base~IT~O~S/
/Base~IT~O/	/Base~ITA~S/	/Base~CITO/

Si bien hay diferentes patrones morfotácticos, todos son consistentes en representar diminutivos. Ninguna otra secuencia parecida (/Base~T~OS/, /Base~T~AS/, /Base~T~A/, /Base~TA/, /Base~TO/) se asocia con diminutivos. Además palabras parecidas se asocian a otros paradigmas, como /EXIT~O~S/.

ii) Hay una tendencia por separar la consonante /T/ debido al cambio consonántico que sufre la base de muchos tipos de palabras, observable principalmente en de-

rivación nominal, por ejemplo /ARISTOKRA~***T***~A/ contra /ARISTOKRA~***S***IA/.

iii) El método descubre, de manera afortunada, patrones morfotácticos con sufijos derivativos intermedios tanto para derivación nominal como verbal, por ejemplo /Base~AL~IDAD/, /Base~AL~MENTE/, /Base~E~AR/, /Base~IK~AMENTE/, /Base~IS~AR/ y /Base~ISA~SIÓN/.

iv) El autómata no representa la morfotáctica del encadenamiento de enclíticos. La tendencia es a separar sólo el clítico final.

v) Las mejores secuencias de transiciones tienden a ser las que están asociadas a más bases, aunque esto no es una regla ya que también se descubrieron secuencias afortunadas asociadas a relativamente pocas bases.

vi) Sufijos muy económicos son segmentados en palabras donde no son sufijos, como en /MAR~SO/ y /KAM~IÓN/.

vii) Algunas veces se proponen sufijos más cortos, regulares y económicos que los esperados, como /~GO/ en lugar de /~ASGO/, /~ÓN/ en lugar de /~SIÓN/ o /~SO/ en lugar de /~OSO/ y /~ASO/. Aunque también se descubren los sufijos largos /~SIÓN/, /~OSO/ y /~ASO/.

viii) Dependiendo del paradigma de palabras semejantes involucradas en los patrones morfotácticos, algún segmento, generalmente vocálico, se une al sufijo o a la base. Por ejemplo la /A/ en /~AMENTE/ y la /A/ en la base de /SALIBA~SO/, en lugar de /SALIB~ASO/, y /DOCTRINA~L/ en lugar de /DOCTRIN~AL/.

El método descubre de manera afortunada bastantes regularidades morfológicas (algunas morfofonológicas) que explican el surgimiento de sufijos y patrones morfotácticos pertinentes, aunque a veces no coincidan con lo esperado. Entre estas regularidades están:

i) Aparición de consonante /g/ en verbos irregulares, por ejemplo /INTERPONER/-/INTERPON~***G***A/.

ii) Pérdida de vocal final de la base de derivación, por ejemplo /SILBAT***O***/-/SILBAT~AZO/ o /AMBIENT***E***/-/AMBIENT~AL/.

iii) Presencia de vocales temáticas de las tres conjugaciones /Base~A~R/, /Base~E~R/ y /Base~I~R/.

iv) Aparición de vocal /i/ en pretérito de indicativo: /Base~I~Ó/ y /Base ~IÓ/.

v) Cambios consonánticos en derivados, por ejemplo /ADOP***T***AR/-/ADOP~***S***IÓN/, /DESKRI***B***IR/-/DESKRI***P***~SIÓN/ o /ESOFA~***G***O/-/ESOFA~***J***IKA/.

vi) Cambio de acento en presencia de marca de plural, por ejemplo /TRIPULA~SI~ÓN/-/TRIPULA~SI~ONES/.

A mi juicio, la mayoría de los patrones morfotácticos descubiertos por el método son pertinentes, por lo que considero al autómata como una buena primera representación de la morfotáctica del corpus y por tanto del español de México. En el anexo B se pueden ver algunos autómatas adicionales generados por el método. Además, pongo en un disco compacto adjunto a este trabajo todos los autómatas generados y sus bases asociadas. La descripción del contenido de este disco se encuentra en el Anexo D.

Por otro lado, en el Anexo C se listan los cien patrones morfotácticos más frecuentes del corpus (Tabla 7.2) y se muestra la curva que relaciona la posición en la lista y la

frecuencia de los 422 patrones encontrados (Figura 7.1). La lista completa de patrones morfotácticos se incluye también en el disco compacto.

En la siguiente sección puntualizo el método propuesto para el descubrimiento de la morfotáctica del español.

5.4. Método para descubrir la morfotáctica

En esta sección describo de manera resumida el método que propongo para generar automáticamente una descripción morfológica del español mediante el descubrimiento de su morfotáctica. Generar una descripción de este tipo se puede ver como un procedimiento con dos grandes fases:

(i) Descubrir las unidades morfológicas.

(ii) Descubrir los patrones morfotácticos que describan su orden y secuencialidad.

Después de la investigación realizada puedo proponer el siguiente método para describir automáticamente la morfología del español. Esta descripción incluye el descubrimiento de sus bases, sufijos y patrones morfotácticos:

4. Cuantificar la afijalidad de segmentos: calcular un índice de afijalidad para cada posible corte al interior de todos los tipos de palabras del corpus mediante el promedio de tres medidas de afijalidad: entropía, economía y cuadros.
5. Descubrir las bases y sufijos: segmentar cada tipo de palabra mediante cortes sucesivos hacia la izquierda en el valor máximo del índice calculado siempre que sea mayor a 0.5.
6. Descubrir los patrones morfotácticos: generar automáticamente un autómata de estados finitos que describa el orden y secuencialidad de las bases y sufijos descubiertos.

De manera esquemática, la Figura 5.25 muestra los pasos del método que se propone. Se puede observar que la entrada del método es un corpus, que para efectos de esta tesis fue el CEMC. La salida es el autómata de estados finitos representado como un diagrama de estados en forma de grafo. El autómata, como ya se mostró en la sección anterior, incluye los patrones morfotácticos descubiertos.

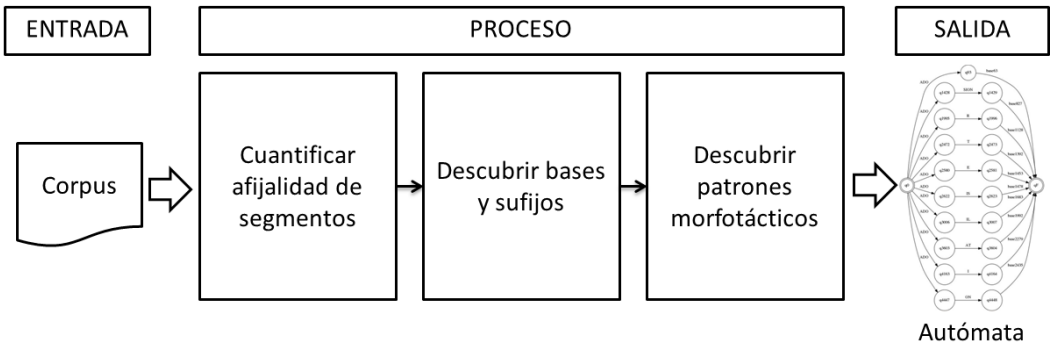


Figura 5.25 Esquema general del método propuesto

La representación del autómata es opcional, se podría también utilizar una tabla de transiciones. Es más, se podría generar la gramática de estados finitos equivalente. En esta investigación se decidió usar el autómata en forma de diagrama de estados por las razones expuestas anteriormente.

Cabe resaltar que el corpus de entrada puede ser distinto, ya sea de español o de otra lengua con morfología predominantemente sufijal. Lo importante es que sea, en la medida de lo posible, un corpus representativo de la lengua de estudio. Por otro lado, a futuro sería factible pensar en combinar sufijos y prefijos haciendo modificaciones al método propuesto. Incluso se podría pensar en ampliar el método para abarcar más fenómenos concatenativos, como la composición, y otras unidades no afijales, como los enclíticos, que en esta investigación no fueron descubiertos.

6. Conclusiones

Presento en estas conclusiones un resumen de cada capítulo de la tesis y de los experimentos realizados. Además, reviso los objetivos y preguntas de investigación establecidas en el capítulo introductorio. Después, consigno el método que propongo para describir la morfotáctica del español, sus problemas, ventajas y trabajo futuro. Al final expongo las conclusiones finales de este trabajo de investigación.

Esta tesis comienza con un capítulo introductorio donde presenté los problemas que la motivaron, las preguntas y objetivos de investigación, la delimitación de su alcance y la metodología para desarrollar el trabajo.

El capítulo uno estuvo dedicado a la morfotáctica. Se incluyó su definición y algunas posturas que intentan explicar su naturaleza. Con la idea de conocer la descripción morfológica que generaría automáticamente, se consignó gran parte de la morfotáctica sufijal del español, tanto verbal como nominal. El capítulo cerró con la revisión de un procedimiento para descubrir esquemas morfotácticos que ayudó a orientar mi investigación.

En el capítulo dos se revisaron algunos métodos de descubrimiento de unidades morfológicas en corpus, cuatro de ellos en detalle porque son métodos no supervisados de segmentación morfológica. Uno de ellos fue el método que utilicé para desarrollar mi trabajo de investigación. Al final del capítulo se hizo una comparación entre los cuatro métodos que sirvió, entre otras cosas, para puntualizar mi perspectiva de trabajo y justificar la selección del método que calcula la glutinosidad (afijalidad) como método para realizar mi investigación.

Presenté en el capítulo tres los fundamentos de las gramáticas formales y de los autómatas de estados finitos. Sobre las gramáticas se incluyeron sus antecedentes, definición, tipos y los lenguajes que generan. Sobre los autómatas se expusieron su definición, tipos y representaciones. Revisé también la equivalencia entre gramáticas y autómatas, aspecto que fue clave para decidir construir un autómata en lugar de una gramática. Al final hice un repaso de cómo al morfología computacional ha tratado la morfotáctica, especialmente bajo los rubros de morfotáctica de estados finitos y morfología de dos niveles. Este repaso ayudó a confirmar que era mejor estrategia crear el autómata.

El capítulo cuatro incluyó la información sobre los experimentos de segmentación morfológica automática. Se consignaron los resultados del primer acercamiento al problema de descubrir todos los sufijos de una palabra. Luego se plasmó una discusión detallada del método de segmentación, lo que permitió establecer distintas variantes del mismo para formar un grupo de experimentos. A final se expuso la manera de evaluar estos experimentos y se dieron los resultados de esa evaluación. Con los resultados obtenidos fue posible seleccionar una estrategia para descubrir las bases y sufijos del corpus de estudio.

Finalmente, en el capítulo cinco se presentó el algoritmo que se desarrolló para generar el autómata de estados finitos y los experimentos que se llevaron a cabo. Además, se plasmó la idea general del método para descubrir la morfotáctica. Finalmente se ofreció la evaluación de algunos patrones morfotácticos inmersos en el autómata para determinar la pertinencia del método propuesto. El detalle de los experimentos de este capítulo y del anterior se resume en la siguiente sección.

6.1. Resumen de experimentos

En esta sección resumo el trabajo de experimentación realizado durante la investigación. Se pueden distinguir dos grupos de experimentos, primero, los relacionados con la segmentación morfológica automática, y segundo, los experimentos de generación del autómata de estados finitos.

Se estableció en el capítulo uno que para describir la morfotáctica de una lengua era necesario determinar los morfemas de las palabras. Para esto se adoptó el método que calcula un índice de afijalidad para cada posible corte dentro de una palabra. En investigaciones anteriores ya se había determinado que el valor máximo de este índice corresponde a una frontera morfológica que separa la base de los sufijos. Mi tarea consistió en usar este índice para obtener todos los cortes posibles en la palabra que correspondieran con fronteras morfológicas entre la base y entre secuencias de sufijos individuales.

El primer acercamiento que tomé fue cortar la palabra en los picos de afijalidad, esto es, donde un índice de afijalidad dentro de la palabra es más alto que el anterior y el posterior. Para evaluar si las segmentaciones obtenidas con esta estrategia eran regulares, se implementó un truncador de palabras que se acopló con un resumidor automático de documentos. Se realizaron experimentos en español, francés e inglés, truncando la palabra en el primer pico a la izquierda, en el primer pico a la derecha, en el valor máximo de afijalidad y con otras estrategias de regularización de palabras.

En español y francés, el mejor resumidor fue el que truncaba en el primer pico de afijalidad a la izquierda de la palabra. Esta evaluación extrínseca, mediante un programa que usó la estrategia de segmentación, demostró que la segmentación era regular, pero no decía si era morfológicamente pertinente (si descubría unidades morfológicas). Después de

observar algunos archivos con palabras truncadas observé que esta estrategia segmentaba al interior de las bases y generaba segmentaciones muy cuestionables.

Se tomó la decisión de analizar más a fondo el método de segmentación poniendo especial interés en el comportamiento de las medias de cuadros, entropía, economía y del mismo índice de afijalidad. Se identificaron cuatro condiciones involucradas en el cálculo del índice de afijalidad: qué medidas se combinan, con qué operación matemática se combinan, la direccionalidad hacia donde se hacen los cortes y el uso del umbral de 0.5 como valor mínimo para un corte.

Las combinaciones de estas condiciones arrojaron un total de dieciséis experimentos que se llevaron a cabo en un corpus de español. El corpus estuvo constituido por un listado de palabras proporcionado por el Laboratorio de Lenguaje Natural y Procesamiento de Texto del IPN, los vocablos del DEM y los tipos de palabras del CEMC. Se decidió hacer una evaluación intrínseca de estos experimentos mediante un corpus segmentado a mano que incluyó flexión y derivación, tanto nominal como verbal (1,600 tipos de palabras).

La estrategia de segmentación automática con mayor cantidad de aciertos fue la que segmenta hacia la izquierda en el valor máximo del índice de afijalidad mayor a 0.5, calculado mediante un promedio de las tres medidas (cuadros, entropía y economía). Esta estrategia obtuvo mejores resultados en la parte de flexión verbal del corpus de evaluación, aunque en términos generales presentó la tendencia a obtener menos segmentos que los esperados.

Una vez determinada la estrategia de segmentación se realizaron los experimentos para obtener el autómata de estados finitos a partir del corpus de estudio: el CEMC. Se realizaron dos experimentos que consistieron en crear un autómata basado en una representación fonológica del corpus y otro sin cambiar la representación ortográfica. Ya que fue im-

práctico el manejo del autómata de manera completa dada la cantidad de estados y transiciones, para su representación y análisis se dividió en tantos autómatas (grafos) como segmentos finales.

La comparación de algunos autómatas generados con ambos experimentos para los mismos segmentos finales mostró que el autómata obtenido a partir de la representación fonológica mostraba mejores patrones morfotácticos; especialmente por la presencia de sufijos derivativos intermedios que no aparecieron en el otro autómata. Además, éste tenía patrones morfotácticos pertinentes que no tenía el autómata de la representación ortográfica. Entonces, se decidió como descripción morfotáctica del corpus el autómata generado de la representación fonológica y se procedió a su evaluación.

Se realizó una evaluación cualitativa del autómata mediante el análisis de algunos grafos de segmentos finales. Se encontraron patrones morfotácticos no pertinentes debido a que son errados, como /Base~D/, /Base~GO/, o porque no dan cuenta del encadenamiento de sufijos, como /Base~ISASIÓN/. También se observó que la mayoría de los patrones fueron pertinentes ya que daban cuenta de regularidades del sistema morfológico del español. Algunos ejemplos son los patrones /Base~ASIÓN/, /Base~SIÓN/, /Base~IS~ASIÓN/ y /Base~ISA~SIÓN/; los dos últimos dan cuenta de la morfotáctica del sufijo derivativo –izar.

Otros ejemplos son los patrones /Base~AMENTE/, /Base~AD~AMENTE/, /Base~OS~AMENTE/ y /Base~IK~AMENTE/, éstos muestran la morfotáctica del sufijo derivativo –amente que da cuenta de tres sufijos derivativos intermedios que crean adjetivos a partir de los cuales se generan los adverbios, esto son –ad(a), –os(a) e –ik(a).

Es justo decir que no todas las palabras asociadas a los patrones pertinentes presentaron segmentaciones válidas. Esto sucede porque el final de la palabra coincide con sufijos

muy económicos y se generaliza la segmentación. Afortunadamente, estos casos fueron la minoría.

Algunas observaciones generales obtenidas de la evaluación fueron que se presentaron varios patrones morfotácticos equivalentes donde uno separaba en varios segmentos y otro los concatenaba, como /Base~ITA/ y /Base~IT~A/. Hubo tendencia a separar un segmento /~T~/ que dio cuenta de los cambios vocálicos que se dan en algunas bases como /ARISTOKRA~***T***~A/ contra /ARISTOKRA~***S***IA/. También se descubrieron un buen número de patrones morfotácticos con sufijos derivativos intermedios. Finalmente, se dio cuenta de distintas regularidades morfológicas asociadas a los patrones morfotácticos, algunas de ellas de carácter morfofonológico.

Con base en este resumen de la experimentación realizada, puedo hacer una revisión de las preguntas y objetivos de investigación planteados al inicio.

6.2. Revisión de objetivos

Fueron dos las interrogantes planteadas al inicio de este trabajo. Sobre la primera, que cuestionaba la posibilidad de generar automáticamente un aparato formal de descripción morfológica a partir de corpus, que diera cuenta de los sufijos y sufitáctica del español, considero que los resultados muestran que sí fue posible hacerlo.

El aparato formal de descripción morfológica fue un autómata de estados finitos inferido automáticamente a partir del Corpus del Español Mexicano Contemporáneo. Se considera una descripción morfológica pertinente porque sus transiciones, estados y alfabeto de símbolos representan de manera afortunada bases, sufijos y sufitáctica del español. La revisión de los patrones morfotácticos inmersos en el autómata permitió ver la tendencia del autómata para descubrir y presentar regularidades morfológicas pertinentes. Claro que que-

dan asuntos pendientes por resolver; sin embargo, el autómata obtenido es una primera descripción morfológica del español mexicano inferida automáticamente.

La segunda pregunta era si una gramática de estados finitos es suficiente como aparato formal de descripción morfológica de los sufijos y sufitáctica del español. Primero hay que aclarar que no se generó una gramática, sino un autómata; sin embargo, la teoría de autómatas y gramáticas formales han demostrado que son equivalentes. Por tanto, creo que puedo discutir esta pregunta a partir de la construcción del autómata.

Dado que el autómata construido automáticamente resultó una buena primera descripción de la morfología del español, específicamente de sus bases, sufijos y morfotáctica, y no fue necesario ningún mecanismo auxiliar que no fuera parte de la definición formal del autómata, considero que el autómata fue un aparato de descripción formal suficiente para esta descripción morfológica. Sin embargo, es necesario mencionar algunas cuestiones.

Restringir solamente a la morfología sufijal, dejando de lado los fenómenos de parasíntesis y composición, ayudó enormemente a que el autómata fuera suficiente. La morfología sufijal es un ejemplo claro de la morfología concatenativa secuencial, por lo que un autómata, donde un estado depende sólo del estado anterior, resultó una representación suficiente.

En un fenómeno de parasíntesis el autómata tendría problemas para representar que un estado depende de estados anteriores no adyacentes. En la composición se esperaría que el autómata represente cierta jerarquía de segmentos para describir varias bases con sus respectivos sufijos. En este caso desconozco si será suficiente.

Otro de los aspectos que permitió que el autómata generado fuera suficiente, es que no involucré cambios morfofonológicos (vocálicos, consonánticos, etcétera) en las bases o sufijos, por lo que el autómata no incluye información de este tipo. Al respecto, fue intere-

sante que el método de segmentación diera cuenta de manera indirecta de algunos de estos fenómenos.

Por otro lado, la manera de construir el autómata aún tiene debilidades, que se discutirán en la siguiente sección, por lo que si bien fue una buena primera descripción morfológica, aún no refleja toda la información morfológica esperada.

Ahora revisaré el cumplimiento de los objetivos planteados. Para este trabajo de investigación establecí el objetivo de desarrollar un método automático no supervisado para generar, a partir de corpus y mediante una gramática de estados finitos, una descripción morfológica del español, acotada al descubrimiento de sus sufijos y su morfotáctica.

El objetivo se cumplió dado que ahora se cuenta con un método con las características requeridas a pesar de que no se genera una gramática, sino un autómata que, como ya se ha establecido reiteradamente, es equivalente a la gramática.

Los pasos del método propuesto son los siguientes:

1. Cuantificar la afijalidad de segmentos: calcular un índice de afijalidad para cada posible corte al interior de todos los tipos de palabras del corpus mediante el promedio de tres medidas de afijalidad: entropía, economía y cuadros.
2. Descubrir las bases y sufijos: segmentar cada tipo de palabra mediante cortes sucesivos hacia la izquierda en el valor máximo del índice calculado siempre que sea mayor a 0.5.
3. Descubrir los patrones morfotácticos: generar automáticamente un autómata de estados finitos que describa el orden y secuencialidad de las bases y sufijos descubiertos.

En consecuencia, también se lograron los dos objetivos específicos planteados. El primero fue el descubrimiento, a partir de corpus y mediante un método no supervisado de

segmentación morfológica automática, los sufijos y sufitáctica de la lengua española. Esto se realiza en el paso dos del método propuesto, que descubre las bases y sufijos de los tipos de palabras del corpus. Al descubrir cada sufijo, se está descubriendo también su secuencialidad.

El segundo objetivo fue generar, a partir de los sufijos y sufitáctica descubiertos, una gramática de estados finitos que describa la morfotáctica sufijal del español. Esto se realiza en el segundo paso del método, aunque en lugar de gramática se genera un autómata. Una vez generado el autómata, un programa de computadora puede seguir algunos pasos establecidos en la teoría de autómatas y gramáticas formales para obtener la gramática equivalente.

En resumen, se han cumplido con los objetivos planteados y se han resuelto favorablemente las preguntas de investigación formuladas al inicio de esta tesis. Revisaré ahora los problemas, ventajas y propuestas de mejora de esta investigación.

6.3. Problemas del método y trabajo futuro

En esta sección describo los problemas y trabajo futuro del método propuesto. Como se dijo en la evaluación del autómata generado, hay patrones morfotácticos que deben ser revisados con mayor detenimiento para buscar mejorarlos. No se trata de forzar el método a obtener lo que las gramáticas consignan, pero si es necesario revisar si hay algún aspecto del método que pueda ser mejorado.

Si bien el autómata generado ya es una buena primera descripción de la morfotáctica del español, presenta al menos una carencia. Ésta es la falta de agrupamiento de patrones morfotácticos a manera de paradigmas morfológicos. Esto es, el método propuesto fue capaz de generar los paradigmas de género y número tanto para bases simples (/Base~A/,

/Base~A~S/, /Base~O/ y /Base~O~S/), como para bases con sufijos derivativos como las siguientes:

/Base~TIB~A/	/Base~ID~A/	/Base~ER~A/	/Base~AD~A/
/Base~TIB~AS/	/Base~ID~A~S/	/Base~ER~A~S/	/Base~AD~A~S/
/Base~TIB~O/	/Base~ID~O/	/Base~ER~O/	/Base~AD~O/
/Base~TIB~OS/	/Base~ID~O~S/	/Base~ER~O~S/	/Base~AD~O~S/

Sin embargo, estos patrones fueron generados separadamente, es decir, las transiciones del autómata no dan cuenta de que forman un paradigma y el análisis humano es el que reconoce en ellos su carácter paradigmático. Sería conveniente que el autómata generado refleje automáticamente estos paradigmas.

Para ello, dado que actualmente cada patrón morfotáctico está representado por una secuencia de transiciones separada, sería necesario combinarlas tomando en cuenta sólo las bases que aparecen en todos los elementos del paradigma, lo que se puede ver como una operación de intersección entre conjuntos de bases. Para el caso de los patrones morfotácticos del sufijo –tiv(o), el autómata hipotético podría ser como el de la Figura 6.1.

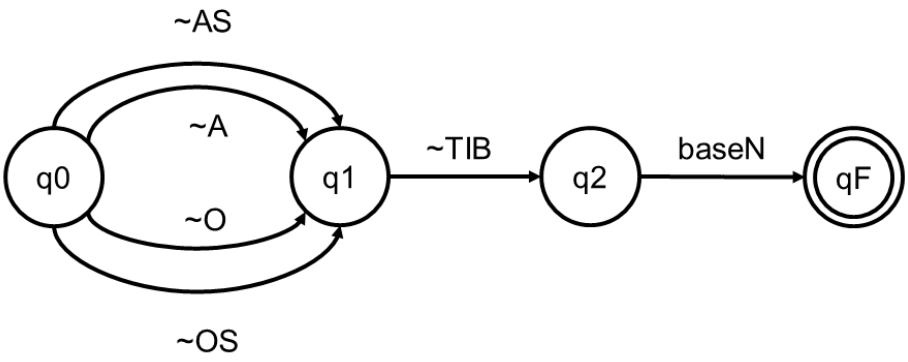


Figura 6.1 Ejemplo de autómata hipotético para sufijo –tiv(o)

Representar los paradigmas de los sufijos –id(o), –er(o) y –ad(o) requeriría también de modificaciones a las secuencias de transiciones. En la Figura 6.2 se puede ver el autóma-

ta hipotético que se esperaría obtener para el sufijo –er(o) una vez que se agrupen las bases presentes en todos los patrones morfotácticos.

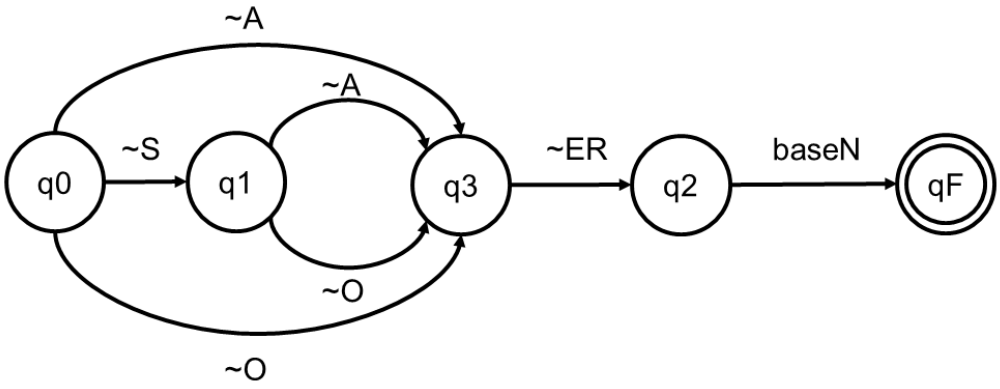


Figura 6.2 Ejemplo de autómata hipotético para sufijo –er(o)

A la larga sería posible pensar en la presencia de estados finales intermedios, lo que conlleva cambiar la dirección de las secuencias de transiciones del autómata actualmente generado. Comenzar a construir automáticamente el autómata por las bases y no por los sufijos finales no es un problema trivial, ya que éstas forman a un conjunto muy grande y variado de segmentos en comparación con los sufijos. Haber empezado a construir el autómata por los sufijos dio la ventaja de contar al principio del procesamiento con relativa menos variedad de segmentos.

Para los patrones morfotácticos del sufijo –er(o), otro autómata hipotético sería el de la Figura 6.3. Nótese el estado final *q3* que marca el final del paradigma de género. Además, véase cómo el estado final es ahora el final de la palabra.

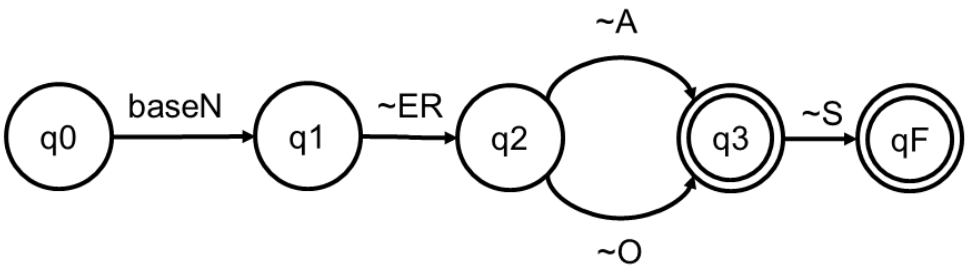


Figura 6.3 Ejemplo de otro autómata hipotético para sufijo –er(o)

En resumen, como trabajo futuro se propone agregar, al menos, un paso al método propuesto en este trabajo. Éste se podría ver como una simplificación del autómata y consistiría en la búsqueda de paradigmas morfológicos mediante la unión de secuencias de transiciones a partir de la intersección de conjuntos de bases. No son pocas las posibilidades de nuevos diseños de autómatas que se pueden idear, pero se debe tener cuidado en no perder de vista que el objeto de estudio es la lengua y no el autómata.

Un problema que queda por resolver en este trabajo de investigación es encontrar otras opciones de evaluación del autómata. Por un lado es necesaria una evaluación completa del mismo, lo que llevará bastante tiempo. Por otro lado, sería bueno encontrar a futuro una forma de evaluación que no dependa de un evaluador.

Esta idea de buscar otras alternativas de evaluación también debe considerarse para la segmentación morfológica. Esto se debe a que la evaluación realizada fue estricta y el corpus de evaluación contenía un inventario muy variado de sufijos y alomorfos derivativos. Además, no es lo mismo evaluar contra una lista de tipos de palabras (como se hizo aquí) que contra ocurrencias de palabras. La segunda es la más común en trabajos publicados de segmentación automática.

De hecho, quedaría pendiente una evaluación de los resultados de este trabajo con resultados generados con otras propuestas de segmentación morfológica, por ejemplo la de

Goldsmith o la basada en algoritmos genéticos, aunque éstas no descubren varios sufijos por palabra. Incluso se puede pensar en comparar en un futuro este método con el de Creutz y Lagus para ver qué proponen ambos métodos en cuanto a la morfotáctica del español. Es más, se podría pensar en combinar lo mejor de todos los métodos en la búsqueda de un nuevo método no supervisado.

Sobre los patrones morfotácticos obtenidos quedan varios aspectos que revisar, de los cuales sólo mencionaré algunos de ellos. El primero es un análisis sobre el grupo de patrones morfotácticos equivalentes, pero que presentan diferentes segmentaciones. Véanse por ejemplo los siguientes patrones del sufijo –ad(o).

/Base~AD~A/	/Base~AD~AS/	/Base~ADA~S/
/Base~AD~A~S/	/Base~AD~OS/	/Base~ADO~S/
/Base~AD~O/	/Base~ADA/	/Base~ADAS/
/Base~AD~O~S/	/Base~ADO/	/Base~ADOS/

Puede observarse que los patrones van desde los más segmentados (/Base~AD~A~S/) hasta los que no presentan segmentación (/Base~ADAS/). Quedaría pendiente una revisión a mayor detalle de estos grupos de patrones para identificar lo que motiva su aparición y tal vez proponer un cambio en el método para reunir los que son equivalentes.

Otro pendiente es la incorporación de prefijos a la descripción morfológica propuesta. Aunque en español el estatus de los prefijos tiende más hacia la composición, sería importante tomarlos en cuenta para la descripción de la morfotáctica. De hecho, los fenómenos de composición serían otro aspecto a considerar en un futuro.

Para tomar en cuenta prefijos y palabras compuestas sería necesario modificar el método de segmentación de manera que se pueda descubrir no solo la secuencialidad de afijos, sino también de bases. Un método así se acerca más al de Creutz y Lagus para len-

guas aglutinantes. Entonces sería interesante llevar a sus límites al método del cálculo de afijalidad para verificar si puede dar cuenta de mayor complejidad morfológica.

Algo parecido sucede con los enclíticos. Según los resultados obtenidos, el índice de afijalidad no sirvió para dividirlos debido a su naturaleza distinta. Entre otras cosas, son menos económicos que los sufijos. Podría utilizarse el índice de cliticidad que propone Medina o buscar una variante del índice afijalidad que dé mayor peso a alguna de las medidas.

Por otra parte, a pesar de que se probaron dieciséis posibilidades de segmentación para descubrir la secuencialidad de sufijos, aún queda espacio para mayores experimentos. Una posibilidad sería modificar el umbral de 0.5 para los valores máximos. Se podría tomar como umbral el promedio de afijalidad al interior de la palabra. También sería posible incrementar el umbral a medida que el corte es más cercano a la base, con la idea de prevenir segmentaciones dudosas.

Es más, se podría pensar en utilizar la idea de los conjuntos difusos (Zadeh, 1965) para estudiar unidades y fenómenos morfológicos mediante las medidas de afijalidad. Esta propuesta permite describir objetos de manera imprecisa permitiendo que los elementos pertenezcan de forma parcial a un conjunto. Tal vez así se pueda tratar el problema de distinguir entre afijos flexivos y derivativos.

Un experimento futuro interesante sería reunir las bases asociadas a los patrones morfotácticos en grandes grupos, que se podría pensar que corresponderían toscamente con clases de palabras. La idea detrás es que si no todas las bases tienen los mismos patrones morfotácticos asociados, tal vez se puedan agrupar en conjuntos, que si bien se intersectarían, podrían ayudar a separar bases nominales de verbales.

Lo anterior hace pensar en la conveniencia de generar un autómata morfológico que dé cuenta de la separación de bases nominales y verbales o que identifique sufijos flexivos

y derivativos. Incorporar estas cuestiones sería un gran avance en la generación no supervisada de una descripción morfológica más completa que permitiría estudiar lenguas poco estudiadas.

Un último experimento para trabajo futuro sería probar el método propuesto en otros corpus. Primero, se podría pensar en corpus de distintas épocas o de distintas regiones, lo que permitiría comparaciones dialectales a partir de los autómatas generados de ambos corpus. Segundo, ya que en investigaciones anteriores se había usado el índice de afijalidad para describir unidades morfológicas en distintas lenguas, es factible pensar que el método propuesto para descubrir patrones morfotácticos pueda utilizarse también en corpus de otras lenguas sufijales.

6.4. Conclusiones finales

Sobre las ventajas que ofrece este método, la más relevante es la posibilidad de describir mediante un método automático no supervisado, con el mínimo de información lingüística *a priori*, la morfotáctica del español, al menos en lo que respecta a las bases y la sufitáctica. El carácter no supervisado del método es lo que da pie a futuros experimentos en corpus de otras lenguas, lo que también es una ventaja del método.

También se modificó el método del cálculo de afijalidad para contar ahora con un método que divida las palabras en varios sufijos. Esto trajo varios beneficios. Uno de ellos es que se sentaron las bases para futuras investigaciones, como las expuestas en el apartado anterior. Además, fue un logro para ese método que pudiera utilizarse con escasas modificaciones también para realizar varios cortes en la palabra. Esto habla de la pertinencia del acercamiento lingüístico de ese método.

Otra virtud del método propuesto es que sienta las bases para investigaciones más profundas en el descubrimiento de la morfotáctica de lenguas afijales mediante la inferencia del aparato de descripción y no mediante su construcción manual. Inferir la descripción del corpus es importante porque permite estudiar el lenguaje sin presuponer sus unidades y su secuencialidad. Esta es la gran ventaja de este método y de métodos que no parten de la idea de que existe una morfología ideal y única.

Así, el presente trabajo de investigación ha logrado desarrollar un método no supervisado para inferir automáticamente la morfotáctica del español. Específicamente se logró que a partir de un corpus representativo de esta lengua se descubrieran sus bases y secuencias de sufijos, para con ellos elaborar una descripción de su orden y secuencialidad mediante un autómata de estados finitos. En este sentido, esta tesis ha desarrollado un método que descubra automáticamente parte de la morfológica del español.

Ya que el método desarrollado se basa fundamentalmente en un método ya existente que calcula la afijalidad (glutinosidad) de algunas unidades morfológicas, este trabajo también ha sido un intento por explorar los límites de ese método. Dados los resultados obtenidos, se ve prometedor que dicho método pueda seguirse ampliando para abarcar cada vez más terreno de la morfología concatenativa del español y otras lenguas.

Finalmente, este trabajo también ha sido un esfuerzo por brindar la posibilidad de estudiar la morfología del español desde una mirada imparcial, dejando que las regularidades emerjan del corpus. En este sentido, los corpus electrónicos son herramientas idóneas para el estudio empírico de las lenguas. Así, este trabajo se inserta en el conjunto de estudios que tratan de explicar la lengua a partir de datos empíricos y no de la introspección de un analista.

7. Anexos

A. Inventario de sufijos derivativos

Consigno en este anexo el inventario de sufijos derivativos que incorporé al corpus de evaluación y que fueron tomados de la recopilación que hace Moreno de Alba (1986). La Tabla 7.1 incluye una columna con el sufijo y sus alomorfos, otra columna con una breve descripción del sufijo, y una tercera columna con un ejemplo tomado de ese autor (excepto 'parisiense' que tomé del CEMC).

Tabla 7.1 Inventario de sufijo de Moreno de Alba

SUFIJO ALOMORFO	DESCRIPCIÓN	EJEMPLO
(V)(C)ión	Sufijo que forma sustantivos de acción o efecto a partir de verbos	
-ACIÓN (-A-CIÓN)		*eliminar > elimin~ación*
-CIÓN		*inscribir > inscrip~ción*
-ICIÓN		*definir > defin~ición*
-IÓN		*reunir > reun~ión*
-SIÓN		*dividir > divi~sión*
-UCIÓN		*evolutivo > evol~ución*
-V	Sufijo que forma sustantivos a partir de verbos con significado general de acción o efecto	
-A		*probar > prueb~a*
-E		*combatir > combat~e*
-O		*consolar > consuel~o*
-(V)(C)it-	Sufijo que expresa diminutivo	
-ECITO(A)		*padre > padr~ecito*
-CITO(A)		*canción > cancion~cita*
-ITITO(A)		*chico > chiqu~itito*
-ITO(A)		*palabra > palabr~ita*
-(V)al	Sufijo que forma principalmente adjetivos con significado de relación o caracterización a partir de sustantivos	
-AL		*sentimiento > sentiment~al*
-IAL		*editor > editor~ial*
-UAL		*texto > text~ual*

Tabla 7.1 Inventario de sufijo de Moreno de Alba (continuación)

SUFIJO ALOMORFO	DESCRIPCIÓN	EJEMPLO
(V)(C)(C)ic-	Sufijo que da lugar a sustantivos y adjetivos con sentido técnico o científico a partir de sustantivos	
-ÁTICA		*problema > problem~ática*
-ÁSTICO(A)		
-ICO(A)		*electrón > electrón~ica*
-ÍFICO(A)		*ciencia > cient~ífico*
-ÍSTICO(A)		*carácter > caracter~ístico*
-TICO(A)		*poema > poé~tico*
(V)(C)(C)ad	Sufijo que crea sustantivos abstractos que indican cualidad, acción o conducta a partir generalmente de adjetivos	
-AD		*amistoso > amist~ad*
-ALDAD (-AL-DAD)		*frío > fri~aldad*
-DAD		*desigual > desigual~dad*
-EDAD		*ansia > ansi~edad*
-IDAD		*materno > matern~idad*
-TAD		*libre > liber~tad*
(V)Vnte	Sufijo que crea adjetivos que significan agentes a partir de verbos, principalmente de la primera conjugación	
-ANTE		*alarmar > alarm~ante*
-ENTE		*absorber > absorb~ente*
-IENTE		*corresponder > correspond~iente*
Vd-	Sufijo que forma sustantivos con diversos significados (acción, resultado de la acción, conjunto, duración, golpe) a partir de verbos o sustantivos	
-ADA		*tiempo > tempor~ada*
-ADO		*estudiante > estudiant~ado*
-IDA		*comer > com~ida*
-IDO		*tejer > tej~ido*
Vncia, -anza	Sufijo que forma sustantivos abstractos con significado de acción o resultado de la acción a partir de verbos de la primera conjugación	
-ANCIA		*constar > const~ancia*
-ANCIO		*cansar > cans~ancio*
-ENCIA		*decadente > decand~encia*
-IENCIA		*eficiente > efic~iencia*
-ANZA		*confiar > confi~anza*
(u)os-	Sufijo que forma adjetivos que indican cualidades o defectos a partir de sustantivos, adjetivos o verbos	
-OSO(A)		*grande > grandi~oso*
-UOSO(A)		*defecto > defect~uoso*

Tabla 7.1 Inventario de sufijo de Moreno de Alba (continuación)

SUFIJO ALOMORFO	DESCRIPCIÓN	EJEMPLO
(Vd)er-	Sufijo que forma sustantivos y adjetivos con significado de agente, nombres de objetos, instrumentos, alimentos, etcétera, a partir de verbos, sustantivos o adjetivos	
-ADERA		*tapar > tap~adera*
-ADERO		*pasar > pas~adero*
-ERA		*sordo > sord~era*
-ERO		*sombra > sombr~ero*
-ERO(A)		*compañía > compañ~ero*
Vm(i)ent-	Sufijo que forma sustantivos con significado de acción, resultado de la acción, colectivo y de lugar, a partir de verbos	
-AMENTO		*acampar > camp~amento*
-AMIENTA		*hierro > herr~amienta*
-AMIENTO		*relajar > relaj~amiento*
-IMIENTO		*descubrir > descubr~imiento*
(Vt)iv-	Sufijo que forma adjetivos que caracterizan personas o cosas, a partir de sustantivos, adjetivos y verbos	
-ATIVO(A) -AT-IVO(A)		*informar > inform~ativo*
-ITIVO(A) -IT-IVO(A)		*primo > prim~itivo*
-IVO(A)		*intenso > intens~ivo*
Vble	Sufijo que forma adjetivos con significado de capacidad o aptitud, a partir principalmente de verbos	
-ABLE		*respetar > respet~able*
-IBLE		*entender > entend~ible*
(V)Cor-	Sufijo que forma adjetivos con significado de agentes, a partir de verbos	
-ADOR(A)		*colaborar > colabor~ador*
-EDOR(A)		*conmover > conmov~edor*
-IDOR(A)		*corregir > correg~idor*
-SOR(A)		*anteceder > antece~sor*
-TOR(A)		*satisfacer > satisfac~tor*
a(ta)ri-	Sufijo que forma sustantivos con significados diversos y adjetivos caracterizadores de personas o cosas, a partir principalmente de sustantivos	
-ARIA		*refacción > refaccion~aria*
-ARIO(A)		*hospital > hospital~ario*
-ATARIO		*mandar > mand~atario*

Tabla 7.1 Inventario de sufijo de Moreno de Alba (continuación)

SUFIJO ALOMORFO	DESCRIPCIÓN	EJEMPLO
í-	Sufijo que forma sustantivos abstractos y adjetivos, a partir de sustantivos, adjetivos y verbos	
-ÍA		*maestro > maestr~ía*
-ÍO		
-ÍO(A)		*tarde > tard~ío*
(V)(C)ón(-)	Sufijo que forma sustantivos y adjetivos aumentativos, atenuativos, de acción contundente o golpe, a partir de sustantivos, adjetivos o verbos	
-ERÓN		*casa > cas~erón*
-ÓN		*batalla > batall~ón*
-ONA		*casa > cas~ona*
-ÓN(A)		*llorar > llor~ón*
-OTÓN -OT-ÓN		*pisar > pis~otón*
Vría	Sufijo que forma sustantivos abstractos y concretos, a partir de sustantivos y adjetivos	
-ARÍA		*secretario > secret~aría*
-ERÍA		*lavandero > lavand~ería*
-ORÍA		*auditor > audit~oría*
-URÍA		*tenedor > tened~uría*
(V)(C)ura	Sufijo que forma sustantivos abstractos y concretos, a partir de sustantivos o verbos	
-ADURA -AD-URA		*diente >dent~adura*
-ATURA		*colegio > colegi~atura*
-IDURA -ID-URA		*vestir > vest~idura*
-TURA		*lección > lec~tura*
-URA		*hermoso > hermos~ura*
(V)(C)ez(-)	Sufijo que forma sustantivos abstractos a partir de adjetivos	
-ALEZA		*fuerte > fort~aleza*
-EZ		*niño > niñ~ez*
-EZA		*bello > bell~eza*

Tabla 7.1 Inventario de sufijo de Moreno de Alba (continuación)

SUFIJO ALOMORFO	DESCRIPCIÓN	EJEMPLO
(V)(C)ori-	Sufijo que forma sustantivos femeninos abstractos, sustantivos masculinos con significado de lugar y adjetivos caracterizadores de cosas, a partir principalmente de verbos	
-ATORIA		*escapar > escap~atoria*
-ATORIO		*observar > observ~atorio*
-ATORIO(A)		*rotar > rot~atorio*
-ITORIO		*audiencia > aud~itorio*
-ORIA		*trayecto > trayect~oria*
-ORIO		*consultar > consult~orio*
-ORIO(A)		*irrisión > irris~orio*
-TORIO		*satisfacción > satisfac~torio*
Vdor(A)	Sufijo que forma sustantivos con diversos significados (objetos, instrumentos, lugares) a partir de verbos	
-ADOR		*tocar > toc~ador*
-ADORA		*incubar > incub~adora*
-EDOR		*correr > corredor*
in-	Sufijo que forma sustantivos con diversos significados y adjetivos caracterizadores o de semejanza, a partir de sustantivos o adjetivos	
-INA		*estudiante > estudiant~ina*
-INO		*plata > plat~ino*
-INO(A)		*cervantes > cervant~ino*
t-	Sufijo que forma sustantivos abstractos con significado de acción o efecto de la acción y adjetivos, a partir de verbos o sustantivos	
-TA		*aristocracia > aristócra~ta*
-TE		*morir > muer~te*
-TO		*instituir > institu~to*
-TO(A)		*atender > aten~to*
(i)(t)ud	Sufijo que forma sustantivos abstractos con significado de acción, conducta o cualidad, a partir de adjetivos y sustantivos	
-ITUD		*exacto > exact~itud*
-TUD		*joven > juven~tud*
-UD		*quieto > quiet~ud*
(c)ill-	Sufijo que forma sustantivos y adjetivos diminutivos o despectivos, a partir de sustantivos y adjetivos	
-CILLO(A)		*joven > joven~cilla*
-ILLO		*cera > cer~illo*
-ILLA		*cama > cam~illa*
-ILLO(A)		*chico > chiqu~illo*

Tabla 7.1 Inventario de sufijo de Moreno de Alba (continuación)

SUFIJO ALOMORFO	DESCRIPCIÓN	EJEMPLO
(i)ci-	Sufijo que forma sustantivos concretos y abstractos, además adjetivos, a partir de adjetivos	
-CIA		*infante > infan~cia*
-ICIA		*justo > just~icia*
-ICIO		*alimentar > aliment~icio*
i-	Sufijo que forma sustantivos abstractos a partir de verbos o sustantivos	
-IA		*molestar > molest~ia*
-IO		*armonía > armon~io*
-ACERO	Sufijo que forma sustantivos a partir de sustantivos	*agua > agu~acero*
-ACIA	Sufijo que forma sustantivos abstractos a partir de sustantivos	*diploma > diplom~acia*
-ACÍA	Sufijo que forma sustantivos abstractos a partir de sustantivos	*abogado > abog~acía*
-ACO(A)	Sufijo que forma adjetivos con significado de 'relativo a', despectivo y gentilicio a partir de sustantivos	*policía > polici~aco*
-ACHO	Sufijo que forma sustantivos con significado despectivo a partir de sustantivos	*popular > popul~acho*
-ADO(A)	Sufijo que forma adjetivos con significado activo a partir de verbos, sustantivos y adjetivos	*criar > cri~ada*
-AJE	Sufijo que forma sustantivos con diversos significados, a partir de verbos o sustantivos	*persona > person~aje*
-ALLA	Sufijo que forma sustantivos con significado colectivo, a partir de sustantivos	*muro > mur~alla*
-ÁN	Sufijo que forma adjetivos o sustantivos con significado de persona y gentilicios, a partir de sustantivos	*alemania < alem~án*
-ANDA	Sufijo que forma sustantivos a partir de verbos	*propagar > propag~anda*
-ANO(A)	Sufijo que forma sustantivos y adjetivos gentilicios o que indican procedencia o pertenencia, a partir de verbos o sustantivos	*lejos > lej~ano*
-IANO(A)	Sufijo que forma sustantivos y adjetivos gentilicios o que indican procedencia o pertenencia, a partir de verbos o sustantivos	*cristo > crist~iano*
-AÑA	Sufijo que forma sustantivos a partir de sustantivos	*monte > mont~aña*
-AÑO	Sufijo que forma sustantivos a partir de sustantivos	*ermita > ermit~año*

Tabla 7.1 Inventario de sufijo de Moreno de Alba (continuación)

SUFIJO ALOMORFO	DESCRIPCIÓN	EJEMPLO
-AR	Sufijo que forma sustantivos con significado de colectivo o de lugar donde abundan plantas, o adjetivos caracterizadores de personas o de objetos, a partir de sustantivos, adjetivos y verbos	*célula > celul~ar*
-ATIVA	Sufijo que forma sustantivos abstractos femeninos, a partir de sustantivos o verbos	*negar > neg~ativa*
-ATO	Sufijo que forma sustantivos con diversos significados, a partir de sustantivos o verbos	*bachiller > bachiller~ato*
-AVO(A)	Sufijo que forma adjetivos fraccionarios, a partir de sustantivos	*ciento > cent~avo*
-AZ	Sufijo que forma adjetivos, a partir de sustantivos	*> ver~az*
-AZGO	Sufijo que forma sustantivos abstractos, a partir de sustantivos y verbos	*novio > novi~azgo*
-AZO	Sufijo que forma sustantivos con significado de acción contundente, golpe o aumentativos, a partir de sustantivos	*bala > bal~azo*
-CIO	Sufijo que forma adjetivos gentilicios, a partir de sustantivos	*egipto > egip~cio*
-ECO(A)	Sufijo que forma adjetivos gentilicios, a partir de sustantivos	*mazatlán > mazatl~eco*
-EJO(A)	Sufijo que forma sustantivos con significado atenuativo o despectivo, a partir de sustantivos	*animal > animal~ejo*
-EL	Sufijo que forma sustantivos a partir de sustantivos	*planta > plant~el*
-ELA	Sufijo que forma sustantivos a partir de sustantivos	*cliente > client~ela*
-ENA	Sufijo que forma sustantivos a partir de sustantivos	*nueve > nov~ena*
-ENIO	Sufijo que forma sustantivos a partir de sustantivos	*diez > dec~enio*
-ENO(A)	Sufijo que forma adjetivos gentilicios a partir de sustantivos	*chile > chil~eno*
-ENSE	Sufijo que forma adjetivos gentilicios a partir de sustantivos	*parís > parisi~ense*[85]
-EÑO	Sufijo que forma sustantivos y adjetivos con significado de gentilicio, semejanza o cualidad, a partir de sustantivos o adjetivos	*brasil > brasil~eño*
-EO(A)	Sufijo que forma adjetivos con significado de ' relativo a', a partir de sustantivos o adjetivos	*árbol > arbór~eo*

[85] Este ejemplo fue tomado del CEMC, no lo propone Moreno de Alba.

Tabla 7.1 Inventario de sufijo de Moreno de Alba (continuación)

SUFIJO ALOMORFO	DESCRIPCIÓN	EJEMPLO
-ERNO(A)	Sufijo que forma adjetivos, a partir de sustantivos	*madre > mat~erno*
-ÉRRIMO(A)	Sufijo que forma adjetivos superlativos, a partir de adjetivos	*pauperismo > paup~érrimo*
-ÉS(A)	Sufijo que forma sustantivos y adjetivos gentilicios o de procedencia, a partir de sustantivos	*francia > franc~és*
-ESCO(A)	Sufijo que forma sustantivos y adjetivos caracterizadores de personas o cosas, a partir de sustantivos, adjetivos o verbos	*pariente > parent~esco*
-ESTRE	Sufijo que forma adjetivos a partir de sustantivos	*tierra > terr~estre*
-ETA	Sufijo que forma sustantivos con significado de objetos, instrumentos o semejante pero pequeño, a partir de sustantivos y verbos	*trompa > tromp~eta*
-ETE	Sufijo que forma sustantivos a partir de sustantivos	*juego > jugu~ete*
-ICIDA	Sufijo que forma sustantivos a partir de sustantivos	*insecto > insect~icida*
-ICHE	Sufijo que forma sustantivos y adjetivos despectivos, a partir de sustantivos	*bolo > bol~iche*
-IDUMBRE	Sufijo que forma sustantivos a partir de sustantivos y adjetivos	*cierto > cert~idumbre*
-EDUMBRE	Sufijo que forma sustantivos a partir de sustantivos y adjetivos	*mucho > much~edumbre*
-ADUMBRE	Sufijo que forma sustantivos a partir de sustantivos y adjetivos	*pesar > pes~adumbre*
-ÍFERO	Sufijo que forma adjetivos o sustantivos a partir de sustantivos	*mama > mam~ífero*
-IJO	Sufijo que forma sustantivos con significado de acción, resultado de la acción o de diminutivo, a partir de sustantivos o verbos	*acertar > acert~ijo*
-IL	Sufijo que forma adjetivos con significado de pertenencia o relación, a partir de sustantivos	*mercante > mercant~il*
-ÍN	Sufijo que forma sustantivos con significado de objeto o adjetivos caracterizadores, diminutivos y gentilicios, a partir de sustantivos	*maleta > malet~ín*
-ÍNEO	Sufijo que forma adjetivos a partir de sustantivos	*sangre > sangu~íneo*
-IÑO	Sufijo que forma adjetivos a partir de sustantivos	*caro > car~iño*
-ISCO(A)	Sufijo que forma sustantivos a partir de sustantivos	*mar > mar~isco*
-ÍSIMO(A)	Sufijo que forma adjetivos superlativos a partir de adjetivos y adverbios	*hermoso > hermos~ísimo*

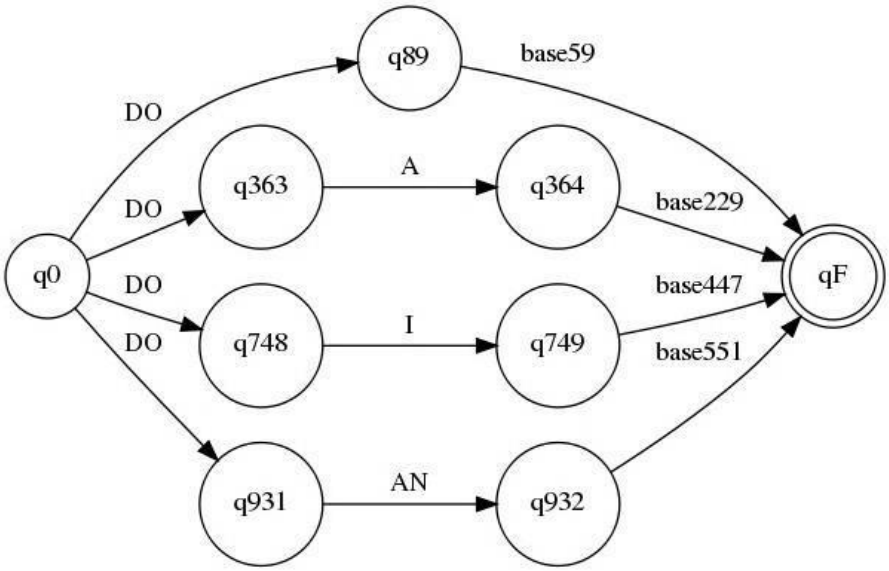
Tabla 7.1 Inventario de sufijo de Moreno de Alba (continuación)

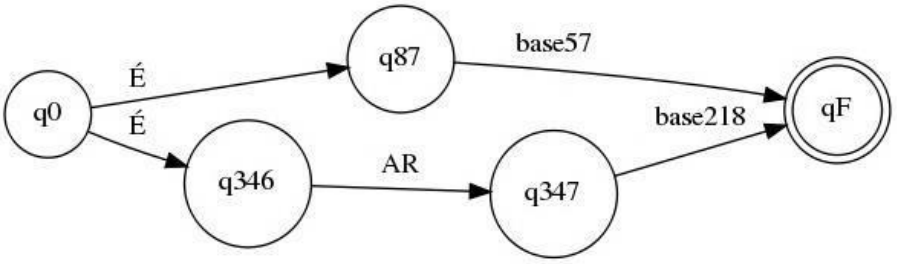
SUFIJO ALOMORFO	DESCRIPCIÓN	EJEMPLO
-ISMO	Sufijo que forma sustantivos con significado de doctrina, secta o calidad, a partir de sustantivos o adjetivos	*comunista > comun~ismo*
-ISTA	Sufijo que forma sustantivos con significado de profesión, oficio y que caracterizan personas o cosas, a partir de sustantivos o adjetivos	*análisis > anal~ista*
-ITA	Sufijo que forma adjetivos gentilicios, a partir de sustantivos	*israel > isreal~ita*
-ITE	Sufijo que forma el sustantivo ESCONDITE a partir de verbo	*esconder > escond~ite*
-ITIS	Sufijo (seudosufijo) que forma sustantivos con significado de inflamación, a partir de sustantivos	*colon > col~itis*
-ÍVORO	Sufijo (seudosufijo) que forma adjetivos, a partir de sustantivos	*insecto > insect~ívoro*
-IZ	Sufijo que forma sustantivos, a partir de sustantivos	*cara > car~iz*
-IZA	Sufijo que forma sustantivos con significado de golpe repetido, a partir de sustantivos o verbos	*palo > pal~iza*
-IZO	Sufijo que forma adjetivos con significado de semejanza, a partir de sustantivos o adjetivos	*rojo > roj~izo*
-O(A)	Sufijo que forma adjetivos de diversos significados, a partir de sustantivos	*fotografía > fotógraf~o*
-OIDE	Sufijo que forma sustantivos y adjetivos con significado de semejanza, a partir de sustantivos	*estrella > aster~oide*
-OL(A)	Sufijo que forma sustantivos y adjetivos diminutivos o gentilicios, a partir de sustantivos	*españón > españ~ol*
-ÓNEO(A)	Sufijo que forma adjetivos, a partir de sustantivos o verbos	*error > err~óneo*
-OR	Sufijo que forma sustantivos abstractos o que designan objetos o instrumentos, a partir de sustantivos o verbos	*temblar > tembl~or*
-OR(A)	Sufijo que forma sustantivos de oficios u ocupaciones, a partir de sustantivos o verbos	*supervisar > supervis~or*
-ORO(A)	Sufijo que forma adjetivos, a partir de sustantivos o verbos	*sonar > son~oro*
-OTA	Sufijo que forma sustantivos femeninos o adverbios aumentativos, a partir de sustantivos o adverbios	*araña > arañ~ota*
-OTE	Sufijo que forma sustantivos masculinos o adverbios aumentativos, a partir de sustantivos o adverbios	*abajo > abaj~ote*

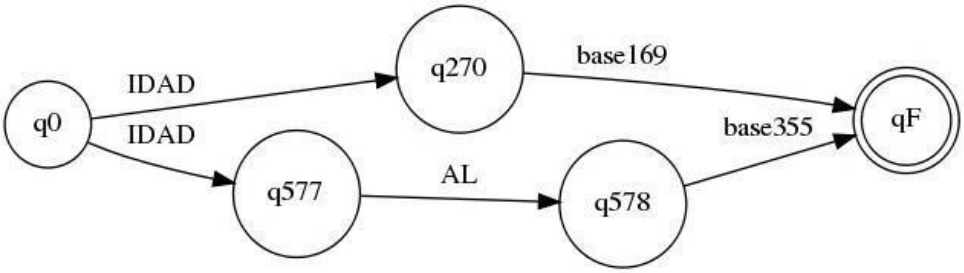
Tabla 7.1 Inventario de sufijo de Moreno de Alba (continuación)

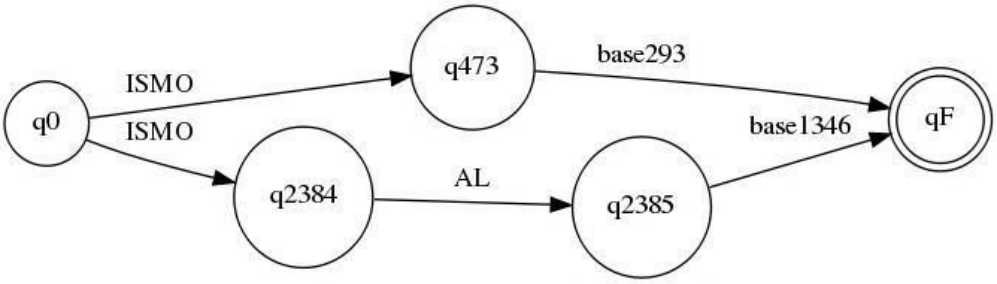
SUFIJO ALOMORFO	DESCRIPCIÓN	EJEMPLO
-OTE(A)	Sufijo que forma sustantivos o adjetivos aumentativos, a partir de sustantivos o adjetivos	*fuerte > fuert~ote*
-SA	Sufijo que forma sustantivos, a partir de sustantivos o adjetivos	*defensivo > defen~sa*
-SO	Sufijo que forma sustantivos abstractos, a partir de verbos de la segunda y tercera conjugación	*ascender > ascen~so*
-TECA	Sufijo (seudosufijo) que forma sustantivos, a partir de sustantivos	*bibliografía > biblio~teca*
-UDO(A)	Sufijo que forma adjetivos caracterizadores de personas o cosas, a partir de sustantivos	*pelo > pel~udo*
-UELO(A)	Sufijo que forma sustantivos diminutivos, a partir de sustantivos	*paño > pañ~uelo*
-UNA	Sufijo que forma sustantivos, a partir de sustantivos	*lago > lag~una*
-URNO(A)	Sufijo que forma sustantivos o adjetivos, a partir de sustantivos	*noctámbulo > noct~urno*
-UZ(A)	Sufijo que forma adjetivos, a partir de sustantivos	*andalucía > andal~uz*

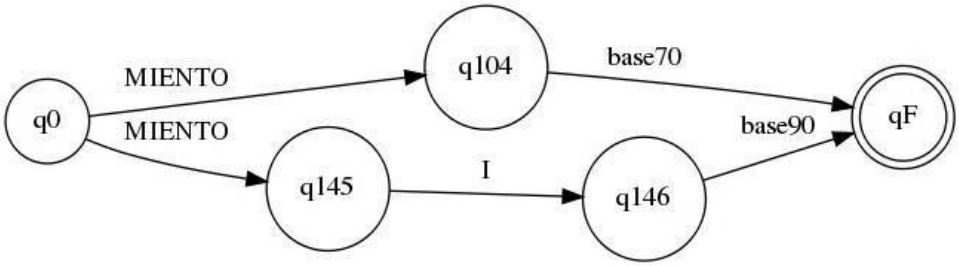
B. Ejemplos de autómatas

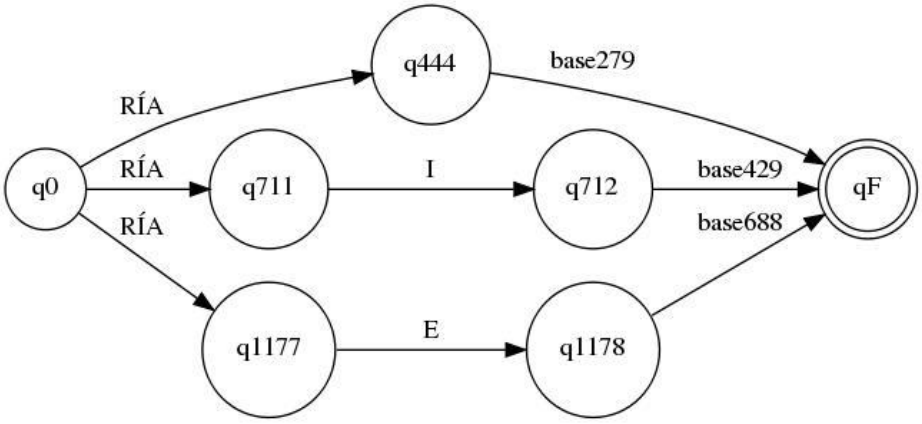


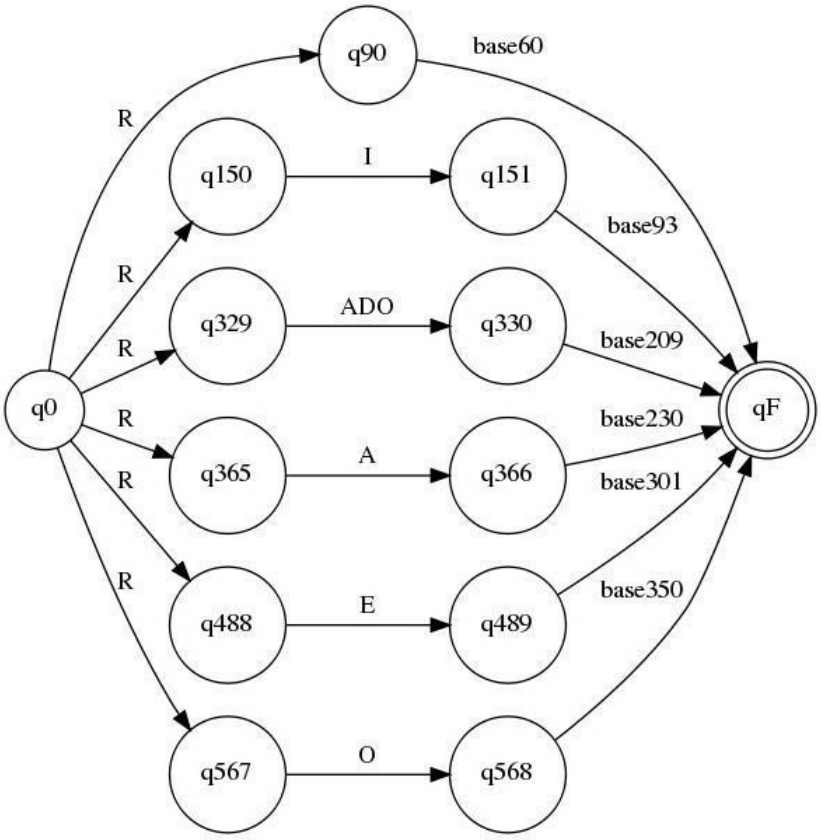


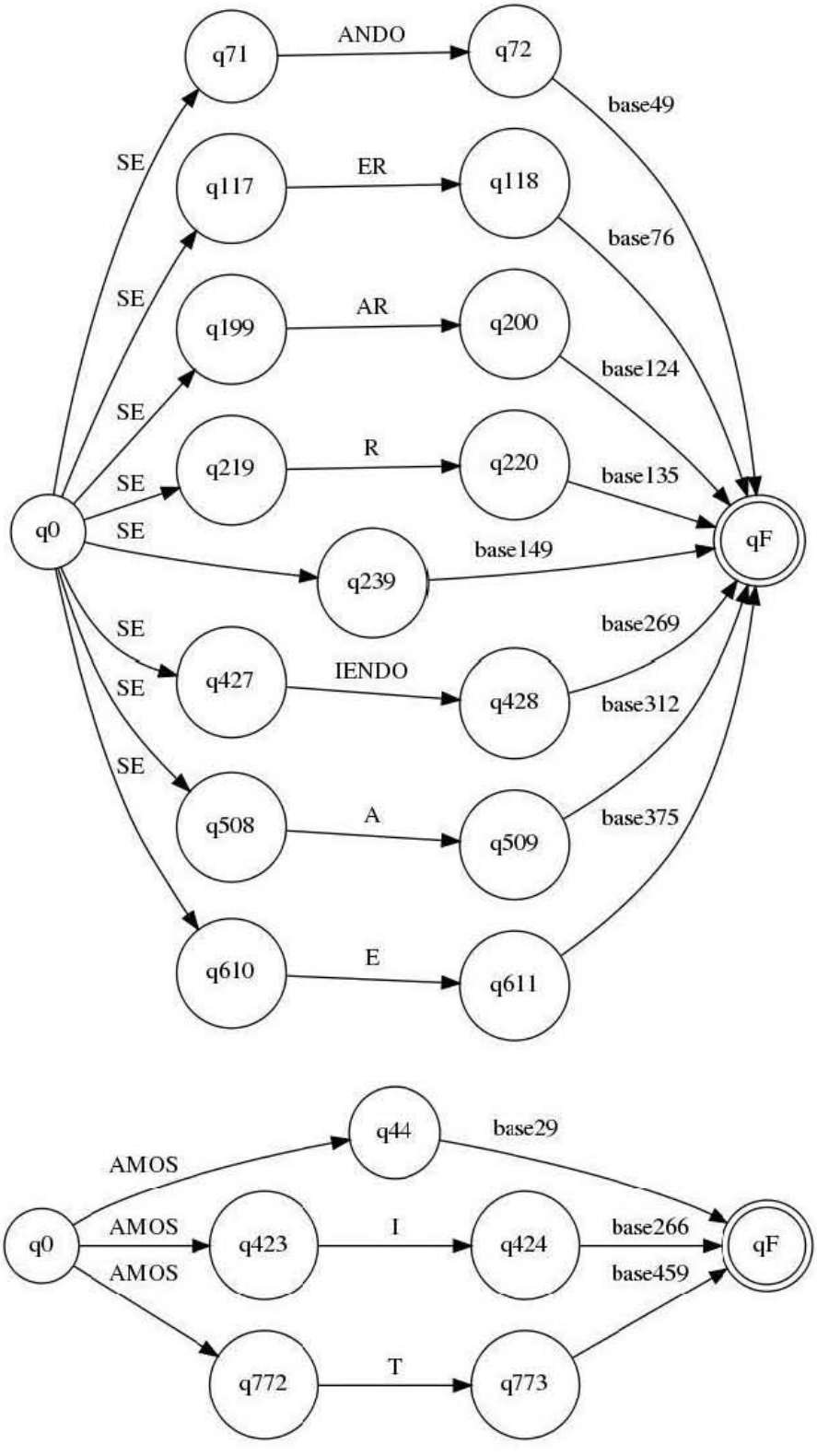


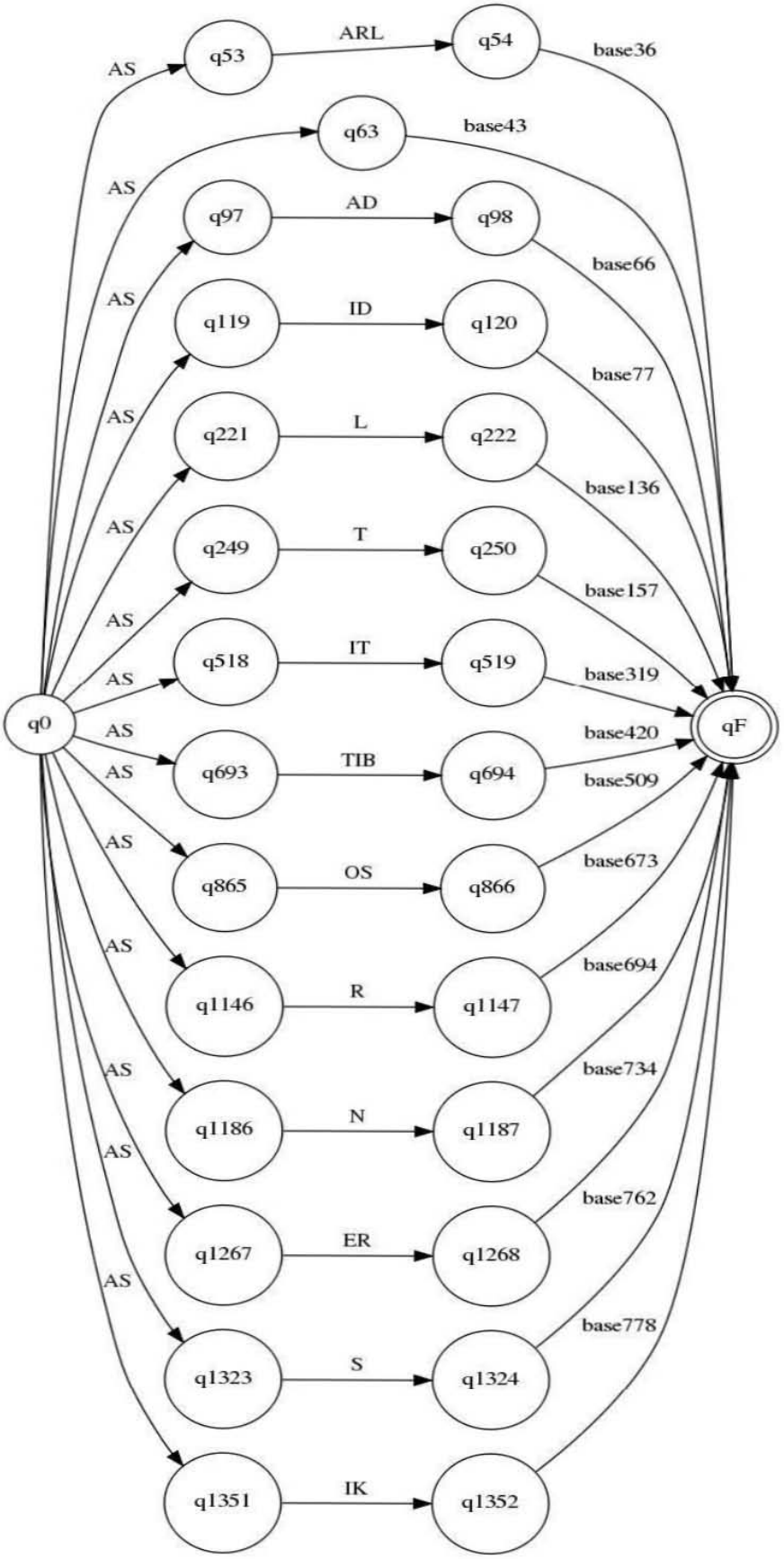


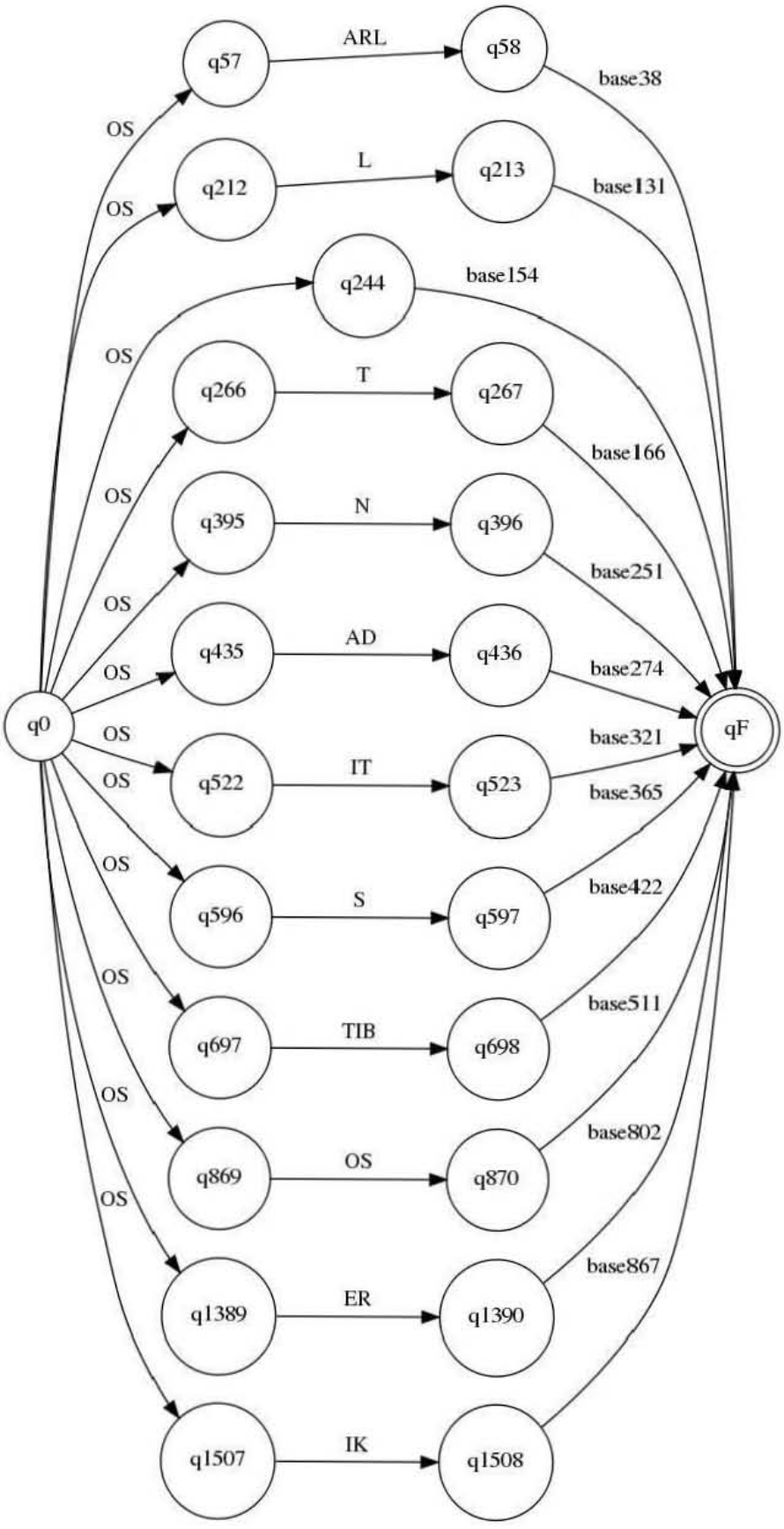

q0
MIENTO
MIENTO
q104
q145
I
q146
base70
base90
qF

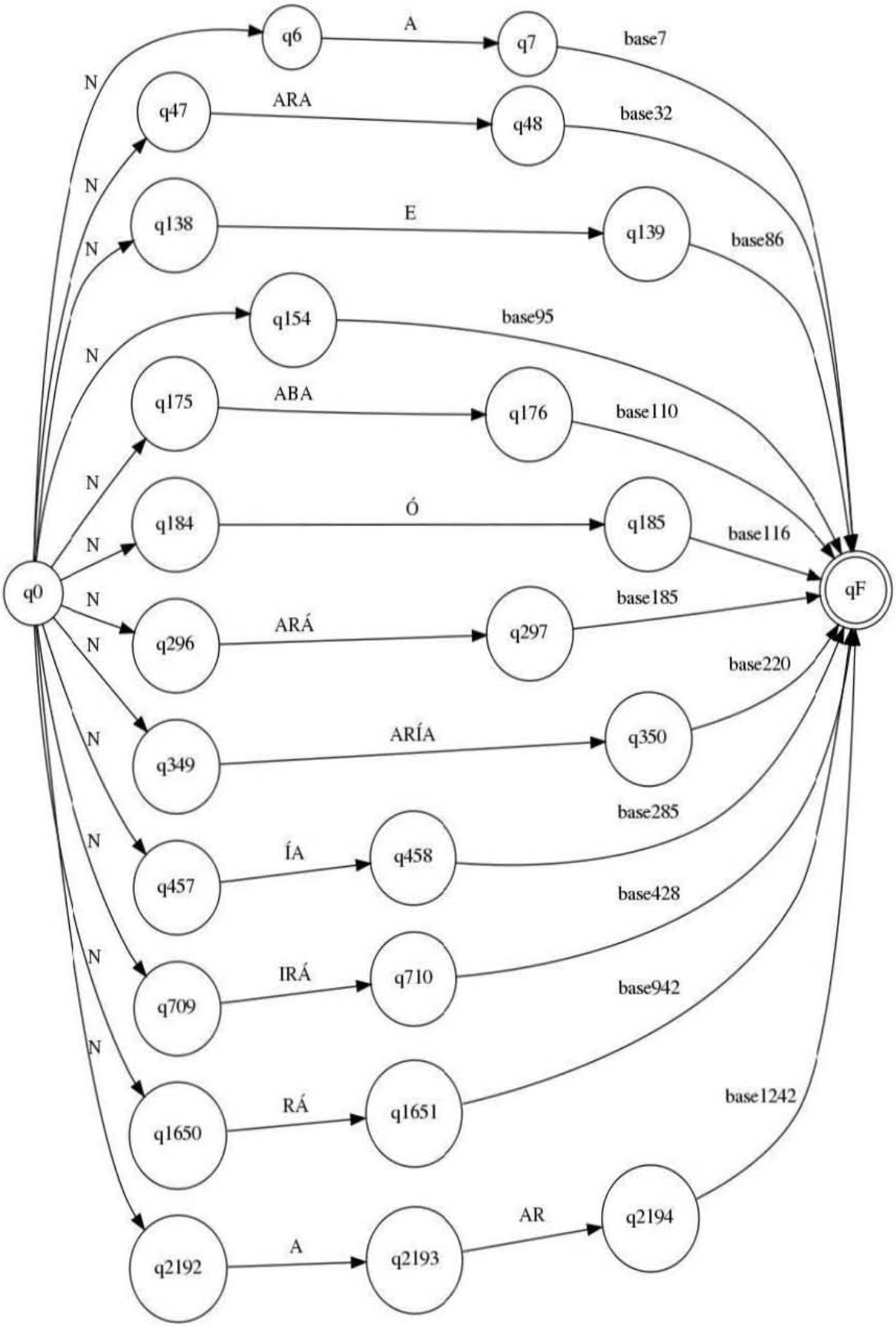

q0
RÍA
RÍA
RÍA
q444
q711
I
q712
q1177
E
q1178
base279
base429
base688
qF


q0
R
R
R
R
R
R
q90
q150
I
q151
q329
ADO
q330
q365
A
q366
q488
E
q489
q567
O
q568
base60
base93
base209
base230
base301
base350
qF


q71
ANDO
q72
base49
SE
q117
ER
q118
base76
SE
q199
AR
q200
base124
SE
q219
R
q220
base135
q0
SE
qF
SE
q239
base149
SE
base269
q427
IENDO
q428
SE
base312
SE
q508
A
q509
base375
q610
E
q611
q44
base29
AMOS
q0
AMOS
q423
I
q424
base266
qF
AMOS
base459
q772
T
q773


q0
q53
ARL
q54
base36
AS
q63
base43
q97
AD
q98
base66
q119
ID
q120
base77
q221
L
q222
base136
q249
T
q250
base157
q518
IT
q519
base319
q693
TIB
q694
base420
base509
q865
OS
q866
base673
q1146
R
q1147
base694
q1186
N
q1187
base734
q1267
ER
q1268
base762
q1323
S
q1324
base778
q1351
IK
q1352
qF


q0
q57
ARL
q58
base38
OS
q212
L
q213
base131
OS
q244
base154
OS
q266
T
q267
base166
OS
q395
N
q396
base251
OS
q435
AD
q436
base274
OS
q522
IT
q523
base321
OS
base365
q596
S
q597
base422
OS
q697
TIB
q698
base511
OS
q869
OS
q870
base802
OS
q1389
ER
q1390
base867
q1507
IK
q1508
qF


q0
q6
A
q7
base7
N
q47
ARA
q48
base32
q138
E
q139
base86
q154
base95
q175
ABA
q176
base110
q184
Ó
q185
base116
qF
q296
ARÁ
q297
base185
base220
q349
ARÍA
q350
base285
q457
ÍA
q458
base428
q709
IRÁ
q710
base942
q1650
RÁ
q1651
base1242
q2192
A
q2193
AR
q2194

C. Los cien patrones morfotácticos más frecuentes

En este anexo se presentan los 100 patrones morfotácticos más frecuentes del corpus (véase Tabla 7.2) tomados de los 422 patrones descubiertos. La lista, ordenada de mayor a menor frecuencia, está encabezada por patrones con sufijos flexivos nominales que marcan género y número. Entre los más frecuentes también aparecen patrones con marcas verbales tanto de verboides (por ejemplo /Base~AR/) como de formas conjugadas (por ejemplo /Base~Ó/). Además, pueden apreciarse pares de patrones equivalentes, donde uno tiene los sufijos separados y el otro los sufijos concatenados, como /Base~O~S/ y /Base~OS/.

Tabla 7.2 Los cien patrones morfotácticos más frecuentes

Posición	Patrón morfotáctico	Frecuencia	Frecuencia x posición
1	/Base~A/	3200	3200
2	/Base~O/	2811	5622
3	/Base~S/	2324	6972
4	/Base~OS/	1318	5272
5	/Base~AR/	1276	6380
6	/Base~E/	1261	7566
7	/Base~AS/	1224	8568
8	/Base~Ó/	1094	8752
9	/Base~A~S/	1022	9198
10	/Base~AN/	938	9380
11	/Base~O~S/	875	9625
12	/Base~AD~A/	812	9744
13	/Base~ANDO/	810	10530
14	/Base~AD~O/	768	10752
15	/Base~ABA/	738	11070
16	/Base~ARON/	685	10960
17	/Base~ES/	599	10183
18	/Base~ARSE/	590	10620
19	/Base~EN/	545	10355
20	/Base~E~S/	541	10820
21	/Base~AD~O~S/	532	11172
22	/Base~ÍA/	495	10890
23	/Base~AD~A~S/	478	10994
24	/Base~AMOS/	463	11112
25	/Base~É/	448	11200
26	/Base~ASIÓN/	436	11336

Tabla 7.2 Los cien patrones morfotácticos más frecuentes (continuación)

Posición	Patrón morfotáctico	Frecuencia	Frecuencia x posición
27	/Base~AMENTE/	435	11745
28	/Base~ADO/	418	11704
29	/Base~R/	411	11919
30	/Base~ABAN/	408	12240
31	/Base~N/	384	11904
32	/Base~ÓN/	370	11840
33	/Base~ARÁ	365	12045
34	/Base~SIÓN/	364	12376
35	/Base~AL/	311	10885
36	/Base~MENTE/	293	10548
37	/Base~E~N/	292	10804
38	/Base~ITA/	270	10260
39	/Base~ONES/	262	10218
40	/Base~ITO/	259	10360
41	/Base~IENDO/	243	9963
42	/Base~DO/	239	10038
43	/Base~ASIONES/	234	10062
44	/Base~ARME/	232	10208
45	/Base~IDAD	229	10305
46	/Base~ID~A/	228	10488
47	/Base~A~N/	222	10434
48	/Base~IÓ/	222	10656
49	/Base~TE/	220	10780
50	/Base~ID~O/	217	10850
51	/Base~ARÍA/	210	10710
52	/Base~IERON/	210	10920
53	/Base~SE/	204	10812
54	/Base~ADO~R/	202	10908
55	/Base~IK~O/	199	10945
56	/Base~ADA/	186	10416
57	/Base~IK~A/	185	10545
58	/Base~ID~O~S/	184	10672
59	/Base~ISTA/	184	10856
60	/Base~ISMO/	180	10800
61	/Base~T~A/	180	10980
62	/Base~AL~ES/	178	11036
63	/Base~ANDO~SE/	176	11088
64	/Base~L/	172	11008
65	/Base~AR~A/	168	10920
66	/Base~ARL~A/	164	10824
67	/Base~I~R/	164	10988
68	/Base~ARL~O/	163	11084
69	/Base~ARÁ~N/	162	11178
70	/Base~ERO/	155	10850

Tabla 7.2 Los cien patrones morfotácticos más frecuentes (continuación)

Posición	Patrón morfotáctico	Frecuencia	Frecuencia x posición
71	/Base~ID~A~S/	154	10934
72	/Base~E~R/	153	11016
73	/Base~L~A/	148	10804
74	/Base~ADO~S/	146	10804
75	/Base~ARA/	145	10875
76	/Base~ADOR~ES/	139	10564
77	/Base~T~O/	139	10703
78	/Base~ANTE/	137	10686
79	/Base~ANTE~S/	137	10823
80	/Base~ARNOS/	136	10880
81	/Base~D~O/	135	10935
82	/Base~LO/	134	10988
83	/Base~AMIENTO/	130	10790
84	/Base~Í	129	10836
85	/Base~OS~O/	129	10965
86	/Base~S~A/	128	11008
87	/Base~ÍAN/	126	10962
88	/Base~LE/	126	11088
89	/Base~EMOS/	124	11036
90	/Base~IT~A/	123	11070
91	/Base~AD~AS/	119	10829
92	/Base~TA/	116	10672
93	/Base~OS~A/	115	10695
94	/Base~L~O/	112	10528
95	/Base~ARLO/	111	10545
96	/Base~I/	108	10368
97	/Base~L~ES/	106	10282
98	/Base~R~A/	106	10388
99	/Base~SI~ONES/	106	10494
100	/Base~I~MOS/	105	10500

En la Figura 7.1, se brinda una gráfica que muestra el comportamiento de la frecuencia (f) de los 422 patrones con relación a su posición (r) en una lista ordenada de mayor a menor frecuencia (los valores están expresados en escala logarítmica). Como exponen Manning y Schütze (1999, págs. 23-24), la relación entre la frecuencia de un elemento lingüístico y su posición en una lista ordenada fue explorada por Zipf, quien propuso una ley empírica al respecto. Según esta ley existe una constante k tal que $f \times r = k$. Para el caso

de los patrones morfotácticos descubiertos, la constante se puede establecer alrededor de 10,000. La cuarta columna de la Tabla 7.2 muestra el resultado de $f \times r$ para los 100 patrones más frecuentes.

El que dicha curva muestre este comportamiento es una caracterización empírica de la economía presente en el sistema morfológico. Esto es, existen pocos patrones muy frecuentes, algunos regularmente frecuentes y la gran mayoría de baja frecuencia.

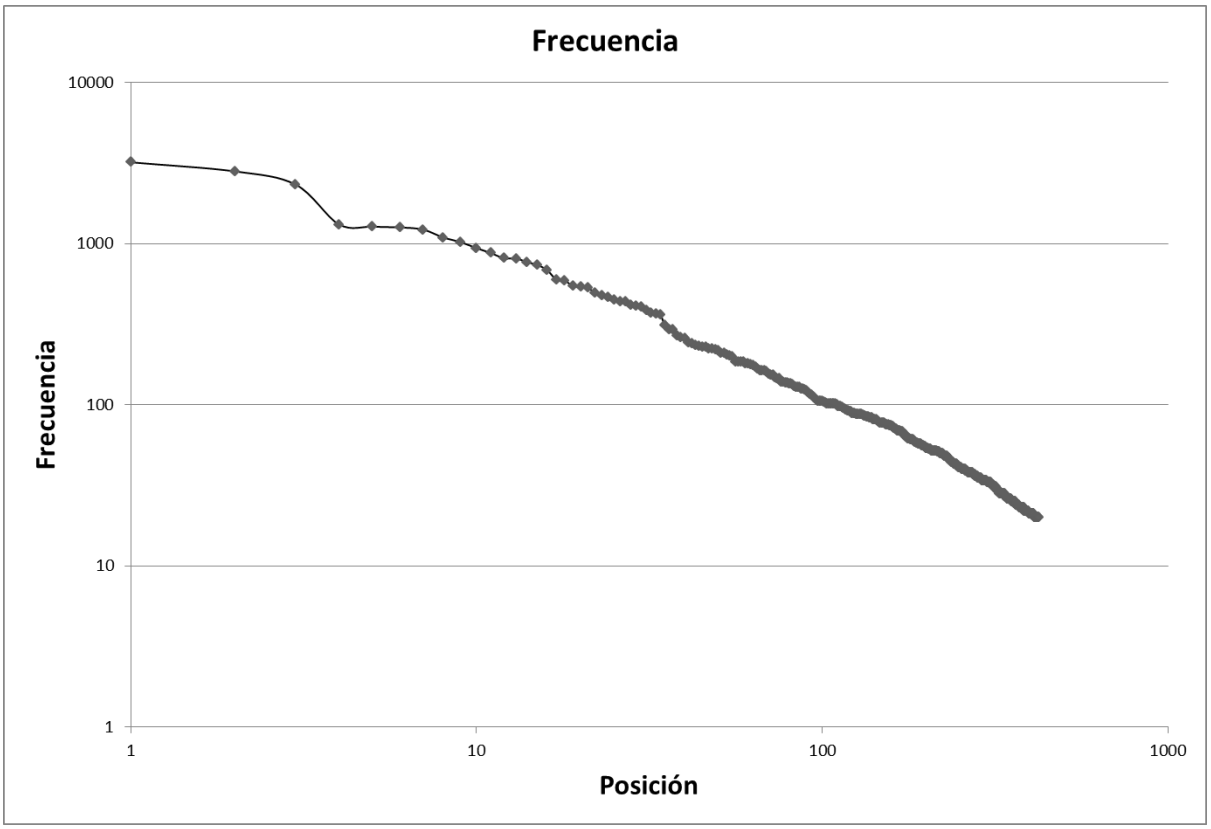
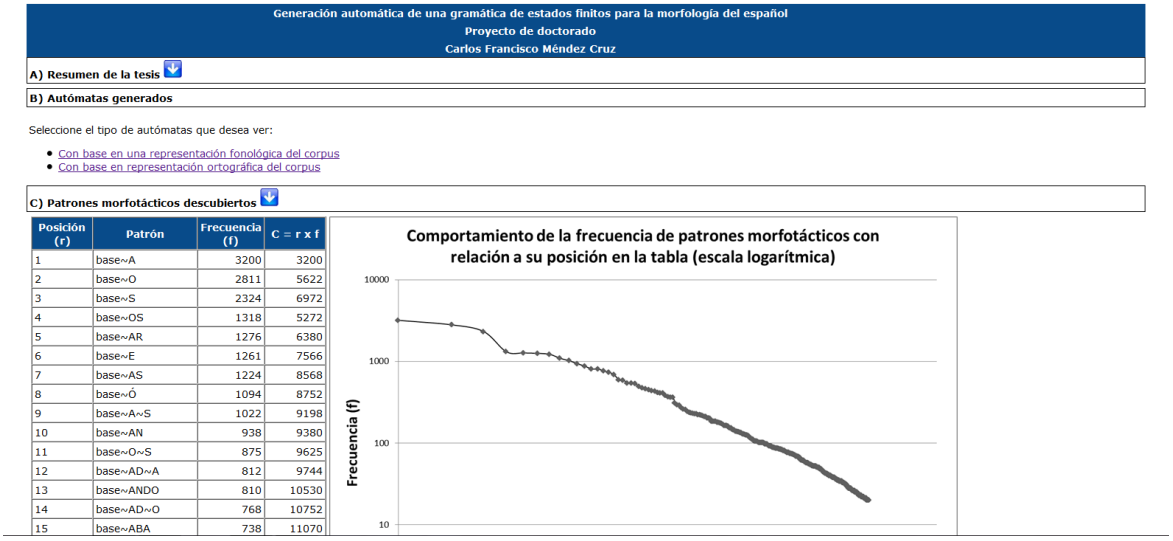


Figura 7.1 Comportamiento de la frecuencia de patrones morfotácticos

D. Descripción del disco compacto

En este anexo se describe el contenido del disco compacto que acompaña a esta tesis, así como la manera de visualizarlo. En términos generales, el disco contiene el conjunto de grafos que representan los autómatas de estados finitos generados y sus bases asociadas. Con la idea de facilitar la visualización de estos autómatas, mostrando las bases que se asocian a cada uno, se elaboraron tres páginas web. La primera (Figura 7.2), punto de partida de la visualización del contenido del disco, incluye:

1. El resumen de la tesis.
2. Dos enlaces que llevan a una página cada uno para visualizar los autómatas generados a partir de la representación fonológica y ortográfica del corpus.
3. La lista completa de patrones morfotácticos descubiertos, su frecuencia, posición en la tabla y una gráfica que representa la relación entre la frecuencia y la posición del patrón (ley de Zipf).



Posición (r)	Patrón	Frecuencia (f)	C = r x f
1	base~A	3200	3200
2	base~O	2811	5622
3	base~S	2324	6972
4	base~OS	1318	5272
5	base~AR	1276	6380
6	base~E	1261	7566
7	base~AS	1224	8568
8	base~Ó	1094	8752
9	base~A~S	1022	9198
10	base~AN	938	9380
11	base~O~S	875	9625
12	base~AD~A	812	9744
13	base~ANDO	810	10530
14	base~AD~O	768	10752
15	base~ABA	738	11070

Figura 7.2 Página principal del disco compacto

Para visualizar la página principal del disco, es necesario llevar a cabo los siguientes pasos:

1. Coloque el disco compacto en el lector de discos de la computadora.
2. Utilice un programa para explorar el contenido del disco.
3. Seleccione y abra el archivo index.html con un programa navegador de Internet. Bastará con hacer doble clic sobre el nombre del archivo.

Las otras dos páginas web permiten visualizar los autómatas y sus bases; ambas tienen la misma estructura. La diferencia entre ellas radica en que una muestra los autómatas generados a partir de la transcripción fonológica y la otra muestra aquellos generados a partir de la transcripción ortográfica. Por lo anterior explicaré sólo el contenido de la primera.

Como se puede ver en la Figura 7.3, la zona (1) muestra la lista de letras con las que comienzan los segmentos finales descubiertos. Si se hace clic sobre una letra, se despliegan los segmentos finales que comienzan con esa letra. Por ejemplo, si se hace clic sobre la A, se muestran los segmentos A, ABA, ABAMOS, etcétera.

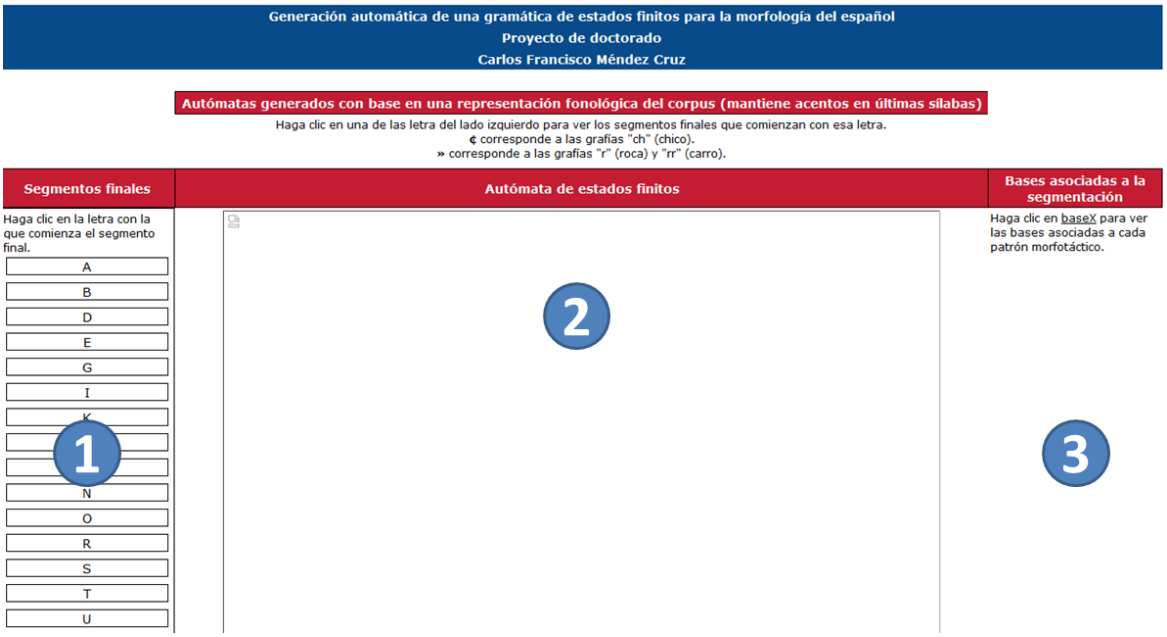


Figura 7.3 Página que permite visualizar autómatas

La zona (2) muestra el grafo del autómata asociado al segmento final seleccionado en la lista de la zona (1); al mismo tiempo, la zona (3) despliega los grupos de bases asociados al autómata. Véase la Figura 7.4 que muestra el grafo y grupos de bases del segmento final ABA.

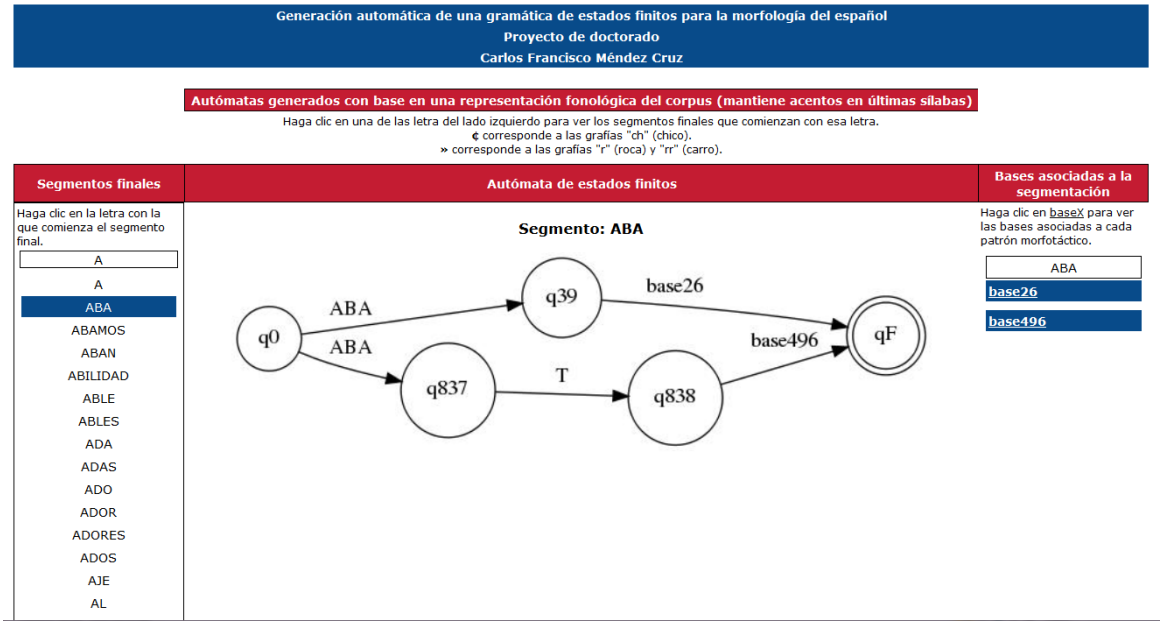


Figura 7.4 Página con autómata asociado al segmento ABA

En la zona del lado derecho (3), aparecen sólo los grupos de bases asociados al autómata que se está visualizando. Para ver las bases de cada grupo, basta con hacer clic en el nombre del grupo, por ejemplo *base26*. Debajo del nombre del grupo aparecerá la lista de bases. La Figura 7.5 muestra el grafo y lista de bases asociadas al patrón morfotáctico Base26~ABA.

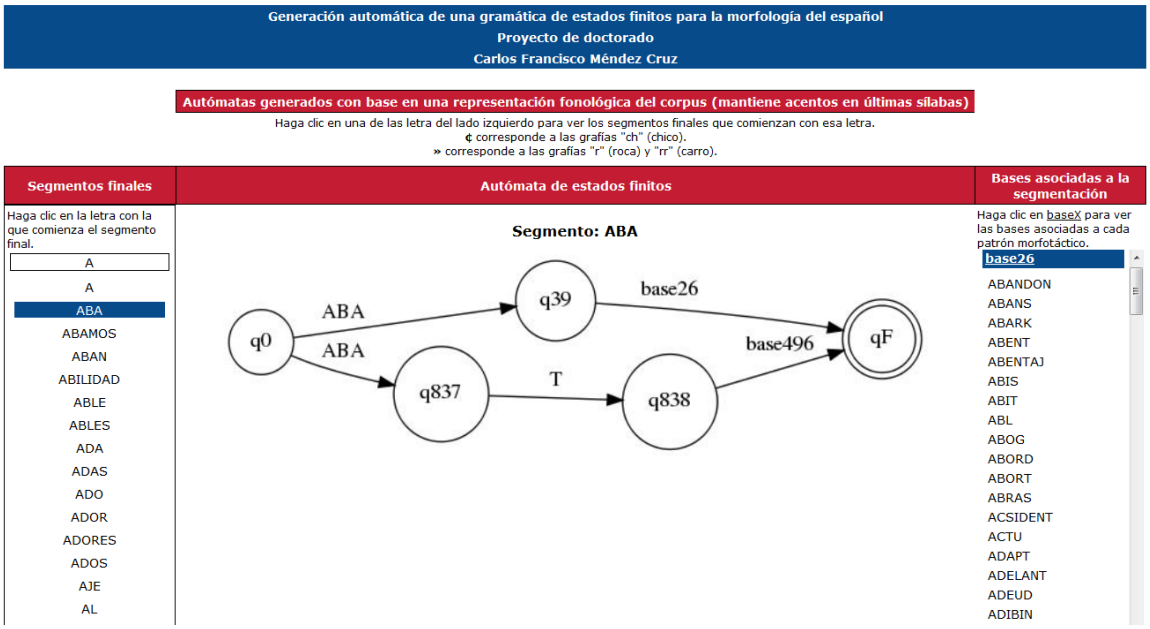


Figura 7.5 Página con autómata y lista de bases del segmento ABA

Bibliografía

Alcoba, S. (1999). La flexión verbal. En I. Bosque y V. Demonte, *Gramática descriptiva de la lengua española* (Vol. 3, págs. 4915-4991). Madrid: Espasa-Calpe y RAE.

Allen, J. (1995). *Natural Language Understanding.* Redwood City, California: Benjamin/Cummings.

Ambadiang, T. (1999). La flexión nominal. Género y número. En I. Bosque y V. Demonte, *Gramática descriptiva de la lengua española* (Vol. 3, págs. 4843-4913). Madrid: Espasa-Calpe y RAE.

Anderson, S. (1985). Typological Distinction in Word Formation. En T. Shopen (Ed.), *Language Typology and Syntactic Description. Grammatical Categories and the Lexicon* (Vol. III, págs. 3-56). Cambridge: Cambridge University Press.

Anderson, S. (1992). *A-Morphous Morphology.* Cambridge: Cambridge University Press.

Ando, R. K. y Lee, L. (2000). Mostly-unsupervised Statistical Segmentation of Japanese. Applications to Kanji. En *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (NAACL)* (págs. 241-248).

Antworth, E. L. (1990). *PC-KIMMO: A Two-level Processor for Morphological Analysis.* Texas: Summer Institute of Linguistics.

Aronoff, M. H. (1976). *Word Formation in Generative Grammar.* Cambridge, Mass.: The MIT press.

Baroni, M., Matiasek, J. y Trost, H. (2002). Unsupervised Discovery of Morphologically Related Words Based on Orthographic and Semantic Similarity. En *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* (págs. 48-57). Association for Computational Linguistics.

Beard, R. (1998). Derivation. En A. Spencer y A. M. Zwicky (Edits.), *The Handbook of Morphology* (págs. 44-65). Oxford y Malden, Mass.: Blackwell.

Beniers, E. (2000). *Lecturas de morfología.* México: UNAM.

Beniers, E. (2004). *La formación de verbos en el español de México.* México: El Colegio de México, UNAM.

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing, 8*(4), 243-257.

Bloomfield, L. (1961). *Language.* London: George Allen.

Brent, M. R. (1999). An Afficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning, 34*, 71-105.

Bybee, J. L. (1985). *Morphology: A Study of the Relation between Meaning and Form.* Amsterdam, Philadelphia: John Benjamins Publishing.

Charniak, E. (1996). *Statistical Language Learning.* Cambridge, Massachusetts: The MIT Press.

Chomsky, N. (1984). *Estructuras sintácticas.* México: Siglo XXI.

Creutz, M. (2003). Unsupervised Segmentation of Words Using Prior Distributions of Morph Length and Frequency. En E. Hinrichs y D. Roth (Edits.), *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (págs. 280-287). Sapporo, Japan.

Creutz, M. y Lagus, K. (2002). Unsupervised Discovery of Morphemes. En *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02, SIGPHON-ACL* (págs. 21–30). Philadelphia.

Creutz, M. y Lagus, K. (2004). Induction of a Simple Morphology for Highlyinflecting Languages. En *Proceedings of 7th Meeting of the ACL Special Interest Group in Computational Phonology SIGPHON-ACL* (págs. 43–51).

Creutz, M. y Lagus, K. (2005). 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 (AKRR'05)* (págs. 106–113). Finlandia: Espoo.

Crystal, D. (2003). *A Dictionary of Linguistics & Phonetics.* Oxford, UK: Blackwell.

De Kock, J. y Bossaert, W. (1974). *Introducción a la lingüística automática en las lenguas románicas.* Madrid: Gredos.

De Kock, J. y Bossaert, W. (1978). *The Morpheme. An Experiment in Quantitative and Computational Linguistics.* Amsterdam, Madrid: Van Gorcum.

De Marcken, C. (1995). *The Unsupervised Acquisition of a Lexicon from Continuous Speech. Technical Report A.I. Memo 1558.* Cambridge, Massachusetts: MIT Artificial Intelligence Lab.

Déjean, H. (1998). Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora. En D. Powers (Ed.), *Workshop on Paradigms and Grounding in Language Learning, ACL* (págs. 295-298).

Deligne, S. y Bimbot, F. (1997). Inference of Variable-length Linguistic and Acoustic Units by Multigrams. *Speech Communication, 23*(3), 223-241.

Diccionario del español de México (DEM). (s.f.). Recuperado el 15 de noviembre de 2012, de http://dem.colmex.mx

Diccionario del español de México. (2010). México: El Colegio de México, CELL.

Gelbukh, A., Alexandrov, M. y Han, S. (2004). Detecting Inflection Patterns in Natural Language by Minimization of Morphological Model. En A. Sanfeliu, J. F. Martínez y J. A. Carrasco (Edits.), *CIARP 2004.* (págs. 432-438). Heidelberg: Springer.

Gelbukh, A. y Sidorov, G. (2003). 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, N 2588* (págs. 215–220). Verlag: Springer.

Gelbukh, A., Sidorov, G. y Velásquez, F. (2003). Análisis morfológico automático del español a través de generación. *Escritos*(28), 9-25.

Gelbukh, A., Sidorov, G., Lara-Reyes, D. y Chanona-Hernández, L. (2008). Division of Spanish Words into Morphemes with a Genetic Algorithm. En E. Kapetanios, V. Sugumaran y M. Spiliopoulou (Edits.), *Natural Language and Information Systems* (págs. 19-26). Berlin, Heidelberg: Springer.

Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics, 27*(2), 153-198.

Goldsmith, J. (2006). An Algorithm for the Unsupervised Learning of Morphology. *Natural Language Engineering, 12*(4), 353-371.

Goldsmith, J. (2010). Segmentation and Morphology. En A. Clark, C. Fox y S. Lappin (Edits.), *The Handbook of Computational Linguistics and Natural Language Processing* (págs. 364–393). Oxford: Wiley-Blackwell.

González Calvo, J. M. (1998). *Estudios de morfología española.* Cáceres: Universidad de Extremadura.

Greenberg, J. (1963). Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. En J. H. Greenberg, *Universals of language* (Vol. 2, págs. 73-113). Oxford, England: MIT Press.

Greenberg, J. (1967). *Essays in Linguistics.* Chicago: The University of Chicago Press.

Grishman, R. (1991). *Introducción a la lingüística computacional.* Madrid: Visor.

Hammarström, H. y Borin, L. (2011). Unsupervised Learning of Morphology. *Computational Linguistics, 37*(2), 309-350.

Harris, Z. S. (1955). From Phoneme to Morpheme. *Language, 31*(2), 190–222.

Haspelmath, M. (2002). *Understanding Morphology.* New York: Oxford University Press.

Hockett, C. F. (1971). *Curso de lingüística moderna.* (E. Gregores y J. A. Suárez, Trads.) Buenos Aires: EUDEBA.

Holland, J. H. (1992). Genetic algorithms. *Scientific american, 267*(1), 66-72.

Hopcroft, J. E. y Ullman, J. D. (1969). *Formal Languages and their Relation to Automata.* Reading, Massachusetts: Addison-Wesley.

Hopcroft, J. E., Motwani, R. y Ullman, J. D. (2001). *Introduction to Automata Theory, Languages and Computation* (2 ed.). New York: Addison-Wesley.

Hull, D. A. (1996). Stemming Algorithms. A Case Study for Detailed Evaluation. *Journal of the American Society for Information Science, 47*(1), 70-84.

Hyman, L. M. y Mchombo, S. (2012). Morphotactic Constraints in the Chichewa Verb Stem. En *Proceedings of the Eighteenth Annual Meeting of the Berkeley Linguistics Society. General Session and Parasession on The Place of Morphology in a Grammar* (Vol. 18, págs. 350-364).

Jurafsky, D. y Martin, J. H. (2009). *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, N.J.: Pearson Prentice Hall.

Kageura, K. (1999). Bigram Statistics Revisited. A Comparative Examination of some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences. *Journal of Quantitative Linguistics, 6*(2), 149-166.

Karttunen, L., Chanod, J.-P. y Grefenstette, G. (1996). Regular Expressions for Language Engineering. *Natural Language Engineering, 2*(4), 305-328.

Katamba, F. y Stonham, J. (2006). *Morphology.* Houndsmills, Basingstoke, Hampshire: Palgrave Macmillan.

Kay, M. (1987). Nonconcatenative Finite-State Morphology. En *Proceedings of the third conference on European chapter of the Association for Computational Linguistics* (págs. 2-10). Association for Computational Linguistics.

Kay, M. (2003). Introduction. En R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (págs. XVII-XX). Oxford: Oxford University Press.

Kiparsky, P. (1983). Word Formation and the Lexicon. En F. Ingemann, *Proceedings of the 1982 Mid-America Linguistics Conference* (págs. 3-29). Lawrence, Kansas: University of Kansas.

Kit, C. y Wilks, Y. (1999). Unsupervised Learning of Word Boundary with Description Length Gain. En *Proceedings of CoNLL99 ACL Workshop.* Bergen.

Koskenniemi, K. (1983). Two-Level Model for Morphological Analysis. En *Proceedings of the 8th International Joint Conference on Artificial Intelligence* (págs. 683-685).

Koskenniemi, K. (1984). A General Computational Model for Word-Form Recognition and Production. En *Proceedings of the 10th International Conference on Computational Linguistics, Association for Computational Linguistics* (págs. 178-181). Helsinki, Finland: University of Helsinki.

Lara Reyes, D. (2008). *Sistema de segmentación automática de palabras en morfemas para el español (Tesis de maestría inédita).* México: CIC-IPN.

Lara, L. F. (2004). ¿Es posible una teoría de la palabra? *Lexis, XXVII*(1-2), 401-427.

Lara, L. F. (2006). *Curso de lexicología.* México: El Colegio de México.

Lara, L. F. y Ham Chande, R. (1974). Base estadística del Diccionario del español de México. *Nueva Revista de Filología Hispánica, 23*(2), 245-267.

Lara, L. F., Ham Chande, R. y García Hidalgo, M. I. (1979). *Investigaciones lingüísticas en lexicografía.* México: El Colegio de México.

Lin, C. Y. (2004). Rouge: A Package for Automatic Evaluation of Summaries. En *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop* (págs. 74-81).

Manning, C. D. y Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, Mass.: The MIT Press.

McEnery, T. y Wilson, A. (1996). *Corpus Linguistics. An introduction.* Edinburgh: Edinburgh University Press.

Medina, A. (2000). Automatic Discovery of Affixes by means of a Corpus. A Catalog of Spanish Affixes. *Journal of Quantitative Linguistics, 7*(2), 97–114.

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

Medina, A. (2007). Affix Discovery by Means of Corpora: Experiments for Spanish, Czech, Ralámuli and Chuj. En *Aspects of Automatic Text Analysis* (págs. 277-299). Berlin, Heidelberg: Springer.

Medina, A. (2008). Affix Discovery based on Entropy and Economy Measurements. *Computational Linguistics for Less-Studied Languages*(10), 99-112.

Medina, A. y Alvarado, M. (2006). Un experimento de reconocimiento automático de la derivación léxica en el ralámuli. En *La lengua y la antropología para un conocimiento global del hombre.* México: Conaculta/INAH.

Medina, A. y Buenrostro, C. (2003). Características cuantitativas de la flexión verbal del chuj. *Estudios de Lingüística Aplicada, 38*, 15-31.

Medina, A. y Hlaváčová, J. (2005). Automatic Recognition of Czech Derivational Prefixes. En A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Pro-cessing. CICLing 2005* (págs. 189-197). Berlin: Springer.

Medina, A., Herrera, J. A. y Alvarado, M. (2009). Towards the Speech Synthesis of Raramuri. A Unit Selection Approach based on Unsupervised Extraction of Suffix Sequences. En A. Gelbukh (Ed.), *Advances in Computational Linguistics, Re-search in Computing Science* (Vol. 41, págs. 243-256). Berlín: Springer.

Méndez-Cruz, C. F., Torres-Moreno, J. M., Medina, A. y Sierra, G. (2013). Extrinsic Evaluation on Automatic Summarization Tasks. Testing Affixality Measurements for Statistical Word Stemming. En I. Batyrshin y M. González Mendoza (Edits.), *MICAI 2012, Part II, LNAI 7630* (págs. 46-57). Heidelberg: Springer.

Méndez-Cruz, C.-F., Soriano-Morales, E.-P. y Medina, A. (2011). Testing a Statistical Word Stemmer based on Affixality Measurements in INEX 2012 Tweet Contextualization Track. En *INEX 2011 Workshop Pree-Proceedings* (págs. 194-200). Hofgut Imsbach, Saarbrücken, Germany: IR Publications.

Moreno de Alba, J. G. (1986). *Morfología derivativa nominal en el español de México.* México: UNAM.

Moreno de Alba, J. G. (1996). *La prefijación en el español mexicano.* México: UNAM, Instituto de Investigaciones Filológicas.

Neuvel, S. y Fulop, S. (2002). Unsupervised Learning of Morphology without Morphemes. En *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* (págs. 31-40). Association for Computational Linguistics.

Nida, E. A. (1949). *Morphology. The Descriptive Analysis of Words* (2 ed.). Ann Arbor: The University of Michigan.

Paik, J. H., Mitra, M., Parui, S. K. y Järvelin, K. (2011). GRAS: An Effective and Efficient Stemming Algorithm for Information Retrieval. *ACM Transactions on Information Systems (TOIS), 29*(4), 19-24.

Pena, J. (1999). Partes de la morfología. Las unidades del análisis morfológico. En I. Bosque y V. Demonte, *Gramática descriptiva de la lengua española* (Vol. 3, págs. 4305-4366). Madrid: Espasa-Calpe y RAE.

Piera, C. y Varela, S. (1999). Relaciones entre morfología y sintáxis. En I. Bosque y V. Demonte, *Gramática descriptiva de la lengua española* (Vol. 3, págs. 4367-4422). Madrid: Espasa-Calpe y RAE.

Porter, M. F. (1980). An Algorithm for Suffix Stripping. *Program, 14*, 130–137.

Redlich, A. N. (1993). Redundancy Reduction as a Strategy for Unsupervised Learning. *Neural Computation, 5*(2), 289-304.

Saggion, H., Torres-Moreno, J. M., da Cunha, I. y SanJuan, E. (2010). Multilingual Summarization Evaluation without Human Models. En *23rd Int. Conf. on Computational Linguistics. COLING '10* (págs. 1059-1067). Beijing, China: ACL.

SanJuan, E., Moriceau, V., Tannier, X., Bellot, P. y Mothe, J. (2011). Overview of the INEX 2011 Question Answering Track (QA@INEX). En *INEX 2011 Workshop*

Pree-Proceedings (págs. 145-153). Hofgut Imsbach, Saarbrücken, Germany: IR Publications.

Santiago, R. y Bustos, E. (1999). La derivación nominal. En I. Bosque y V. Demonte, *Gramática descriptiva de la lengua española* (págs. 4505-4594). Madrid: Espasa-Calpe y RAE.

Sapir, E. (1954). *El lenguaje. Introducción al estudio del habla.* México: Fondo de Cultura Económica.

Schone, P. y Jurafsky, D. (2000). Knowledge-Free Induction of Morphology using Latent Semantic Analysis. En *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning* (págs. 67-72). Association for Computational Linguistics.

Schone, P. y Jurafsky, D. (2001). Knowledge-Free Induction of Inflectional Morphologies. En *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (págs. 1-9). Association for Computational Linguistics.

Serrano-Dolader, D. (1999). La derivación verbal y la parasíntesis. En I. Bosque y V. Demonte, *Gramática descriptiva de la lengua española* (Vol. 3, págs. 4684-4755). Madrid: Espasa-Calpe y RAE.

Shannon, C. y Weaver, W. (1964). *The Mathematical Theory of Communication.* Chicago: The University of Illinois.

Spärck-Jones, K. y Galliers, J. (1996). *Evaluating Natural Language Processing Systems:*. New York: Springer-Verlang.

Spencer, A. (1991). *Morphological Theory. An Introduction to Word Structure in Generative Grammar.* Cambridge: Cambridge University Press.

Sproat, R. (1992). *Morphology and Computation.* Cambridge, London: The MIT Press.

Sproat, R., Shih, C., Gale, W. y Chang, N. (1996). A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. *Computational Linguistics, 22*(3), 377-404.

Stump, G. T. (1998). Inflection. En A. Spencer y A. M. Zwicky (Edits.), *The handbook of morphology* (págs. 13-43). Oxford y Malden, Mass.: Blackwell.

Swadesh, M. (1966). *El lenguaje y la vida humana.* México: FCE.

Teahan, W. J., Wen, Y., McNab, R. y Witten, I. H. (2000). A Compression-based Algorithm for Chinese Word Segmentation. *Computational Linguistics, 26*(3), 375-393.

Torres-Moreno, J. M. (2011). *Résumé automatique de documents.* Paris: Lavoisier.

Torres-Moreno, J. M., St-Onge, P. L., Gagnon, M., El-Bèze, M. y Bellot, P. (2009). Automatic Summarization System coupled with a Question-Answering System (QAAS).

Torres-Moreno, J., Saggion, H., da Cunha, I., SanJuan, E. y Velázquez-Morales, P. (2010). Summary Evaluation with and without References. *Polibits, 42*, 13-19.

Tzoukermann, E. y Mark, L. (1990). A Finite-State Morphological Processor For Spanish. En *Proceedings of the 13th International Conference on Computational Linguistics: COLING.* Helsinki, Finland.

Val Álvarado, J. F. (1999). La composición. En I. Bosque y D. Violeta, *Gramática descriptiva de la lengua español* (Vol. 3, págs. 4757-4841). Madrid: Espasa-Calpe y RAE.

Varela, S. y García, J. M. (1999). La prefijación. En I. Bosque y V. Demonte (Edits.), *Gramática descriptiva de la lengua española* (Vol. 3, págs. 4993-5040). Madrid: Espasa-Calpe y RAE.

Wall, R. (1972). *Introduction to Mathematical Linguistics.* Englewood Cliffs, New Jersey: Prentice Hall.

Zacarías, R. (2011). Formación de diminutivos con el sufijo/—ít—/. Una propuesta desde la morfología natural. *Anuario de Letras*(44), 77-103.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control, 8*(3), 338-353.