UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN LINGÜÍSTICA

IDENTIFICACIÓN AUTOMÁTICA DE CATEGORÍAS GRAMATICALES EN ESPAÑOL DEL SIGLO XVI

TESIS QUE PARA OBTENER EL GRADO DE
MAESTRO EN LINGÜÍSTICA HISPÁNICA
PRESENTA:

CARLOS FRANCISCO MÉNDEZ CRUZ

ASESOR: DR. ALFONSO MEDINA URREA

MÉXICO, D. F.
2009

Inamovible e inconmensurable es mi amor

AGRADECIMIENTOS

Agradezco en especial a mi asesor, el Dr. Alfonso Medina, por su gran apoyo, continuo entusiasmo y por compartir conmigo ideas y conocimientos que dieron guía a esta investigación. De igual manera, quiero agradecer al Dr. Gerardo Sierra, jefe del Grupo de Ingeniería Lingüística (GIL), por su constante y generoso apoyo, desde mi ingreso al GIL, hasta la culminación de este trabajo.

Muchas gracias a la Dra. Celia Díaz, el Dr. Luis Fernando Lara, la Dra. Jeanett Reynoso y el Dr. Gerardo Sierra por su lectura y valiosos comentarios a esta tesis.

Finalmente, agradezco el apoyo brindado por el grupo de etiquetadoras, la infraestructura del Instituto de Ingeniería de la UNAM, sede del GIL, y el soporte económico de los proyectos DGAPA-PAPIIT, IN 400905, “Constitución del corpus histórico del español de México”; DGAPA-PAPIIT, IN 402008, “Glutinometría y variación dialectal”; CONACYT, “Extracción de conceptos en textos de especialidad a través del reconocimiento de patrones lingüísticos y metalingüísticos”; y CONACYT, “Inducción automática de patrones de afitáctica de diversas lenguas”.

Índice general

Índice de tablas

Introducción

Los avances tecnológicos en materia computacional han traído, desde hace tiempo, nuevas posibilidades para la recolección, procesamiento, estudio y difusión de materiales lingüísticos. Así, en las últimas décadas, ha crecido el interés por el uso de corpus lingüísticos electrónicos como un recurso útil para llevar a cabo análisis lingüísticos, manuales, automáticos o semiautomáticos. Un corpus encierra los datos que el lingüista recopila para el estudio y descripción de una lengua. Se trata de un conjunto de textos orales o escritos seleccionados bajo criterios lingüísticos con el fin de ser usados como una muestra, preferentemente representativa, de cierta lengua, estado de lengua o dialecto. De esta manera, los corpus electrónicos se forman de textos en formato digital[1], y requieren de herramientas computacionales para su procesamiento.

Hoy en día, algunos de éstos se encuentran disponibles en Internet. Por ejemplo, dos de los corpus electrónicos de carácter diacrónico más utilizados son el Corpus Diacrónico del Español de la Real Academia Española (CORDE) y el Corpus del Español de Mark Davies (2003a, 2003b)[2]. Una de las ventajas de estos corpus es la posibilidad de realizar búsquedas de información de forma rápida y flexible.

La UNAM, por medio del Instituto de Ingeniería, ha desarrollado investigación en corpus lingüísticos electrónicos desde hace tiempo. En particular, llevó a cabo un proyecto (DGAPA PAPIIT IN400905 2005-2007) para la constitución de un corpus diacrónico del español novohispano y mexicano de los siglos XVI al XIX: el *Corpus Histórico del Español en México* (CHEM)[3]. El objetivo de este proyecto fue desarrollar un corpus textual representativo de los cuatro siglos y las herramientas necesarias para analizarlo.

La constitución de un corpus lingüístico electrónico, desde la perspectiva lingüística y computacional, es una tarea compleja (cf. Medina y Méndez 2006). De hecho, para poder hacer operativas búsquedas de información automáticas en los documentos del corpus es necesario desarrollar herramientas computacionales particulares: generadores de concordancias, etiquetadores automáticos y semiautomáticos, por mencionar algunos.

A propósito de lo anterior, las posibilidades de búsqueda en un corpus electrónico son muy variadas. Un ejemplo es la selección de contextos donde aparece una palabra completa o la

[1] Los textos en formato digital son archivos computarizados que pueden o no llevar información lingüística codificada mediante, por ejemplo, XML.

[2] Es posible consultar estos corpus electrónicos en http://www.rae.es y http://www.corpusdelespanol.org.

[3] Una primera versión de este corpus electrónico puede verse en http://www.iling.unam.mx/chem.

combinación de dos o más palabras; también, se pueden obtener contextos a partir de parte de una palabra. Otro ejemplo es la recuperación de contextos a partir de construcciones gramaticales del tipo *verbo + determinante + sustantivo* equivalentes a *echar un ojo* o *metió la pata*.

Esta tesis responde precisamente a la inquietud por el desarrollo de este último tipo de búsqueda. Con un corpus en el que cada palabra tenga asociada su categoría gramatical, el lingüista tendría la posibilidad de buscar no sólo contextos gramaticales, sino también los diversos usos de esas palabras y averiguar cuántas veces una palabra es usada, por ejemplo, como sustantivo o adjetivo; labor que resulta importante para un lexicógrafo, entre otros.

Con la idea de identificar las categorías gramaticales en el CHEM, se decidió investigar maneras automáticas para realizarlo. A este proceso se le conoce comúnmente como *etiquetado de partes de la oración* (en inglés, *Part of Speech Tagging*). Se llama etiquetado porque a cada palabra se le adhiere un código (etiqueta) que indica la categoría a la que pertenece. Además, el proceso es asociado al concepto de partes de la oración debido a que los métodos tradicionales identifican únicamente las categorías mayores de palabras, es decir, verbo, sustantivo, adjetivo, etc., sin tomar en cuenta rasgos gramaticales como género, número o tiempo. Para esta tesis, sin embargo, prefiero usar el término identificación automática de categorías gramaticales ya que se trata de identificar, mediante diferentes criterios, las categorías gramaticales generales (verbo, nombre, adjetivo) y específicas (tiempo, modo, género, número) de las palabras de un corpus.

Este tipo de trabajo no es nuevo en la lingüística mexicana. De hecho, en El Colegio de México, en el marco del proyecto del Diccionario del Español de México, se comenzó a mediados de los setenta una tradición en el desarrollo de herramientas computacionales para análisis lingüístico (cf. Lara, Ham y García 1979). Hasta donde tengo noticia, fue aquí donde se desarrolló el primer analizador gramatical de la lengua española, para procesar el Corpus del Español Mexicano Contemporáneo (cf. García Hidalgo 1979). Dicho analizador permitió la identificación automática de categorías gramaticales en ese corpus haciendo uso de reglas elaboradas por lingüistas y de un árbol de caracteres para el análisis al interior de la palabra. Siguiendo esta tradición, propongo el presente trabajo de investigación.

Planteamiento del problema

En este apartado, presento la problemática que deberé resolver para alcanzar los objetivos de investigación. Identificar categorías gramaticales, de forma manual o automática, requiere de solucionar diversos problemas lingüísticos. Por una parte, dependiendo de la lengua que se analice, y específicamente de la morfología de esa lengua, será pertinente estudiar los fenómenos

morfológicos o sintácticos para encontrar las pautas que den indicios de cómo identificar las categorías de tal lengua. Por ejemplo, para lenguas más sintéticas será la morfología la que posiblemente brinde mejores pistas en esta identificación. Además, lenguas con pocas categorías gramaticales tendrán pautas distintas a las que cuentan, por ejemplo, con cinco o más (español, inglés).

Otro problema de carácter lingüístico es la ambigüedad categorial. Lenguas como el español cuentan con palabras que pueden estar asociadas a distintas categorías; por lo general, el contexto sintáctico determina cuál de ellas es la correcta. Además, respecto a las lenguas en general, si bien las marcas de flexión y derivación brindan pistas para la determinación de categorías, no se puede olvidar que en muchas lenguas existen paradigmas defectivos y fenómenos morfológicos, derivativos y flexivos, que no son concatenativos o que no cambian la forma de la palabra. Los cambios de tono y acento para marcar flexión o derivación en distintas lenguas son ejemplos de ello (cf. Bloomfield 1961/1933: 221 y Stump 1998: 32).

Hablando en particular del español, se pueden mencionar dificultades adicionales. Algunos sustantivos españoles suelen reflejar alguna categoría gramatical a través del determinante y no por la marca flexiva (*el colega* frente a *la colega; los análisis* frente a *el análisis*) y muchos cuentan con categorías arbitrarias, principalmente de género. Un problema adicional es la falta de distinción formal entre sustantivos y adjetivos; de hecho, la gramática tradicional no hace distinción entre las dos categorías y habla sólo del nombre (cf. RAE 1973).

Para desarrollar la tesis, utilizaré un corpus textual del siglo XVI[4], por lo que se agregan algunos problemas. Entre ellos está la variación de formas ortográficas del español de la época y la aparición de léxico indígena. También, ya que el corpus está constituido en su mayoría por cartas y juicios inquisitoriales, hay una abundancia de fórmulas de cortesía, títulos nobiliarios y nombres propios que será interesante examinar en el contexto de la identificación automática de categorías.

[4] Conviene enfatizar que, por obvias razones, esta investigación se enfoca en la lengua escrita. Más adelante indicaré con detalle los textos que conforman el corpus de estudio. Por ahora, es importante aclarar que se trata de ediciones actualizadas de diversos documentos de la época. Éstas han sido elaboradas por filólogos o lingüistas que tomaron decisiones sobre la modernización de la puntuación, separación de palabras, acentuación y empleo de mayúsculas y minúsculas; así como sobre el manejo de grafías, omisiones y abreviaturas. Al respecto, para ver los criterios utilizados para esta labor, se puede consultar la introducción de Company a los *Documentos lingüísticos de la Nueva España. Altiplano Central* (1994: 1-19) y la aclaración de Lope Blanch a *El habla de Diego de Ordaz. Contribución a la historia del español americano* (1985: 191-192).

En otros aspectos, hoy en día existen procesos computacionales capaces de realizar la identificación de categorías gramaticales de forma automática o semiautomática con niveles aceptables de precisión[5]. Sin embargo, la mayoría de dichos programas de computadora han sido desarrollados para lengua inglesa y, en menor medida, otras lenguas indoeuropeas (cf. Voutilainen 1999b). A pesar de esto, es posible encontrar programas de etiquetado de categorías para el español, pero gran parte de ellos están orientados al español contemporáneo peninsular y ciertamente no al español novohispano.

Los programas para identificar categorías pueden agruparse en dos grandes clases de acuerdo con los métodos y modelos que utilizan. Por un lado, aquellos basados en métodos estadísticos como los modelos de Markov (cf. Charniak 1996/1993, Church 1988, DeRose 1988, Brants 2000) y por otro, los basados en reglas, siendo el más representativo el desarrollado por Eric Brill (1992, 1993, 1994, 1995)[6]. Tradicionalmente, los primeros no hacen análisis al interior de la palabra y se ayudan de la dependencia que guarda una palabra con sus palabras anteriores y posteriores, expresada mediante probabilidades[7]. Los segundos hacen uso de reglas léxicas (de análisis muy simple al interior de la palabra) o contextuales (mirando las palabras aledañas) establecidas previamente o adquiridas de forma automática.

A grandes rasgos, el tipo de análisis que realiza el método de Brill para etiquetar una palabra se hace mediante reglas basadas en uno, dos, tres o cuatro caracteres iniciales y finales. Un ejemplo de este tipo de reglas para el español sería:

- *Si los dos caracteres finales de la palabra son "er" la palabra es etiquetada como verbo*.

Este es el tipo de regla "morfológica" que utiliza este método y que podría no ser suficiente para algunas lenguas. De hecho, el número de caracteres al principio o al final es un parámetro que necesitaría ser modificado de una lengua a otra dependiendo del tamaño de los afijos y cadenas de afijos de cada una. Aunque parece una buena aproximación a la parte "morfológica" que determina

[5] La precisión tendrá en esta tesis dos sentidos. Por un lado, se referirá a la exactitud de un proceso o método; por el otro, a una medida de evaluación de sistemas de recuperación de información obtenida de dividir el número de predicciones correctas entre el número total de predicciones. Así, por ejemplo, Brill (1995) ha obtenido para el inglés hasta el 97% de precisión (97 palabras de cada 100 son etiquetas correctamente).

[6] Ya que basaré el desarrollo de esta tesis en el método de Brill, en un apartado especial lo explicaré con detalle.

[7] Al respecto, el método TnT, a diferencia de los métodos más tradicionales, sí utiliza partes finales de palabra (cf Brants 2000).

rasgos categoriales, no es suficiente para ganarse esa denominación ni para que los pedazos de palabra encontrados puedan llamarse morfemas.

A propósito del problema anterior, existe la posibilidad de utilizar un método automático de descubrimiento de afijos, que permita determinar unidades morfológicas que puedan involucrarse en el método de identificación de categorías. Existen propuestas que han probado ser útiles en esta labor. Una de ellas es la de Medina (2000, 2003), quien ha experimentado con distintos métodos de segmentación de palabras e índices de afijalidad[8] para varias lenguas.

Un problema muy importante es la cantidad de corpus etiquetado requerido para generar, ya sean modelos estadísticos o reglas. Aunque existen posturas en contra (cf. Brants 2000), parece válido pensar que los métodos estadísticos requieren de mayor cantidad de texto para lograr un modelo más productivo que sirva para etiquetar corpus adicional. En el caso del método de Brill, es posible comenzar con muy poco texto etiquetado y generar con él reglas que etiqueten más corpus. Para mi investigación, ésta es una ventaja del método, ya que no es fácil contar con corpus electrónicos etiquetados de español novohispano.

La tradición anglosajona tiene la tendencia a la generación de recursos lingüísticos ampliamente compartidos. En este sentido, existen grandes corpus electrónicos etiquetados disponibles para su análisis (*Brown Corpus*, *Penn Treebank, Wall Street Journal*); desafortunadamente no sucede lo mismo para el español y mucho menos para el español antiguo. Por ello, tuve que producir un corpus etiquetado del español del siglo XVI[9] con los problemas que esto conlleva, principalmente: el etiquetado manual es lento y al final hay inconsistencias de criterios.

En resumen, el problema al que me enfrento es que los métodos de identificación de categorías gramaticales apenas toman en cuenta unidades morfológicas y, si lo hacen, su tratamiento es muy simple. Además, no encontré desarrollos de este tipo de métodos para el español novohispano.

---

[8] El índice de afijalidad, como veremos más adelante, es una medida obtenida de cuantificar las características de un afijo: capacidad combinatoria, cantidad de información en relación a otras unidades y la economía que aporta el sistema lingüístico.

[9] Como quedó implícito arriba, siempre que me refiera al español del siglo XVI, lo haré en el entendido de que se trata del español transcrito a partir de manuscritos de ese siglo que, como ya dije, han sido modernizados (véase *supra* nota 4).

Objetivos

En esta sección pongo los objetivos que intentaré alcanzar con el desarrollo de este trabajo. Ya que los métodos de identificación automática de categorías gramaticales han sido desarrollados sobre todo para la lengua inglesa, toman muy poco en cuenta la morfología. Sin embargo, hablantes de otras lenguas han tenido la preocupación de desarrollar métodos de análisis morfológico que rindan cuenta de la complejidad de sus lenguas (por ejemplo, el finlandés). Para el español del siglo XVI, creo que es indispensable utilizar un método que tome en cuenta el interior de la palabra mediante un análisis morfológico basado en criterios lingüísticos.

Por ello, considero que el método de Brill resulta apropiado para esta propuesta de investigación. Por un lado, expresa en reglas cierto conocimiento para la identificación de categorías gramaticales. Por otro, no requiere de un gran corpus etiquetado para comenzar la generación de reglas. Además, divide la generación de reglas en léxicas y contextuales, situación que se adapta a la propuesta de inclusión de afijos. Por tanto, el presente trabajo de investigación tiene los siguientes objetivos:

i. Profundizar en el conocimiento del método de Brill para identificación automática de categorías gramaticales con el fin de agregar la generación de reglas morfológicas basadas en afijos descubiertos automáticamente.
ii. Obtener un inventario de afijos, descubiertos automáticamente a partir del corpus, utilizando el método de Medina basado en un índice de afijalidad.
iii. Etiquetar con categorías gramaticales textos del español del siglo XVI de manera automática para comparar el método que usa reglas morfológicas con el método que no lo hace.
iv. Construir un etiquetador de categorías gramaticales que haga uso de reglas morfológicas basadas en afijos descubiertos automáticamente aplicable al siglo XVI del CHEM.

En resumen, intento proponer un método de identificación de categorías gramaticales, basado en el de Brill y usable en textos de español del siglo XVI, que tome en cuenta unidades morfológicas descubiertas a partir de criterios lingüísticos cuantificables mediante un índice de afijalidad.

Hipótesis

Como se había mencionado antes, los métodos tradicionales para identificación automática de categorías gramaticales no toman en cuenta la morfología. Métodos más recientes, como el de Brill, incluyen en su análisis el procesamiento de caracteres iniciales y finales de palabras. Esta situación podría deberse a la falta de desarrollos en lingüística computacional para lenguas no anglosajonas,

por lo que los desarrollos se han enfocado tradicionalmente a lenguas con poca morfología o ignorando su riqueza morfológica. La falta de corpus electrónicos, e inclusive textuales, de lenguas poco estudiadas y de morfologías complejas puede ser otra causa. Sin embargo, hay que tomar en cuenta que existen tradiciones en lingüística computacional que han estudiado el alemán, el ruso e, incluso, el finlandés; esto es, lenguas con morfología más compleja que la inglesa.

Tal vez otra razón sea el enfoque de donde han partido muchos trabajos en lingüística computacional. Es decir, muchas de las investigaciones en este campo parten de la computación hacia la lingüística y pocos de la lingüística hacia la computación. En otras palabras, no es lo mismo que un científico de cómputo pruebe un algoritmo con datos lingüísticos a que un científico de la lingüística pruebe una teoría mediante un algoritmo computacional. Resultan importantes, en este sentido, los trabajos de lingüistas que se han preocupado por generar modelos computacionales del lenguaje.

Así, teniendo en mente lo anterior, pongo a continuación las hipótesis de trabajo que serán puestas a prueba durante mi investigación:

i. La identificación automática de categorías gramaticales que toma en cuenta unidades morfológicas tiene mejores resultados que aquella que se concreta a utilizar caracteres iniciales o finales de palabras.
ii. El método de Brill con la incorporación de reglas morfológicas deducidas a partir de afijos descubiertos automáticamente con el método de Medina será mejor que el método de Brill solo.

Estas hipótesis serán examinadas mediante experimentación en textos de español del siglo XVI. En el siguiente apartado, describiré la metodología propuesta para llevar a cabo la investigación.

Metodología

He introducido algunos problemas teóricos y prácticos referentes a la identificación automática de categorías gramaticales. Además, establecí los objetivos que intento alcanzar y las hipótesis que serán probadas. Ahora, expongo la metodología que seguiré para llevar a cabo la investigación. Ésta se divide en varios pasos que muestro a continuación.

Preparación del corpus de estudio

Serán seleccionados textos del Corpus Histórico del Español en México (CHEM) del siglo XVI. Éstos provendrán principalmente de: *Los procesos inquisitoriales contra indígenas*[10], los *Documentos lingüísticos de la Nueva España, Altiplano Central*[11], y *El habla de Diego de Ordaz: contribución a la historia del español americano*[12].

Determinación del conjunto de etiquetas de categorías gramaticales

Ya que existe una gran variedad de formas de etiquetar categorías gramaticales en textos, revisaré algunos estándares de codificación para determinar si adoptar alguno de forma completa o alguno modificado. El objetivo de esta fase es obtener unas guías generales de etiquetado para la tesis.

Etiquetado del corpus de estudio

Los textos serán etiquetados con categorías gramaticales de forma manual con la ayuda de herramientas computacionales desarrolladas para tal efecto. En esta etapa participarán estudiantes de la carrera de Lengua y Literaturas Hispánicas de la Facultad de Filosofía y Letras de la UNAM. El corpus será dividido en tres partes siguiendo los requerimientos del método de Brill: un corpus de entrenamiento para generar reglas, un corpus adicional y uno de evaluación.

Descubrimiento automático de afijos en el corpus con el método de Medina

Se utilizará el método desarrollado por Medina para el descubrimiento de afijos en textos del CHEM. De esta forma, se obtendrá un inventario de unidades morfológicas que usaré en la generación de reglas de etiquetado.

Adaptación del método de Brill para incluir el uso de plantillas de reglas morfológicas

Se modificarán las plantillas de reglas léxicas propuestas por Brill para que incluyan el inventario de unidades morfológicas descubiertas. Las plantillas del método original toman en cuenta sólo cuatro caracteres al inicio y final de la palabra, por tanto, el objetivo de esta etapa es que el programa de

---

[10] E. Buelna Serrano. 2005. *Los procesos inquisitoriales contra indígenas que realizó Fray Juan de Zumárraga en Nueva España (1536-1543)*, UAM-A, (manuscrito electrónico).

[11] Concepción Company Company. 1994. *Documentos lingüísticos de la Nueva España. Altiplano Central,* México: UNAM.

[12] J. M. Lope Blanch. 1985. *El habla de Diego de Ordaz. Contribución a la historia del español americano.* Publicaciones del Centro de Lingüística Hispánica, 20. México: UNAM, IIF.

Brill genere reglas usando los afijos y cadenas de afijos previamente descubiertos. A estas nuevas reglas se les podría llamar reglas morfológicas.

Generación de reglas y comparación de métodos

Serán generadas las reglas léxicas con el método original a partir del corpus de estudio y con ello se obtendrá un primer corpus etiquetado y un conjunto de reglas. Después, con los mismos textos, serán generadas las reglas morfológicas a partir de los afijos descubiertos. Esto nos dará como resultado un segundo corpus etiquetado y un nuevo conjunto de reglas comparables con los primeros.

Resultados y evaluación del método propuesto

Haré un análisis de los resultados obtenidos con la aplicación de ambos métodos en el corpus. Además, evaluaré con la medida de precisión los logros de cada uno para obtener una comparación cuantitativa. Finalmente, obtendré conclusiones.

Delimitación y alcance

En este apartado, expongo el alcance de la tesis, los aspectos que atenderé y los que quedarán fuera de mi investigación. Nunca será por falta de interés que deje de lado temas lingüísticos o computacionales, lo haré para lograr los fines propuestos.

Si bien el corpus que etiquetaré es del siglo XVI, tengo en mente la aplicación del método en otras lenguas y en distintas variantes (temporales, espaciales, sociales y temáticas) del español, por ejemplo: corpus de español contemporáneo, corpus especializados y corpus de habla infantil. Por esta razón, es muy importante el uso de un método automático de descubrimiento de unidades morfológicas no supervisado, es decir, sin información lingüística a priori[13]. La intención de usar un método de este tipo es evitar darle al proceso de identificación de categorías una lista preestablecida

[13] En este sentido, el método de Medina es no supervisado, ya que no requiere de información para realizar el descubrimiento. Medina explica que dar información lingüística previa marcada en el corpus sería: "Lo que en inglés se conoce como entrenamiento supervisado del sistema (*supervised training*) y que se refiere al hecho de proporcionarle manualmente al programa la información gramatical de los datos de una muestra. Cuando la clasificación de dichos datos no está disponible, se habla de aprendizaje o entrenamiento no supervisado (*unsupervised training*)" (2003: 4, nota 7). En Manning y Schütze podemos encontrar una distinción entre aprendizaje supervisado y no supervisado: "The distinction is that with supervised learning we know the actual status (here, sense label) for each piece of data on which we train, whereas with unsupervised learning we do not know the classification of the data in the training sample" (1999: 232).

de afijos, y en su lugar descubrir los afijos de la lengua de forma automática. De hecho, poder descubrir afijos para el siglo XVI me permite no depender de los consignados en otros trabajos de investigación, lo que significa introducirme empíricamente a la morfología del XVI. Así, corpus de lenguas con morfología concatenativa, flexiva y derivativa, serán susceptibles de ser etiquetados con el método.

Un reto interesante sería descubrir no sólo las unidades morfológicas de la lengua, sino el conjunto de categorías gramaticales mismo. Es decir, mediante un método computacional, determinar las palabras con comportamiento similar y agruparlas[14]. En esta tesis no atiendo este problema.

Es importante resaltar que tomo el método de Brill (1992, 1993, 1994, 1995) y lo modifico para integrar en él el descubrimiento automático de afijos propuesto por Medina (2000, 2003). Es decir, no desarrollo un método desde cero, sino que propongo uno nuevo a partir del de Brill. Está modificación pondrá en relieve dos aspectos.

El primero es el uso de unidades morfológicas, determinadas mediante métodos estadísticos de carácter lingüístico, para la identificación de categorías gramaticales tal y como la tradición lingüística lo ha sostenido en innumerables bibliografías. El segundo es un cuestionamiento sobre la posibilidad de que el uso de unidades morfológicas mejore la identificación de categorías gramaticales en oposición al mero uso de caracteres finales o iniciales.

Con el resultado de esta tesis, se obtendrá una propuesta de generación de reglas de identificación de categorías basadas en unidades morfológicas descubiertas y no de simples caracteres. El conjunto de reglas podrá ser usado en español del siglo XVI por cualquier etiquetador basado en la idea de Brill. Como se dijo arriba, este método genera reglas léxicas que analizan el interior de la palabra, y contextuales, que ven su relación con otras palabras. En esta tesis sólo propondré modificaciones a las primeras.

Por otra parte, esta tesis no incluye una comparación entre los métodos estadísticos y los basados en reglas para la identificación de categorías gramaticales. La comparación será hecha entre un método basado en reglas y el mismo método con la incorporación del descubrimiento automático de afijos. A propósito del método de Brill, éste fue concebido desde la perspectiva de la inteligencia artificial, específicamente en el campo del aprendizaje de computadora. Esta área del conocimiento ha estado muy ligada al estudio del lenguaje, principalmente a través del procesamiento de lenguaje

[14] De hecho, ya existen propuestas para realizar esta labor, el mismo Eric Brill (1993: 53-54) prueba un método basado en la distancia de Kullback y Leibler.

natural. Pero por la amplitud del tema y porque esta tesis no es de inteligencia artificial, no investigaré aquí los fundamentos del aprendizaje de computadora.

A propósito del método automático de identificación de afijos, no realizo una comparación entre distintas propuestas. Sino que tomo la propuesta de Medina ya que ha sido probada en distintas lenguas con buenos resultados, entre ellas el español del siglo XX (cf. Medina 2000), la lengua checa (cf. Medina y Hlaváčová 2005), el chuj (cf. Medina y Buenrostro 2003), y el tarahumara o rarámuri (cf. Medina y Alvarado 2006; y Medina, Camacho y Alvarado 2009).

Ya que la tesis aborda un problema que cae en el terreno de la morfología, tanto por el descubrimiento de afijos como por las reglas para determinar categorías gramaticales, será necesario revisar la literatura al respecto. Sin embargo, los estudios de morfología, al igual que la cantidad de fenómenos que éstos abarcan, son muy amplios. Por ello, para efectos de este trabajo de investigación, sólo atenderé la morfología concatenativa y en especial la afijal. Ya que el método de descubrimiento de afijos propone fragmentos con características afijales, creo pertinente revisar los fenómenos que involucren adición de segmentos incluyendo, además de flexión y derivación, incorporación y composición. Es pertinente decir que estos fenómenos serán revisados desde una perspectiva descriptiva y sin tomar una postura teórica, ya que mi objetivo es mostrar la complejidad morfológica de las lenguas.

Tampoco es mi intención hacer un inventario exhaustivo de los afijos de flexión y derivación del español del siglo XVI y discutir a profundidad su pertinencia. Sólo haré referencia a los que sean pertinentes para el desarrollo de la tesis, básicamente, aquellos que formen parte de las reglas generadas. Al respecto, no se resolverán problemas teóricos sobre estos afijos ni se discutirán aspectos semánticos. Finalmente, puedo decir que la tesis no se inserta en el plano diacrónico, sino en el sincrónico, ya que se analiza un corpus del español en un momento específico del tiempo.

Estructura de la tesis

En esta sección, describo la estructura de la tesis y comento brevemente el contenido de cada capítulo. Ya que este trabajo tiene que ver con los fenómenos morfológicos, en el primer capítulo, "Morfología", hago una revisión de bibliografía lingüística para determinar su concepto y objeto de estudio. Recupero el concepto de morfema y la manera de identificarlo en corpus. En seguida, describo la realización de distintos fenómenos morfológicos: reduplicación, afijación, cambios fonéticos y otros. También, pongo en distintos apartados la flexión, derivación, composición e incorporación. Además, agrego secciones sobre la pertinencia de distinguir entre flexión y derivación y los tipos de lenguas según su morfología.

El segundo capítulo está dedicado a "Las categorías gramaticales". En él se revisan los conceptos lingüísticos de clases de palabras, partes de la oración y categorías gramaticales, y de manera breve abordo una discusión sobre los problemas terminológicos al respecto. También, indago sobre los criterios utilizados para identificarlas en distintas lenguas. Para terminar el capítulo, expongo una discusión sobre la posibilidad de que algunas de las categorías puedan ser universales.

En el tercer capitulo, "Identificación automática de categorías gramaticales", hablo de los métodos y modelos que se han propuesto para resolver este problema desde el punto de vista computacional. Básicamente reviso dos: los estadísticos y los basados en reglas. Al respecto, incluyo una breve historia de los programas de etiquetado de categorías. También, dedico varias páginas para explicar el método que utilizaré en la tesis y que fue propuesto por Brill. Además, menciono algunos aspectos sobre el analizador gramatical desarrollado en El Colegio de México para analizar el Corpus del Español Mexicano Contemporáneo. Una sección especial tratará de los conjuntos de etiquetas utilizados para marcar corpus.

El cuarto capítulo, "Descubrimiento de afijos por computadora", tiene el propósito de dar cuenta de las posibilidades de descubrir unidades lingüísticas con métodos computacionales. Por ello, expongo una serie de propuestas de segmentación de palabras, supervisadas y no supervisadas. Después, con mayor nivel de profundidad, abordo el método propuesto por Medina que, como ya se dijo, utiliza un índice de afijalidad como medida para determinar afijos en corpus.

El siguiente capítulo, "Identificación automática de categorías gramaticales en español del siglo XVI", describe cómo se aplicaron los métodos automáticos de descubrimiento de afijos y de identificación de categorías al corpus de estudio. Doy cuenta de las características del conjunto de etiquetas y del corpus utilizado. Al respecto, expongo cómo se dividió para la realización de los experimentos. También, reporto cómo se llevó a cabo el proceso de descubrimiento de afijos y hago una sencilla evaluación del mismo. En una sección de este capítulo, presento el resultado de la aplicación de los métodos automáticos de identificación de categorías con y sin el uso de afijos. Esto dio paso a una evaluación cuantitativa que permitió comparar ambos métodos.

El capítulo final, "Conclusiones", presenta los aspectos significativos de la investigación realizada, posibles aportaciones y expectativas de investigación a futuro. También, reviso las hipótesis y los objetivos planteados al inicio de la tesis. Al final, brindo unas conclusiones generales.

El capítulo de "Apéndices", incluye principalmente la lista de afijos descubiertos y las listas de reglas obtenidas por el método original y por el método propuesto. Agrego también información de carácter más computacional, como los detalles para ejecutar el método de generación de reglas basado en transformaciones de Brill.

De esta manera, en esta introducción he detallado los problemas involucrados en la identificación de categorías gramaticales de forma automática. Además, expuse los objetivos e hipótesis de la presente investigación, los alcances que ésta tendrá y la metodología para llevarla a cabo. El siguiente capítulo, como ya se había mencionado, estará dedicado al conjunto de fenómenos lingüísticos encerrados en lo que se conoce como morfología.

1. Morfología

Antes de comenzar a exponer diversos aspectos de la morfología, quiero justificar la necesidad de abordar este nivel de lengua para mi investigación. En primer lugar, para la identificación de categorías gramaticales son útiles los criterios morfológicos, por ejemplo, la presencia o ausencia de afijos flexivos y derivativos permite determinar la pertenencia de una palabra a cierta categoría. En segundo lugar, el método de identificación automática utiliza segmentos de palabras de uno, dos, tres o cuatro caracteres. Mi intención es mejorar esta situación y utilizar afijos descubiertos automáticamente con métodos lingüísticos.

Por lo anterior, creo necesario indagar sobre la morfología de las lenguas, conocer los fenómenos que intervienen en la formación de palabras y establecer los procedimientos generales para la determinación de unidades morfológicas en corpus. Se propuso que el uso de unidades morfológicas mejorará en algún aspecto el método de identificación automática de categorías gramaticales, por consiguiente, es importante entender cómo se comporta el nivel morfológico dentro del sistema lingüístico. En caso de que existiera un mecanismo computacional que descubriese la morfología de una lengua, éste podría ser usado para determinar las categorías gramaticales de manera más efectiva.

Si bien el método automático será probado en español novohispano, tengo en mente su aplicación en otras lenguas con morfología concatenativa. Por esto, creo pertinente mirar la morfología desde una perspectiva general y no hispanista. Además, en la medida en que conozca la variedad de fenómenos que ocurren en diferentas lenguas, seré más consciente de la complejidad que encierra este nivel del lenguaje. En este sentido, es común encontrar propuestas computacionales que resuelven problemas específicos de una lengua sin tomar en cuenta aspectos tipológicos.

En este capítulo reviso conceptos básicos como el de palabra, morfema y morfología. Además, examino algunas propuestas generales para aislar morfemas en corpus, básicamente de tipo estructuralista distribucional. Asimismo, describo diversos fenómenos morfológicos concatenativos y no concatenativos en distintas lenguas.

Más adelante, indago sobre los mecanismos de formación de palabras: flexión, derivación, composición e incorporación, por ser fenómenos que pueden realizarse mediante adición o transformación de material fonológico o, en lengua escrita, cadenas de caracteres. Entre las secciones de flexión y derivación, reviso la distinción entre ellas, ya que trabajaré en la tesis específicamente con afijos de estos dos tipos y sería importante conocer sus diferencias.

La siguiente sección trata de la definición y pertinencia de varias unidades involucradas en el análisis morfológico. Parto de la unidad más grande en la morfología, la palabra, hasta las más pequeñas, como raíz y afijos.

1.1. La palabra y el morfema

Los estudios de morfología responden al interés científico de explicar los fenómenos que sufren ciertos segmentos de las emisiones lingüísticas. Estos segmentos han sido tradicionalmente llamados palabras; uno de varios términos de la lingüística que son muy útiles y a la vez muy criticados (Cf. Anderson 1985b: 150-165, González Calvo 1998: 11-37, Lara 2004: 401-408, Lara 2006 primera parte, y Pena 1999: 4327-4328).

Para el Esbozo (cf. RAE: 1973: 163) las palabras son identificables especialmente por su carácter de separabilidad, un concepto "virtual" en cuanto que no es perceptible en la emisión sonora. En el caso de lenguas con sistemas de escritura como la del español, que han heredado el uso de espacios para delimitar palabras gráficas, estos sistemas reafirman la separabilidad de palabras. Por otra parte, los métodos automáticos para descubrir afijos y para etiquetar categorías gramaticales utilizados en esta tesis están basados en el procesamiento de palabras gráficas, es decir, unidades separadas por espacios[1].

En opinión de Pena (1999: 4327), la palabra en español sí es una entidad propia e identificable, a diferencia de otras lenguas. Pero para definirla son necesarias pruebas adicionales a la separabilidad. Propone entonces criterios al exterior de la palabra y a su interior. Entre los primeros están: posibilidad de cambiar su posición en la secuencia del sintagma, y que el hablante puede hacer una pausa antes o después de una palabra. Los segundos son: inseparabilidad y orden fijo de los morfemas que la constituyen.

Los criterios anteriores ya habían sido discutidos por otros autores, entre ellos Anderson (1985b). Este autor, desde su visión tipológica, agrega otros. Uno de ellos es el acento, que permite identificar límites entre palabras. En ciertas lenguas éste aparece siempre en una sílaba, por ejemplo en la lengua checa siempre se acentúa la primera. Un criterio adicional es el tipo de modificación en las palabras, en algunas lenguas, no ocurren los mismos cambios al interior de la palabra que en los límites entre una y otra. También es posible que determinadas lenguas restrinjan las clases de sonidos agrupables al final de la palabra. Finalmente, otro criterio al interior de la palabra se da en

[1] La palabra gráfica no necesariamente corresponde a la palabra como concepto lingüístico, pero este recurso nos permite aplicar las técnicas que se presentarán en el capítulo correspondiente al descubrimiento de afijos.

lenguas donde las palabras deben llevar flexión obligatoria; así, estas marcas servirían de límites entre ellas.

Según Lara (2006), las características fonológicas y morfológicas mencionadas no son suficientes para determinar la existencia de la unidad palabra. Por ello, propone que son necesarias además características semánticas de la palabra como unidad de denominación (cf. Lara 2006: 37-51). Así, las tres características juntas serían condiciones necesarias y suficientes para determinar la existencia de la palabra como unidad lingüística en cualquier lengua.

El estructuralismo definió la lengua como un sistema formado por un conjunto de unidades lingüísticas y, a partir de esta idea, fueron numerosos los esfuerzos para delimitarlas en los distintos niveles lingüísticos: fonología, morfología, sintaxis y semántica. Las unidades del primer nivel, desprovistas de significado, fueron llamadas fonemas. Varios de éstos forman segmentos con significado, que fueron conocidos como morfemas.

Bloomfield (1961/1933: 160), sin hablar aún de morfema, propuso dos tipos de formas lingüísticas clasificadas a partir de su separabilidad y semejanza con otras. Por un lado, están las que pueden aparecer aisladas (*free forms*), y por otro, las que son inseparables y siempre forman parte de un segmento más grande (*bound forms*).

Esta observación permitió plantear en la tradición lingüística estructuralista el concepto de morfema y los procedimientos para identificarlo mediante la comparación de segmentos. Así, se definió a los morfemas como las formas lingüísticas con significado que no pueden dividirse en otras más pequeñas. Por ejemplo, para Bloomfield, el morfema es definido por su falta de semejanza parcial con otras formas lingüísticas: "A linguistic form which bears no partial phonetic-semantic resemblance to any other form, is a *simple* form or *morpheme*" (1961/1933: 161).

Por su parte, Hockett define los morfemas como "elementos mínimos con significado individual de las emisiones de una lengua" (1971/1958: 125) y Nida, en la misma dirección, propone que: "Morphemes are the minimal meaningful units which may constitute words or parts of words" (1949/1946: 1).

Con las definiciones anteriores se puede caracterizar el morfema dentro del estructuralismo como un segmento con las siguientes características: (i) tiene significado, (ii) es indivisible, (iii) es una unidad mínima y (iv) forma palabras.

La gramática generativa trajo nuevas propuestas para la definición de morfema. Como ejemplo, tenemos a Aronoff: "A morpheme is a phonetic string which can be connected to a linguistic entity outside that string. What is important is not its meaning, but its arbitrariness" (1976: 15). Este autor pone el aspecto semántico en lugar secundario y le da énfasis a la estructura formal y

a la distribución. La idea de Aronoff fue hacer más amplia la definición estructuralista de morfema y permitir la entrada en ellas de unidades que funcionan como tales pero sin significado constante.

Para Anderson (1985b: 161) el concepto de morfema tiene problemas en la correspondencia uno a uno entre forma y significado. Propone entonces usar el término: formativo, para caracterizar segmentos mínimos de palabras. Otros nombres usados para sustituir al morfema son: formante, alternante y exponente (cf. Pena 1999: 4313, nota 5).

Al respecto, Moreno de Alba (1986: 26) propone que no conviene negarle valor significativo al morfema; y en el Esbozo (cf. RAE: 1973: 164) se encuentra nuevamente que un morfema es la forma lingüística mínima dotada de significado. En posición un tanto contraria, Pena (1999) comenta que el morfema ha sido definido en términos generales ya sea como "unidad significativa mínima" o como "unidad gramatical mínima" y afirma que para el español la segunda definición es más adecuada. La distinción fundamental entre estas dos unidades es que las primeras tienen significado constante y las segundas pueden no tenerlo, de ahí que basándose en los interfijos y vocales temáticas del español, entre otros fenómenos, este autor proponga la segunda (cf. Pena 1999: 4318-4326).

Por su parte, Lara comenta que "*el morfema no 'tiene significado' por sí mismo, sino en conexión con otro morfema, con el que forma las unidades de denominación a las que significa*"[2] (2006: 62). Para este autor, el morfema es una unidad mínima de valor significativo.

Los morfemas pueden tener una o varias formas alternas distintas, las cuales dependen de determinadas condiciones como los cambios producidos por la combinación con otros. Cuando se trata de una sola se conoce como morfo. En caso de que sean varias representaciones se llaman alomorfos (cf. Hockett 1971/1958: 274 y Pena 1999: 4313). Aronoff explicó estas alteraciones con lo que llamó reglas de reajuste y las clasificó en: eliminación de un morfema final, anterior al sufijo; y ajuste de su forma fonológica cuando está junto a otro (cf. Aronoff 1976).

Los conceptos de morfema y alomorfo se han llevado a límites interesantes, abarcando al menos dos posibilidades: un morfema o alomorfo cero, y morfemas no segmentales. El primero es asociado a palabras que se forman a partir de otras sin ningún cambio en ellas. Por ejemplo, el singular del español ha sido asociado a un morfema cero (Ø); el plural para terminaciones en vocal no acentuada seguida de *s*, por ejemplo *análisis*, ha sido formulado igual (cf. Ambadiang 1993). La segunda posibilidad incluye unidades morfológicas que no son necesariamente segmentos de una

[2] La tipografía cursiva es del original.

palabra, esto sucede en algunas lenguas donde, por ejemplo, los cambios suprasegmentales pueden ser considerados como un morfema.

Se han propuesto otras unidades, adicionales al morfema, como constituyentes de la palabra: tema, raíz, base y afijo. En seguida, basándome en Pena (1999), describo cada una y sus posibles diferencias. El tema es el segmento que queda en una palabra después de la eliminación de las marcas de flexión (afijos flexivos); en otras palabras, la parte inmutable en un paradigma de flexión regular (*blancuzc-o, blancuzc-o-s blancuzc-a* y *blancuzc-a-s*). La raíz es lo que queda después de quitar las marcas de flexión y de derivación (*blanc-* en *blanc-uzc-o-s*), es decir, los afijos. Estos pueden ser: derivativos (*-uzc*), si forman parte del tema (*blancuzc-*), o flexivos (*-o* y *-s*), también llamados desinencias, si se unen a él (*blancuzc-o-s*).

Así, los afijos derivativos generan nuevos temas mientras que los flexivos no lo hacen. Puede ser que en una palabra el tema coincida con la raíz (*blanc-* en *blanc-o*) o que esté formado por una raíz y afijos (*blancuzc-* en *blanc-uzc-o-s*). También, puede suceder que un tema coincida con una palabra (*casa-* en *casa*).

Por un lado, la base es la parte de la palabra sobre la que operan los procesos morfológicos de flexión, derivación y composición. Su pertinencia es de tipo más bien metodológico ya que algunas veces ni raíz ni tema coinciden con la base de un proceso de formación de palabras. En opinión de Pena, en *inconfesable* la raíz sería *confes-* y el tema *inconfesable*, mientras que *confesa-* sería la base de sufijación con *–ble* y *confesable* la base de prefijación con *in-* (cf. Pena 1999: 4312-4318).

Los morfemas son entonces unidades que constituyen palabras y se vuelven importantes en nuestra investigación porque permiten ver a éstas como segmentables. Así, surgieron propuestas de procedimientos para determinar segmentos asociados a cambios de significado. Éstos, basados principalmente en la comparación de formas, han dado paso a métodos computacionales como el que usaré en este trabajo. Por lo anterior, en el siguiente apartado exploraré algunos de los primeros acercamientos a la determinación de morfemas.

1.2. Determinación de morfemas

Revisaré en esta sección propuestas que coinciden en el uso primordial de la comparación de segmentos para la determinación de morfemas. Éstas forman parte del estructuralismo y vienen bien a la tesis ya que los descubren a partir de corpus y lo hacen con el uso de criterios distribucionales.

La primera es la de Hockett (1971/1958: 125-130). Su procedimiento está basado en la comparación formal y semántica de segmentos de formas lingüísticas tantas veces como sea

necesario. A partir de porciones de la emisión del hablante se verifica si cada una aparece en otra emisión con significado aproximado. De ser cierto lo anterior, sería una forma gramatical, pero no necesariamente un morfema. Para serlo, es necesario corroborar que el segmento no esté formado por otras formas gramaticales. Entonces, será necesario dividirlo en partes más pequeñas y buscar si coinciden con otras emisiones del hablante con significado aproximado. Si el segmento no es divisible, se tratará de un morfema.

La segunda es la de Nida (1949/1946: 7-58) y está expresada en seis principios. Se puede decir que su propuesta abarca la de Hockett, pero va más allá de la simple comparación de segmentos. Como el mismo autor lo expresa, el procedimiento consiste en descubrir los patrones recurrentes de los enunciados y clasificarlos. Las partes mínimas con significados recurrentes en los enunciados son morfemas. Un aspecto que debo resaltar de la propuesta de Nida, además de ser muy completa, es que está basada en trabajo con múltiples lenguas, de las cuales buena parte son de América.

El primer principio descubre, a la manera de Hockett, morfemas basados en la mera comparación formal de segmentos de palabras. El segundo comprende diferencias definibles fonológicamente de un mismo morfema, es decir, alomorfos. Si la diferencia entre formas no puede ser explicada de manera fonológica, el tercer principio permite agrupar estas formas gracias a sus relaciones semánticas y a pesar de sus diferencias en forma.

El cuarto principio determina morfemas que no pueden ser descubiertos mediante la comparación de segmentos, sino a través de cambios en las palabras o en el orden de fonemas, por ejemplo, la sustitución de vocales en el inglés (*foot* > *feet*). El quinto principio clasifica formas homófonas como morfemas distintos cuando tienen distinto significado. El sexto y último principio ayuda a determinar unidades morfológicas que nunca ocurren de manera aislada, siempre y cuando al menos uno de los elementos con los que ocurren pueda aparecer aislado (*fraga-* en *fragante*, *fragancia*; *-sipid-* en *insípido* (cf. Pena 1999: 4318)).

Estas propuestas que parten de la comparación formal entre palabras han sido la base de muchos métodos computacionales de análisis morfológico. Sin embargo, como se verá en el capítulo dedicado al descubrimiento de afijos, no basta la mera comparación formal para determinar unidades morfológicas, razón por la cual se han elaborado métodos como el que usaré aquí desarrollado por Medina (2000, 2003).

Una vez presentados los aspectos anteriores sobre el morfema: su definición y algunos procedimientos para aislarlos en corpus, es posible revisar la definición de morfología.

1.3. Definición de morfología

En esta sección, expongo varias definiciones de morfología establecidas por algunos representantes de la lingüística estructural y distribucional. Me apego a este tipo de posturas teóricas porque brindan mejores bases para el trabajo computacional que desarrollaré, esencialmente porque trabajan con corpus y le dan preferencia al nivel formal de las palabras[3]. Agrego también las definiciones de la Academia Española para conocer su postura en el español.

Para el estructuralismo lingüístico, el concepto de morfología tenía como base principal el morfema, por ejemplo, para Hockett, "la morfología comprende el repertorio de morfemas segmentales y las maneras en que se forman las palabras a partir de ellos" (1971/1958: 178). Nida tiene una definición parecida: "Morphology is the study of morphemes and their arrangements in forming words" (1949/1946: 1).

Con la llegada de la gramática generativa transformacional, la mirada hacia la morfología cambió. Las primeras posturas de esta corriente negaban la independencia de la morfología como un nivel de lengua, para ellas, la gramática consistía sólo de sintaxis y fonología. Al parecer, el surgimiento de la morfología como estudio independiente en el modelo generativo transformacional se dio con el artículo *Remarks on Nominalization* de Chomsky[4]. Un ejemplo de la nueva visión generativa fue el de Aronoff (1976: 1), quien propone que la morfología debe atender el análisis de palabras existentes y la formación de nuevas palabras.

Según el Esbozo, "al estudio de los morfemas trabados, sus clases y su organización en el cuerpo de las palabras atiende en lo esencial la morfología" (RAE: 1973: 165). Además, la última gramática descriptiva de la Real Academia Española establece con mayor detalle sus objetivos: "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" (Pena 1999: 4307).

Los fenómenos de formación de palabras pueden llegar a ser muy complejos. Lenguas semíticas, por mencionar un conocido ejemplo, forman palabras a partir de grupos consonánticos a

[3] No debe entenderse que sean las únicas posturas teóricas que puedan dar paso a metodologías de tipo computacional. De hecho, suele ser más común asociar el trabajo en lingüística computacional con el generativismo transformacional que con otras escuelas.

[4] El artículo se puede encontrar en Noam Chomsky. 1970. *Remarks on Nominalization*, en Roderick A. Jacobs y Peter S. Rosenbaum (eds.), Waltham, Mass: Ginn.

los que se insertan vocales (**ktb-* 'escribir' > *ktaab* 'libro'). Además, muchas veces la explicación de un fenómeno morfológico obliga a entrar en el terreno de otros niveles de lengua, por ejemplo, los de tipo morfosintáctico.

Concluyo que la morfología estudia la estructura de las palabras y los procesos involucrados en su formación, los cuales serán revisados en apartados especiales más adelante. Por el momento, incluyo una pequeña referencia a la clasificación de lenguas según la estructura de sus palabras.

1.4. Tipos de lenguas según su morfología

Para que un método computacional sea aplicado a distintas lenguas, como sería mi intención con el que presento en esta tesis, es necesario tomar en cuenta aspectos tipológicos. Por tal razón, expongo en este apartado, y de manera corta, dos clasificaciones de lenguas por su complejidad morfológica. Los tipos de clasificación se han basado generalmente en dos criterios. El primero es la complejidad en términos del número de elementos que forman la palabra. El segundo, la transparencia de los límites entre esos elementos.

Una de las primeras clasificaciones fue dada, como anota Hockett (1971/1958: 182), por los lingüistas del siglo XIX, quienes trataron de clasificar las lenguas en: analíticas, sintéticas y polisintéticas. Creyeron que las palabras del chino constaban de un solo morfema y clasificaron esa lengua como analítica. Otras con varios elementos por palabra, como el griego, latín y español, fueron llamadas sintéticas. Cuando conocieron lenguas con un número aún mayor de morfemas que las anteriores, inventaron el término polisintético.

Otra clasificación, basada en los límites entre morfemas, divide las lenguas en cuatro tipos: aislantes, aglutinantes, polisintéticas y flexivas. Algunos de estos términos coinciden con la primera clasificación. Así, las aislantes se refieren a lenguas donde las palabras se forman con un solo morfema, lo que equivale a una lengua analítica (el chino). Las aglutinantes tienen múltiples morfemas al igual que las flexivas, pero con un morfema por cada significado (como el turco). En las flexivas, por otra parte, la correspondencia es entre un morfema y varios significados (como el latín). Por último, las polisintéticas son similares a las aglutinantes, pero algunos segmentos pueden tener el mismo significado reservado para palabras independientes (cf. Anderson 1985a: 8-10, Oflazer 1999: 177).

Para Bloomfield (1961/1933: 207-208) las dos clasificaciones no resultan claras. La primera es muy relativa y la segunda carece de criterios rígidos de diferenciación entre los tres últimos tipos. En lo personal, creo que el problema de ambas clasificaciones puede resumirse en que terminan siendo rebasadas por la realidad lingüística, es decir, las lenguas no se ajustan a un solo tipo de

morfología sino que echan mano de varios. En este sentido, la propuesta de Hockett resulta pertinente: "Una escala continua es más útil. Podemos usar los términos analítico y sintético en forma relativa. El inglés es más sintético que el chino, pero más analítico que el fox" (1971/1958: 182)[5].

Si bien esta tesis no atiende los problemas de comparación de lenguas, sí resulta interesante anotar que los métodos cuantitativos de análisis morfológico tienen la posibilidad de brindar medidas de comparación de complejidad morfológica; complejidad que no necesariamente debe medirse en número de morfemas. Por ejemplo, el método de descubrimiento de afijos que utilizo en esta investigación brinda escalas de afijalidad comparables con las de otras lenguas o con la misma lengua en un eje temporal, espacial o social distinto.

He mostrado que la clasificación de lenguas no es un problema acabado, lo que refleja la variedad y complejidad de los fenómenos morfológicos. Ahora, es necesario revisar a detalle en qué consisten estos y cómo se manifiestan en distintas lenguas. Por tanto, enseguida expongo sus posibilidades de realización.

1.5. Realización de fenómenos morfológicos

La morfología se realiza de diversas maneras según la lengua que se estudie. Al respecto, es posible dividir estos fenómenos de realización morfológica en concatenativos, cuando hay adherencia de segmentos, y no concatenativos, si la modificación de las unidades lingüísticas es fonológica o prosódica.

Las dos modificaciones de carácter concatenativo que sufren las palabras son básicamente la afijación y la reduplicación. Cuando no hay adición de segmentos se pueden encontrar cambios vocálicos, consonánticos, eliminación de fonemas y modificaciones prosódicas. Dos fenómenos adicionales que no involucran concatenación son las formas que sustituyen completamente una palabra, llamadas supletivas; y la ausencia de cambio, situación asociada al llamado morfema cero.

En los apartados siguientes detallaré estas modificaciones. Mi principal interés es, por un lado, mostrar la complejidad morfológica de las lenguas y, por otro, revisar el alcance que tendría el método automático que propongo en la tesis. Si se conocen las realizaciones de los fenómenos morfológicos en las lenguas, se puede saber cuánto de esta complejidad lingüística abarca el método de descubrimiento de afijos que utilicé.

---

[5] Existen otras clasificaciones que toman en cuenta varios criterios adicionales. Una de ellas es la de Sapir, expuesta por Anderson (1985a: 10-11).

Por lo anterior, a continuación expongo los fenómenos de afijación, reduplicación, cambios fonológicos, cambios prosódicos y otras realizaciones.

1.5.1. Afijación

La afijación es una manera común de marcar los fenómenos morfológicos de flexión y derivación mediante afijos. Un afijo es la forma lingüística que no puede aparecer aislada y que se agrega a una base para formar nuevas palabras (afijos derivativos) o para marcar una categoría gramatical (afijos flexivos). Al respecto, como mencioné antes (véase *supra* p. 26), los de flexión se unen a un tema para crear un paradigma flexivo. Los de derivación se unen a lo que se llama raíz y forman parte del tema.

De manera general, se puede llamar base a la parte de la palabra a la que se concatenan los afijos. Así, es posible clasificarlos por su orden de aparición. Si preceden a la base, entonces se habla de prefijación (*in-confesable*), si aparecen después es sufijación (*blanc-o*) y si se insertan al interior de la base se llama infijación. Otro tipo de afijos son los circunfijos de los fenómenos de parasíntesis. Estos son discontinuos y rodean la base, son la combinación de un prefijo y un sufijo dependientes entre sí.

Según Pena (1999), en español hay presencia de infijos en la derivación apreciativa, por ejemplo, con el infijo –it– en: *lej-it-os, azuqu-ít-ar*. También de circunfijos en derivaciones como: *sombra* > *en-sombr-ec-er*, *rojo* > *en-roj-ec-er*.

1.5.2. Reduplicación

El otro fenómeno concatenativo es la reduplicación de palabras o partes de palabras, situación que conlleva un cambio de significado, por ejemplo en (1).

(1) Fox (Bloomfield 1961/1933: 218)

(a) [wa:pamɛ:wa] ‘lo mira’
[wa:-wa:pamɛ:wa] ‘lo examina observándolo’
[wa:pa-wa:pamɛ:wa] ‘lo sigue mirando’

Tanto los fenómenos de derivación como de flexión pueden ser marcados por reduplicación. Por citar algunos casos, el indonesio forma adverbios (2) y plurales (3)(a) por reduplicación completa, esta situación ocurre también en el pápago (3)(b). Otro ejemplo es la verbalización de adjetivos del dakota en (4).

(2) Indonesio (Beard 1998: 63)

(a) kira ‘adivinación’ > kira-kira ‘aproximadamente’

(b) pagi ‘mañana’ > pagi-pagi ‘en la mañana’

(3) Pápago e indonesio (Stump 1998: 32)

(a) Indonesio babi ‘cerdo’ > babibabi ‘cerdos’

(b) Pápago bana ‘coyote’, kuna ‘esposo’ > baabana ‘coyotes’, kuukuna ‘esposos’

(4) Dakota (Beard 1998: 48)

(a) puza ‘seco’> puspuza ‘estar seco’

(b) čʰepa ‘gordo’ > čʰepčʰepa ‘estar gordo’

1.5.3. Cambios fonológicos

Dentro de los fenómenos de realización morfológica no concatenativos, se pueden encontrar casos donde un cambio fonológico en alguna vocal hace el papel de morfema, como en algunos plurales del inglés (5). Además, modificaciones consonánticas como en (6) pueden marcar flexión.

(5) Inglés (Bloomfield 1961/1933: 217)[6]

(a) man [mæn] ‘hombre’ > men [men] ‘hombres’

(b) foot [fut] ‘pie’ > feet [fi:t] ‘pies’

(6) Fula (Stump 1998: 32)

(a) yiite ‘fuego’ > giite ‘fuegos’

Otro tipo de cambio, que involucra fonemas vocálicos, se da en lenguas semíticas. En estas lenguas, los elementos léxicos se componen sólo de consonantes y las vocales son usadas para marcar fenómenos morfológicos. Por ejemplo, las derivaciones de (7).

(7) Árabe (Beard 1998: 62)

(a) *ktb- ‘escribir’

(b) ktaab ‘libro’

(c) kaatəb ‘escritor’

[6] La trascripción fonética es mía.

La eliminación de un segmento puede estar asociada también al cambio de significado. Bloomfield pone el ejemplo del género en francés, el cual formaría el masculino a partir del femenino y mediante una pérdida de la consonante final (8).

(8) Francés (Bloomfield 1961/1933: 217)

(a) laide [lɛd] > laid [lɛ]
femenino > masculino 'feo'

(b) longue [lõg] > long [lõ]
femenino > masculino 'largo'

También se puede encontrar otro caso de eliminación como marca de flexión en el huichol. En esta lengua la forma completiva de los verbos surge de la pérdida de la sílaba final (9).

(9) Huichol (Stump 1998: 32)

(a) pïtiuneika 'él bailaba' > pïtiunei 'el bailó (con aspecto completivo)'

1.5.4. Cambios prosódicos

Un fenómeno morfológico podría estar marcado por el cambio de acento en una palabra o por la modificación en su entonación. En lenguas como el sueco, el cambio de tono sería un morfema, como puede verse en (10).

(10) Sueco (Bloomfield 1961/1933: 221)

(a) [le:s-] 'leer'
[ˇle:ser] 'lector'
['le:ser] 'él lee'

Fenómenos similares suceden en otras lenguas como el somalí (11)(a).

(11) Somalí (Stump 1998: 32)

(a) Somalí èy 'perro' (con tono descendente) > éy 'perros' (con tono alto)

1.5.5. Otras realizaciones

En otras ocasiones, la modificación morfológica que sufre la palabra no sigue un patrón definido, sino que se utiliza una palabra totalmente distinta (forma supletiva) para sustituirla (*good > better)*.

En otros casos, no hay una forma supletiva, afijo, reduplicación o cambio fonológico que marque el cambio de significado. En esta situación se habla de un morfema cero (muchas veces simbolizado con ∅)[7]. Un ejemplo en español podría ser el plural de palabras terminadas en –s del tipo *análisis*, *síntesis* o *lunes*.

Creo que esta gama de fenómenos ha dado muestra de la complejidad que encierra el sistema morfológico. Imaginemos ahora lo que representa tratarlo de manera automática. Por esta razón, en esta tesis sólo trato con la morfología concatenativa afijal. En seguida, describo los fenómenos de formación de palabras, que hacen uso principalmente de los fenómenos concatenativos reportados en esta sección.

1.6. Formación de palabras

En esta sección, abordaré los fenómenos de formación de palabras con el fin de describir el funcionamiento del sistema lingüístico en el nivel morfológico. En primer lugar, definiré mi postura ante la interrogante de cuáles son estos fenómenos y dejaré asentado que existen otros fenómenos que podrían agregarse a los anteriores, pero que no son tomados en cuenta en mi investigación. Una vez hecho esto, expondré en apartados independientes los que sí tomo en cuenta.

Es común asociar el término *formación de palabras* únicamente con los fenómenos de derivación y composición; y algunas veces sólo con el primero. Sin embargo, me interesa incluir en él todos los procesos que permitan generar una nueva palabra mediante la modificación de una ya existente. En este sentido, caben también, además de los fenómenos ya mencionados, la flexión y la incorporación.

La incorporación es generalmente incluida en la composición, pero la separo con el interés de conocerla y revisar su funcionamiento. La flexión y la derivación casi siempre son separadas, aunque las propuestas para considerarlas un solo fenómeno son, como se verá adelante, comunes[8]. En conclusión, la formación de palabras incluirá: flexión, derivación, composición e incorporación. Anderson (1985a: 15) propone una clasificación así, pero dividida en dos grupos: modificación de bases, que incluye flexión y derivación; y composición, en la que incluye la incorporación.

---

[7] Algunas propuesta teóricas de la morfología no aceptan la existencia de este tipo de morfema, véase al respecto Aronoff (1976) o Ambadiang (1993).

[8] En un apartado especial revisaré la pertinencia de dividir flexión y derivación. Por el momento me quedo con la visión más tradicional y las trato separadas.

Existen otros fenómenos que caben en la definición de formación de palabras. Es pertinente decir que no los abordo en la tesis principalmente por dos razones. Por un lado, algunos son escasos y poco productivos, al menos en español. Por otro, a diferencia de la flexión, derivación, composición e incorporación, tienden a ser más voluntarios que gramaticales (el hablante es más consciente del cambio que está haciendo) y muchas veces irregulares. Por ejemplo, los fenómenos de entrecruzamiento o mezcla (*cantante* + *autor* > *cantautor*), acortamiento (*profesor* > *profe*) y formación de acrónimos, entre otros[9].

Desde el aspecto computacional, el método de descubrimiento de afijos que utilizaré está basado en medidas obtenidas a partir de la comparación de porciones de palabras. Por tal razón, el resultado es una lista de segmentos concatenativos tanto flexivos como derivativos. Hay evidencia de que, si el método se aplica a lenguas con fenómenos recurrentes de composición o de incorporación, se producen como salida segmentos propios de estos fenómenos y no sólo cadenas afijales; por ejemplo, en la lengua checa muchos sustantivos son composiciones de adjetivos y nombres, por lo que los primeros se descubren como prefijos en dicho método (cf. Medina y Hlaváčová 2005).

En los siguientes apartados, expongo cada uno de los tipos de formación de palabras que permitirían una posible segmentación automática: flexión, derivación, composición e incorporación.

1.6.1. Flexión

En esta sección, revisaré algunas definiciones del término flexión, las funciones que realiza a nivel sintáctico y las categorías gramaticales que codifica. Para esta tesis, es muy importante estudiar la flexión porque es una manera común de codificar información categorial y gramatical, sobre todo mediante afijos. En este sentido, el método propuesto descansa en la idea de que descubrir estos afijos flexivos nos llevará a mejor la manera de identificar las categorías gramaticales de un corpus.

1.6.1.1.Definición

Es común definir el concepto de flexión a partir del cambio que sufre una unidad cuando entra en relación sintáctica con otras. Como dice Anderson: "inflection on the other hand serves to 'complete' a word by marking its relations within larger structures" (1985b: 162).

Para Hockett (1971/1958: 213), la parte de la palabra que sufre el fenómeno flexivo es el tema. Sin embargo, para Stump (1998: 13) la flexión es definida a partir de un lexema, una unidad

[9] Para más detalle sobre estos fenómenos véase Beard (1998), Aronoff (1976: 20) y Pena (1999: 4332, nota 16).

lingüística que pertenece a una categoría sintáctica, tiene un significado particular o función gramatical, y puede entrar en combinaciones sintácticas como palabra sola.

A grandes rasgos, la flexión se caracteriza por crear paradigmas muy regulares, que mantienen sin modificación parte de la palabra. Pero los paradigmas presentan al menos dos particularidades. Una es la aparición de huecos o palabras faltantes, por lo que se habla de paradigmas defectivos. Otra es cuando aparecen palabras irregulares al interior de los paradigmas, en otras palabras, es posible encontrar más de una raíz en un paradigma. Éstas son llamadas formas supletivas, como las del paradigma del verbo *ir* en español: *vamos*, *iremos, fuimos, íbamos*[10].

La estructura de un paradigma depende de las propiedades morfosintácticas de la lengua. Puede haber paradigmas verbales o nominales y dentro de cada uno darse las combinaciones de distintas categorías flexivas como número, género, tiempo, modo, aspecto, etc. Éstas serán revisadas en un apartado posterior.

En algunas lenguas, los paradigmas son representados por alguno de sus integrantes, el cual participa en los fenómenos de composición o derivación. Por ejemplo, en inglés, sería el infinitivo: *play* > *playground, player*. En otras, estarían representados por una parte de la palabra, llamada base y que no necesariamente debe coincidir con la raíz (12).

(12) Alemán (Bloomfield 1961/1933: 225)

(a) lachen ['lax-en] ‘reír’

(b) (ich) lache ['lax-e] ‘yo río’

(c) (er) lacht ['lax-t] ‘el ríe’

(d) (er) lachte ['lax-te] ‘el rió’

(e) Lacher ['lax-er] ‘humor’

En conclusión, la flexión encierra mecanismos usados para deducir palabras a partir de ciertas bases formando un paradigma generalmente regular. La derivación y composición, por el contrario, involucran mecanismos para deducir nuevas bases. Pongo enseguida las funciones que cumple la flexión.

[10] Esto representa un problema para el proceso conocido como lematización automática, que consiste en la asignación de un lema o palabra de diccionario a cada palabra de un corpus. En el trabajo que presento, no me involucro con este problema lingüístico.

1.6.1.2.Funciones

Una vez definida la flexión, voy a describir una de sus características más importantes: la función que cumple a nivel sintáctico. Al respecto, no es difícil encontrar que ésta deja de ser considerada como parte de la morfología y se pasa al terreno de la sintaxis; piénsese, por ejemplo, en las gramáticas generativas transformacionales. Para mí, es importante revisar estas funciones ya que son necesarias para la identificación de categorías y por lo general son contempladas en alguna parte de los métodos automáticos.

La flexión codifica relaciones entre miembros de una construcción sintáctica; una de estas relaciones es la concordancia. Las lenguas muestran amplia variedad de relaciones de concordancia morfológica. Entre las principales están: concordancia del modificador o especificador con su elemento principal, concordancia del verbo con alguno de sus argumentos, y concordancia de la expresión anafórica con su antecedente.

En español, sustantivo y adjetivo concuerdan en género y número. Esta relación es muy estricta y no depende de la presencia de afijos de flexión en el sustantivo. Tómense por ejemplo (13)(a), en donde el adjetivo tiene flexión de género de acuerdo al sustantivo femenino *mano*, a pesar de que éste no tenga la terminación *–a*. Para el caso del número sucede algo parecido, en (13)(b) se puede ver la flexión de número del adjetivo sin cambio en el sustantivo. Relaciones similares se presentan entre los nombres y los determinantes.

(13) Español
- (a) buena mano
- (b) análisis correcto vs. análisis correctos

En el caso de la identificación automática de categorías, retomando los ejemplos de (13), debe notarse la complicación para determinar solamente con aspectos morfológicos el género de *mano* y el número de *análisis*. Por esta razón, los métodos de identificación cuentan con módulos que toman en cuenta aspectos de la sintaxis. Así, la única manera para que, en (13)(a), *mano* quedará etiquetada como nombre femenino sería porque *buena* tendría la etiqueta adjetivo femenino. De hecho, el método seleccionado para esta tesis funciona así.

Es posible encontrar otro tipo de relación de concordancia en la que un miembro impone restricciones específicas sobre las características morfosintácticas de otro, aunque no necesariamente compartan esta propiedad. Esta relación es conocida como régimen o rección. Un ejemplo de este tipo se da en verbos o preposiciones que rigen el caso de su objeto nominal, como en alemán, donde

helfen ‘ayudar’ y *mit* ‘con’ gobiernan caso dativo, mientras que *sehen* ‘ver’ y *ohne* ‘sin’ gobiernan acusativo (cf. Stump 1998: 24).

A continuación revisaré las categorías gramaticales que codifica la flexión, llamadas por ello: categorías flexivas. Estas características morfosintácticas suelen ser codificadas mediante afijos flexivos, aunque no siempre.

1.6.1.3.Categorías flexivas

En esta parte, indago sobre las propiedades morfosintácticas que expresa un sistema flexivo. Hockett las llamó *categorías gramaticales*, aunque también son referidas como *categorías flexivas* (cf. Stump 1998). Decido usar este último para hacer hincapié en que son expresadas por el mecanismo de la flexión.

Para mi trabajo, resulta de suma importancia conocer las categorías gramaticales que son expresadas por medio de la flexión ya que mi objetivo es identificarlas automáticamente. Así, cuando ésta sea realizada mediante concatenación de afijos, tendré la posibilidad de identificar las categorías con el método propuesto. Además, podré evaluar las reglas de etiquetado producidas por éste.

Otro aspecto que reafirmaré con esta revisión será el conjunto de etiquetas de categorías y su especificidad. Comúnmente, los métodos de etiquetado sólo toman en cuenta las categorías generales sin rasgos morfosintácticos, es decir, sólo etiquetan sustantivos, pero no sustantivo masculino singular; sin embargo, quiero dar cuenta de características gramaticales más particulares.

Como ya lo he comentado, no tomo una perspectiva totalmente hispanista, por lo tanto, revisaré las categorías flexivas desde una mirada general para conocer sus posibilidades en varias lenguas. Primero me ocuparé del nombre, luego del verbo y al final del adjetivo. En Anderson (1985b) se puede encontrar una discusión desde la perspectiva tipológica de estas categorías.

1.6.1.3.1. Categorías flexivas del nombre

En este apartado, mediante pequeñas secciones, explicaré las categorías asociadas al nombre. Generalmente, éstas comprenden el género, número, definición, caso y posesión. A continuación abordaré cada una de ellas.

Género

En muchas lenguas, el género es una categoría flexiva asociada a los sustantivos, que no necesariamente está relacionada con la oposición femenino-masculino. Según Anderson (1985b: 175-176), los sistemas de género están basados principalmente en sexo (masculino, femenino y

neutro), carácter animado (humano, no humano) y otros como los géneros comestible y bebible de las lenguas fiyi. Algunas lenguas suelen expresar el género sólo a través de la flexión de sus palabras concordantes, como en español. Por ejemplo en la construcción: *la sal*, la marca de género está en el determinante y no en el sustantivo. Otras lenguas como el kikuyu marcan simultáneamente el género y número (cf. Stump 1998: 26).

Según Hockett (1971/1958) para que se pueda considerar un sistema de género, cada sustantivo debe pertenecer a una de las clases y muy pocos a más de una. Por lo general, el sistema presenta algún elemento de consistencia semántica, ya sea el sexo, naturaleza animada, tamaño, forma, grado de abstracción o cualquier otro concepto por el estilo; sin embargo, algunas asignaciones son siempre arbitrarias.

Los géneros masculino y femenino del español, francés, italiano y portugués presentan cierta consistencia en cuanto a su asignación. Muchos sustantivos claramente masculinos se refieren a individuos de sexo masculino y lo mismo para los femeninos, pero el resto son arbitrarios.

Si se toman en cuenta otras lenguas como el alemán, latín, griego o sánscrito se encuentran tres géneros: masculino (14)(a), femenino (14)(b) y neutro (14)(c).

(14) Alemán (Hockett 1971/1958: 236)

(a) der Mann ‘el hombre’, der Tisch ‘la mesa’

(b) die Frau ‘la señora’

(c) das Weib ‘la mujer’, das Kind ‘el niño’, das Blut ‘la sangre’

Es posible también que los sustantivos de una determinada declinación pertenezcan a un determinado género. Como se logra ver, las lenguas varían en el número de géneros que codifican, se pueden encontrar desde dos (francés), tres (latín), cuatro (lengua dyirbal), hasta veinte o más (lengua fula) (cf. Anderson 1985b: 175).

Número

El número es otra categoría asociada al sustantivo y es tal vez su categoría inherente en la mayoría de las lenguas del mundo (cf. Anderson 1985b: 174). Una gran cantidad de ellas sólo marcan dos tipos: singular y plural. Otras incluyen más, como dual y trial, por ejemplo, en sánscrito hay singular (15)(a), dual (15)(b) y plural (15)(c).

(15) Sánscrito (Stump 1998:27)

(a) aśvas ‘caballo’

(b) aśvāu ’dos caballos’

(c) aśvās ‘más de dos caballos’

Aunque en español el número parece tener mucha consistencia semántica, existen sustantivos con asignaciones arbitrarias que no siguen una regularidad morfológica y que expresan el número a través de adjetivos o determinantes: *los análisis* frente a *el análisis*, *las caries* frente a *la caries*.

Definición o especificidad

Otra categoría flexiva de los sustantivos es lo que llamaré definición (*definiteness*). Esta característica marca el nombre si el referente está definido o es identificable en el contexto. Por ejemplo, en árabe de Siria la frase nominal puede llevar el prefijo de definición (l-) si la entidad es identificable de manera única, como en (16)(a), de lo contrario no aparece como en (16)(b).

(16) Árabe de Siria (Stump 1998:27)

(a) l-madīne l-ʾkbīre ‘la gran ciudad’

(b) madīne kbīre ‘una gran ciudad’

Caso

El caso es otra categoría flexiva común. Los casos son formas flexionadas de los nombres que los adecuan para participar en construcciones con los verbos. Pueden ser relaciones directas (nominativo, acusativo, ergativo, absolutivo, dativo y genitivo) y oblicuas (instrumental, ablativo y locativo). El número de casos de un sistema puede ir desde dos hasta veinte o treinta (cf. Hockett 1971/1958: 238).

Posesión

El nombre también puede tener flexión para marcar la relación con su poseedor, en uyghur (uigur) un sustantivo concuerda en persona con su frase sustantiva poseedora (17). En menomini se observa flexión de sustantivos para indicar persona y número del posesor.

(17) Uigur (Stump 1998:27)

(a) Nuriyi-niŋ yoldiš-i

Nuriyä-GEN esposo-3ª.POSESIÓN ‘El esposo de Nuriyä’

Hasta aquí, he revisado las categorías flexivas del nombre y parece ser amplia la variedad de rasgos marcados en ellas. Me llama la atención las diferencias entre lenguas porque dan muestra de que el lenguaje es también una expresión de la cultura y la concepción de mundo. Enseguida describo las categorías verbales.

1.6.1.3.2. Categorías flexivas del verbo

Las categorías flexivas del verbo son tiempo, aspecto, polaridad, voz y, en algunos casos, modo. Al igual que para los sustantivos, es importante conocer cómo se manifiestan éstas. Los siguientes apartados las describen brevemente.

Tiempo

El tiempo indica diferentes localizaciones de un acontecimiento con respecto a la enunciación. Los verbos del español, por ejemplo, expresan básicamente tres tiempos: presente, pasado y futuro. En lenguas germánicas y eslavas los verbos sólo se flexionan para el contraste entre presente y pasado (cf. Hockett 1971/1958: 240). El kikuyu, según reporta Stump, tiene seis tiempos: pasado lejano, pasado cercano, pasado reciente, presente, futuro cercano y futuro lejano (cf. Stump 1998: 28).

Aspecto

El aspecto es una categoría que define los diferentes sentidos en los que el evento puede ser situado en un intervalo de tiempo particular. El español sólo tiene dos aspectos flexivos: *Juan cantó* (perfectivo) y *Juan cantaba* (imperfectivo).

Por otra parte, y siguiendo con el ejemplo del kikuyu, esta lengua muestra seis aspectos para el presente: aspecto continuo, el evento está en progreso a través del intervalo (18)(a); aspecto habitual, el evento es rutinario en el intervalo (18)(b); aspecto proyectado, indica la intención de realizar el evento en el intervalo de tiempo (18)(c); aspecto completivo, indica que el evento ha tenido lugar en el intervalo (18)(d); aspecto iniciativo, indica que el estado resultante de llevar a cabo el evento sigue realizándose en el intervalo (18)(e); y de experiencia (*experiential*), que indica que el evento ha sucedido en algún punto indefinido anterior al intervalo (18)(f).

(18) Kikuyu (Stump 1998: 28)

(a) tũraagũra nyama ‘estamos comprando carne’

(b) tũgũraga nyama ‘compramos carne’

(c) tũũkũgũra nyama ‘compraremos carne’

(d) twagũra nyama ‘hemos comprado carne’

(e) tũgũrĩĩte nyama ‘(casi) hemos comprado carne’

(f) twanagũra nyama ‘hemos comprado (alguna vez) carne’

Polaridad

La polaridad distingue sentencias afirmativas de negativas. En kikuyu las sentencias negativas marcan el verbo con el prefijo *ti-* como puede verse en (19).

(19) Kikuyu (Stump 1998: 29)

(a) tũ-kaagwata ‘(lo) atraparemos’

(b) tũ-ti-kaagwata ‘no (lo) atraparemos’

Voz

La voz indica relaciones temáticas que pueden existir entre el verbo y su sujeto. Los verbos latinos tienen dos voces: activa y pasiva. El griego y el sánscrito tienen tres: activa, pasiva y media. Por ejemplo, en sánscrito, el verbo aparece en voz activa si el sujeto es agente y no beneficiario de la acción (20)(a), en voz media si el sujeto es agente y beneficiario (20)(b), y en voz pasiva si el sujeto es tema (20)(c). En español la voz no es una categoría flexiva sino sintáctica.

(20) Sánscrito (Stump 1998: 29)

(a) odanam āpnoti ‘obtiene avena (para alguien más)’

(b) odanam āpnute ‘obtiene avena (para sí mismo)

(c) odana āpyate ‘(la) avena es obtenida’

Modo

Las maneras como una oración se relaciona con la realidad del hablante se pueden codificar con flexión de modo. Por ejemplo, en sánscrito clásico, el modo indicativo se usa para afirmar hechos verdaderos, el modo optativo para expresar proposiciones cuya realidad es deseada, y el modo imperativo como orden (cf. Stump 1998: 29). En español se encuentra algo parecido ya que cuenta con tres modos: indicativo, subjuntivo (proposiciones hipotéticas) e imperativo.

Persona

Esta categoría clasifica entidades con relación al hablante (primera persona) y al oyente (segunda persona), las demás son tercera persona.

En conclusión, en esta sección he examinado las categorías flexivas típicas del verbo, entre otras cosas, para conocer la especificidad de las etiquetas que usaré para el corpus. Toca el turno a las del adjetivo.

1.6.1.3.3. Categorías flexivas de adjetivos

En comparación con el nombre y el verbo, el adjetivo tiene muy pocas categorías flexivas (cf. Anderson 1985b: 198). El grado es una de ellas y permite distinguir la extensión con la cual un referente evidencia alguna característica. Algunas lenguas como el inglés tienen grado positivo *tall*, comparativo *taller* y superlativo *tallest*. Otras lenguas como el irlandés también tienen flexión para la igualdad (cf. Anderson 1985b: 199).

En otros casos, un adjetivo puede tener flexión distinta si es atributivo o predicativo. En ruso, por ejemplo, el femenino nominativo singular de 'nuevo' es *nóvaja* en uso atributivo (*nóvaja kníga* 'nuevo libro') y *nová* en uso predicativo (*kníga nová* 'el libro es nuevo') (cf. Stump 1998: 31). Muchas veces, en relación de concordancia, los adjetivos tienen la misma flexión que su nombre, éste es el caso del español.

He revisado hasta el momento la definición de flexión, sus funciones en la sintaxis y las categorías gramaticales asociadas a nombres, verbos y adjetivos. En el siguiente apartado presentaré los aspectos relacionados a la derivación.

1.6.2. Derivación

En la sección anterior revisé diversos aspectos de la flexión morfológica con el fin de conocer su pertinencia en la identificación de categorías gramaticales. Ahora, toca el turno a la derivación, mecanismo muy parecido a la flexión y generalmente asociado al cambio de clase de palabra.

Esencialmente, expondré su definición y presentaré una clasificación de ella. Además, abordaré brevemente una cuestión interesante respecto a las palabras derivadas: cuál de sus partes, base o afijos, determina su categoría. También hablaré de que la correspondencia entre el significado y los afijos derivativos no siempre es de uno a uno.

Tanto la derivación como la flexión son aspectos fundamentales en nuestro trabajo. Una de las razones es que se realizan típicamente mediante concatenación de material fonológico, o letras en la escritura, principalmente por afijación. Como ya lo he mencionado, el método automático para descubrir unidades morfológicas permitirá extraer una lista de afijos derivativos y flexivos, los cuales serán usados para identificar las categorías.

1.6.2.1.Definición

La derivación, de manera tradicional, ha sido definida como el proceso que genera una nueva palabra con un nuevo significado a partir de una ya existente. Esta definición tan amplia permitiría incluir la composición y la incorporación, pero la diferencia con éstas radica en la manera de crear la nueva palabra: estos fenómenos lo hacen con la concatenación de otra y no con la adición de una marca derivativa (esto es, mediante afijación, reduplicación, etc.). Además, la derivación ha sido asociada al cambio de categoría gramatical a diferencia de los otros fenómenos.

La derivación también ha sido vista como el estudio de los segmentos (también llamados temas) que quedan después de eliminar los morfemas flexivos (cf. Hockett 1971/1958). En este sentido, Hockett deja ver que sí hay distinción entre flexión y derivación. En relación a esto, es común que la primera sea definida a partir de sus diferencias con la segunda. Por tal situación, en una sección aparte presentaré criterios de diferenciación. Por ahora, expondré una clasificación de este fenómeno.

1.6.2.2.Clasificación

Con el objeto de comprender más el fenómeno de derivación, voy a revisar aquí una propuesta de clasificación de este mecanismo de formación de palabras. Como ya lo mencionaba, por lo general se habla de derivación cuando existe cambio de categoría de la palabra: "*Derivational morphology* produces a new word usually of a different part-of-speech category by combining morphemes" (cf. Oflazer 1999:177). Aunque parece que existen algunos otros tipos de derivación (cf. Beard 1998: 57), presento los dos principales: transposición y derivación expresiva (*expressive derivation*).

La transposición es un tipo de derivación en donde opera un cambio de categoría. Simplemente se cambia una palabra de una categoría a otra algunas veces marcando el proceso con afijación y en otras ocasiones no. Algunos ejemplos del español se pueden ver en (21).

(21) Español

(a) barba > barb-udo
N > A

(b) almacenar > almacen-aje
V > N

La derivación expresiva no cambia la categoría gramatical de la palabra. El cambio en la palabra está asociado a una percepción subjetiva del hablante. Un buen ejemplo son los grados que hablantes rusos utilizan para describir la lluvia y que son marcados con afijos, véase los ejemplos de (22).

(22) Ruso (Beard 1998: 46)

(a) dožd' 'lluvia'
(b) dožd-ik 'lluvia ligera'
(c) dožd-ič-ek 'lluvia muy ligera'

En español también se encuentra este tipo de derivación: *pintar* > *pintarrajear*, *licenciado* > *licenciadillo*, *licenciadito*. La última gramática de la Academia se refiere a ella como derivación apreciativa y es asociada con tres grupos de afijos: diminutivos, aumentativos y peyorativos (cf. Lázaro 1999).

La derivación es un proceso muy importante en esta investigación porque permite la identificación de la categoría de una palabra derivada. Cuando ésta se haga mediante afijación, el método de identificación automática debe ser capaz de representar en una regla la asociación del afijo con su categoría. Así, una vez aplicados los métodos de Medina y Brill, espero captar en varias reglas información gramatical relacionada con la identificación de categorías.

Una vez revisados los tipos de derivación, enseguida expongo algunas propuestas que tratan de explicar qué parte de un derivado (base o afijo) determina su categoría gramatical.

1.6.2.3. Elemento que determina la categoría de un derivado

En este pequeño apartado, quiero hacer referencia al cuestionamiento de conocer qué parte del derivado determina sus características léxicas y por tanto su categoría gramatical. Sobre esto es posible encontrar al menos dos posturas, la que opta porque la base determina la categoría y la que dice que es el afijo derivativo.

Si los afijos son elementos léxicos regulares, entonces podrían determinar la categoría. Por el contrario, si son el resultado de procesos, entonces no podrían hacerlo, en cuyo caso lo haría las bases. Por ejemplo, Dixon (2000/1982) explica que las palabras pertenecen primero a un tipo

semántico universal y después a una categoría gramatical, por tanto, pueden estar asociadas a otra mediante una marca morfológica o sintáctica.

Una idea común es que el afijo más externo de una palabra está asociado con su categoría. En muchas lenguas esto significa el elemento más a la derecha. Si esta postura es válida, se esperaría que el método automático de identificación de categorías funcione bastante bien ya que se basará en los sufijos finales de las palabras. Por eso, si se descubriera automáticamente la morfología derivativa de una lengua, entonces se esperarían mejores resultados en los sistemas de identificación de categorías. Sin embargo, no es claro qué parte del derivado determina la categoría.

En la siguiente sección hago una exposición de los problemas para determinar afijos derivativos, y que bien puede extenderse a los flexivos. Sobre este problema son varios los aspectos a considerar y tienen mucho que ver con la segmentación morfológica, la asignación de una categoría gramatical y por tanto con la tesis.

1.6.2.4. Problemas para determinar afijos

Hasta el momento no he reparado en los problemas semánticos y estructurales que tiene la identificación de afijos flexivos y derivativos. No quiero pasar por alto una reflexión al respecto ya que impactan en los métodos automáticos de segmentación morfológica. Por tanto, en esta sección expondré algunos de los problemas más conocidos.

Una de las dificultades a las que me enfrentaré, en la identificación de categorías gramaticales mediante afijos, es el hecho de que los significados de los derivados y la afijación con que se marcan no son siempre de uno a uno. Se puede tomar como referencia los afijos que marcan genitivo en ruso: *-i*, *-a*, *-u*, ya que cada uno tiene además múltiples funciones. La terminación *–a* marca también nominativo femenino singular y neutro plural; y la terminación *–i* marca femenino y masculino nominativo plural, así como genitivo, dativo y locativo singular en la declinación III. En otras palabras, es claro que los afijos son cofuncionales y multifuncionales (cf. Beard 1998: 54)[11].

En español hay otros problemas como la presencia de afijos que realizan derivación nominal y derivación adjetival como *–dor, –ero, –ario*. Además suele ser difícil determinar si dos sufijos son diferentes o alomorfos de uno mismo. Al respecto se pueden tomar en cuenta dos posibilidades:

[11] A raíz de este tipo de observaciones, diversos autores propusieron lo que se llamó la *Separation Hypothesis*, que propone como independientes las operaciones funcionales y fonológicas de la derivación. Esta hipótesis también formula un conjunto de operaciones léxicas abstractas separadas de las operaciones en las representaciones semántica y fonológica. Algoritmos de un componente morfológico autónomo modificarían la representación fonológica de las base derivadas (cf. Beard 1998 y Spencer 1991: 428-434).

diferenciar por mera distinción de su forma o identificar como alomorfos los que se comportan con distribución complementaria y parecido formal y semántico (cf. Santiago y Bustos 1999: 4507).

Es bien sabido que una de las complicaciones más grandes para un método computacional que intente procesar el lenguaje humano es la ambigüedad. Por esto, es importante reconocer que los métodos usados en esta tesis (el método de identificación de categorías y el de descubrimiento de afijos) tendrán problemas para manejar los aspectos mencionados arriba. Pero al mismo tiempo, es necesario decir que ambos métodos han reportado muy buenos resultados en su aplicación. Con respecto al método de identificación de categorías, es posible que las reglas contextuales, de carácter sintáctico, logren resolver las complicaciones, pero es de esperarse que no en todos los casos.

Cierro con esta sección lo referente a la derivación morfológica. Ahora sería interesante ofrecer una discusión sobre los esfuerzos en poner criterios para distinguir flexión de derivación. Muchas veces se define una en oposición a la otra. Por tanto, en el siguiente apartado abordo dicha temática.

1.6.3. Distinción entre flexión y derivación

En esta sección, resaltaré aspectos importantes sobre la relación entre flexión y derivación. El primero tiene que ver con la duda que han planteado algunos autores sobre la pertinencia de separar estos dos fenómenos o de tratarlos como uno solo. El segundo aspecto comprende los criterios que se han postulado para distinguirlos. El tercero, y último, contempla la posibilidad de que exista un orden de aparición de los afijos flexivos en relación a los derivativos.

1.6.3.1. Pertinencia de la distinción

Es común definir la derivación como aquella donde se generan nuevas palabras asociadas a nuevos significados. La flexión, por su parte, expresaría sobre todo categorías gramaticales sin alterar el significado de las palabras. En términos de Beard (1998: 64), la morfología derivativa difiere de la flexiva en que produce nuevos nombres léxicos para objetos, relaciones y atributos en el mundo. A pesar de lo anterior no es difícil encontrar discusiones que atacan esta diferenciación.

Stump (1998: 19) menciona que se ha puesto en duda, por diversos autores, la pertinencia de separar la flexión de la derivación. Anderson (1985b) es uno de ellos y defiende la idea de que ambas usan los mismos tipos de realizaciones morfológicas (afijación, reduplicación, etc.) y que ninguno de estos está restringido para uno u otro fenómeno.

Además, para Anderson tampoco se puede hacer distinción tomando como base las categorías gramaticales que cada una expresa, ya que en una lengua una categoría puede ser flexiva

y, en otra lengua, derivada. Pone como ejemplo la categoría de diminutivo, que en la lengua fula se produce con flexión, pero en alemán es generalmente considerada derivación (cf. Anderson 1985b: 162-163).

1.6.3.2. Criterios para distinguir flexión y derivación

A pesar de lo mencionado en el apartado anterior, no han sido pocos los esfuerzos por establecer criterios que distingan la flexión de la derivación. Expongo enseguida algunos de ellos. Desde la visión estructuralista, Hockett ya proponía algunos indicios para clasificar un afijo como flexivo o derivativo: creación de paradigma, carácter endocéntrico o exocéntrico (cf. Hockett 1971/1958: 214-215). En la gramática generativa transformacional también ha existido preocupación por esta distinción. De hecho, ésta formaba parte fundamental de su modelo teórico ya que, de acuerdo con dicha propuesta, la flexión es parte de la sintaxis mientras que las palabras derivadas tienen una estructura interna a la cuál ésta no tiene acceso.

Algunos autores como Stump (1998: 14-19) y Beard (1998: 44-46), a partir de la propuesta lexicalista de la sintaxis, proponen varios criterios comúnmente usados para distinguir flexión y derivación. Estos criterios son independientes unos de otros, así que no es necesario que se presenten todos.

Cambio en el significado léxico o categoría gramatical

La derivación conlleva un cambio en el significado léxico, un cambio de categoría gramatical o ambos (23)(a). Por otra parte, la flexión no implicaría cambio alguno (23)(b).

(23) Español

(a) barba > barb-udo
N > A

(b) barba > barba-s
N > N

Este criterio tiene limitaciones, ya que no toda derivación implica cambio de categoría (24)(a) ni cambio de significado léxico (24)(b). Además, los participios son generalmente vistos como parte del paradigma verbal aunque en muchas lenguas funcionan como adjetivos.

(24) Inglés (Stump 1998: 15)

(a) read > reread

(b) cyclic/cyclical

Determinación sintáctica

En el caso de la flexión, el contexto sintáctico suele exigir un elemento particular de un paradigma flexivo. En español, por ejemplo, las palabras de una construcción deben aparecer con los sufijos flexivos requeridos por el fenómeno de concordancia. Sin embargo, sería difícil que el contexto sintáctico requiera que algunas palabras tengan una marca derivativa. Véase el caso de (25)(a) y (b) donde el cambio de flexión de número del sujeto hace obligatorio el cambio de marcas flexivas de otros elementos de la construcción. En cambio, en (c) la presencia de la derivación apreciativa (*muchachito*) no provoca los mismos cambios.

(25) Español
  (a) El pequeño muchacho corrió.
  (b) Los pequeños muchachos corrieron.
  (c) El pequeño muchachito corrió.

Productividad

La flexión es generalmente más productiva que la derivación. Por ejemplo, si se selecciona una palabra, es casi seguro que se aplique todo el paradigma de flexión correspondiente. En cambio, es más probable que no sea posible aplicar alguna derivación, véase por ejemplo (26) en donde no es posible derivar verbos con todos los afijos derivativos. Este criterio suele ser inconsistente ya que en algunas lenguas la derivación resulta ser muy productiva, además, es posible encontrar paradigmas flexivos defectivos (cf. Anderson 1985b: 163-164, Pena 1999: 4330 y Stump 1998: 15).

(26) Español (Pena 1999: 4330)
  (a) frágil > fragilizar, *fragilar, *fragilear, *fragilificar, *fragilecer

Regularidad semántica

La flexión es semánticamente más regular que la derivación en el sentido de que una palabra flexionada mantendrá el mismo significado (27)(a) y (b), pero un elemento derivado ofrecerá significados más variados o se quedará sólo con alguna acepción de su palabra base de derivación (27)(c) y (d). Para Stump (1998: 15), este criterio no suele coincidir con los anteriores ya que es posible encontrar, aunque es raro, diferencias semánticas en elementos que con otros criterios serían flexivos. De igual manera, hay elementos derivativos con un significado muy regular.

(27) Español (Pena 1999: 4310)

(a) botar > boto, botas, botamos, botan

(b) alto > alta, altos, altas

(c) botar > bote, botadura

(d) alto > alteza, altura, altitud

Cierre o cancelación

Al parecer, la flexión cierra o cancela las palabras para futuras derivaciones. En muchas lenguas las palabras no pueden derivarse a partir de formas flexionadas, pero sí de otras formas derivadas. Pero ciertas lenguas muestran inconsistencia con este criterio, por ejemplo, el ruso: en (28)(a) está el verbo del que deriva la forma de (28)(b), esta forma derivada cambia el sufijo de flexión internamente y mantiene el de derivación externamente (28)(c).

(28) Ruso (Stump 1998: 18)

(a) [stučat'] 'tocar la puerta'

(b) [stučat'-sja] 'tocar la puerta deliberadamente'

(c) [stučím-sja] 'nosotros tocamos la puerta deliberadamente'

Para el español, Pena (1999: 4329-4331) recupera los criterios de distinción más importantes. Además de incluir el de cambio en la categoría gramatical, productividad, determinación sintáctica y regularidad semántica, agrega dos más. El primero es que los sufijos de flexión en español son más externos que los de derivación: *nub-os*-o-s, *escol-ar-iz-á*-ba-mos. El segundo, que los procesos derivativos pueden repetirse (*Europa* > *europeo* > *europeizar* > *europeización*) y los de flexión no (**cantá-ba-ba-mos*, **cantá-ba-se-mos*).

No parece fácil la decisión de separar la flexión de la derivación, principalmente por los argumentos tipológicos, que me parecen bastante fuertes. Lo que sí es claro es que para la descripción de una lengua resulta útil la separación. Al respecto del orden de los sufijos flexivos y derivativos, si unos son más externos que otros, en la siguiente sección reviso esta temática.

1.6.3.3.Orden de la flexión y la derivación

Anoto en esta pequeña sección referencias a posturas que apoyan y rechazan la afirmación de que los sufijos flexivos son más externos que los derivativos. Se veía en el apartado anterior que Pena (1999: 4329) proponía esta situación como criterio de distinción entre ambos fenómenos para el español. Bloomfield (1961/1933: 222) había señalado ya que en muchas lenguas los cambios

morfológicos que sufren las palabras parecen tener una tendencia: los fenómenos flexivos están más afuera (lejos) de la base o raíz y los de derivación están más adentro (cerca).

Aronoff (1976: 2) también da como cierta esta idea sobre el orden. Al respecto, la gramática generativa explicó la pertinencia de este ordenamiento con base en la idea de que las operaciones léxicas preceden a las sintácticas. Sin embargo, Beard insiste en que los marcadores de flexión ocurren ampliamente en posición más interna que los de derivación:

> "However, inflectional markers occur widely inside derivational markers. For example, the derivation of verbs by *preverbs*, prefixes which often share the form of an adverb or adposition, is considered derivational, since these derivates often lexicalize semantically. In English these derivations are marked with discontinuous morphemes: for example, *bring* (someone) *around*. In Sanskrit, however, similar derivations prefix the base: for example, pari=nayat, literally 'around he.leads', the present active for 'he marries'. The imperfect is derived by inserting a marker between the idiomatized prefix and stem: that is, pary=a-nayat" (Beard 1998: 45-46).

Stump (1998: 18), por su parte, muestra un ejemplo del ruso (Véase el ejemplo (28) en el apartado anterior).

He terminado con esto, no sólo la descripción de la derivación, sino también su distinción de la flexión. Estos dos fenómenos, como he ido anotando, son claves para el proceso de identificación automática de categorías gramaticales. El de flexión porque la aparición de una marca flexiva permite discriminar entre categorías, en español verbos contra nominales y adjetivos. En el caso de la derivación, los sufijos nos indican la pertenencia a una categoría.

Ya que usaré un método para descubrir afijos, parecería suficiente la descripción de la flexión y la derivación para nuestros fines (ambas se realizan en español principalmente de forma afijal). Sin embargo, existen dos fenómenos más que se manifiestan con la adición de segmentos de palabra: composición e incorporación. Además de que el método de descubrimiento de afijos es susceptible de identificar en cierta medida elementos de estos tipos (como se verá adelante) y con el fin de completar el entendimiento de la complejidad que encierra la morfología, sobre todo la concatenativa, explicaré estos fenómenos en las siguientes secciones. Primero el de composición, que es algo común en el español y al final el de incorporación.

1.6.4. Composición

Como ya decía, en esta sección intentaré explicar el fenómeno de composición de forma sucinta. Éste se caracteriza por aglutinar varias palabras y tratarlas como una sola. El principal aspecto que

me interesa de este mecanismo morfológico es que un método basado en la comparación de segmentos de palabras puede dar cuenta de los elementos de un compuesto. En el caso de esta tesis, el método de descubrimiento automático de afijos propone cortes al interior de ellas por lo que cabe la posibilidad de encontrar miembros de compuestos[12].

Otro aspecto que atrae mi interés es la manera de determinar la categoría gramatical de un compuesto, situación que revisaré en esta sección. Así pues, a continuación expongo algunas particularidades sobre este mecanismo de formación de palabras. Entre ellas están su definición, clasificación y la manera de identificar compuestos en comparación con otros fenómenos.

1.6.4.1.Definición

Enseguida revisaré algunas definiciones propuestas para un compuesto y más adelante examinaré este tipo de palabras en español. En términos de Bloomfield (1961/1933), una palabra compuesta tiene dos o más formas libres. Por ejemplo, el elemento léxico del malayo *mata-hari* 'sol' es un compuesto formado de dos palabras: *mata* 'ojo' y *hari* 'día' (cf. Fabb 1998: 66).

Sin embargo, es posible encontrar compuestos formados por una palabra unida a un morfema que no se encuentra de manera independiente y que tampoco es un afijo. Ejemplos del inglés serían: *church-goer, ironmonger, television, cranberry*. Los segmentos *–goer*, *–monger*, *tele–* y *cran–* no están identificados en inglés como palabras independientes y, aunque algunas veces son localizados en otros elementos léxicos, no hay evidencia suficiente como para llamarlos afijos (cf. Fabb 1998: 69).

Otras veces, según Anderson (1985a: 40), ningún elemento aparece de manera independiente, por lo que propone que este proceso está basado en la combinación de dos o más miembros de una clase léxica abierta, que generalmente aparecen como palabras independientes. Aún si una o ninguna parte del compuesto aparece libre, se toma como tal cuando forma un patrón estructural.

En la gramática del español, Pena establece que "si el elemento añadido a la base es otra base, hablamos del proceso de composición" (1999: 4335). Hockett propuso tres tipos: formados por dos palabras que pueden encontrarse de manera libre (29)(a); aquellos donde el primer elemento presenta una forma especial (29)(b); y los que tienen un primer elemento flexionado (29)(c).

[12] Será necesario esperar al capítulo dedicado al método de descubrimiento de afijos para confirmar si los elementos de algún compuesto pudieran aparecer en la lista de afijos. De cualquier manera, adelantándome un poco, el método mide niveles de afijalidad de segmentos, situación que nos permitiría distinguir entre afijos flexivos, derivativos y compuestos (para cuestiones del método véase Medina 2000 y 2003).

(29) Español (Hockett 1971/1958: 247)

(a) boca-calle

(b) peli-rrojo, ali-corto, cabiz-bajo, pati-zambo

(c) saca-corcho, quita-mancha, baja-mar

Val expone dos tipos de compuestos: léxicos o propios, por ejemplo *pelirrojo*; y sintagmáticos, sintácticos o impropios, por ejemplo *bienmesabe* y *fin de semana*. Explica este autor que el primero consiste en la combinación de dos palabras y que el segundo requiere "la fijación de una estructura sintáctica en una forma determinada, lo que conlleva la pérdida de propiedades sintácticas y la hace hábil para expresar conceptos unitarios" (1999: 4760).

Es claro, después de lo visto anteriormente, que para definir un compuesto es necesario tomar en cuenta aspectos tanto morfológicos como sintácticos. Resulta interesante que para algunos autores (Val 1999, Anderson 1985a) una construcción sintáctica sea vista como un compuesto (*fin de semana*). De hecho, para Anderson, lo importante no es la distinción entre un compuesto y los otros procesos de formación de palabras, sino entre éste y las estructuras sintácticas. En este sentido, resulta importante conocer los criterios para identificar un compuesto. Por esto, en el siguiente apartado expongo algunos de ellos.

1.6.4.2.Criterios de identificación

Examino aquí algunos criterios para identificar compuestos. Me interesa dejar claro dos cosas: la distinción entre éstos y los procesos de flexión y derivación; y cuáles serían las posibilidades de que el método de descubrimiento de afijos descubra elementos de un compuesto. Recordemos que el método está basado principalmente en la comparación de segmentos.

Los criterios se han propuesto primordialmente para distinguirlos de construcciones sintácticas. De cualquier forma, su aplicación puede variar de una lengua a otra y en algunas los compuestos son más productivos. Sobre esto, dice Val (1999) que en lenguas romances, como el español, son menos productivos que en otras como las germánicas.

Los compuestos sufren modificaciones fonológicas y morfológicas que permiten su identificación. Muchas veces estas modificaciones no suceden en construcciones sintácticas equivalentes, aunque puede ser que si lo hagan. Una de las modificaciones más reconocida es la pérdida de acento en alguna de las palabras, ciertas lenguas dejan sin acento el primero y otras al

segundo miembro. Ejemplo de las primeras es el italiano y el español[13], y de las segundas el inglés y danés.

Cuando se unen las dos palabras en un compuesto, es posible la elisión o inserción de elementos fonológicos y morfológicos. Por ejemplo, Fabb (1998: 80) reporta que en la lengua dakota, a diferencia de las frases, se inserta una *a* epentética al final de la primera palabra, y en malayo, existe elisión de nasal y alargamiento vocálico en la última palabra. En español también suceden estos fenómenos de elisión e inserción como muestran los ejemplos de (30). De hecho, puede darse la eliminación de la sílaba final como en (30)(d).

(30) Español (Val 1999: 4761)

(a) Pelo + rojo > *pelirrojo*

(b) Col + flor > *coliflor*

(c) Mano + atar > *maniatar*

(d) Norte + coreano > *norcoreano*

En algunas lenguas, es posible encontrar un morfema sin significado que se inserta entre dos palabras de un compuesto. Anderson (1985a: 41-42) reporta este fenómeno en el alemán y comenta que en el mandarín existe una partícula separada que aparece con modificadores en construcción sintáctica, por lo que su ausencia identificaría un compuesto.

Otra característica distintiva de un compuesto es el cambio en el comportamiento de las marcas de flexión. Bloomfield presenta un ejemplo del griego antiguo para hacer notar que en construcción sintáctica las marcas flexivas aparecen en ambas palabras (31)(a), pero en el compuesto sólo en la segunda (31)(b). En español sucede lo mismo. Compárese por ejemplo *girasol* > *girasoles* contra **giransol* o **giransoles*; o *abrecartas* frente a **abrencartas*.

(31) Griego antiguo (Bloomfield 1961/1933: 230)

(a)	Nominativo	[ne'a: 'polis]
	Acusativo	[ne'a:n 'polin]
	Genitivo	[neˇa:s 'poleo:s]
(b)	Nominativo	[ne'a:polis]
	Acusativo	[ne'a:polin]
	Genitivo	[nea:'poleo:s]

[13] En español Val (1999: 4761) reporta como excepciones: *épico-lírico* y *salón-comedor*.

También, el orden y estructura en un compuesto son más rígidos que en una construcción. Por esta razón, no puede ser dividido, modificado por otros elementos ni sufrir cambio de orden de sus elementos: **abrebiencartas*, **girasiempresol*. Compuestos más cercanos a estructuras sintácticas como *carta bomba* no cumplen todos los criterios, por ejemplo, son permitidas flexiones del tipo: *cartas bomba*, *cartita bomba*; pero no **cartas bombas* (cf. Val 1999: 4762).

El último criterio es que los compuestos son semánticamente más especializados. En otras palabras, el sentido de un compuesto no se forma por la suma de rasgos semánticos de sus constituyentes, es común que brinden un significado con matices distintos (*agua ardiente* frente a *aguardiente*; y *tela de araña* frente a *telaraña*). Puede ser también que con el tiempo, una construcción vaya sufriendo cambios hasta producir un compuesto y luego desaparezca, aunque puede darse la presencia de ambos.

He repasado los criterios para diferenciar compuestos de construcciones sintácticas y creo que ha quedado clara, al mismo tiempo, su distinción con los procesos de flexión y derivación. Por lo que expuse arriba, existe la posibilidad de que el método de descubrimiento de afijos obtenga partes de un compuesto, pero es de suponerse que serán muy pocas ya que estos segmentos no son tan productivos como los flexivos y derivativos, al menos en español. Además, en todo caso, los compuestos presentarán marcas flexivas por medio de las cuales se obtendrá su categoría. En el corpus de estudio encontré *malogrados* y *susodicho*, que tienen sufijos finales –o y –s, lo que permitirá con el método propuesto etiquetarlos como sustantivos o adjetivos[14].

1.6.4.3.Clasificación

La revisión de las distintas clases de compuestos tiene importancia ya que nos ofrecerá idea de cómo determinar su categoría gramatical. En seguida, expongo las clasificaciones.

1.6.4.3.1. Compuestos sintácticos y no sintácticos

Bloomfield (1961/1933: 233-235) propone dos maneras de clasificar los compuestos. La primera está basada en la relación entre sus miembros. Así, existirían compuestos sintácticos, que guardarían las mismas relaciones entre sus miembros que si fueran palabras en una frase; y los no sintácticos, cuyos miembros no serían combinables en construcciones sintácticas. Sería posible, sin embargo,

[14] Un aspecto que no tomo en cuenta en este apartado, pero del cuál estoy consciente, es la discusión que despierta tomar la prefijación del español como composición, o aún más, como incorporación de preposiciones. Para revisar este asunto véase Val (1999: 4775-4776), y Varela y Martín (1999).

encontrar compuestos que se localizarán en una parte intermedia entre los tipos anteriores, ya que sus miembros tendrían un paralelo sintáctico, pero con alguna modificación.

Val (1999: 4772-4774), de manera parecida, reconoce que los compuestos en español guardan proximidad con las estructuras sintácticas. En este sentido, coincide con el primer tipo de Bloomfield. Subdivide, además, los compuestos sintácticos de acuerdo a su relación interna ya sea subordinativa o coordinativa.

En relación subordinativa puede darse que el núcleo tenga rección sobre el elemento no nuclear (*vasodilatación*, *maniatar*) o sólo que sea modificado o complementado (*altiplanicie*, *malgastar*). La relación coordinativa se da entre elementos equivalentes, que compartirían la característica de ser núcleo. Fabb (1998: 67) expone que pueden darse por combinación de sinónimos, antónimos o por la combinación de cosas paralelas. Ejemplos del español serían: *agridulce* y *sordomudo*.

Otra clasificación a partir de las relaciones sintácticas al interior del compuesto es la propuesta por Anderson (1985a) para el chino mandarín. Su clasificación, que puede verse en (32), está fundada en relaciones de modificación (32)(a), verbo-objeto (32)(b), sujeto-predicado (32)(c), coordinación (32)(d) y de verbos resultativos (32)(e).

(32) Chino mandarín (Anderson 1985a: 46-52)

(a) gāng-bǐ
acero-pluma ‘bolígrafo’
(b) xiū-xǐng
cultivar-conducta ‘convertirse en budista’
(c) tóu-téng
cabeza-dolor ‘tener dolor de cabeza’
(d) chē-mǎ
vehículo-caballo ‘tráfico’
(e) xiě-cuǒ
escribir-equivocado ‘escribir erróneamente’

1.6.4.3.2. Compuestos endocéntricos y exocéntricos

El otro modo de clasificación de Bloomfield se basa en la relación del compuesto, como un todo, con sus miembros. De esta manera habría dos tipos: endocéntricos y exocéntricos.

Los primeros serían aquellos en los que el compuesto tiene la misma función gramatical que sus miembros. Para Fabb (1998: 66-67) y Val (1999: 4765-4772), las palabras compuestas que tienen una cabeza o núcleo son llamadas endocéntricas. Éste tendría características similares al de una frase: representa el centro del significado del constituyente y determina categoría. En

zarzamora, por ejemplo, el núcleo es *mora*, ya que el compuesto es un tipo de mora y es un sustantivo.

En los exocéntricos, los miembros o el miembro principal tendrían una función y significado diferente a la que tendría el compuesto en su conjunto. Fabb (1998) establece que los compuestos sin núcleo son de este tipo. Según Val (1999), estos se dan mediante procedimientos tropológicos como la metáfora, metonimia o sinécdoque. Ejemplos de este autor serían *gallocresta*, una planta medicinal con hojas parecidas a la cresta del gallo, y *cascarrabias*.

Según Fabb la distinción entre los dos tipos no es siempre clara y cuando hay un núcleo identificable en el compuesto, éste parece estar en las lenguas hacia la derecha (inglés) o hacia la izquierda (vietnamita, francés). Val ejemplifica en español compuestos endocéntricos con núcleo a la izquierda, *pez espada*, a la derecha, *zarzamora*, y en ambos de manera conjunta, *aguamiel*.

Como se ha visto, la determinación del tipo de compuestos no es simple y entran en juego criterios semánticos para la identificación del núcleo. Afortunadamente, en español son más frecuentes los compuestos endocéntricos, los cuales presentan marcas de flexión. Como dice Val: "La más frecuente es la flexión marginal [en el segundo constituyente]. En la mayoría de los tipos, en el núcleo, localizado en segunda posición, radican los rasgos y, en su caso, sus correspondientes marcas flexivas; de ahí son asignados a la voz compleja" (1999: 4771), ejemplos de ello son: *altiplanicies, zarzamoras, malsanas* y *patizambas*.

Ha sido expuesto, aunque con brevedad, que los compuestos son fenómenos de formación de palabras muy interesantes, principalmente por su cercanía con las estructuras sintácticas. Al respecto, un tipo especial de fenómeno, que varios autores separan de la composición, es la incorporación. Su funcionamiento es muy parecido al que ya mostré, pero tiene peculiaridades que resultan interesantes revisar, al menos de manera muy general, en el siguiente apartado.

1.6.5. Incorporación

El último fenómeno de formación de palabras que abordaré es la incorporación, la cual conlleva la unión de dos palabras condicionadas por su categoría gramatical. Como se trata nuevamente de un método que aglutina palabras en otras, es posible que el método automático de descubrimiento de afijos identifique estos segmentos en la palabra compuesta. Para resolver el cuestionamiento y conocer este mecanismo morfológico, a continuación expongo su definición y características.

Una interrogante adicional es si éste se presenta en español, por lo que haré alguna anotación al respecto. Dado que se trata también de un fenómeno más alejado de mis objetivos, lo revisaré de manera concisa y enfocándome sólo a la incorporación nominal.

1.6.5.1.Definición

Para definir la incorporación en términos generales, puedo decir que consiste en la combinación de una palabra, generalmente un verbo o una preposición, con otro elemento, generalmente un nombre, pronombre o adverbio. Un tipo especial es la incorporación nominal basada en la combinación de un verbo y un nombre, el cual aparecería en construcción sintáctica como objeto, para producir una forma compuesta que sirva como predicado de una frase.

Un ejemplo que se ha hecho clásico para mostrar este tipo de fenómeno es el de (33). En él se puede corroborar que no se incorpora la palabra completa, sino su base. Esto es claro ya que el sufijo *–tl* aparece en sustantivos de colocación libre (33)(a), pero se pierde cuando son incorporados (33)(b).

(33) Náhuatl (Sapir 1911, citado en Gerdts 1998)

(a)	ni-c-qua	in nacatl	(b)	ni-*naca*-qua
	Yo-esto-comer	la carne		Yo-carne-como
	‘(Yo) como la carne’			‘(Yo) como carne’

En este ejemplo del náhuatl, el sustantivo incorporado funciona como objeto de verbo transitivo, pero según reporta Anderson (1985a: 54), esta lengua también tienen ejemplos de incorporación de sujeto de verbo intransitivo y sustantivo en función de instrumental.

Un aspecto interesante de la incorporación es el hecho de que las lenguas imponen restricciones a los nombres que pueden ser incorporados (cf. Fabb 1998: 69). Por ejemplo, es común que sustantivos que surjan a través de nominalización o composición no se incorporen. En la lengua tiwa del sur, los nombres propios no se incorporan y, en general, los sustantivos inanimados se incorporan con mayor facilidad que los animados (Gerdts 1998: 85). Este mismo autor menciona que en muchas lenguas la incorporación tiene la función de marcar el nominal menos prominente en el discurso, por lo tanto, entre más preponderante a nivel discursivo sea el sustantivo, es menos probable que sea incorporado

En español, la pertinencia de aceptar la presencia de la incorporación en la lengua ha sido discutida en casos como *maniatar*, pero al menos la Real Academia parece resistirse y tratar estos ejemplos como composición. Val dice al respecto: “Otro aspecto que cabe valorar es en qué medida es adecuado proponer que la gramática del español cuenta con un procedimiento como la incorporación, cuya productividad se ha revelado notablemente en otras lenguas, para dar cuenta de una lista cerrada y de escasos elementos” (1999: 4760).

Después de haber definido este fenómeno, creo que a nivel de concatenación de segmentos no hay diferencia entre incorporación y composición. La peculiaridad de la primera se define a nivel sintáctico y semántico. Ya que es un fenómeno poco frecuente en español, será muy difícil que aparezcan elementos incorporados en la lista de afijos.

2. Las categorías gramaticales

En el capítulo anterior, dedicado a la morfología, describí los fenómenos relacionados con la manera en que están compuestas las palabras. Lo hice siempre desde una perspectiva general sin concentrarme en el español ni tomar una postura teórica. Esto me permitió tener una visión más amplia sobre la morfología y ver si el método de identificación automática de categorías gramaticales puede aplicarse a otras lenguas con fenómenos concatenativos.

A pesar de que el concepto de palabra no termina de ser completamente aceptado por todas las corrientes de estudio lingüístico, su uso como término y las referencias a él parecen inevitables (Cf. Anderson 1985b: 150-165, González Calvo 1998: 11-37, Lara 2004: 401-408, Lara 2006 primera parte, y Pena 1999: 4327-4328, entre otros). De tal suerte que estudiar el comportamiento de las palabras es importante para describir y entender el funcionamiento del sistema lingüístico. Para efectos del presente trabajo de investigación, lo que me interesa sobre estas palabras es la manera de clasificarlas en distintas categorías de acuerdo con ciertos criterios.

Importante resulta aclarar que no busco entender la naturaleza de tal clasificación, en el sentido de que no me ocuparé del descubrimiento de las clases a las que podrían asociarse las palabras. Siendo más específico, no estudio la naturaleza y organización del sistema de categorías o partes de la oración. Así, este trabajo se resume en realidad a la asignación de una palabra a una clase predefinida.

En este capítulo, por tanto, estableceré con más detalle la definición de categoría gramatical. También hablaré sobre las categorías que heredamos de la tradición grecolatina y cómo fueron identificadas en la antigüedad. Además, discutiré los criterios más aceptados que permiten asociar una palabra a una categoría determinada. Esta parte es importante porque me permitirá conocer la naturaleza de esta asignación y relacionarla con el método automático.

El capítulo cierra con un breve recorrido por varios problemas muy interesantes sobre la universalidad de las clases de palabras. Por ejemplo, si existen categorías “universales” o presentes en todas las lenguas, qué criterios tipológicos permitirían definirlas y cómo hace una lengua para expresar el mismo significado que otra sin compartir las mismas clases de palabra.

2.1. Definición

Desde el principio de la tesis he hablado de categorías gramaticales y ahora es el momento justo para definir el término. Por ello, en esta sección, no sólo estableceré tal definición y su relación con las partes de la oración, término más tradicional, sino que expondré un problema terminológico

importante al respecto. Éste se ha derivado del hecho de que cada postura teórica, al dar prioridad a distinto nivel de lengua, cambia el nombre, por ejemplo a categoría sintáctica o categoría funcional o categoría léxica. Por lo anterior, expongo concisamente las miradas de varias corrientes lingüísticas al respecto de la agrupación de palabras en categorías.

Un hecho aceptado por cualquier estudio lingüístico es que las palabras de cualquier lengua pueden agruparse en categorías de acuerdo con su comportamiento (cf. Trask 1999), aunque este agrupamiento puede ser "complejo y opaco" (cf. Anward 2000). Al respecto, está igualmente aceptado que no todas las lenguas tienen el mismo número de ellas, pero no está aún confirmado del todo si existen algunas categorías universales.

Imaginar la inexistencia de estas agrupaciones de palabras, es decir, que cada una se comportara de manera distinta, nos llevaría a graves problemas en el análisis lingüístico; la sintaxis, por ejemplo, sería totalmente distinta a la que conocemos. Además, esta falta de economía implicaría que el humano tendría recursos ilimitados de memoria para recordar el uso de cada elemento léxico.

La clasificación de palabras ha sido propuesta desde la antigüedad. Griegos como Platón y Aristóteles dieron paso a propuestas que llegaron a reconocer hasta ocho clases de palabras por Dionisio de Tracia, discípulo de Aristarco: sustantivos (incluidos los adjetivos), verbos, participios, artículos, pronombres, preposiciones, adverbios y conjunciones[1].

Los romanos intentaron disminuirlas a cuatro categorías, pero finalmente prevalecieron ocho, eliminando artículos y separando interjecciones. Estas fueron llamadas *partes orationis* y se deben a Apolonio Díscolo y Prisciano. Durante la Edad Media, se mantuvieron, aunque fueron discutidas las diferencias entre sustantivos y adjetivos. Las partes de la oración para el inglés quedaron asentadas para la posteridad por Lindley Murray en 1795 como: sustantivo, verbo, adjetivo, adverbio, pronombre, preposición, conjunción e interjección.

Menciona González Calvo (1998: 64) que para la tradición español se han establecido distintos conjuntos de categorías, por ejemplo, doce entre 1771 y 1847. Posteriormente, la Academia de la lengua tuvo vacilaciones en aceptar entre nueve (1870) y diez (1870-1917) partes de la oración. Al contrario de la tradición inglesa, en español fue común mantener la unión de sustantivos y adjetivos en una sola clase (cf. RAE 1973). Sin embargo, Bello (1984/1848) reconoció siete

[1] En otras tradiciones antiguas también es posible encontrar indicios del surgimiento del concepto de categoría. En la India antigua, 350 años a. C., Panini propuso al menos nominales y verbos. Lo mismo en la tradición árabe con Sibawaihi, quien habló de sustantivos, verbos y partículas (cf. Voutilainen 1999a: 3).

oficios: sustantivo, adjetivo, verbo, adverbio, preposición, conjunción e interjección. La última gramática de la Academia (cf. Bosque y Demonte 1999 v. 1) ya separa, en dos clases, nombres y adjetivos, y brinda amplia discusión no sólo sobre las principales categorías gramaticales, sino también sobre subclases de éstas.

Con lo anterior, se nota la influencia que la tradición grecolatina tiene y ha tenido sobre los estudios de lingüística contemporáneos, al menos en lo que se refiere a la agrupación de palabras. Es notorio que las clases se han mantenido prácticamente con los mismos nombres y, como se verá adelante, con los mismos criterios de clasificación.

Parece ser que fue Sapir quien en su momento despertó cierto interés en estudiar las clases de palabras. Sus estudios con lenguas americanas y su observación de que no todas las lenguas tenían el mismo número de partes de la oración llamaron mucho la atención. Más adelante, en la lingüística general, resaltan las descripciones de lingüistas como Bloomfield (1961/1933); Charles Fries, que propone para el inglés cuatro clases mayores y quince menores, entre ellas los determinantes; y Hockett (1971/1958) sobre los sistemas de categorías gramaticales.

El conjunto de categorías clasificadoras ha recibido diversos nombres dependiendo de la tradición, escuela o criterio de clasificación (morfológico, sintáctico, semántico). Los más clásicos son los de *partes de la oración* para la tradición española, *parts of speech* en inglés y *parties du discours* en francés. Otros nombres son: clases de palabras, categorías gramaticales, categorías sintácticas, categorías léxicas y clases funcionales.

Como puede verse, no es menor el problema que la lingüística tiene ante esta falta de consenso terminológico (cf. González Calvo 1998). A pesar de ello, los estudios lingüísticos continúan y generalmente adoptan el término más enraizado: *partes de la oración*, no obstante el término oración también es ampliamente criticado.

Los estudios más tradicionales adoptaron los términos de partes de la oración y clases de palabra debido a que consideraban a la palabra como unidad básica de la gramática y definían la oración en términos de la palabra. Los criterios de clasificación que desde la antigüedad han estado presentes son: formales, funcionales y semánticos. Los tres se traslapaban para realizar la clasificación y, como se verá en la siguiente sección del capítulo, generalmente uno ganaba primacía. Cuando vino la discusión de qué criterio era realmente preponderante se dispararon las distinciones terminológicas.

Por otra parte, posturas modernas que incluyen como unidades de la gramática al morfema y al sintagma, en lugar de palabra y oración, ponen en serio cuestionamiento el sentido de partes de la oración. González Calvo atribuye la explosión de términos a una falta de distinción entre unidades,

categorías y funciones (cf. González Calvo 1998). De hecho, concluye que el término *partes de la oración* es inadecuado.

Ahora bien, el término categoría gramatical ha sido asociado comúnmente a los rasgos gramaticales (número, género, tiempo, modo, aspecto, etc.) marcados por fenómenos morfológicos sobre las palabras y por tanto como complementos de las partes de la oración (Hockett 1971/1958: 234-242). Sin embargo, también ha sido utilizado como término equivalente.

Para esta tesis, me interesó resaltar que el método automático permitirá identificar tales rasgos gramaticales en el corpus. Por tanto, decidí usar el término *categoría gramatical* para referirme a las tradicionales clases de palabras y partes de la oración, junto con las propiedades gramaticales (rasgos). Estas propiedades son las que revisé en el capítulo de morfología bajo el rubro de categorías flexivas.

Entonces, usaré los tres términos como equivalentes. Este uso lo consigna también Brown (1999: xiii): "Grammatical categories, that is the parts of speech, the word classes themselves, and the categories traditionally associated with them such as case, mood, tense, aspect and voice" [2].

Sigo ahora con distintas perspectivas lingüísticas y su postura ante la clasificación de palabras en categorías gramaticales. Por ejemplo, para el distribucionalismo americano, los criterios de clasificación morfológicos y semánticos fueron puestos en lugar secundario. Esto permitió hablar más bien de clases funcionales, en las que son asignados constituyentes, más que palabras.

Los primeros trabajos en gramática generativa adoptaron la nomenclatura de partes de la oración y con ella categorías como verbo, nombre y adjetivo. Pero las propuestas de Chomsky y sus seguidores cambiaron el sentido de la clasificación. En general, propusieron nuevas clases como: auxiliares, determinantes (artículos, demostrativos), cuantificadores, complementantes, modificadores, partículas y subordinadores (Cf. Trask 1999).

En recientes teorías sintácticas como la Gramática de Rol y Referencia, y de corte generativo como Principios y Parámetros, es posible encontrar también prácticamente las mismas categorías mayores de palabras como sustantivo o verbo y las categorías gramaticales como tiempo y número. Los criterios para definir éstas son morfológicos, sintácticos, semánticos y pragmáticos. En el caso de la postura de la Gramática de Rol y Referencia, se propone una estructura de cláusula en la que se

[2] La lingüística computacional, por ejemplo, tiene muy arraigado el nombre de *part-of-speech tagging* para describir el proceso de asignar etiquetas de categorías gramaticales a palabras de un corpus. Como la tesis es en español, he podido darle otro nombre al proceso: identificación automática de categorías gramaticales, pero estoy consciente de que para una traducción al inglés, tendría que usar el término *part-of-speech tagging*.

involucran distintas categorías de palabras. Por su parte, en el modelo de Principios y Parámetros existe una distinción entre categorías léxicas (partes de la oración), y categorías funcionales (categorías gramaticales de tiempo, persona, género, número, caso) (Cf. Brown 1999: xxi-xii).

Un aspecto interesante lo constituye el hecho de que una palabra pueda ser clasificada en varias clases o que cuente con características de varias clases. Por ejemplo, el participio del español, que se refiere a procesos temporales como verbo, pero que tiene rasgos de género y número como adjetivo. De hecho, la decisión de cuándo es participio y cuándo es verbo puede llegar a requerir de criterios de diferenciación muy específicos (cf. Bosque 1999)

Por tanto, realizar una asignación uno a uno entre palabras y categorías es muchas veces difícil y se vuelve, como dice Anward (2000), "complejo y opaco". Es más, cada lengua cuenta con mecanismos que permiten el cambio de una categoría a otra; recuérdese la derivación morfológica, por ejemplo. Lo anterior ha llevado a posturas teóricas a proponer elementos lingüísticos centrales y periféricos sin límites entre categorías en lo que se puede llamar un *continuum* categorial. Un tipo de gramática así es conocida como gramática no discreta.

La gramática o lingüística cognitiva, por ejemplo, sigue esta idea de categorías "radiales" y continuas. Acepta los dos subsistemas lingüísticos tradicionales: léxico y gramatical, pero a manera de un continuo que abarca desde las clases más gramaticales, que determinan la estructura del sistema lingüístico, hasta las más léxicas, que establecen el contenido. Los elementos más gramaticales forman categorías pequeñas y cerradas, mientras que las categorías más léxicas (nombres, verbos y adjetivos, al menos en español) son consideradas abiertas.

Las categorías, en la lingüística cognitiva, no están definidas con base en la forma o distribución, ni siquiera en términos de la semántica. Más bien, están determinadas a partir de la configuración de dominios cognitivos que se proyectan de alguna manera. Así, el nombre "perfila" cosas: objetos estáticos, no graduables y permanentes; el verbo, cambios en un escenario, un proceso; y a la mitad de los dos, el adjetivo, que tiene lo estático del nombre y lo gradual del verbo (Cf. Delbecque 2008).

Se observa que, a pesar de los cambios de nombre que cada escuela pueda darle a las partes de la oración, los nombres de las clases en sí no cambian. Al menos tres clases de palabras: nombre, adjetivo y verbo, se mantienen de escuela en escuela y sólo cambia la motivación lingüística que permite justificar la asociación de las palabras en ellas. Es quizás solamente la gramática generativa la que introdujo nuevos nombres a la clasificación.

Hasta aquí, hice un breve repaso sobre el surgimiento del concepto de clasificación de palabras en categorías[3] y sobre la variada nominación que de ellas se hace. Generalmente, estas distintas maneras de llamarle a las tradicionales partes de la oración dependen del criterio de clasificación que predomine. Por esto, en el siguiente apartado discutiré los criterios que han sido utilizados para clasificar las palabras de una lengua.

2.2. Criterios de identificación

Ya que mi objetivo de investigación tiene que ver con la identificación de categorías gramaticales en un corpus, creo que es importante conocer y discutir los criterios que permiten esta identificación. En otras palabras, una vez aceptado el hecho de que las palabras se agrupan por su comportamiento, es necesario describirlo. Por lo anterior, en esta sección recopilaré las ideas de algunos autores sobre los criterios que consideran necesarios para identificar categorías gramaticales.

Como primera observación, puedo decir que los criterios constantemente cuestionados siguen siendo los que heredamos de la primera clasificación 100 años antes de Cristo, la de Dionisio de Tracia. En ella se incluían criterios morfológicos, sintácticos y semánticos. Parece que cada uno ha tenido su apogeo en distintas épocas o escuelas lingüísticas, algo de ello se veía en el apartado anterior.

La importancia de los criterios morfológicos radica en que permiten distinguir entre categorías que contienen palabras flexionadas de aquellas que contienen formas invariables. En español, por ejemplo, se distinguiría verbo y sustantivo de preposiciones y conjunciones. Darle supremacía a este criterio sobre los demás apoya propuestas que, por ejemplo, no distinguen adjetivos como una categoría aparte, sino como subclase de nombre, como en español.

Criterios de tipo morfológico aplicados al español definirían al sustantivo por su flexión de género y número en oposición al verbo por su flexión de tiempo y persona. En otras palabras, se encontrarían definiciones como: “lo que flexiona para tiempo y persona es verbo”, y “lo que flexiona para género y número es sustantivo”. Claro que en lenguas de poca morfología flexiva, estos criterios serán insuficientes. Además, lenguas como el sueco, que cuentan con categorías no nominales asociadas a marcas de flexión nominal son otro problema para estos criterios (cf. Anward: 2000)

Dentro del criterio morfológico también está la derivación, que cumple funciones de cambio de categoría de una palabra mediante el uso, principalmente, de afijos derivativos. Entonces, es

[3] Para leer más detalles al respecto véase Trask (1999) y la introducción de Brown (1999).

posible usar la presencia de estas marcas morfológicas para determinar la categoría de una palabra. Los criterios que usan flexión y derivación serán muy importantes en este trabajo de tesis ya que el descubrimiento de unidades morfológicas nos dará una lista de afijos, flexivos y derivativos. El método de identificación automática de categorías gramaticales, como se verá más abajo, incluirá estas unidades. Entonces, trataré de confirmar si el uso de ellas mejora en algo el método de identificación.

En segundo lugar se encuentran los criterios sintácticos. Éstos están asociados con la distribución y función de la palabra. La distribución es el conjunto de ambientes, posiciones sintácticas, en los que ésta aparece, la función describe la naturaleza de la dependencia entre la palabra y otras con las que es concurrente. El estructuralismo americano favoreció este criterio, ya que definía una clase de palabra por el "hueco" que podía llenar en una sentencia, es decir, si dos palabras podían aparecer en la misma posición, éstas pertenecían a la misma clase de palabra. La realidad es que esta prueba no es de aplicación general y no brinda muestra contundente de la pertenencia a una clase[4], pero tiene alguna utilidad.

Desde la perspectiva tipológica, surgió una propuesta de clasificación de lenguas por su sistema de categorías gramaticales basada en la función sintáctica de las palabras: el modelo Ámsterdam. La clasificación fue hecha a partir de cuatro funciones o usos, que fueron asociados a clases de palabras: predicado (verbo), término (nombre), función exclusiva de modificador de término (adjetivo) y función exclusiva de modificador de predicado (adverbio). A las anteriores se agregaron funciones compuestas como adjetivo-adverbio o nombre-adjetivo-adverbio. Las lenguas entonces serían clasificadas dependiendo de las funciones-categorías que presentaren (cf. Anward 2000).

El último criterio es el semántico. Según Trask (1999: 280), fue popular en el pasado, pero hoy en día es rechazado debido a que produce problemas en su interpretación. Piénsese en la definición de sustantivo como "lo que significa una cosa o persona", entonces la palabra *vida* no pertenecería a esta categoría. Bloomfield (1961/1933: 265-279) ya había hecho esta anotación al criticar los criterios de las gramáticas escolares y decir que es necesario determinar las partes de la oración del inglés no por su correspondencia con el mundo práctico, sino solamente por sus

[4] En el método computacional que usaremos, se toman en cuenta relaciones contextuales entre las categorías de las palabras. Como veremos después, éstas se expresan en condiciones que revisan una, dos o tres categorías anteriores o posteriores a la palabra a etiquetar.

funciones en la sintaxis. La falta de solidez del criterio semántico es expresada por Croft (1991: 38): "This purely semantic approach, intuitively attractive as it is, is inadequate as it stands"[5].

Con todo lo anterior, es posible notar que ningún criterio es suficiente para la identificación de categorías gramaticales, por lo tanto, la mayoría de las veces es necesario utilizar varios. Ahora, voy a presentar algunas propuestas teóricas que intentan descifrar cómo funciona un sistema de categorías gramaticales.

Desde una perspectiva estructuralista, Hockett (1971/1958: 225) establecía que: "El sistema de partes de la oración de una lengua es la clasificación de todos sus temas sobre la base de similitudes y diferencias en el comportamiento flexional y sintáctico de los mismos". Como puede verse, para este autor, el significado no estaba presente en la clasificación. Hockett presenta la siguiente descripción del sistema categorial del español basado únicamente en el criterio de flexión:

1. Nombres. Tienen flexión de número.
    a. Sustantivos. Pertenecen a un género.
        i. Masculino: *hombre*.
        ii. Femenino: *mujer*.
        iii. Indiferente (masculino o femenino): *pianista*.
    b. Adjetivos. Flexionan para género.
        i. Masculino: *lindo*.
        ii. Femenino: *linda*.
        iii. No flexiona para género (masculino o femenino): *ágil*.
    c. Pronombres. Sustituyen nombres y tienen:
        i. Flexión de número y no de género: *cual, usted*.
        ii. Flexión de género y pertenecen a un número: *ninguno, nosotros*.
        iii. Flexión de género y número: *mío, aquel*.
        iv. Invariantes: *yo, alguien*.
2. Verbos. Tienen flexión de persona y número.
3. Partículas. Son invariantes.
    a. Interjecciones: *hola, cáspita, chau*.
    b. Adverbios: *bien, despacio, lejos*.
    c. Que sustituyen adverbios: *hoy, nunca, aquí*.
    d. Preposiciones y conjunciones.

En el mismo sentido, Bloomfield (1961/1933) tomó en cuenta la forma más que el significado y propuso que la asignación a una categoría (*form-class*) está determinada por los hablantes a través de: la estructura y constituyentes de la forma (para el caso de compuestos); la inclusión de un constituyente especial (una marca de flexión o derivación); o la identidad de la forma en sí misma (cuando la palabra o forma pertenece a una categoría arbitraria).

[5] En el método que ocupo en esta tesis (de Brill), no es tomado en cuenta ningún aspecto semántico en lo que respecta a los criterios de identificación de categorías.

Dixon menciona que “el reconocimiento de las clases de palabras al interior de una lengua dependen de criterios morfológicos y sintácticos” (2000/1982: 87). Es de resaltar su explicación sobre el hecho de que diversas lenguas con características morfológicas y sintácticas distintas compartan las mismas clases de palabras. Según su postura teórica, esto se debe a que existen criterios semánticos universales y que la semántica tiene prioridad sobre la sintaxis.

Para Dixon, las palabras pertenecen primero a un “tipo semántico” y después a una clase de palabra. Los tipos son agrupados de acuerdo con la lengua en las distintas clases, esto permitiría que en una lengua las relaciones familiares (tipo semántico universal) pertenezcan a los verbos y en otra a los sustantivos. Este autor también menciona que una palabra puede estar asociada a otra clase de palabra, situación que llama “derivación extensional” y que estaría marcada morfológica o sintácticamente.

Dixon no sólo deja un legado importante con sus estudios sobre los adjetivos en numerosas lenguas, sino también parece ser uno de los primeros en dar pie a propuestas que intentan explicar los sistemas de categorías desde una visión tipológica. Bajo criterios meramente formales no sería posible comparar sustantivos de dos lenguas: lo que formalmente es un sustantivo en español no lo es en el inglés. Por lo tanto, es necesaria una base semántica que sea compartida[6].

Al respecto de usar propiedades formales y criterios distribucionales para definir las categorías, Delbecque (2008: 43) comenta: “Pero entonces se las convierte en categorías de una lengua particular, y se vuelve más difícil la comparación y la tipología de lenguas”.

Otra mirada es la de Croft (1991: 37) que dice que los criterios semánticos (criterios externos) son plausibles porque podrían aplicarse a cualquier lengua, pero no se cumplen en muchos casos. Además, los criterios formales y distribucionales (criterios internos) resultan bien en lenguas individuales, pero dificultan la comparación entre lenguas. Su postura es que ambos son necesarios y que, de hecho, deben combinarse.

Este autor propone una teoría sobre las categorías sintácticas, que desde su perspectiva resuelven el problema de definirlas. De esta manera, propone una caracterización universal de tales categorías basada en la teoría de la marcación, la teoría de prototipos y el uso de criterios pragmáticos y discursivos que apoyen una clasificación semántica tradicional de estas categorías (cf. Croft 1991).

Una propuesta adicional que intenta explicar los problemas de categorización es la de Anward (2000). Está basada en dos premisas. La primera es que el lenguaje es aprendido como

[6] Ya anotaba cómo Dixon soluciona así el problema de las categorías entre lenguas.

resultado de una expansión sucesiva a partir de un sistema lingüístico muy simple; no se aprende "de golpe". La segunda es que este proceso de expansión está dirigido, sintagmática y paradigmáticamente, por la necesidad de aumentar la capacidad expresiva y es restringido por consideraciones de economía y contraste. Un principio de economía muy importante es el principio verde, *green principle*, el cuál dice que reciclar los recursos ya existentes es preferible a introducir nuevos.

Entonces, en este modelo de lenguaje basado en las premisas referidas, este autor propone:

> The "deep" organizing factors of part-of-speech systems are not motivated by properties of such system. They are instantiations of factors which drive language development in general: maximization of meaning, minimization of effort. Speakers do not set out to acquire part-of-speech systems, well-designed or not. Part-of-speech systems are what "happen", as language users engage in processes of successive syntagmatic and paradigmatic expansion (Anward 2000: 4).

En español, existen afijos derivativos que indican la pertenencia a una categoría, por ejemplo: -ura, -ción indican sustantivo; -oso, -ble indican adjetivo. Las marcas de flexión permiten una primera distinción entre nombre, adjetivo y verbo frente a adverbios, preposiciones y conjunciones. Después, la presencia de categorías flexivas ayudaría a la distinción entre sustantivos y adjetivos frente a verbos. En seguida, los fenómenos de concordancia podrían dar cuenta de la distinción entre nombres y adjetivos. Pero como dice Pena: "Aun así, las propiedades formales internas de la palabra no son suficientes para definir la totalidad de las clases de palabras. Hay que acudir también a las propiedades sintácticas o combinatorias de la palabra en el marco de las unidades superiores e incluso, en una fase posterior, a determinadas características de tipo semántico" (1999: 4311).

He reconocido hasta aquí los criterios generalmente usados para la categorización de palabras: morfológicos, sintácticos y semánticos. Todo indica que deben complementarse. Ahora expongo algunos aspectos sobre la universalidad de las categorías.

2.3. Universalidad

Es totalmente aceptado el hecho de que las partes de la oración difieren de una lengua a otra. Esta situación lleva a dos preguntas: ¿cuáles categorías están presentes en todas las lenguas? Y ¿cómo hace un sistema lingüístico para no tener una categoría y expresar lo mismo que otra que sí la tiene? A pesar de que los estudios para responder estas interrogantes son numerosos y muy interesantes, no profundizaré en ello. Por eso, a continuación presento de manera muy general algunos aspectos relacionados con la universalidad de las categorías gramaticales.

Dos posturas están en pugna al respecto de lo anterior. Por un lado, se establece que algunas de las categorías gramaticales son universales (nombre, verbo, adjetivo) y aparecen en todas la lenguas del mundo. Por otro, estas categorías no existen en todas las lenguas y por tanto son específicas de cada una de ellas (cf. Croft 2000).

Un hecho reconocido es que algunas clases de palabras no existen en ciertas lenguas y en otras sí. Ejemplo de esto son las llamadas adposiciones, preposiciones y posposiciones. Lenguas como el japonés o el turco cuentan con elementos pospuestos a los objetos con el mismo significado que las preposiciones en el español. Otras lenguas parecen no contar con una clase separada de adjetivos y expresar el mismo significado con verbos, sustantivos o partículas (cf. Dixon 2000/1982). Los adverbios tampoco aparecen en todas las lenguas y su función puede ser expresada con adjetivos, verbos o sustantivos (cf. Van der Auwera 1999: 10).

Sobre la universalidad de ciertas categorías, se han propuesto al menos dos clasificaciones que intentan responder a esta situación. La primera divide las palabras en dos clases: *clases léxicas* con pleno contenido semántico y *clases gramaticales* dedicadas a funciones específicas de tipo gramatical. Lo universal, al parecer, es este contraste. La segunda distinción es entre *clases abiertas*, en donde hay muchos miembros y regularmente se adhieren nuevas palabras, y *clases cerradas* en las que difícilmente se integran otras y son escasas en número. En este caso, parece que ambas clases son universales, aunque lenguas donde los afijos cumplen todas las funciones gramaticales ponen en duda esta posición (cf. Trask 1999).

Una propuesta común ha sido que sólo verbos y sustantivos son clases universales. Sin embargo, varios lingüistas, entre ellos Hockett, han puesto ejemplos en contra, principalmente de lenguas del norte de América. Para Hockett (1971/1958: 230), la distinción en el nootka es entre palabras flexionadas y no flexionadas, más que entre sustantivos y verbos. Algo similar reporta Trask (1999: 282) cuando escribe que “Kinkade (1992: 391) concludes that parts of speech in the Salishan languages are generally absent or at least only weakly developed”.

Pero la discusión al respecto no ha logrado consenso y las posturas opuestas siguen presentes (cf. Bhat 2000: 54-58, Brown 1999: xx-xxi, Dixon 2000/1982: 88-89 nota 1y Trask 1999: 283, entre otros).

3. Identificación automática de categorías gramaticales

La clasificación de palabras en categorías gramaticales ha sido una preocupación constante en los estudios lingüísticos. Tal situación ha quedado descrita en el capítulo anterior. Al respecto, la lingüística computacional y el procesamiento de lenguaje natural han desarrollado métodos estadísticos y computacionales que ayudan a resolver esta labor. Si bien los primeros esfuerzos se dieron en la década de los 50, aún se encuentran trabajos que buscan mejorar los resultados obtenidos hasta ahora, mediante la incorporación de nuevos modelos y procedimientos.

En sentido práctico, el uso de estos programas, también llamados *etiquetadores de categorías gramaticales*, es valioso, ya que hoy en día existe la necesidad de grandes corpus lingüísticos etiquetados[1] y su creación resulta costosa y requiere mucho tiempo. Estos corpus son utilizados para diversos estudios de carácter lingüístico como el análisis gramatical, la lexicografía, la adquisición de habla y escritura, los estudios de variación histórica y dialectal, entre otros. También, son usados para análisis y traducción automática, reconocimiento de términos por computadora, minería de textos[2], por citar algunos. Por otra parte, estos programas se han vuelto recursos fundamentales en el procesamiento de lenguaje natural (cf. Leech y Smith 1999).

El etiquetado de categorías gramaticales no es la única manera de anotar un corpus. También es posible enriquecerlo con información de cualquier otro nivel de lengua o con marcas de fenómenos observados dentro de él. La información codificada en las etiquetas dependerá de los objetivos de investigación. Por ejemplo, existen hoy en día corpus anotados con estructura de frases (frase nominal, frase verbal), dependencias sintácticas, relaciones léxicas (homonimia, hiperonimia) o relaciones anafóricas.

El estudio del lenguaje mediante grandes cantidades de ejemplos obtenidos a partir de corpus es conocido como lingüística de corpus. Uno de sus primeros procedimientos para procesar información es la obtención de concordancias. Éstas consisten en contextos lingüísticos que rodean

[1] Los corpus etiquetados son conjuntos de textos en los que se incluye información adicional de tipo lingüístico codificada en símbolos llamados etiquetas. Estas etiquetas son incluidas en los mismos textos. Otra forma común de llamarle a un corpus etiquetado es "corpus anotado" o "corpus marcado". De hecho, la anotación, marcado y etiquetado suelen usarse de manera indistinta. Sin embargo, es posible encontrar una diferencia entre las dos primeras y la última, ya que la manera más común y óptima de anotar o marcar un corpus es mediante etiquetas, es decir mediante etiquetado.

[2] La minería de textos es el proceso de descubrimiento y extracción de información nueva (conocimiento) mediante el análisis de textos. Este proceso permite encontrar relaciones, patrones o tendencias inmersas en los textos, que ayudan a describir o predecir algún fenómeno. Para una caracterización más completa de la minería de textos véase Ananiadou y McNaught (2006: 1-11) y Weiss, Indurkhya, Zhang y Demerau (2005: 1-13), entre otros.

un patrón de búsqueda, el cuál puede ser una sola palabra, parte de ella o la combinación de varias palabras. Además, es posible rescatar concordancias basadas en categorías gramaticales, por ejemplo, *verbo + determinante + sustantivo* equivalentes a *echar un ojo* o *meter la pata*. Precisamente, esta tesis responde a la inquietud por el desarrollo de este último tipo de búsqueda en un corpus del siglo XVI, para lo cual será necesario un etiquetador de categorías gramaticales.

Por lo tanto, en este capítulo hablaré de los programas para identificar de manera automática categorías gramaticales. En primer lugar, daré una definición de este proceso. Después, explicaré un aspecto importante: las implicaciones del conjunto de etiquetas a utilizar. Por último, describiré los principales métodos para llevar a cabo esta identificación. Algunos de ellos los trataré de manera muy general y otros de forma más específica, pero ahondaré bastante en el método clave para esta tesis, propuesto por Eric Brill (1992, 1993, 1994, 1995).

Es necesario decir que la mayoría de los primeros programas de identificación automática fueron desarrollados para lenguas anglosajonas, por lo que me parece interesante incluir algunos párrafos sobre el analizador gramatical para español desarrollado en El Colegio de México, ya que fue un esfuerzo muy valioso para su tiempo.

3.1. Definición

A continuación explico en qué consiste el proceso de identificación automática de categorías gramaticales. Además, doy cuenta de sus componentes principales y el objetivo de cada uno.

El problema de la identificación automática de categorías gramaticales puede entenderse como el problema de asignar cierta etiqueta, que corresponde a una categoría, a cada palabra de un corpus (cf. Charniak 1996/1993). Para Voutilainen (1999a: 3) las etiquetas son símbolos descriptivos que se asignan a palabras en un texto ya sea de forma manual o automática. Entonces, estas etiquetas codificarían el nombre de la categoría y los rasgos morfosintácticos asociados.

Dado que trabajo con documentos en formato digital, dicha asociación se lleva a cabo a través de la concatenación a cada palabra de ciertos caracteres o letras (etiqueta). Entre la palabra y la etiqueta es utilizado un símbolo separador, que pueden ser una diagonal (/) o un guión bajo (_). Por ejemplo, dada la frase de (1)(a) se pueden etiquetar las categorías de sus palabras como se muestra en (1)(b). En este caso, el conjunto de etiquetas utilizado y su categoría asociada son: {V: verbo, C: conjunción, R: adverbio, P: pronombre, E: preposición, D: determinante, N: nombre}.

(1) Ejemplo de etiquetado

(a) Dixo que sy, e que la tiene en la casa

(b) Dixo/V que/C sy,/R e/C que/C la/P tiene/V en/E la/D casa/N

Si se quisiera tener mayor detalle en los rasgos de alguna palabra, sería necesario cambiar el conjunto de etiquetas. Cada una, entonces, tendría que codificar más detalle como por ejemplo: casa/NCFS que equivaldría a nombre común femenino singular. También sería posible que las etiquetas no estuvieran concatenadas sino separadas por espacios, lo que generalmente sucede cuando cada palabra está en un renglón del corpus, véase por ejemplo (2).

(2) Ejemplo de etiquetado separado por renglones

(a) Dixo V
que C
sy, R
e C
…

Los etiquetadores de categorías gramaticales suelen contar por lo general con tres módulos, que realizan distintas tareas (van Halteren y Voutilainen 1999: 109):

1. Separación de palabras gráficas y signos de puntuación.
2. Asignación de una o varias categorías probables para cada palabra. Esta tarea generalmente comienza con la búsqueda de la palabra en un lexicón, que contiene las posibles categorías. Si la palabra no es encontrada en él, se utiliza algún proceso de asignación de una o más categorías probables.
3. Proceso de desambiguación que elimina las categorías menos apropiadas y deja una sola para cada palabra. Para realizar esto se usa información contextual.

En la práctica, esta división no se muestra de manera transparente, pero conceptualmente es útil para estudiar los diferentes modelos y realizar comparaciones entre ellos. Al respecto, cuando una palabra no es encontrada en el lexicón[3], se pueden usar métodos de análisis morfológico o lexicones de bases de palabras para proponer categorías probables. En el tercer módulo, para

[3] Es tradicional en la literatura llamar a éstas: palabras desconocidas (*unknown words*).

resolver el problema de desambiguación, suelen usarse modelos lingüísticos o estadísticos obtenidos de corpus o escritos por lingüistas.

Para terminar esta sección, quiero resaltar algunos problemas que deben ser considerados en el proceso de identificación. El primero tiene que ver con las contracciones de palabras, ya que el programa deberá asignarles categorías pertinentes. En estos casos es posible dividir la palabra en sus distintos elementos o que la contracción reciba varias etiquetas. Las palabras *del* y *al* del español son ejemplos de esta situación[4].

Otro problema son los compuestos sintácticos, ya que será necesario definir si sus elementos serán etiquetados de forma separada o se unirán para asignarles una sola etiqueta. Piénsese en *sin embargo*, *a través de*, por citar algunos ejemplos. El último problema, más severo en algunas lenguas que en otras, son los morfemas discontinuos o palabras cuyos componentes están separados por otros (cf. Cloeren 1999: 44-46).

Hasta aquí, he explicado en qué consiste el proceso de identificación automática de categorías gramaticales y su funcionamiento general. Una parte importante de éste es el conjunto de etiquetas a utilizar, ya que de su selección depende la información codificada y añadida al corpus. Por esta situación, hablaré enseguida de las características a considerar en los conjuntos de etiquetas.

3.2. Conjuntos de etiquetas

A las etiquetas para codificar categorías gramaticales se les ha llamado también *etiquetas de partes de la oración* o *etiquetas morfosintácticas*. Cuando en el proyecto de etiquetado de un corpus se han decidido las categorías que serán identificadas y sus respectivas etiquetas se habla de un conjunto de etiquetas del corpus. Sobre éste hay al menos dos aspectos fundamentales que atender. Por una parte, la especificidad de las categorías, es decir, el nivel de detalle en los rasgos que serán identificados. Por otra, la manera de codificar esos rasgos en las etiquetas. Expongo a continuación estos dos aspectos.

Es común encontrar etiquetadores de categorías gramaticales que dan cuenta solamente de las clases mayores de palabras sin rasgos morfosintácticos. Por ejemplo, el conjunto de etiquetas utilizado en el Corpus del Español Mexicano Contemporáneo (CEMC) fue de trece y pueden verse

[4] En el corpus de español del XVI encontré varios casos como preposición + artículo: *del* (*de el*), preposición + pronombre: *dellas* (*de ellas*), conjunción + pronombre: *quel* (*que él*), y conjunción + verbo: *porques* (*porque es*). Más adelante abordaré la solución tomada al respecto.

en la Tabla 3.1[5]. Si se observa el conjunto de etiquetas del CEMC, resalta que se utilizaron números y letras para las etiquetas, y no se indican rasgos morfosintácticos.

Tabla 3.1: Conjunto de etiquetas del CEMC

Etiqueta	Categoría
0	Ambigua
1	Adverbio
2	Adjetivo
3	Conjunción
4	Preposición
5	Pronombre
6	Artículo
7	Contracción
8	Nominal
9	Verbo
A	Apoyos conversacionales
B	Nombres propios
C	Otros (cifras, errores y palabras que comenzaban con mayúscula)

Si se agrega mayor detalle al conjunto de etiquetas será necesario tomar en cuenta algunos aspectos. Primeramente es necesario conocer las características de la lengua, ya se establecía en capítulos anteriores que no todas las lenguas expresan de la misma manera las distintas marcas gramaticales. Conviene, por tanto, conocer cuáles de ellas se expresan mediante la morfología (concatenativa o no) y cuales a través de la sintaxis.

Otro aspecto es la postura teórica a partir de la cuál se decide el conjunto. Por ejemplo, para el español es necesario tomar partido entre considerar los vocablos *esta, este, esa, ese* como determinantes demostrativos o como adjetivos demostrativos. Por otra parte, en el caso de los verbos será necesario preguntarse si es requerida la distinción entre verbos principales y auxiliares, o si conviene marcar algún verbo en especial. Por ejemplo, en esta tesis resalté con una etiqueta especial el verbo *ser*. Algo que se debe tener en mente es que el conjunto de etiquetas, cuando incluye rasgos, se vuelve más grande.

Como parte del conjunto de etiquetas, es posible encontrar una etiqueta adicional para la puntuación. Aunque no es una categoría gramatical, la razón de su uso es más bien práctica. Según Cloeren (1999: 38) contemplar esta etiqueta le da consistencia al corpus ya que todo elemento dentro de él está etiquetado. Además, según este autor, hay lenguas donde la puntuación tiene

[5] Más información sobre este conjunto de etiquetas puede verse en García Hidalgo (1979), Medina (2003: 359) y Anguiano (2007: 33).

repercusiones en la gramática. Finalmente, es importante por los beneficios que aporta a otros programas de cómputo que procesen el corpus. Otras posibles etiquetas adicionales son las usadas para palabras extranjeras, símbolos matemáticos, abreviaturas y hasta una para elementos no clasificables.

Con lo anterior he mostrado algunos aspectos que intervienen en la selección de las categorías que serán identificadas en el corpus. Una vez decidido lo anterior, será requerida su codificación en etiquetas. Existen al menos dos caminos a seguir: adoptar el conjunto de etiquetas de un corpus ya existente, o tomar algún estándar de codificación producido por un organismo dedicado a ello.

En caso de que no se adopte un conjunto preestablecido de etiquetas, es necesario pensar cómo representar las categorías. En general, hay tres posibilidades: usar números, letras o los nombres completos. La segunda tiene ventaja sobre la primera ya que permite una lectura más o menos natural, se esperaría que las letras utilizadas tengan relación con el nombre de la categoría (N para nombre, V para verbo, por ejemplo). La tercera, aunque es la más clara de todas, por hacer explícito el nombre de la categoría, tiene el problema de que aumenta el tamaño de los archivos electrónicos de manera considerable. La decisión, insisto, deberá ser tomada de acuerdo con los objetivos de investigación y uso del corpus.

Para la tradición anglosajona existen varios corpus que se han vuelto paradigmáticos y generalmente de ellos se copian los conjuntos de etiquetas. Los más conocidos son el *Brown University Corpus* (Brown), el *Lancaster-Oslo/Bergen Corpus* (LOB), el *British National Corpus* (BNC) y el *University of Pennsylvania Treebank* (Penn Treebank). Ejemplos de sus conjuntos de etiquetas se pueden encontrar en los anexos de van Halteren (1999) y en los de Brill (1993). En español algunos corpus con etiquetas de categorías gramaticales son el CRATER (*Corpus Resources and Terminology Extraction*) y el corpus del Instituto Universitario de Lingüística Aplicada (IULA) de la Universidad Pompeu Fabra.

Adoptar un estándar en lugar de copiar el conjunto de etiquetas de un corpus ya constituido tiene ventajas. El corpus se vuelve un recurso intercambiable dentro de la comunidad científica, ya que cualquier entidad puede tomar el mismo estándar y procesarlo. Adicionalmente, si se desarrolla tecnología para corpus con el mismo estándar, el corpus se vuelve reutilizable. Sin embargo, también presenta al menos una desventaja. Un estándar puede no contemplar toda la variedad de fenómenos lingüísticos en distintas lenguas. Sirva de ejemplo que para el español del siglo XVI, los lineamientos de etiquetado que adopté tomaban en cuenta las contracciones *del* y *al*, pero no otras como *dellas* (*de ellas*), *quel* (*que él*), o *porques* (*porque es*).

La Unión Europea (UE), por su carácter multilingüe, ha desarrollado iniciativas para estandarizar el trabajo en ingeniería lingüística y corpus. Una de éstas, llamada EAGLES (*Expert Advice Group for Language Engineering Standards*), ha trabajado desde 1994 en la propuesta de lineamientos de etiquetado de categorías gramaticales. Estos incluyen el danés, alemán, inglés, francés, holandés, griego, italiano, portugués y español, razón por la cual han tenido buena difusión y aceptación al menos en la Unión Europea.

La motivación detrás del trabajo de EAGLES puede leerse en el siguiente fragmento: "In the interests of interchangeability and reusability of annotated corpora and particularly for the development of multi-lingual corpora, it is important to avoid a free-for-all in tagging practices" (Leech y Wilson 1999: 58). El trabajo de esta iniciativa ha dado como resultado lineamientos que explicaré a continuación y que tienen la bondad de considerar su aplicación a diversas lenguas y a la vez brindar la posibilidad de extenderse o adaptarse a ciertas particularidades de una sola.

La codificación de EAGLES se basa en un conjunto de letras para cada categoría (atributos obligatorios), rasgos morfosintácticos (atributos recomendados) y particularidades de cada lengua (atributos opcionales), ordenadas en posiciones consecutivas. Para explicar esto voy a tomar el ejemplo de la codificación para nombres. El rasgo obligatorio para nombre es la propia categoría representada por el código N, los demás rasgos se muestran en la Tabla 3.2.

Tabla 3.2: Tabla de codificación de rasgos para nombres de EAGLES

(a)	Atributos recomendados					
	Tipo:	1. Común	2. Propio			
	Género:	1. Masculino	2. Femenino	3. Neutro		
	Número:	1. Singular	2. Plural			
	Caso:	1. Nominativo	2. Genitivo	3. Dativo	4. Acusativo	5. Vocativo
(b)	Atributos opcionales					
	Carácter contable:	1. Contable			2. Masa	

Cada rasgo tiene una letra asociada, por lo general la letra inicial, que lo representará en la secuencia final de la etiqueta. La ausencia de un rasgo sería indicado por un 0. Así, para un nombre común femenino plural, sin caso y de tipo contable se obtendría la siguiente codificación: NCFP0C. Una ventaja de esta representación es que si fuera necesario agregar rasgos, sólo bastaría agregar letras en las siguientes posiciones. Por ejemplo, si se incluye el rasgo de definición en el ejemplo

anterior, situación que se marca morfológicamente en algunas lenguas, a partir de los valores 1. Definido y 2. Indefinido, el nuevo código sería: NCFP0CD[6].

Otra iniciativa de la UE, fue el proyecto PAROLE (*Preparatory Action for Linguistic Resources Organization for Language Engineering*). Éste se unió a los esfuerzo de EAGLES y convirtió sus lineamientos en una realidad operativa cuando los aplicó en el desarrollo de un lexicón en varias lenguas (cf. Monachini y Calzolari: 1999: 172). El principal aspecto que atrae de PAROLE es que cuenta con una propuesta de etiquetado específica para el español, obtenida a partir de los lineamientos de EAGLES[7].

Una vez que se han decidido las categorías gramaticales, su especificidad y el conjunto de etiquetas para codificarlas en el corpus, será necesaria la selección de un método de etiquetado que tome esas etiquetas y las asigne a cada palabra. Por esto, en la siguiente sección, revisaré cuales son los principales métodos automáticos que se han propuesto para llevar a cabo tal proceso.

3.3. Métodos

Pongo enseguida la descripción de varios métodos que han sido desarrollados para identificar automáticamente categorías gramaticales en corpus. Los clasifico en estadísticos y basados en reglas. Los primeros utilizan preferentemente métodos y modelos estadísticos para asignar las categorías más probables y resolver los problemas de ambigüedad. Se basan en tablas o modelos de probabilidad obtenidos de corpus. Los segundos utilizan reglas de etiquetado ya sea escritas manualmente o producidas automáticamente a partir de textos previamente etiquetados[8].

Voutilainen (1999b: 9) propone una clasificación que coincide en algunos aspectos con la anterior. Los divide en métodos lingüísticos y conducidos por los datos (*data-driven*). A los primeros los caracteriza por que expertos en gramática escriben reglas basadas en su conocimiento lingüístico. En los segundos se crea un modelo del lenguaje a partir del análisis estadístico de gran

[6] Los lineamientos completos de EAGLES pueden verse en http://www.ilc.cnr.it/EAGLES/home.html (última visita 01/02/09) y un resumen en Leech y Wilson (1999).

[7] Esta propuesta de PAROLE es precisamente la que uso en la tesis y que describiré a detalle en el capítulo **¡Error! No se encuentra el origen de la referencia.**.

[8] Un aspecto muy importante que debo aclarar es que una regla de etiquetado puede ser una regla lingüística en el sentido de que represente conocimiento gramatical, pero también puede referirse a reglas heurísticas, que son efectivas en la práctica aunque lingüísticamente falsas; por lo general, para un método automático, siempre se complementan (cf. Voutilainen 1999b: 9). Cuando hablo de regla, no hago referencia a las reglas de tipo generativo transformacional.

cantidad de texto etiquetado manualmente; el modelo estadístico obtenido se puede representar en matrices de colocación, modelos de Markov, reglas locales o redes neuronales. Entonces, primero doy una breve historia del desarrollo de programas para etiquetado y luego abordo en secciones separadas los métodos estadísticos y basados en reglas.

Los primeros trabajos en etiquetadores automáticos fueron desarrollados a finales de los años cincuenta. Estaban basados en reglas escritas manualmente y lexicones que incluían todas las posibles categorías de una palabra. Las palabras que no aparecían en el lexicón eran analizadas con base en ciertos caracteres al inicio o final de la palabra y algunas otras características gráficas. La desambiguación se realizaba con reglas escritas por lingüistas, éstas tomaban en cuenta el contexto a la izquierda y derecha de la palabra. Al final, cualquier mal etiquetado era corregido manualmente. En México, un etiquetador de este tipo fue desarrollado en la década de los sesenta en El Colegio de México para analizar el Corpus del Español Mexicano Contemporáneo. Este analizador gramatical, del cuál hablaré más adelante en un apartado especial (véase *infra* p. 86), estaba basado en reglas elaboradas por un grupo de lingüistas. Tales reglas eran de carácter morfológico y contextual.

Después llegaron los programas que usaron modelos de estados finitos para representar una pequeña gramática usada en el proceso de desambiguación, según Voutilainen (1999b: 11) el primero se desarrolló en la Universidad de Pensilvania. Uno de los primeros en utilizar gran cantidad de texto fue TAGGIT y usó el corpus Brown para obtener un modelo de desambiguación.

En los años setenta surgió un nuevo modelo de etiquetador basado en estadísticas de bigramas y desarrollado con ayuda del corpus LOB, su nombre fue CLAWS. Éste abandonó prácticamente el uso de reglas y en su lugar utilizó bigramas de palabras y etiquetas asociados a ciertas probabilidades a partir de las cuales decidía la etiqueta a utilizar. Sus resultados fueron de hasta 97% de precisión (cf. Voutilainen 1999b: 13). Posteriores etiquetadores fueron desarrollados basados en CLAWS, los más conocidos hoy en día son los de Church (1988) y DeRose (1988).

A finales de los ochenta, surgió otra propuesta que derivaba reglas de desambiguación automáticamente de un corpus etiquetado (cf. Hindle 1989). Las reglas estaban ordenadas de acuerdo con su efectividad y podían referirse a palabras o etiquetas. Unas reglas podían corregir errores de otras. El programa generaba reglas de desambiguación a partir de reglas preestablecidas, si su efectividad era mejor se ponían antes de las preestablecidas de lo contrario se ordenaban abajo. A partir de 350 reglas hechas manualmente, se generaron al final 35,000 reglas ordenadas (cf. Hindle 1989).

Actualmente, es posible encontrar una gran variedad de métodos para etiquetar. Una práctica común es adoptar técnicas estadísticas o de inteligencia artificial y aplicarlas a esta labor. Entre ellos, puedo mencionar los que usan redes neuronales, árboles de decisión y los basados en casos. Sin embargo, existen también los de tipo híbrido, que realizan unas labores con técnicas estadísticas y otras mediante el uso de reglas. Algunos de éstos usan varios métodos para la misma tarea y comparan cuál resulta mejor o permiten una "votación" por una determinada solución[9].

Para terminar esta sección, quiero resaltar que en los años noventa también se trabajaron métodos basados en reglas escritas manualmente. Uno de ellos utilizó una gramática de restricciones. Ésta se trata de una serie de reglas que aceptan o rechazan cierto análisis en el texto. Si el análisis pasa todas las restricciones (reglas) entonces se toma como correcto. Según Voutilainen (2003: 228) estos métodos logran reducir los niveles de error principalmente en lenguas con morfología compleja y de orden libre de palabras[10].

He presentado arriba la evolución de los métodos de identificación de categorías y los métodos principales que han sido usados. Por la importancia que tienen los de tipo estadístico, explicaré ahora, de manera general, su funcionamiento.

3.3.1. Métodos estadísticos

Como se vio anteriormente, no pasó mucho tiempo desde las primeras propuestas de etiquetado para que los modelos estadísticos fueran integrados a este tipo de programas. Al principio se utilizaron estadísticas de bigramas y posteriormente se incorporó el uso de modelos ocultos de Markov. Estaban basados en la idea de hacer un pequeño modelo lingüístico a partir de un corpus y así proponer una etiqueta como más probable. Hoy en día estos métodos siguen siendo utilizados, aunque con ciertas variantes, por lo que expondré en los siguientes apartados las bases de funcionamiento de estos métodos.

3.3.1.1.El método más simple

Bajo una mirada estadística, el problema de la asignación de categorías gramaticales puede entenderse, en un principio, como la asignación de la etiqueta más probable para una palabra. El procedimiento más simple para llevar a cabo esta afirmación consistiría en asignar a cada palabra

[9] Para más información sobre estos sistemas véase Van Halteren (1999) especialmente el capítulo 17.

[10] Puede leerse más acerca de estos métodos en Voutilainen (1999b: 18, 2003: 226-230) y en el capítulo 14 de Van Halteren (1999).

del corpus la etiqueta más probable de manera aislada, es decir, sin tomar en cuenta ningún contexto lingüístico. Esto es, a partir del conteo de etiquetas asociadas a cada palabra en un corpus, se asigna siempre a determinada palabra la etiqueta más frecuente. Allen (1995) dice que con este método se obtiene un noventa por ciento de exactitud para la lengua inglesa[11].

Por ejemplo, en el Corpus del Español Mexicano Contemporáneo (cf. Ham 1979: 54-55) la palabra *la* es 87,827 veces un artículo y 4,259 veces un pronombre. Con estos datos, y tomando el sencillo método explicado antes, se etiquetaría cada palabra *la* de nuestro corpus con la etiqueta de artículo. El gran problema con este método es que la ambigüedad categorial es recurrente en la lengua y el desarrollo de un etiquetador es interesante precisamente por su proceso de desambiguación.

Una mejor estrategia sería tomar en cuenta el contexto lingüístico, es decir, las palabras que acompañan a la palabra en cuestión y sus respectivas etiquetas de categorías. Además, como ya se dijo, es un hecho que la identificación de categorías debe atender aspectos sintácticos. Por ejemplo, la categoría de la palabra que sigue a un determinante es más probable que sea un sustantivo. Para esto sirve la probabilidad condicional que explico en seguida.

3.3.1.2.La probabilidad condicional

La probabilidad de que suceda un evento cuando otro evento ha sucedido es conocida como probabilidad condicional. La fórmula para determinarla se encuentra en (3). De esta fórmula se nota que la probabilidad de que ocurra el evento A, dado que ha ocurrido el evento B, se obtiene de dividir la probabilidad de que ocurran ambos entre la probabilidad de que ocurra el primero.

(3) $$\text{PROBABILIDAD(evento B} \mid \text{evento A)} = \frac{\text{PROBABILIDAD(evento A y evento B)}}{\text{PROBABILIDAD(evento A)}}$$

Regresando a la identificación de categorías en corpus, el problema entonces consistiría en encontrar la secuencia de categorías más probable para una determinada secuencia de palabras. Si pensamos en C_1, C_2, C_3…,C_T como una secuencia de categorías y en w_1, w_2, w_3…,w_T como una secuencia de palabras, se puede expresar el problema con la fórmula de (4). Este planteamiento

[11] Hay que tomar esta afirmación con reserva, ya que, entre otros factores, la especificidad del conjunto de etiquetas repercute en la exactitud del etiquetado: a mayor detalle de rasgos la precisión suele disminuir.

permitiría conocer cuál es la probabilidad de una secuencia de categorías dada una secuencia de palabras en un corpus.

(4) $\text{PROBABILIDAD}(C_1,\ldots,C_T \mid w_1,\ldots,w_T)$

Pero existe un problema con la solución propuesta en (4) ya que el número de datos requeridos para lograr estimaciones razonables debe ser muy grande (cf. Allen, 1995). Por lo tanto, es necesario tomar una aproximación al problema. Gracias a la regla de Bayes[12], se reduce el problema a la búsqueda de una secuencia de categorías que brinde la mayor probabilidad para la fórmula de (5). Esto es, la probabilidad de la secuencia de categorías por la probabilidad de que dada esa secuencia, le corresponda una determinada secuencia de palabras.

(5) $\text{PROBABILIDAD}(C_1,\ldots,C_T) * \text{PROBABILIDAD}(w_1,\ldots,w_T \mid C_1,\ldots,C_T)$

Sin embargo, nuevamente es obligado manejar muchos datos para obtener probabilidades razonables de secuencias de palabras y cadenas tan largas. Por lo cual, es necesario utilizar aproximaciones adicionales al problema. La que ha demostrado ser muy útil es la basada en modelos de bigramas o trigramas, que explico a continuación.

3.3.1.3.Modelo de bigramas y trigramas

Ya que una de las limitantes del método presentado hasta ahora es la gran cantidad de operaciones para calcular la probabilidad de una secuencia de categorías, resulta necesario simplificar esta cantidad de operaciones mediante ciertas aproximaciones. Para la primera parte de la fórmula de (5), la aproximación consistiría en limitar el número de categorías necesarias para obtener la probabilidad de una secuencia de ellas.

Es decir, en lugar de tomar en cuenta todas las categorías para determinar la probabilidad, se tomarán en cuenta sólo algunas. Así, el modelo de bigramas propone usar dos elementos y el modelo de trigramas propone usar tres. Estos modelos pueden generalizarse como un modelo de n-gramas, donde *n* representa el número de elementos a tomar en cuenta.

---

[12] La regla de Bayes o ley de Bayes establece una relación entre la probabilidad condicional de un evento A cuando ha ocurrido un evento B, con la probabilidad condicional del evento B dado el evento A:

$$\text{PROBABILIDAD(evento A | evento B)} = \frac{\text{PROBABILIDAD(evento B | evento A)} * \text{PROBABILIDAD(vento A)}}{\text{PROBABILIDAD(evento B)}}$$

Para el problema de la identificación de categorías, el modelo de bigramas establecería que la probabilidad de una secuencia de categorías estaría determinada por el producto de las probabilidades condicionales de una categoría y su categoría precedente[13]. Así, la primera parte de la expresión de (5) puede ahora aproximarse de la siguiente manera (6).

(6) $$\prod_{i=1}^{T} \text{PROBABILIDAD}(Ci \mid Ci\text{-}1)$$

Para la segunda parte de la función de (5) es necesaria otra aproximación. Por la utilidad que esto brinda, se asume que una palabra aparece asociada a una categoría de forma independiente de las palabras y categorías anteriores, situación que es falsa, pero útil. Así, es posible reformular la operación como el producto de las probabilidades de que una palabra tenga determinada categoría (7).

(7) $$\prod_{i=1}^{T} \text{PROBABILIDAD}(wi \mid Ci)$$

Con todo lo anterior, se obtiene una aproximación que cumple con dos características. Por un lado, el número de operaciones para calcularla es razonable, y por otro, las probabilidades pueden ser obtenidas a partir de un corpus ya etiquetado. Esta aproximación sería la de (8).

(8) $$\prod_{i=1}^{T} \text{PROBABILIDAD}(Ci \mid Ci\text{-}1) * \text{PROBABILIDAD}(wi \mid Ci)$$

La fórmula de (8) permite calcular la probabilidad de todas las secuencias de categorías posibles para un enunciado con el fin de seleccionar la mayor. A pesar de la simplificación hecha mediante aproximaciones, todavía se tiene un número de posibles secuencias igual al número de categorías (N) elevado al número de palabras (T), esto es, N^T posibles secuencias.

[13] Charniak menciona que: "This model makes the drastic assumption that only the previous n – 1 words have any effect on the probabilities for the next word. While this is clearly false, as a simplifying assumption it often does a serviceable job" (1993: 39).

Ya que se asumió que la probabilidad de aparición de una categoría depende solamente de la categoría anterior, es posible modelar el problema como un modelo oculto de Markov, el cual ha demostrado ser muy útil en procesamiento de corpus. En la siguiente sección explico en qué consiste.

3.3.1.4.Modelos ocultos de Markov

Un modelo oculto de Markov involucra una serie de tuplas de cuatro elementos: un conjunto de estados, un estado inicial, un conjunto de símbolos de salida y un conjunto de transiciones. A su vez, las transiciones se forman de un estado inicial, un estado final, un símbolo de salida y la probabilidad de que la transición sea realizada. Un estado inicial puede tener varias transiciones que produzcan un mismo símbolo de salida, pero que vayan a distintos estados finales. Por tanto, no es posible conocer la transición que tomó el modelo a partir del símbolo de salida, esto es lo que le da el carácter de oculto. El modelo asume que la probabilidad de un símbolo de salida o de un estado siguiente depende sólo del estado anterior.

Entonces, cada estado correspondería a una categoría y las transiciones tendrían la probabilidad de que una categoría siga a otra. Además, habría una probabilidad de salida correspondiente a la probabilidad de que esa categoría sea llenada por una determinada palabra. Para obtener la probabilidad de que una secuencia de categorías genere una secuencia de palabras, se multiplicarían las probabilidades del camino seguido en el modelo de Markov por las probabilidades de salida de las palabras en cuestión.

Una ventaja de estos modelos ocultos de Markov, es que se pueden aplicar procedimientos que recorran de manera óptima el conjunto de estados. Precisamente, en el caso del modelo anterior, evaluar todos los caminos posibles, es decir, todas las combinaciones de etiquetas, no es óptimo. Por tal razón, se puede usar el algoritmo de Viterbi, que permite encontrar un camino de forma optimizada. Enseguida presento de manera breve su funcionamiento.

3.3.1.5.El algoritmo de Viterbi

Como se dijo antes, obtener todas las posibles secuencias de todas las categorías y rescatar la mayor no es tan óptimo. Por ejemplo, si hubiera cuatro categorías y cuatro palabras, la cantidad de combinaciones de categorías sería de 4^4=256. Haciendo uso del algoritmo de Viterbi es posible disminuir el número de combinaciones y cálculos necesarios para determinar la secuencia de categorías más probable.

Para cada palabra, el algoritmo obtiene una probabilidad por cada categoría posible tomando en cuenta la categoría que se presentó en el estado anterior. Del conjunto de probabilidades anteriores, se obtiene la máxima para ser multiplicada por la probabilidad de que esa palabra aparezca con una determinada categoría. Esto evita generar todas las combinaciones posibles, usando sólo aquellas que son máximas, lo que reduce los caminos posibles de solución del problema, es decir, las posibles combinaciones de categorías[14].

3.3.1.6.Ejemplos de etiquetadores estadísticos

He descrito, de manera muy general, el método fundamental para etiquetado estadístico. A partir de él han surgido otros que incluyen distintas variantes. Entre los más conocidos está el llamado *Trigrams'n'Tags* (TnT), que ha reportado niveles de etiquetado de 96.7%. Fue desarrollado para el alemán por Brants (2000) y tiene la peculiaridad de utilizar análisis al interior de la palabra para determinar la categoría de las palabras que no están en el lexicón: las palabras menos frecuentes eran analizadas con base en sus diez letras finales. Además, está basado en trigramas de etiquetas.

Según Charniak (1996/1993), los métodos estadísticos permiten lograr niveles de exactitud de hasta el 95% para lengua inglesa. Además, han sido probado en otras lenguas como inglés, sueco, chino, francés y alemán (cf. Voutilainen 1999b). Para español, es posible encontrar propuestas de tipo estadístico para identificar categorías gramaticales. Una de ellas es la de Pla, Molina y Prieto (2001) que utiliza bigramas lexicalizados y obtiene una precisión de 97.42%. Otra, que toma el método de Brants visto arriba y lo aplica al español, es la de Morales y Gelbukh (2003).

Revisé hasta aquí los métodos estadísticos para etiquetado de categorías gramaticales, los cuales siguen siendo producto de innovaciones metodológicas. Por lo anterior y porque, como se ha visto, obtienen muy buenos niveles de precisión, siguen siendo una fuerte tendencia para resolver los problema del etiquetado. En la siguiente sección abordaré los métodos basados en el uso de reglas, ya sea escritas manualmente o inferidas de un corpus.

3.3.2. Métodos basados en reglas

Ya he mencionado las tendencias en el uso de reglas para identificar categorías gramaticales. Al respecto, existen tres posibilidades: hacerlas manualmente aprovechando el conocimiento lingüístico, generarlas a partir de un corpus etiquetado o hacer una combinación de ambas. La

[14] Una explicación más amplia del algoritmo de Viterbi para el problema de asignación de categorías puede encontrarse en Allen (1995: 202-203).

creación de reglas por parte de lingüistas, al menos para Voutilainen (1999b, 2003), no debe ser olvidada y según este autor brinda resultados satisfactorios.

Para el caso de la lengua española, Jiménez y Morales (2002) propusieron un programa llamado SEPE fundamentado en varias técnicas de etiquetado, entre ellas, la incorporación de reglas morfosintácticas. Parte de estas reglas utilizan una lista de sufijos, los nominales obtenidos de Moreno de Alba (1986) y los verbales de un programa de computadora ya existente. SEPE combina eficientemente, reglas, diccionarios y procedimientos computacionales de clasificación basados en árboles de decisión construidos a partir de medidas estadísticas.

En los dos apartados siguientes, expongo otros métodos basados fundamentalmente en reglas. En primer lugar recupero una propuesta valiosa desarrollada en El Colegio de México en los años 70. En segundo lugar, y de manera amplia, expongo el método sobre el que baso mi trabajo de investigación.

3.3.2.1.El analizador gramatical del DEM

Quiero recuperar en esta sección un trabajo que representa el inicio de las investigaciones en lingüística computacional en México, y probablemente en el mundo de habla hispana. Se trata de un analizador gramatical utilizado en los estudios lexicográficos para la elaboración del Diccionario del Español de México (DEM) en la década de los setenta. Lo primero que llama la atención es que el objetivo de procesar corpus electrónicamente no sólo era ahorrar trabajo, sino dar objetividad y regularidad al análisis lingüístico.

El objetivo de este analizador fue asignar automáticamente categorías gramaticales a las palabras del Corpus del Español Mexicano Contemporáneo (CEMC), con la idea de producir listas de palabras, concordancias y estadísticas. El método de solución para tal consigna fue el desarrollo de varios tipos de reglas: reglas basadas en la morfología de las palabras, reglas de precedencia y reglas de postcedencia (cf. García Hidalgo 1979: 89).

El procedimiento general del analizador, omitiendo muchos detalles, es el siguiente. Las palabras fueron asociadas a clases dependiendo de su comportamiento distribucional y mediante criterios gramaticales elaborados por el grupo de lingüistas del DEM. Si dos palabras ocurrían en los mismos contextos se clasificaban como iguales. Luego, esas clases fueron asociadas a categorías gramaticales de acuerdo con ciertas generalizaciones.

Palabras ambiguas, con dos o tres categorías posibles, eran analizadas por un algoritmo morfológico definido a partir de una lista de sufijos verbales. Este análisis determinaba la pertenencia de una palabra a la categoría de verbo y además obtenía una raíz para asociar otras

palabras a manera de un paradigma flexivo. Finalmente, para palabras que se mantenían como ambiguas, las reglas de precedencia y postcedencia permitían etiquetarlas a partir de la categoría de la palabra anterior o posterior. Los resultados reportados en el etiquetado del CEMC fueron de un 55% y de 67.5% para el género de literatura.

No obstante que sólo muestro a grandes rasgos el procedimiento de este analizador, creo que quedan claras las características principales del método[15]. Por un lado, está fundamentado en consideraciones de tipo lingüístico expresadas mediante reglas, lo que brindó la posibilidad de etiquetar el corpus sin ningún tipo de etiquetado previo. Por otro lado, muestra la importancia de tomar en cuenta la morfología y la sintaxis para la identificación de las categorías.

Otro método que también utiliza reglas de tipo morfológico y sintáctico, pero adquiridas automáticamente a partir de un corpus etiquetado, es el que propuso Eric Brill en los noventa. Ya que en este método fundamento el trabajo de la tesis, lo expongo en el siguiente apartado.

3.3.2.2.El método de Brill

En este apartado voy a revisar, de manera general, la propuesta desarrollada por Eric Brill (1992, 1993, 1994, 1995) para la asignación de categorías gramaticales. En el siguiente apartado lo haré con mayor detalle. Éste se ha vuelto uno de los métodos más conocidos y utilizados por su sencillez y efectividad. Lo adopté para mi trabajo de tesis porque, como se verá, incluye el uso de información de carácter morfológico y sintáctico en módulos separados. Esta situación permite incluir la lista de afijos descubiertos de manera más efectiva. Además, capta información gramatical en reglas que si bien no serían propiamente reglas lingüísticas, sí representarían conocimiento empírico para la identificación de categorías.

Aunque, como ya mencioné, los métodos estadísticos son muy efectivos para el etiquetado de categorías gramaticales, generalmente necesitan de corpus de gran tamaño (millones de palabras) para encontrar probabilidades útiles para el etiquetado. En mi caso, cuento con un corpus etiquetado relativamente pequeño[16] (16,500 palabras aproximadamente, que deben repartirse en dos partes, una

[15] Puede verse a detalle el procedimiento, su formalización matemática y su implementación computacional en García Hidalgo (1979).

[16] Para clasificar un corpus como pequeño o grande no se debe tomar en cuenta únicamente el número de palabras que lo conforman. También se debe atender al tipo de textos, la dificultad o facilidad para obtenerlos, el tipo de hablante que los produce, entre otros aspectos. Sin embargo, consideraré el corpus de estudio usado en esta tesis como pequeño en comparación con los corpus comúnmente usados en lingüística computacional y procesamiento de lenguaje natural, que suelen tener cientos de miles, o millones, de palabras.

de entrenamiento y otra para evaluación), pero suficiente para que el método de Brill genere reglas útiles. Ésta es la razón principal por la que seleccioné éste método en lugar de uno estadístico. Otra razón es que el conocimiento lingüístico obtenido con un método estadístico se representa principalmente en tablas de probabilidades, las cuales pueden ser menos fáciles de interpretar que las reglas obtenidas con el método de Brill.

La propuesta de Brill pertenece al ámbito de la inteligencia artificial, específicamente a la parte del llamado aprendizaje de computadora[17]. Así, llama a su método: *aprendizaje basado en transformaciones y dirigido por errores*. Siendo muy breve, éste tiene por objeto obtener reglas de transformación[18] que permitan identificar categorías gramaticales a partir de la generación de todos los cambios de etiquetas posibles y verificando los errores que estos producen. La salida de este método es un conjunto de reglas ordenadas, que indican las condiciones y contextos en los cuales deberá ser asignada una etiqueta. El corpus a procesar comenzaría con un etiquetado simple y la aplicación ordenada de las reglas mejoraría su estado inicial.

Para obtener el conjunto de transformaciones ordenadas, se compara el corpus etiquetado inicialmente con el etiquetado por un experto, esto permite determinar su número de errores. Después, son probadas todas las posibilidades de etiquetado para cada palabra, y medida estadísticamente la mejoría que cada una produce. Una vez obtenidas las medidas de todas las posibilidades, se toma la mejor (mayor reducción de errores) y se coloca al final de la lista de reglas.

El proceso se repite nuevamente hasta obtener una calidad del corpus suficientemente cercana al corpus etiquetado por el experto. Las posibilidades de etiquetado para cada palabra dependen de ciertas condiciones expresadas como plantillas de transformaciones. Una plantilla tiene dos posibles formas que se muestran en (9).

---

[17] El Aprendizaje de Computadora (*Machine Learning*) forma parte de los estudios en inteligencia artificial que tratan de representar, mediante modelos, el conocimiento humano y darle a un programa de computadora la posibilidad de "aprender" de ese modelo. Si un programa puede simular tomar una decisión a la manera de un humano, como la de clasificar un ejemplo en cierta categoría, podemos decir que ha simulado el aprendizaje. Para revisar algunos aspectos del aprendizaje de computadoras relacionado con la identificación de categorías puede verse Walter Daelemans. 1999. "Machine Learning Approaches", en Hans Van Halteren (ed.). *Syntactic Wordclass Tagging*. Dordrecht, Netherlands: Kluwer Academic, entre otros.

[18] Debe entenderse una regla de transformación como una regla de etiquetado que cambiar una etiqueta por otra de acuerdo con una determinada condición. Nada tiene que ver con alguna postura generativa transformacional de la lingüística.

(9) Formas de plantillas de transformación

(a) Cambia la etiqueta X por la etiqueta Y si se da la condición W.

(b) Cambia la etiqueta actual (no importa cuál sea) por la etiqueta Y si se da la condición W.

Un ejemplo para una condición (*W*), que produciría una transformación, podría ser: "si la última letra de la palabra es z". Donde *z* puede ser cualquier letra vista en el corpus. Así, de una plantilla de transformación es posible obtener un conjunto de transformaciones o reglas de transformación aplicables y evaluables para determinar cuál es la mejor. A continuación muestro un ejemplo de esta situación.

Dado el conjunto de etiquetas {NCS: nombre común singular, NCP: nombre común plural}, las últimas letras de las palabras de un corpus {a, s, o, e}, y la plantilla de reglas de transformación: "cambia la etiqueta actual (no importa cuál sea) por la etiqueta Y si la última letra de la palabra es z", se obtiene el conjunto de reglas de transformación posibles puesto en (10).

(10) Conjunto de reglas obtenidas a partir de una plantilla

Cambia la etiqueta actual por NCS si la última letra de la palabra es a
Cambia la etiqueta actual por NCS si la última letra de la palabra es s
Cambia la etiqueta actual por NCS si la última letra de la palabra es o
Cambia la etiqueta actual por NCS si la última letra de la palabra es e
Cambia la etiqueta actual por NCP si la última letra de la palabra es a
Cambia la etiqueta actual por NCP si la última letra de la palabra es s
Cambia la etiqueta actual por NCP si la última letra de la palabra es o
Cambia la etiqueta actual por NCP si la última letra de la palabra es e

Como se puede ver, todas las posibilidades de cambio de etiquetas son obtenidas a partir de la plantilla de transformación, la sustitución de *Y* por cada etiqueta, y de *z* por cada letra final. Tal como lo mencionaba antes, éstas son aplicadas para determinar cuál es la mejor de acuerdo con la comparación con un corpus etiquetado por un experto. Es importante resaltar que la evaluación de las reglas es automática en el sentido de que no interviene el lingüista para decidir cuál es mejor o peor. Más bien, mediante el conteo de errores se determina cuál regla supera a las otras, situación que se refleja en el ordenamiento final de las reglas.

Las reglas que quedan ordenadas en la salida final del método no son todas las generadas a partir de las plantillas, sino únicamente aquellas que mejoraron la calidad del etiquetado del corpus. Más adelante entraré en mayor detalle y esto quedará más claro.

En el analizador gramatical usado en el DEM, por ejemplo, fueron lingüistas los que determinaron previamente las mejores reglas. Se puede decir entonces que en ese caso eran las mejores reglas lingüísticamente hablando, pero en el caso del método de Brill son reglas que mejor se acercan al etiquetado del corpus de comparación; lo interesante es que captan de manera automática el conocimiento implícito en el corpus para el problema del etiquetado de categorías gramaticales.

Para Brill (1993: 14-15) es posible manejar tres variables en su método: la calidad del etiquetado inicial del corpus, los tipos de plantillas de transformación y el grado de anotación del corpus de comparación previamente etiquetado. En este sentido, su propuesta brinda algunas ventajas. El etiquetado inicial del corpus puede ser muy simple, al grado de comenzar con un corpus donde todas las palabras tengan la misma categoría, por ejemplo, nombre común. A partir de aquí, el método obtendrá el conjunto de reglas necesario para mejorar este etiquetado simple.

Al respecto, las plantillas de transformación podrían ser muy variadas. Se pueden comparar distintos números de letras finales o iniciales de palabra (el método original de Brill utiliza uno, dos, tres o cuatro). Además, pueden tomar como base de comparación la palabra misma o las etiquetas o palabras aledañas. En términos generales, Brill propone dos tipos generales de plantillas, aquellas de nivel léxico y otras de nivel contextual. Las primeras no pasan la frontera de la palabra, las segundas sí lo hacen[19].

Con respecto a la tercera variable mencionada arriba, este autor indica que una gran ventaja de su propuesta es que con muy poco corpus etiquetado es posible comenzar a generar reglas. Ya que decidí crear mi propio corpus etiquetado, esta situación nos brindó una ventaja decisiva en la selección de este método; la cantidad de corpus que se obtuvo del proceso de etiquetado manual resultó muy poca en comparación con los enormes corpus etiquetados para el inglés (véase *supra* nota 16).

Después de esta introducción al método basado en reglas de Brill, voy a revisar con mayor detalle su propuesta poniendo énfasis en el procedimiento que sirvió de base para esta tesis.

---

[19] Para ver a detalle las posibles plantillas de reglas propuestas por este autor puede revisarse Brill 2005 (555-559). Más adelante mostraré las usadas en la tesis.

3.3.2.3.El método de Brill a detalle

A continuación detallaré el procedimiento propuesto por Eric Brill para la identificación de categorías gramaticales en corpus. Es muy importante esta sección ya que permitirá entender las modificaciones hechas al método como parte de las propuestas de la presente tesis y para entender los resultados obtenidos de éstas. Explicaré la propuesta de este autor en dos secciones: el proceso de etiquetado de nuevos corpus mediante la aplicación de las reglas de transformación previamente generadas y el proceso de entrenamiento o generación de esas reglas. Creo que explicar primero cómo se etiqueta un corpus permitirá entender mejor el proceso de entrenamiento.

3.3.2.3.1. Proceso de etiquetado

El proceso de etiquetado, como ya expliqué, es equivalente a asignar a cada palabra del corpus una etiqueta que represente su categoría gramatical. La entrada a este proceso es un corpus sin etiquetas, resultado, por ejemplo, del trabajo de recopilación de algún grupo de investigación. La salida será el mismo corpus, pero anotado o etiquetado con la categoría gramatical de cada palabra. Para asignar estas etiquetas se utilizan tres recursos: un lexicón, un conjunto de reglas de transformación de carácter léxico y un conjunto de reglas de transformación de carácter contextual. Enseguida explicaré en qué consiste cada uno y cómo se utiliza.

El lexicón contiene una lista de palabras y su categoría más probable expresada por una etiqueta. Este recurso es creado a partir de un corpus etiquetado previamente. Los dos conjuntos de reglas, léxicas y contextuales, son obtenidos del proceso de entrenamiento o generación de reglas que revisaré en la siguiente sección. Por el momento, daré por hecho que se entiende este proceso y que se cuenta con la lista de reglas ordenadas.

El conjunto de categorías gramaticales que serán asignadas al corpus nuevo dependen directamente del conjunto utilizado en el corpus etiquetado por el experto. En otras palabras, en el proceso de entrenamiento, como se verá adelante, no se infieren nuevas etiquetas.

El procedimiento de etiquetado puede ser dividido en tres pasos: etiquetado de estado inicial, etiquetado léxico y etiquetado contextual. A manera de algoritmo, presento enseguida estos tres pasos:

ETIQUETADO DE ESTADO INICIAL

1. Para cada tipo de palabra (*type*) del corpus nuevo, que aparece en el lexicón:
    1.1. Se le asigna la etiqueta más probable registrada en el lexicón.
2. Para cada tipo de palabra del corpus nuevo, que no aparece en el lexicón:

2.1. Si comienza con mayúscula entonces se le asigna la etiqueta de nombre propio (NP00), si no, entonces se le asigna la etiqueta de nombre común masculino singular (NCMS).

ETIQUETADO LÉXICO

1. Para cada tipo de palabra (*type*) del corpus nuevo etiquetado hasta el momento, que no aparece en el lexicón:
   1.1. Se aplican cada una de las reglas de carácter léxico en el orden establecido.

ETIQUETADO CONTEXTUAL

1. Para todas las palabras (*tokens*) del corpus nuevo etiquetado hasta el momento:
   1.1. Se aplican cada una de las reglas de carácter contextual en el orden establecido.

Como mencioné arriba, el etiquetado de estado inicial es un etiquetado muy simple. De hecho, podría haberse etiquetado todos los tipos de palabra que no estuvieran en el lexicón con una sola etiqueta, por ejemplo, nombre común masculino singular. Pero con el fin de ejemplificar el proceso, optaré por el etiquetado de estado inicial mostrado arriba. Esto produce un nuevo corpus etiquetado a partir del cual se aplican los siguientes procesos. Ya que muchas de las palabras etiquetadas hasta ahora como nombre común masculino singular o nombre propio seguramente pertenecen a otra categoría, se aplican en ellas las reglas de carácter léxico con el fin de darles una categoría más cercana a la correcta.

Para determinar la categoría de las palabras no vistas en el lexicón, son aplicadas las reglas de carácter léxico que miran a la palabra y sus rasgos. El método de Brill propone el uso de los siguientes tipos de reglas Tabla 3.3.

Tabla 3.3: Tipos de reglas léxicas de Brill

	Regla	Interpretación
(a)	Si el tipo de palabra tiene uno, dos, tres o cuatro letras iniciales o finales, por ejemplo: *NCMS a* FHASSUF *1 NCFS*	Si el tipo de palabra tiene etiqueta nombre común masculino singular y la última letra es una "a", entonces cambia su etiqueta por nombre común femenino singular.

Tabla 3.3: Tipos de reglas léxicas de Brill (continuación)

	Regla	Interpretación
(b)	Si agregar o quitar uno, dos, tres o cuatro letras al inicio o final del tipo de palabra resulta en un tipo de palabra del lexicón, por ejemplo: *r* ADDSUF *1 VMIP3S0*	Si agregar la letra "r" al final del tipo de palabra resulta en un tipo de palabra del lexicón, entonces cambia su etiqueta por verbo indicativo presente tercera persona singular (por ejemplo, en la frase *se llama Xicalango*, agregar una "r" al final de *llama* resulta en una palabra del lexicón: *llamar*, y el tipo de palabra es etiquetado como verbo indicativo presente tercera persona singular: *llama/VMIP3S0*).
(c)	Si aparece con cierto tipo de palabra anterior o siguiente, por ejemplo: *NCMS la* FGOODRIGHT *NCFS*	Si el tipo de palabra tiene la etiqueta nombre común masculino singular y a su izquierda aparece el tipo de palabra "la", entonces cambia su etiqueta a nombre común femenino singular.
(d)	Si aparece cierta letra al interior del tipo de palabra, por ejemplo: *Ç* CHAR *NP00*	Si el tipo de palabra tiene la letra mayúscula Ç, entonces cambia su etiqueta a nombre propio.

Obsérvese en la Tabla 3.3 (a) y (b) que la condición de la regla considera uno, dos, tres o cuatro letras iniciales o finales. Esta es la manera como Brill incluye en su procedimiento de etiquetado características morfológicas de las palabras, situación que no es óptima para lenguas de morfología distinta el inglés. Al respecto, la idea de esta tesis es sustituir esta situación por afijos obtenidos con métodos lingüísticos.

Es importante resaltar algunos otros aspectos sobre el conjunto de reglas. Por una parte, recordemos que están en determinado orden, lo que significa que unas se aplican después de otras y puede ser posible que cambien el etiquetado de las anteriores. Por esto, el conjunto de reglas está ordenado de las más generales a las más específicas. En otras palabras, las primeras reglas suelen aplicarse a muchos tipos de palabras y las siguientes tienden a corregir casos específicos de ciertos tipos.

Para continuar con el proceso de etiquetado, enseguida de las reglas léxicas se aplican las reglas de carácter contextual, cuya finalidad es corregir el etiquetado logrado hasta el momento. Las etiquetas de cada una de las palabras del corpus etiquetado con los pasos anteriormente descritos son cambiadas de acuerdo con las reglas contextuales.

Dos aspectos importantes resaltan de este etiquetado contextual. Primero, se basa en el análisis de ocurrencias de palabras (*tokens*) y no de tipos de palabras como el de carácter léxico. Segundo, miran el contexto de la palabra a etiquetar, ya sea a partir de las palabras o de las etiquetas

aledañas. A continuación, muestro algunos de los tipos de reglas que se aplican en esta parte del proceso (Tabla 3.4).

Tabla 3.4: Tipos de reglas contextuales de Brill

	Regla	Interpretación
(a)	Si la palabra aparece antes o después de cierta etiqueta, por ejemplo: *DA0NS0 PP3MS00* NEXTTAG *VMIS3S0*	Si la palabra tiene la etiqueta de determinante adjetivo neutro singular y la siguiente palabra tiene etiqueta verbo indicativo pasado simple tercera persona singular, entonces cambia su etiqueta por pronombre personal tercera persona masculino singular.
(b)	Si la palabra aparece antes o después de cierta etiqueta, por ejemplo: *VMP00SM NCMS* PREVTAG *DP*3C*S0*	Si la palabra tiene la etiqueta de verbo participio singular masculino y la palabra anterior tiene etiqueta determinante posesivo tercera persona género común singular, entonces cambia su etiqueta por nombre común masculino singular.
(c)	Si cierta etiqueta aparece una o dos posiciones antes de la palabra, por ejemplo: *DA0NS0 PP3MS00* PREV1OR2TAG *RN*	Si la palabra tiene la etiqueta de determinante artículo neutro singular y una o dos posiciones antes aparece la etiqueta adverbio de negación, entonces cambia su etiqueta por pronombre personal tercera persona masculino singular.
(d)	Si la palabra es determinada palabra, por ejemplo: *VMIS3S0 NCMS* CURWD ayuno	Si la palabra tiene la etiqueta de verbo indicativo pasado simple tercera persona singular y es la palabra "ayuno", entonces cambia su etiqueta por nombre común masculino singular.

A diferencia de las reglas léxicas, para que la transformación se dé, la ocurrencia de la palabra siempre debe estar etiquetada con una determinada etiqueta. En el tipo de reglas léxicas, había algunas que cambiaban la etiqueta sin importar cuál era la que tenía en ese momento la palabra a etiquetar. Por otra parte, como puede verse en (b) de la Tabla 3.4 , las reglas contextuales miran un contexto de una o dos palabras a la izquierda o derecha del elemento a etiquetar. Otras reglas, no ejemplificadas aquí, pueden incluir hasta una ventana de tres etiquetas o tomar en cuenta tanto la etiqueta anterior como la posterior.

Por la naturaleza del método cabe la posibilidad de que algunas reglas realicen cambios erróneos; sin embargo, una vez aplicadas todas estas reglas contextuales, la calidad de etiquetado del corpus debe aumentar. Eric Brill probó su método en tres corpus del inglés: el *Brown Corpus*, un corpus obtenido del *Wall Street Journal* y el *Helsinki Corpus* de inglés antiguo. En todos los casos el etiquetado con reglas contextuales mejoró al etiquetado con reglas léxicas. Los resultados

tomados de su tesis doctoral (Brill 1993) son resumidos en seguida (Tabla 3.5). En otro experimento, Brill (1995) reporta niveles de exactitud de hasta el 97.2%.

Tabla 3.5: Resultados reportados por Brill

	Exactitud (*accuracy*) *(%)*	
Corpus	Etiquetado léxico	Etiquetado léxico y contextual
Wall Street Journal (4000 enunciados)	92.2	94.4
Brown Corpus	89.9	91.8
Helsinki Corpus	84.2	85.9

Una vez presentado el procedimiento de etiquetado de Brill, ahora voy a revisar el proceso de generación de reglas. En éste se obtienen dos conjuntos: reglas léxicas y contextuales.

3.3.2.3.2. Proceso de generación de reglas

El objetivo del proceso de generación, también llamado de entrenamiento, consiste en obtener dos conjuntos de reglas que sean aplicables a un corpus para asignar a cada palabra una categoría gramatical. El procedimiento de etiquetado lo revisé en la sección anterior. En ésta voy a detallar el proceso para obtener automáticamente las reglas.

Dicho proceso obtiene dos tipos de información a partir del corpus etiquetado por un lingüista: información léxica que indique la posible clase de palabra para un vocablo determinado e información contextual que señale una etiqueta para una palabra en un contexto específico. Las entradas a este proceso son: un corpus etiquetado por un lingüista, una serie de plantillas de reglas de transformación y una estrategia para el etiquetado de estado inicial.

El corpus se divide en varias porciones, las cuales serán utilizadas en distintas etapas del entrenamiento. Describo a continuación cada una de estas:

- Corpus de entrenamiento léxico etiquetado. Es una porción del corpus etiquetado por el lingüista que será utilizado sólo en la parte de la generación de reglas de carácter léxico. Para usarlo será necesario retirarle las etiquetas.
- Corpus de entrenamiento contextual etiquetado. Porción del corpus etiquetado por el experto lingüista que será utilizado sólo para la generación de reglas contextuales. Para usarlo será necesario retirarle las etiquetas.
- Corpus de entrenamiento etiquetado. Se forma de los corpus de entrenamiento léxico y contextual etiquetados.

- Corpus de entrenamiento no etiquetado. Se forma de los corpus de entrenamiento léxico y contextual anteriores, pero sin etiquetas, más una porción adicional del corpus también sin etiquetas.
- Corpus de evaluación etiquetado. Es una porción distinta de las anteriores que será utilizada para evaluar los resultados del entrenamiento, por lo mismo no participa en ninguna etapa del entrenamiento. Para usarlo será necesario retirarle las etiquetas.

Las plantillas de reglas de transformación, de las cuales derivarán las reglas para el proceso de etiquetado, se dividen en dos grupos. El primero, dirigido a generar reglas de carácter léxico, se forma de las plantillas de la Tabla 3.6, junto a cada una incluyo los nombres técnicos usados en el programa de computadora.

Tabla 3.6: Plantillas de reglas léxicas

1. Cambia la etiqueta X por la etiqueta Y si la palabra contiene la letra z.	FCHAR
2. Cambia la etiqueta (cualquiera que sea) por la etiqueta Y si la palabra contiene la letra z.	CHAR
3. Cambia la etiqueta X por la etiqueta Y si al quitar la cadena de letras U (uno, dos, tres o cuatro letras) del final (inicio) de la palabra, el resultado es una palabra del corpus.	FDELETESUF Y FDELETEPREF
4. Cambia la etiqueta (cualquiera que sea) por la etiqueta Y si al quitar la cadena de letras U (uno, dos, tres o cuatro letras) del final (inicio) de la palabra, el resultado es una palabra del corpus.	DELETESUF Y DELETEPREF
5. Cambia la etiqueta X por la etiqueta Y si al agregar la cadena de letras U (uno, dos, tres o cuatro letras) al final (inicio) de la palabra, el resultado es una palabra del corpus.	FADDSUF Y FADDPREF
6. Cambia la etiqueta (cualquiera que sea) por la etiqueta Y si al agregar la cadena de letras U (uno, dos, tres o cuatro letras) al final (inicio) de la palabra, el resultado es una palabra del corpus.	ADDSUF Y ADDPREF
7. Cambia la etiqueta X por la etiqueta Y si la palabra contiene la cadena de letras U (uno, dos, tres o cuatro letras) al final (inicio).	FHASSUF Y FHASPREF
8. Cambia la etiqueta (cualquiera que sea) por la etiqueta Y si la palabra contiene la cadena de letras U (uno, dos, tres o cuatro letras) al final (inicio).	HASSUF Y HASPREF
9. Cambia la etiqueta X por la etiqueta Y si la palabra aparece antes (después) de la palabra W.	FGOODRIGHT Y FGOODLEFT
10. Cambia la etiqueta (cualquiera que sea) por la etiqueta Y si la palabra aparece antes (después) de la palabra W.	GOODRIGHT Y GOODLEFT

Se puede notar que algunas de las plantillas generarán reglas que verificarán primero si la palabra tiene cierta etiqueta, otras reglas cambiarán la etiqueta, cualquiera que ésta sea. Además, algunas de ellas (9 y 10) miran las palabras anteriores o posteriores del contexto.

También hay plantillas que toman en cuenta la composición de las palabras obteniendo cierto tipo de segmentos iniciales o finales a manera de “afijos”. En este sentido, el método de Brill es muy

simple y queda lejos de ser un método de descubrimiento de morfología. Los métodos de Hockett o Nida, por ejemplo, proponían encontrar los morfemas y alomorfos de una lengua, mediante corpus, con base en la comparación de formas. En otras palabras, proponían el estudio de la distribución de segmentos en el sistema lingüístico. El método de Brill, por su parte, sólo examina ciertos segmentos útiles, a manera de pistas, para determinar una clase de palabra.

Brill (1995: 559-560) asume como ventaja de su método que ningún afijo sea especificado previamente y por tanto sea descubierto. Esta situación, menciona el autor, le brinda independencia de la lengua, aunque observa que en el caso de lenguas con morfología muy distinta al inglés cabría la posibilidad de modificar las plantillas. Precisamente esta posibilidad es la que aprovecha la propuesta que se hace en esta tesis. Como se mostrará más adelante, aprovecharé el método automático de descubrimiento de afijos de Medina (2003) para modificar las plantillas de reglas propuestas por Eric Brill.

El otro grupo de plantillas son las usadas para obtener reglas de carácter contextual. La siguiente lista muestra estas plantillas (Tabla 3.7).

Tabla 3.7: Plantillas de reglas contextuales

1. Cambia la etiqueta X por Y si la etiqueta anterior (posterior) es Z.
2. Cambia la etiqueta X por Y si una de las dos etiquetas anteriores (posteriores) es Z.
3. Cambia la etiqueta X por Y si la palabra anterior (posterior) es W.
4. Cambia la etiqueta X por Y si las dos palabras anteriores (posteriores) son W_1 W_2.
5. Cambia la etiqueta X por Y si la palabra es W y la etiqueta anterior (posterior) es Z.
6. Cambia la etiqueta X por Y si una de las dos palabras anteriores (posteriores) es W.
7. Cambia la etiqueta X por Y si una de las tres etiquetas anteriores (posteriores) es Z.
8. Cambia la etiqueta X por Y si las dos etiquetas anteriores (posteriores) son Z_1 Z_2.
9. Cambia la etiqueta X por Y si la palabra está entre las etiquetas Z_1 Z_2.
10. Cambia la etiqueta X por Y si la segunda etiqueta anterior (posterior) es Z.
11. Cambia la etiqueta X por Y si la segunda palabra anterior (posterior) es W.
12. Cambia la etiqueta X por Y si la palabra es W.
13. Cambia la etiqueta X por Y si la palabra es W y la segunda etiqueta anterior (posterior) es Z.
14. Cambia la etiqueta X por Y si la palabra es W y la segunda palabra anterior (posterior) es W.

Lo primero que resalta es que todas las plantillas producirán reglas que revisarán si la palabra está etiquetada con cierta etiqueta antes de hacer la transformación, ésta es una diferencia con las de

carácter léxico. Además, son plantillas que miran el contexto ya sea de palabras o de etiquetas. La ventana del contexto llega hasta tres etiquetas o dos palabras, ambas anteriores o posteriores.

Ya había dicho que una de las variables del método era precisamente el tipo de plantillas de trasformaciones. En este sentido, es posible modificar, agregar o quitar plantillas de acuerdo con el conocimiento lingüístico sobre una lengua o, mejor aún, sobre muchas de ellas para obtener un método más independiente o "universal".

Siguiendo con la misma idea, el analizador gramatical desarrollado para el DEM sería un ejemplo de un sistema más dependiente de una lengua, el español, ya que se establecieron reglas específicas para ella. La intención del método de Brill es tener plantillas independientes de la lengua que, en el proceso de generación, hagan surgir las que serían propias de ella.

Las plantillas para producir transformaciones léxicas y contextuales son utilizadas después de haber aplicado un etiquetado de estado inicial al corpus. Dicho estado consiste en asignar a cada palabra del corpus su categoría más frecuente, de acuerdo con un conteo en todo el corpus de entrenamiento etiquetado. Para las palabras que no aparezcan en el corpus de entrenamiento etiquetado, será necesario elegir una estrategia de asignación. La opción más simple es asignarles siempre una categoría arbitraria, por ejemplo, sustantivo común masculino singular. Otra posibilidad, propuesta por Brill, sería poner a todas las palabras que comienzan con mayúscula: nombre propio, y a las restantes: nombre común masculino singular.

Voy a dividir el proceso de generación de reglas en dos partes: generación de reglas léxicas y generación de reglas contextuales. A continuación pongo los pasos de cada uno a manera de un algoritmo general.

GENERACIÓN DE REGLAS LÉXICAS

Etiquetado de estado inicial

1. Para cada tipo de palabra (*type*) del corpus de entrenamiento léxico sin etiquetas:

1.1. Si comienza con mayúscula entonces se le asigna la etiqueta de nombre propio, si no, entonces se le asigna la etiqueta de nombre común masculino singular.

Generación de reglas

2. Mientras la calificación de las reglas sea mayor a un umbral (valor predefinido):

2.1. Para cada etiqueta del conjunto de etiquetas posibles:

2.1.1. Para cada tipo de palabra del corpus de entrenamiento léxico sin etiquetas:

2.1.1.1. Para cada plantilla de regla de transformación:

2.1.1.1.1. Se generan las reglas de transformación de la plantilla y se registra su calificación.

2.2. Se obtiene la regla con la mejor calificación y se agrega al final de la lista de reglas léxicas.

2.3. Se aplica la regla a la lista de tipos del corpus de entrenamiento léxico sin etiquetas.

GENERACIÓN DE REGLAS CONTEXTUALES

3. Se etiqueta el corpus de entrenamiento contextual sin etiquetas con las reglas léxicas.

4. Mientras la calificación de las reglas sea mayor a un umbral (valor predefinido):

4.1. Para todas las palabras del corpus de entrenamiento contextual sin etiquetas:

4.1.1.1. Para cada plantilla de regla de transformación:

4.1.1.1.1. Se generan las reglas de transformación de la plantilla y se registra su calificación.

4.2. Se obtiene la regla con la mejor calificación y se agrega al final de la lista de reglas contextuales.

4.3. Se aplica la regla a las palabras del corpus de entrenamiento contextual sin etiquetas.

Las dos partes son realizadas de manera separada para obtener dos conjuntos distintos de reglas. En esta investigación sólo me concentro en la primera, es decir, en la generación de reglas léxicas. Enseguida describiré brevemente algunos aspectos de este proceso que quedan sin aclarar.

Cada regla es calificada de acuerdo con su mejoría en la calidad de etiquetado del corpus. Esta medida se obtiene a partir de la siguiente fórmula (Brill 1993: 64):

$$\sum \frac{\text{Freq (W, Y)} - \text{Freq (W, X)}}{\text{Freq (W)}}$$

En la fórmula, *W* es la palabra a etiquetar, *Y* es la nueva etiqueta propuesta y *X* la etiqueta actual asociada a la palabra. Las frecuencias son obtenidas del corpus de entrenamiento léxico etiquetado.

La generación de reglas se repite hasta llegar a un umbral para calificar reglas, que está determinado arbitrariamente desde el comienzo. En el caso de los experimentos de esta tesis, el valor se estableció en 2. Entonces, el proceso se detendrá una vez que la calificación de una nueva regla esté por debajo de dicho umbral (véase anexos C, D y E). Las reglas que van saliendo como ganadoras son agregadas al final de una lista (archivo), lo que implica un orden de aplicación en el proceso de etiquetado.

He revisado en esta sección el método propuesto por Eric Brill para la identificación automática de categorías gramaticales. Sin involucrar aspectos computacionales demasiado técnicos, presenté tanto el proceso de etiquetado de nuevos corpus, como la generación de reglas para ese

etiquetado. Ha quedado claro cómo se utilizan las características morfológicas de las palabras para la asignación de categorías. Creo que tomar cierto número arbitrario de caracteres y utilizar un método tan simple para proponer cadenas "afijales" abren la posibilidad a mejoras de carácter lingüístico.

Por lo anterior, en el siguiente capítulo reviso el proceso de descubrimiento de afijos de manera automática. Esto me permitirá incluir en el método de Brill segmentos afijales para generar reglas de carácter morfológico.

4. Descubrimiento de afijos por computadora

En el capítulo anterior, describí el procedimiento de identificación automática de categorías gramaticales que usaré en esta tesis (método de Brill). Como dije, éste utiliza fragmentos, secuencias de letras al final y al principio de la palabra gráfica, para asignar cierta categoría a una palabra. La longitud de estos fragmentos es arbitraria, predefinida a uno, dos, tres o cuatro caracteres iniciales o finales, al menos en el método original. Tal restricción le otorga un carácter de dependencia hacia la lengua, es decir, para algunas será necesario aumentar la longitud. Incluso cuando la morfología pertinente de una lengua queda dentro de estas pocas letras, el procedimiento genera reglas a ciegas para cada inicio o final de palabra; las menos afortunadas se descartan después, pero no dejan de generarse innecesariamente.

Con respecto a la restricción en la longitud del fragmento, surgen por lo menos dos preguntas: ¿cómo determinar la longitud para cada lengua? Y ¿qué longitud sería "suficiente"? Posibles aproximaciones a las respuestas podrían encontrarse en procedimientos de prueba y error: por ejemplo, a través de varias ejecuciones del programa de computadora con distintas longitudes, se podría llegar a un valor adecuado. Otra opción sería revisar las gramáticas y la bibliografía lingüística de la lengua en busca de listas de afijos, de esta manera se obtendría la longitud del afijo más largo.

Por otra parte, la forma como el método de Brill selecciona esos segmentos es muy simple: sólo revisa si al agregarlos o eliminarlos de una palabra resultan en otra palabra del corpus (véanse las plantillas de reglas de **¡Error! No se encuentra el origen de la referencia.**). Esta situación no es del todo mala y es de reconocer que el método sea tan exitoso, al menos para el inglés. Sin embargo, deja de lado aspectos lingüísticos interesantes que bien vale la pena incluir y que no necesariamente deben ser complejos[1].

Entonces, si en lugar de utilizar una longitud predefinida para obtener estos segmentos e identificarlos apenas por su calidad combinatoria, fueran descubiertas verdaderas unidades afijales con un método lingüístico automático, sería entonces posible utilizar el método de Brill en otras lenguas sin necesidad de variar el parámetro. Claro que el método que descubra los afijos deberá

[1] Estoy consciente de que el objetivo de Brill fue contar con un método que trabaje a partir de información lingüística mínima; además, Brill trabajó desde la postura de la inteligencia artificial. Pero estas situaciones no tienen por qué cancelar la posibilidad de integrar o aprovechar algún tipo de procedimiento lingüístico. En este sentido, estaría reconociendo la aportación de Brill y proponiendo, a partir de ella, un método con características más lingüísticas.

también ser simple, efectivo y no dependiente de la lengua. Además, usar los afijos disminuiría el número de reglas generadas innecesariamente.

Bajo esta perspectiva, presento en este capítulo un método de segmentación y descubrimiento morfológico desarrollado con métodos cuantitativos. Éste ha demostrado ser efectivo en su aplicación a diversas lenguas no emparentadas, por ejemplo: el chuj (lengua maya) (cf. Medina y Buenrostro 2003), la lengua checa (cf. Medina y Hlaváčová 2005), el rarámuri rarámuri (cf. Medina y Alvarado 2006; Medina, Camacho y Alvarado 2009) y el español (cf. Medina 2000 y Medina 2003). Además, está fundamentado en una motivación totalmente lingüística y no requiere de ningún parámetro predefinido, ya que se basa totalmente en procesamiento de corpus

Comenzaré describiendo algunos acercamientos a la segmentación de palabras como vía para descubrir unidades morfológicas. Esto permitirá contar con un panorama sobre los métodos que se han propuesto. Luego explicaré el método de descubrimiento de afijos que usaré en mi experimentación y que se basa en lo que Medina llama: índice de afijalidad.

4.1. Segmentación de palabras

En esta sección presento de manera breve algunas propuestas de segmentación de palabras a partir de la investigación que realizó Medina (2003). Mi objetivo es conocer otras posibles soluciones al problema de descubrimiento de afijos. Separar una palabra en signos más pequeños no es una complicación trivial. Tal vez se vuelve más simple en lenguas de escasa morfología, pero en lenguas de fenómenos más complejos el problema se agudiza. Pensemos, por dar un ejemplo, en la formación de palabras de las lenguas semíticas donde las bases son grupos de consonantes en los que se insertan vocales (véase *supra* ejemplo **¡Error! No se encuentra el origen de la referencia.**).

Además, como también se expuso, la adición de segmentos morfológicos acarrea cambios de tipo fonológico en la base o afijo. Otras veces aparecen segmentos sin significado, que marcan cierta característica del tema, como las vocales temáticas del español. En general, trato de dejar claro que la segmentación de palabras es una cuestión no agotada.

Aprovechando la recopilación que hizo Medina (2003: 62-90) sobre distintos métodos de segmentación, enseguida presento un resumen de éstos.

4.1.1. Segmentación supervisada (manual) y reglas de segmentación

Los métodos de segmentación morfológica (*stemming*, etc.) dominantes en procesamiento del lenguaje natural y minería de textos son supervisados porque presuponen el conocimiento

morfológico de la lengua a segmentar, mismo que se codifica en el programa segmentador. Este tipo de métodos se han desarrollado principalmente desde los años ochenta.

De esta manera, uno de estos métodos de segmentación de palabras, de mediados de los ochenta, fue basado en el aprendizaje de reglas de segmentación (Thurmair 1986). A partir de listas de palabras, en donde el investigador indicaba los puntos de segmentación, un sistema de cómputo generaba (aprendía) las reglas necesarias con los contextos apropiados para hacer la segmentación automática. La desventaja de este método es que el investigador debía proporcionar los ejemplos con los cortes ya realizados.

Luego, a partir de la idea anterior surgió otra propuesta (Meya 1986). En ella, el investigador recopiló una lista de morfemas libres y ligados de la lengua. Además, escribió reglas para una serie de transformaciones asociadas a las formas supletivas, modificaciones vocálicas y fonemas epentéticos. Así, mediante la búsqueda de similitud entre la palabra a analizar y la lista de morfemas (reconocimiento de patrones) se podían obtener los componentes de la palabra. En este método, nuevamente fue el investigador quien codificó el conocimiento lingüístico en un tipo de gramática.

Otro método interesante fue el basado en el conteo de combinaciones de letras (Klenk y Langer 1989). Éste tomó como fundamento la idea de que ciertas letras ocurren predominantemente en ciertas posiciones de la palabra, permitiendo diferenciar límites morfológicos. Los resultados que obtuvo este método fueron bastante buenos, entre 68% y 90% para el español, pero igual que los métodos anteriores incluía demasiado trabajo manual. El investigador tenía que decidir dónde estaban los cortes de muchas palabras y obtener la frecuencia de los pares de letras involucrados y no involucrados en una frontera morfológica.

Finalmente, el método más utilizado hoy en día es el llamado algoritmo de Porter (1980) desarrollado originalmente para el inglés pero con implementaciones para las lenguas europeas dominantes, incluido el español. Éste busca el segmento más largo de una palabra y lo sustituye por otro (que puede ser vacío) para dejar sola la base. Está asentado en reglas que incluyen segmentos predefinidos por el investigador y sus correspondientes sustituciones. Son reglas simples ordenadas y obtenidas del análisis de diccionarios en las que hay poco análisis lingüístico.

Hasta aquí, este conjunto de métodos se caracterizan porque el verdadero trabajo de descubrimiento de segmentos lo hace el investigador, quien codifica reglas para reconocer segmentos en nuevas palabras. Los métodos que abordaré en seguida, por otra parte, intentan descubrir morfemas de forma automática.

4.1.2. Segmentación automática no supervisada

El primer trabajo de segmentación morfológica no supervisada se debe a Zellig Harris. Su método para segmentar palabras se fundamente en la variación de fonemas potenciales anteriores y posteriores a un corte morfológico. Entre más variedad de fonemas potenciales, mayor la probabilidad de una frontera morfológica, ya que esa variedad representa mayor incertidumbre (Harris 1955). Otra manera de medir la incertidumbre es mediante la cantidad de información. Este concepto viene de la teoría matemática de la comunicación formulada por Shannon y Weaver (1949). Para ellos, la información se mide por la libertad de seleccionar un mensaje entre varios posibles, nada tiene que ver con lo que contiene cada mensaje o su significado. Así, la cantidad de información se mide a través del logaritmo del número de mensajes u opciones posibles[2]. Esta noción ha sido asociada al concepto de entropía, surgido en la termodinámica. Un mensaje con alta entropía es un mensaje con gran cantidad de información. Estos conceptos pueden ser utilizados para determinar fronteras morfológicas. Si se mide la entropía de una segmentación, es de esperarse que los afijos lleven menor entropía, ya que son más gramaticales, frecuentes y forman un conjunto finito de opciones (no hay tanta libertad de escoger uno de ellos). En cambio las bases, al tener la carga de contenido semántico y ser menos frecuentes, aunque de un conjunto enorme de opciones, estarían asociadas a medidas mayores de entropía.

Luego, en la primera mitad de los años sesenta, un equipo de investigadores rusos a cargo de N. D. Andreev desarrolló un método para determinar automáticamente afijos de flexión en varias lenguas como el ruso, húngaro y vietnamita. Estaba basado en la idea de que los afijos son más frecuentes que las bases. Así, dicho método consistía en buscar los segmentos de palabras finales o iniciales más frecuentes (posibles afijos). Una vez encontrados, eran combinados con los segmentos restantes de las palabras (posibles bases). Si se daban combinaciones que aparecían en el texto y que permitían intercambiar sus elementos, se establecían como segmentos morfológicos (Cromm 1996).

Otra propuesta para segmentar palabras, probada en francés y español, es la de de Kock y Bossaert (Medina 2003: 87-88). Toman sus fundamentos del principio de economía de signos (rentabilidad del sistema), el cuál establece, entre otras cosas, que la combinación de signos lingüísticos del nivel morfológico (pocos y más frecuentes) produce signos de nivel sintáctico (muchos, pero poco frecuentes), haciendo con esto un sistema económico.

Finalmente, los métodos de estadística de digramas pueden servir para encontrar segmentos morfológicos en palabras. Un digrama es un par de elementos que concurren en un corpus, por

[2] Se recomienda el logaritmo base dos para obtener una unidad de información llamada "bit" (*binary digit*).

ejemplo, una posible base y un posible afijo. Estas medidas permiten determinar, a partir de los segmentos, medidas de independencia o de no asociación entre ellos. Se han usado medidas como: la prueba de independencia de x^2, la razón de semejanza, el coeficiente de coligación de Yule y la estadística de información mutua[3].

Hasta el momento, he presentado de manera general y concisa diversos acercamientos a la segmentación de palabras, estos últimos automáticos de carácter no supervisado. Con base en la combinación de algunos de éstos, Medina propone el cálculo de un índice de afijalidad, que explicaré en la siguiente sección. Otros métodos recientes de descubrimiento de afijos utilizan estadística bayesiana (Creutz y Lagus 2005) y medidas de distancia mínima (Goldsmith 2001), ninguno de éstos son revisados en esta tesis.

4.2. Descubrimiento de afijos basado en un índice de afijalidad

El índice de afijalidad que propone Medina (2003) se fundamenta en la caracterización del afijo como una unidad de la cual es posible medir su propiedad combinatoria, la aportación que hace a la economía del sistema y la cantidad de información que conllevan (en el sentido explicado arriba).

Su propuesta considera situaciones interesantes que van bien con los objetivos de la tesis. Por un lado, las mediciones son basadas en corpus, lo que le da un carácter empírico. Por otro, se trata de un procedimiento independiente de la lengua, ya que no requiere información lingüística explicita para el descubrimiento. Finalmente, utiliza un análisis cuantitativo de las características de las unidades lingüísticas y no la formulación de un modelo de inteligencia artificial, por tanto, su objeto de estudio es el sistema lingüístico y no la simulación de su comportamiento.

Medina (2003: 20) caracteriza a los afijos como: objetos gramaticales desgastados fonológica y semánticamente, que aparecen siempre concatenados a otro objeto, su número es limitado, son muy frecuentes, ocurren en muchas estructuras combinatorias, contienen poca información (en comparación con las bases) y se unen a muchos objetos para darles estructura.

Como decía, la ventaja de tratar un afijo de esta manera es que sus características pueden ser cuantificables y por tanto la afijalidad sería medible: “En suma, la afijalidad de un segmento se puede concebir como una combinación de ciertas dimensiones medibles entre dos segmentos de palabra” (Medina 2003: 338). A partir de esta idea y de su aplicación, de manera similar, a la

[3] Si se desea ver el detalle de las fórmulas estadísticas, éstas se encuentran en Kageura (1999) y Medina (2003: 72-75).

cuantificación de la cliticidad, este autor propone una medida de "glutinosidad" entre segmentos: "Hay una fuerza de enlace o glutinosidad entre los segmentos más gramaticales y los segmentos léxicos de un corpus que es directamente proporcional a la entropía que disparan los primeros y al número de signos del nivel siguiente producidos mediante la combinación de los gramaticales con los léxicos" (Medina 2003: 339)[4].

Las medidas que Medina usó para determinar la afijalidad fueron las siguientes: número de cuadros, entropía, índice de economía y medidas estadísticas de digramas. Las que resultaron mejores para indicar cortes morfológicos fueron los índices de cuadros, la entropía y el cociente de de Kock. Enseguida menciono brevemente en qué consisten.

El concepto de cuadro fue propuesto por Greenberg. Un cuadro consiste en un conjunto de segmentos A, B, C y D, que se combinan de la forma AC, BC, AD y BD, por ejemplo, cas– (A), sill– (B), –a (C) y –ita (D), que generarían el siguiente cuadro: *cas::a, sill::a, cas::ita, sill::ita*. Para decir que existe un cuadro, es necesario que las combinaciones aparezcan en el corpus de estudio y que cada segmento también lo haga; también es posible que alguno de los elementos esté vacío. Por ejemplo, el cuadro: in::teligente, in::tolerable, Ø::*teligente, Ø::tolerable, no contaría ya que el elemento C no existe de manera independiente en el corpus. El índice consiste en el número de cuadros para cada segmentación.

La entropía, desde su propuesta como medida de la cantidad de información, fue considerada para el estudio del lenguaje por lingüistas como Harris o Greenberg. Hablando de afijalidad, es de esperarse que los cambios de entropía al interior de una palabra den indicios de su segmentación. Así, si el afijo tiene menos cantidad de información que la base, una medida baja de entropía indicaría el inicio de un afijo, mientras que una medida alta, el de una base (Medina 2003: 108).

La entropía puede ser medida de izquierda a derecha de la palabra o de derecha a izquierda. Según los experimentos de Medina, ambas direcciones son buenos indicadores de un corte. De hecho, de derecha a izquierda, los indicadores de inicio de un sufijo resultaron mejores.

El principio de economía utilizado por Medina está basado en el trabajo propuesto por de Kock y Bossaert y se puede expresar como sigue: a partir de dos segmentos de una palabra, si uno pertenece a un conjunto pequeño de segmentos muy frecuentes y aparece en muchos vocablos,

[4] De hecho, este autor propone que a partir de este índice de glutinosidad es posible cuantificar la gramaticalidad de las unidades lingüísticas. Su trabajo llega a proponer un sistema de medición basado en la unidad de medida *Varrón*. Por cuestión de espacio, no abordo estos interesantes aspectos y me quedo solamente con lo referente a la afijalidad.

mientras que el otro forma parte de un conjunto muy grande, pero de baja frecuencia, y aparece en pocas palabras, el primero sería un afijo y el segundo una base.

El índice de economía de una segmentación dada se obtendría de cuantificar la diferencia de tamaño entre el conjunto de posibles bases y el conjunto de posibles afijos. Siguiendo esta idea son calculados dos valores: "uno dividiendo el número de segmentos a la izquierda propuestos como raíz entre el número de segmentos a la derecha propuestos como afijo, el otro dividiendo los segmentos a la derecha propuestos como raíz entre los de la izquierda propuestos como afijo. El segmento más probable como raíz será aquél cuyo valor calculado bajo la hipótesis de que es la raíz sea mayor a uno" (Medina 2003: 89-90).

Según Medina, los índices de cuadros, economía y entropía[5] resultaron mejores porque miden las propiedades atribuidas al afijo y se basan en conocimiento lingüístico. Este autor calculó dos probabilidades de que un segmento sea un afijo. La primera (prob. 1) se refiere a la posibilidad de que el segmento sea un afijo del corpus, es decir, que lo sean en una lista de tipos de palabras del corpus. La segunda (prob. 2), a que lo sea en una determinada cadena de palabras, es decir, a nivel de ocurrencias de palabras.

También, de la combinación de las medidas anteriores (por ejemplo, promediándolas) se obtuvo un índice de afijalidad. La combinación de los índices es importante ya que alguno de ellos puede dar un corte erróneo; por ejemplo, en la palabra *aumente*, el índice de entropía propone el afijo ~mente, mientras que el de economía propone correctamente ~e. Éste fue probado en el Corpus del Español Mexicano Contemporáneo dando como resultado un 90.41% de formas bien segmentadas.

Las probabilidades más altas dieron la mayor certeza de afijalidad, aunque no correspondieron a índices de afijalidad también altos. Por ejemplo, el sufijo ~a resultó con alta afijalidad, pero baja probabilidad en comparación con sufijos más largos o con el sufijo ~o. Lo anterior significa que si en una cadena se encuentra el segmento final ~a y en otra ~o, el último sería más probable de ser el afijo de la cadena. Entre más largo el afijo, más alta fue su probabilidad. Estas observaciones son importantes, ya que experimentaré en esta tesis con el orden de los sufijos que integraré al método de identificación de categorías.

[5] Las fórmulas y formalizaciones matemáticas para obtener los índices de cuadros, entropía y economía no se incluyen en esta tesis por cuestión de espacio, pero pueden verse en Medina (2003), junto con una tabla comparativa de los mismos, también pueden verse en Medina (2000).

Enseguida menciono otras características interesantes del método de Medina. La primera tiene que ver con que las medidas utilizadas para el índice de afijalidad representan características de un afijo. Así, se puede pensar que distintas combinaciones darían cuenta de distintos tipos de afijos. Por ejemplo, Medina (2000: 106) propone revisar si un valor bajo de economía con un valor alto de cuadros podría dar cuenta del fenómeno de composición.

Una característica adicional es la posibilidad de usar el método sobre una representación fonológica del corpus. Medina lo hace de esta manera mediante una correspondencia grafema-fonema, siguiendo algunas reglas para el español y estableciendo algunas restricciones para los acentos (Medina 2000). Otra situación interesante es que el catálogo final de afijos incluye tanto de flexión como de derivación, resultando que los primeros obtuvieron valores más altos de afijalidad en comparación con los segundos.

En los experimentos realizados con un corpus de lengua checa (Medina y Hlaváčová 2005) se confirmaron dos cosas. La primera es que en la lista de afijos descubiertos se incluyen segmentos de compuestos. La otra es que el método resultó muy efectivo para descubrir prefijos en esta lengua. Además, en los experimentos con el chuj (Medina y Buenrostro 2003), Medina demostró que un corpus pequeño brinda alentadores resultados para el descubrimiento de afijos flexivos y que en estos casos el índice de entropía es más pertinente.

Creo que a pesar de omitir el aparato formal con el que Medina expone su propuesta, he dejado clara la idea central de su método. Al respecto, considero que uno de los aspectos más afortunados es la cuantificación de características lingüísticas a partir de criterios lingüísticos.

Otro aspecto que ha quedado claro es la diferencia en el descubrimiento de afijos entre el método de Brill y el de Medina. El primero apenas revisa si los segmentos se combinan con otras bases, en cambio, el segundo mide el número de cuadros que produce, su entropía y la economía que aporta al sistema. Esta diferencia es la que me lleva a indagar, entre otras cosas, si la inclusión de los afijos descubiertos automáticamente, con el método de Medina, mejorará el resultado del método de Brill. Esta situación la resolveré en el siguiente capítulo dedicado a la experimentación en un corpus de español del siglo XVI.

5. Identificación automática de categorías gramaticales en español del siglo XVI

En este capítulo, plasmo el desarrollo del trabajo de experimentación que me permitirá resolver si las intuiciones mencionadas en los capítulos anteriores son ciertas. En primer lugar, describiré el corpus de estudio y su etiquetado manual. En seguida, hablaré de los resultados del descubrimiento de afijos en el corpus. Al final, presentaré el resultado de la generación de etiquetas con el método original de Brill y el modificado con la propuesta de Medina.

Como dije, el método de Brill permite obtener, automáticamente, pistas para identificar la categoría gramatical de una palabra. Éstas son usadas para generar reglas de transformación (etiquetado) a partir de ciertas plantillas. Cuando éstas son de carácter léxico, están basadas en una palabra a la izquierda o derecha, en la palabra misma, en alguna letra en su interior o en segmentos de uno, dos, tres o cuatro caracteres iniciales o finales. Algunas plantillas verifican si al eliminar o agregar uno de estos segmentos se genera una palabra atestiguada en el corpus (véase plantillas en **¡Error! No se encuentra el origen de la referencia.**).

Es necesario reconocer que este procedimiento, aunque simple, obtiene muy buenos resultados. Sin embargo, guarda una amplia distancia con un método de descubrimiento de unidades morfológicas. Por tanto, si existiera un método que descubriera los morfemas de una lengua para incluirlos en las plantillas de transformación, sería posible pensar que las reglas obtenidas de esas plantillas serían mejores que las reglas obtenidas con el método original.

Como exponía en su momento, la asignación de categorías gramaticales a las palabras de un corpus toma en cuenta principalmente aspectos morfológicos y sintácticos. En este sentido, el método basado en transformaciones y dirigido por errores de Brill resulta cercano a esta perspectiva lingüística. Este método permite obtener reglas de etiquetado que toman en cuenta el interior de la palabra y su orden de aparición en relación con otras. Por tal razón, seleccioné este método para realizar la tesis.

Así pues, utilizaré el método de descubrimiento de afijos propuesto por Medina, revisado en el capítulo anterior, para obtener una lista de sufijos del español del siglo XVI mediante el procesamiento de textos del Corpus Histórico del Español en México (CHEM). Una vez descubiertas estas unidades, modificaré las plantillas de transformación para obtener reglas morfológicas a partir de ellas. Finalmente, etiquetaré un fragmento del corpus para evaluar resultados. Como la idea es comparar éste método con el original de Brill, también generaré reglas con las plantillas normales y etiquetaré el mismo fragmento para comparar resultados. A continuación detallo los pasos seguidos para llevar a cabo la investigación propuesta.

5.1. Corpus de estudio

El método para la identificación automática de categorías, que usaré en este trabajo, está basado en el análisis de corpus. Su objetivo, como lo había dicho, es procesar automáticamente un corpus ya etiquetado por un experto, para obtener cierta información lingüística útil que permita etiquetar nuevos corpus no etiquetados. Entonces, en este trabajo de investigación, me di a la tarea de recopilar un corpus de estudio.

Ya que esta tesis se desarrolla en el marco del proyecto de investigación para la constitución del Corpus Histórico del Español en México (CHEM) decidí trabajar con español del siglo XVI. Utilizaré al CHEM porque es un corpus electrónico, cuenta con textos de español del siglo XVI producidos en Nueva España y a futuro será el corpus a etiquetar con el producto de mi investigación[1]. A continuación hablo brevemente de él.

5.1.1. El Corpus Histórico del Español en México (CHEM)

El Corpus Histórico del Español en México (CHEM) es un corpus lingüístico electrónico de los siglos XVI al XIX que incluye textos producidos en Nueva España y México. Es un corpus abierto, ya que día con día se agregan nuevos textos. Su principal objetivo es convertirse en herramienta útil para el análisis lingüístico. Los textos que conforman el CHEM han sido obtenidos mediante convenios con distintos autores e instituciones o mediante la recopilación y trascripción de documentos impresos.

Los documentos que fueron utilizados para el corpus de estudio, y que forman parte del CHEM, son[2]:

BUELNA SERRANO, E. 2005. *Los procesos inquisitoriales contra indígenas que realizó Fray Juan de Zumárraga en Nueva España (1536-1543)*. UAM-A, (manuscrito electrónico). (Buelna)

COMPANY COMPANY, CONCEPCIÓN. 1994. *Documentos lingüísticos de la Nueva España. Altiplano Central*. México: UNAM. (DLNE)

---

[1] Se espera que el resultado de esta tesis sea un método para etiquetar el CHEM, pero es necesario aclarar que esto implica dos cosas: el entrenamiento para generar reglas de etiquetado para cada período del corpus, por ejemplo siglos, y el software para realizar ese etiquetado. Mi trabajo atiende a lo primero, ya que es lo lingüísticamente interesante. Realizar lo segundo tiene retos técnicos y computacionales, como tratar con archivos XML, pero es algo que dejaré para después.

[2] El CHEM cuenta también con revistas como *El Mercurio Volante*, documentos periodísticos de José Antonio de Alzate y Ramírez, las constituciones de Yucatán, Veracruz, Chihuahua y Jalisco de 1857 y un recetario del siglo XIX, entre otros.

LOPE BLANCH, J. M. 1985. *El habla de Diego de Ordaz. Contribución a la historia del español americano.* Publicaciones del Centro de Lingüística Hispánica, 20. México: UNAM, IIF. (El habla)

Sobre estos documentos, puedo mencionar que no son literarios, ya que están formados principalmente por juicios y cartas. De hecho, los dos primeros, Buelna y DLNE, intentan ser representativos del habla popular de la época. Esta situación es favorable, ya que estaré trabajando con un tipo de texto que puede resultar más difícil de etiquetar que un texto literariamente cuidado. Por tanto, será más interesante ver los resultados del experimento.

En la siguiente sección describo cuáles son los textos que utilicé y que fueron extraídos de los documentos mencionados.

5.1.2. Textos y división del corpus

Los textos que seleccioné de cada uno de los documentos descritos en la sección anterior son listados a continuación. En primer lugar, pongo los que forman el corpus de entrenamiento etiquetado en la Tabla 5.1.

Tabla 5.1: Textos del corpus de entrenamiento etiquetado

Documento fuente	Texto	Título	Número de ocurrencias de palabras
Buelna	3	Proçeso del Santo Oficio contra Gaspar, yndio idólatra.	729
Buelna	7	Proçesso del Santo Ofiçio de la ynqujsiçion contra los yndjos Descapuçalco.	3,587
Buelna	8	Proceso contra Marcos Atlaucatl de Santiago Tlaltelolco.	1,726
Buelna	14	[Sin título]	994
DLNE	1	Carta autógrafa de Rodrigo de Albornoz al emperador Carlos V, proponiendo mejores formas de gobierno y soluciones a distintos problemas en la Nueva España (fragmento).	1,022
DLNE	2	Carta autógrafa de Alonso de Estrada al emperador Carlos V en defensa de Hernán Cortés.	448
DLNE	39	Carta autógrafa de fray Andrés de Arroyo a su hermano, dándole consejos sobre qué hacer en caso de que viniera a la Nueva España.	379
DLNE	40	Carta autógrafa de Juana Bautista a su hermana, preguntando por los parientes que estaban por viajar a la Nueva España.	871
DLNE	77	Solicitud de justicia y de aceptación de nuevos testimonios, que el licenciado Obregón, corregidor de la ciudad de México, dirige al Consejo de Indias, por haberle hecho el virrey un juicio de residencia injusto.	1,047
DLNE	78	Carta autógrafa de Diego González a su hermano pidiéndole se reúna con él en la Nueva España.	457

Tabla 5.1: Textos del corpus de entrenamiento etiquetado (continuación)

Documento fuente	Texto	Título	Número de ocurrencias de palabras
DLNE	10	Información de Jerónimo López sobre los abusos que cometían los oidores de la Nueva España.	461
DLNE	20	Testimonio de Nuño Méndez en el juicio que se le hizo por haber tenido acceso carnal con madre e hija.	507
DLNE	30	Carta autógrafa de fray Nicolás de Witte, o de San Paulo, en defensa del virrey don Luis de Velasco.	511
DLNE	60	Testimonio de un hombre de treinta y siete años, barbero, en un juicio por cuchilladas frente a la puerta del Santo Oficio.	546
El habla	I	CARTA I: Toledo, 2 de abril de 1529 (fragmento de la parte inicial).	1,919
		TOTAL	**15,204**

Para el corpus adicional requerido por el proceso de generación de reglas de Brill utilicé el siguiente texto, que representa aproximadamente el 9% del corpus total. Éste no fue etiquetado (Tabla 5.2).

Tabla 5.2: Texto del corpus adicional

Documento fuente	Texto	Título	Número de ocurrencias de palabras
DLNE	16	Proceso contra Tezacacoacatl y Ollin, indios de Ocuila.	1,584

Finalmente, para la evaluación de resultados, recurrí al texto que describo en la Tabla 5.3 y que también fue etiquetado manualmente. Éste representó un 8% del corpus de estudio total.

Tabla 5.3: Texto del corpus de evaluación

Documento fuente	Texto	Título	Número de ocurrencias de palabras
El habla	I	CARTA I: Toledo, 2 de abril de 1529 (fragmento de la parte final).	1,386

Todos los textos descritos formaron el corpus de estudio. El total de palabras fue de 18,174, de las cuales el corpus etiquetado representó el 84%, el adicional el 9% y el de evaluación el 8%. Es importante decir que es un corpus pequeño para el tamaño normal de los grandes corpus electrónicos

con los que se realiza la identificación de categorías gramaticales[3]. Sin embargo, como se verá más adelante, fue posible generar reglas y lograr alentadores niveles de precisión en el etiquetado.

La situación mencionada en el párrafo anterior fue una de las razones por las que opté por el método de Brill. Éste tiene la ventaja de generar reglas útiles con poco corpus. En la próxima sección hablaré del trabajo realizado para el etiquetado manual.

5.2. Etiquetado manual del corpus

Ya quedó asentado en el capítulo **¡Error! No se encuentra el origen de la referencia.**, en la sección que detalla el método de Brill, que para generar los dos conjuntos de reglas es necesario partir de un corpus previamente etiquetado. Por tanto, el corpus de estudio fue etiquetado por estudiantes de la Licenciatura en Lengua y Literaturas Hispánicas de la UNAM, mediante un programa de cómputo que facilitó su labor. Dicho programa mostraba en un formulario de captura el conjunto de categorías gramaticales, los estudiantes seleccionaban la categoría correspondiente a la palabra en cuestión y el programa asignaba la etiqueta que la codificaba.

El trabajo de etiquetado manual es arduo y minucioso, situación que lleva a cometer errores. Los principales problemas con el etiquetado del corpus de estudio fueron producto de la falta de criterios unificados en la asignación de etiquetas y de la lentitud. Pero la lentitud del etiquetado no hubiera sido un problema grave si al final no hubiera existido tanta inconsistencia en él. Ante tal situación, se tuvo que revisar nuevamente todo el corpus para solucionar problemas y eliminar el ruido, esto provocó un retraso importante en la investigación. Creo que en proyectos de este tipo es importante imponer un sistema de control de calidad del etiquetado para no repetir este trabajo.

Como en su momento resalté, el conjunto de etiquetas utilizado en el proceso de identificación de categorías es muy importante, ya que de él depende el nivel de especificidad de la información lingüística plasmada en el corpus. Decidí utilizar un conjunto basado en un estándar, el cual detallo en seguida.

5.2.1. Definición del conjunto de etiquetas

Como había dicho, sobre el conjunto de etiquetas hay al menos dos aspectos fundamentales que atender. Por un lado, la especificidad de las categorías, es decir, el nivel de rasgos que serán

[3] Como había dicho (véase *supra* nota **¡Error! Marcador no definido.**), el tamaño de un corpus es relativo. Para esta tesis, consideraré el corpus de estudio como pequeño en comparación con los corpus comúnmente usados en lingüística computacional y procesamiento de lenguaje natural, que suelen tener cientos de miles, o millones, de palabras. Sin embargo, estos corpus son generalmente de inglés o español contemporáneo.

identificados. Por otro lado, la manera de codificar esos rasgos en etiquetas. En esta sección describo las decisiones tomadas al respecto.

Sobre el primer aspecto decidí marcar no sólo las categorías mayores de palabra, sino también sus rasgos gramaticales. Resolví también que convenía apegarse a un estándar y obtener de él los beneficios que ya mencioné. Así, adopté el conjunto de etiquetas de EAGLES ya que esta codificación trata de abarcar diversas lenguas, entre ellas el español[4]. Las codificaciones para las etiquetas son mostradas en las siguientes tablas. Éstas contienen la codificación para cada rasgo y la posición en la etiqueta.

Tabla 5.4: Codificación de etiquetas para adjetivos

Pos.	Atributo	Valor	Código
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
		Ordinal	O
3	Grado	-	0
		Comparativo	C
		Superlativo	S
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Función	-	0
		Participio	P

En la Tabla 5.4 se muestra la codificación utilizada para etiquetar los adjetivos. Tomé como adjetivos sólo aquellas palabras que acompañan al nominal para calificarlo[5]. Siguiendo la propuesta

[4] Como había mencionado, usaré en especial el conjunto de etiquetas propuesto por la iniciativa PAROLE.

[5] Existen perspectivas como la de Dixon (1999: 1) para quien los adjetivos son de varios tipos:

a) Predeterminantes.
b) Determinantes artículos, demostrativos y posesivos.
c) Superlativos y comparativos.
d) Números cuantificadores, ordinales y cardinales.
e) Adjetivos descriptivos.
f) Modificadores.

Sin embargo, en español parece ser más aceptada su separación como una categoría distinta. Al igual que los determinantes, artículos, demostrativos y cuantificadores deben concordar con el nombre, pero a diferencia de ellos sí

de PAROLE, también se marcaron como adjetivos los ordinales, aunque generalmente aparecen a la izquierda a manera de determinantes. El rasgo 6 de participio se puso cuando la forma del adjetivo coincidía con una forma de participio de verbo. Para la marcación de rasgos de género y número, en casos de duda, me guié por la concordancia con el nombre.

Tabla 5.5: Codificación de etiquetas para adverbios

Pos.	Atributo	Valor	Código
1	Categoría	Adverbio	R
2	Tipo	General	G
		Negativo	N
		Tiempo	T
		Modo	M
		Lugar	L

Sobre la codificación de adverbios (Tabla 5.5) sólo puedo decir que los que no eran de tiempo, modo, lugar o negativos fueron marcados como generales. El adverbio negativo se reservó para el vocablo *no*.

Tabla 5.6: Codificación de etiquetas para determinantes

Pos.	Atributo	Valor	Código
1	Categoría	Determinante	D
2	Tipo	Demostrativo	D
		Posesivo	P
		Interrogativo	T
		Exclamativo	E
		Indefinido	I
		Artículo	A
		Numeral	N
3	Persona	-	0
		Primera	1
		Segunda	2
		Tercera	3

cuentan con significado léxico. Otra cosa que los hace diferentes es que no son “referencializadores” de sustantivos sino clasificadores (cf. Demonte 1999).

Tabla 5.6: Codificación de etiquetas para determinantes (continuación)

Pos.	Atributo	Valor	Código
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Poseedor	-	0
		Singular	S
		Plural	P

La Tabla 5.6 muestra la codificación para determinantes. Los atributos persona y poseedor se usaron para los determinantes posesivos. El segundo es un atributo pensado para marcar el número del referente poseedor. Sirve para distinguir *nuestras* de *mías*, ya que ambas palabras son determinantes posesivos primera persona femenino plural: DP1FP, sólo distinguibles porque el poseedor de la primera es plural DP1FP**P** y el de la segunda singular DP1FP**S**. Lo mismo sucede con *nuestra* y *mía* ya que ambas son determinantes posesivos primera persona femenino singular: DP1FS, y sólo se distinguen por el poseedor plural de la primera, DP1FS**P**, y singular de la segunda, DP1FS**S**.

El tipo artículo sólo se aplicó en los casos de los tradicionalmente llamados artículos definidos (*el, la, lo, los, las*). Los artículos indefinidos se trataron como determinantes indefinidos, estos fueron los tradicionales *un, una, uno, unas* y otros como *alguna, alguno, ninguno, otra, otro*. En los determinantes numerales se incluyen los cardinales, partitivos y multiplicativos. El género neutro se reservó únicamente para el artículo *lo* y el común fue utilizado cuando no era posible determinar si era femenino o masculino como el caso de *mi*, *su* o los numerales.

Cuando encontramos cantidades seguidas de *año*(*s*), *día*(*s*) o algún sustantivo, éstas fueron etiquetadas como determinantes numerales, por ejemplo: *treynta días*, *dos esclabas* y *quinze años*. Se decidió marcar los determinantes numerales con género común si no sufrían flexión[6].

[6] Esta decisión es discutible ya que contradice el principio de concordancia de género en español entre determinante y sustantivo, pero no hay flexión en la palabra que marque el cambio de género. Desde el criterio morfológico podrían verse como de género común: *treinta años*, *treinta días*; pero desde la sintaxis tomarían el género del sustantivo. Ya que me basé en un manual de etiquetado de PAROLE, en él se etiquetaban los numerales como de

Las guías de etiquetado de PAROLE tratan como determinantes interrogativos a las palabras: *cuánta, cuántas, cuánto, cuántos* y *qué*; así, se etiquetó: *Preguntado qué/*DT0CN0 *hijas tenja la dicha Leonor Alvarez*.

Tabla 5.7: Codificación de etiquetas para nombres

Pos.	Atributo	Valor	Código
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N

Para el etiquetado de nombres (ver Tabla 5.7) se evitó el uso de género común y número invariable. Cuando fue necesario, se miró el contexto para decidir el conveniente. Se etiquetaron como nombres propios los antropónimos y topónimos[7]. También se etiquetaron así las formas apelativas de entidades divinas como *Nuestra Señora* o *Dios Nuestro Señor*; y nombres de instituciones, en su mayoría religiosas, como *Santo Ofiçio*, *Real Consejo*, *Abdiençia Real*, *Rreal chancillería*, *Santa Madre Iglesia* y *Capitanja General*. En el caso de sobrenombres o apelativos de entidades pertenecientes a la realeza como *Señoría* o *Magestad*, no es claro si deben tratarse como nombres propios por lo que decidí etiquetarlos como nombres comunes.

género común. La reflexión anterior la hice tarde, después de haber procesado el corpus varias veces y decidí dejarlo así, aunque en otras circunstancias daría preferencia a la concordancia.

[7] Estos dos tipos de nombres propios, antropónimos y topónimos, son dos clases que gozan de aceptación general en la gramática como pertenecientes a este tipo de palabra. Otras clases como: nombres de instituciones, períodos temporales, apelativos informales y títulos, son cuestionadas en cuanto a su pertenencia a esta clase de palabra (Fernández Leborans 199: 80-81).

Tabla 5.8: Codificación de etiquetas para verbos

Pos.	Atributo	Valor	Código
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar haber	A
		Verbo ser	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
		Condicional	C
		-	0
5	Persona	Primera	1
		Segunda	2
		Tercera	3
		-	0
6	Número	Singular	S
		Plural	P
		-	0
7	Género	Masculino	M
		Femenino	F
		-	0

La Tabla 5.8 muestra la codificación para verbos. Decidí marcar de manera especial el verbo haber, cuando fue auxiliar en formas compuestas (*an venido*, *a sido*, *hubiese acontesido*), y el verbo ser, en cualquiera de sus formas: *es*, *era*, *fueron*, *syendo*, *fue*, *sean*, *fuese*, *ser*, *estamos*. El atributo de género sólo fue utilizado en los participios, para el resto de los modos se especificó 0. También quedaron en 0 el tiempo, persona y número de infinitivos y gerundios. Para el caso de verbos en pasado de subjuntivo, siempre se etiquetaron como imperfectos (I) y no como pasado simple (S), casos así fueron los de: *dieran*, *fuera*, *fuese* e *hiziese*.

Tabla 5.9: Codificación de etiquetas para pronombres

Pos.	Atributo	Valor	Código
1	Categoría	Pronombre	P
2	Tipo	Personal	P
		Demostrativo	D
		Posesivo	X
		Indefinido	I
		Interrogativo	T
		Relativo	R
		Numeral	N
		Exclamativo	E
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Invariable	N
7	(no utilizado)		0
6	Poseedor	-	0
		Singular	S
		Plural	P

Las posibilidades de etiquetado de pronombres pueden verse en la Tabla 5.9. Sólo los pronombres personales y posesivos fueron marcados con el atributo de persona. Al igual que los determinantes, se utilizó el género común en los pronombres que son usados sin distinción, como *yo*, *mí*; y el neutro sólo para *ello*. El número invariable fue asignado a palabras que no presentan cambio de flexión en singular y plural, como *se*, *sí*, *que*.

En el caso de los pronombres posesivos, el rasgo de poseedor permitiría diferenciar entre *mía* y *nuestra*. Ambas serían pronombres posesivos de primera persona femenino singular, esto es PX1FS0, pero con el rasgo de poseedor se distinguen ya que *mía* tiene poseedor singular PX1FS0**S** y *nuestra* plural PX1FS0**P**. Los pronombres interrogativos fueron marcados sin rasgos PT00000, sólo para *qué* se marcó género común y número invariable PT0CS00, siguiendo la propuesta de

PAROLE. El atributo de la séptima posición no se utilizó porque se refiere al caso, aunque queden reminiscencias de él en los pronombres del español.

Tabla 5.10: Codificación de etiquetas para conjunciones

Pos.	Atributo	Valor	Código
1	Categoría	Conjunción	C
2	Tipo	Coordinada	C
		Subordinada	S

En la Tabla 5.10 muestro la codificación para conjunciones. Encontré en el corpus casos de contracciones de conjunciones con verbos, pronombres y el determinante artículo *el*. Para estos casos decidí mantener la misma etiqueta de conjunción, CS o CC, y unirla a la etiqueta del elemento contraído, en otras palabras, sólo uní las dos etiquetas. Para ejemplificar, la contracción *ques* se etiquetó como CSVSIP3S0, la unión de conjunción subordinante (CS) con un verbo ser indicativo presente en tercera persona singular (VSIP3S0); también se encontraron *porques* y *quera*. En el caso de los pronombres personales, por ejemplo, *quel* (que + él), estos fueron etiquetados como CSPP3MS00 (CS + PP3MS00), la misma contracción apareció como conjunción con determinante y quedó con la etiqueta CSDA0MS0 (CS + DA0MS0). Otra contracción encontrada fue *quello* (que + ello).

Tabla 5.11: Codificación de etiquetas para interjecciones

Pos.	Atributo	Valor	Código
1	Categoría	Interjección	I

Como muestra la Tabla 5.11, las interjecciones fueron marcadas con la etiqueta I. En el corpus de estudio sólo encontré *amén*.

Tabla 5.12: Codificación de etiquetas para preposiciones

Pos.	Atributo	Valor	Código
1	Categoría	Adposición	S
2	Tipo	Preposición	P
3	Forma	Simple	S
		Contraída con artículo	C
		Contraída con determinante	D
		Contraída con pronombre	P

Tabla 5.12: Codificación de etiquetas para preposiciones (continuación)

4	Género	Masculino	M
		Femenino	F
5	Número	Singular	S
		Plural	P
6	Persona	Primera	1
		Segunda	2
		Tercera	3

Ya que el estándar de EAGLES está pensado para muchas lenguas, éste contempla las posposiciones, que no existen en español (véase Tabla 5.12). Por tal razón, la categoría es adposición, que se divide en preposiciones y posposiciones. Cuando la preposición fue simple no puse etiquetado de otros rasgos y quedó así: APS000. Pero en español es posible encontrar contracciones formadas por preposiciones más artículos, como *del* o *al*, las cuales se marcaron con rasgos de número y género: APCMS0.

Para el español del siglo XVI fue necesario incluir otras formas contraídas con determinantes, *desta*, y con pronombres, *dellos*, *della*, *desto*. Así, agregué dos códigos más para el atributo de forma: contraída con determinante (D) y contraída con pronombre (P). El rasgo de persona se utilizó sólo para la contracción con pronombre, por ejemplo *dellos* (APPMP3).

Tabla 5.13: Codificación de etiquetas para signos de puntuación

¡	Faa		«	Fra		--	Fgl
!	Fat		»	Frt		/	Fh
,	Fc		{	Fla		¿	Fia
[	Fca		}	Flt		?	Fit
]	Fct		(	Fpa		...	Fs
:	Fd		)	Fpt		%	Ft
"	Fe		.	Fp		$	Fsp
-	Fg		;	Fx		-, +, =	Fz

Con el fin de mantener la consistencia en el etiquetado del corpus y siguiendo los lineamientos de PAROLE, decidí etiquetar también los signos de puntuación. La Tabla 5.13 muestra las etiquetas correspondientes.

Tabla 5.14: Codificación de etiquetas para cifras

Pos.	Atributo	Valor	Código
1	Categoría	Cifra	Z
2	Tipo	Moneda	m

La codificación de la Tabla 5.14 se utilizó para cantidades aisladas o cantidades seguidas de un indicador de moneda como *ducados*, *rreales* o *pesos*.

Tabla 5.15: Codificación de etiquetas para fechas y horas

Pos.	Atributo	Valor	Código
1	Categoría	Fecha/Hora	W

Ya había indicado que cuando las cantidades estaban antes de palabras como *años* o *días* fueron etiquetadas como determinantes numerales, pero aquellas que no determinaron sustantivo alguno fueron marcadas como fechas, siempre y cuando tuvieran ese significado. Además, los nombres de días y meses también fueron etiquetados como fechas. Un ejemplo claro es el siguiente: *Contra marcos 1539/*W *Mayo/*W.

Los párrafos anteriores, junto con las tablas, describen no sólo la manera de codificación y el conjunto de etiquetas utilizadas para el corpus, sino también los criterios adoptados para su etiquetado manual. Vale la pena volver a decir que este proceso no fue nada fácil de realizar, ya que los criterios del grupo de etiquetadoras llegaron a diferir en algunos aspectos, lo que causó inconsistencias; además, el proceso tedioso de etiquetar muchas palabras provocó errores. Aunque realicé una corrección final sobre todo el corpus que me permitió consolidar criterios y reducir al máximo los errores; de cualquier manera es posible que dejara escapar algunos.

Además del corpus etiquetado, el otro elemento importante para comenzar el proceso de generación de reglas fue la lista de afijos de español de la época. En la siguiente sección describo cómo fue la obtención de éstos mediante la aplicación del proceso de descubrimiento discutido en el capítulo **¡Error! No se encuentra el origen de la referencia.**.

5.3. Descubrimiento de afijos

En esta sección pongo los resultados del proceso de descubrimiento automático de afijos con el método de Medina. Primero, es necesario hacer dos anotaciones. Por un lado, el programa para descubrir estas unidades fue ejecutado sobre un subconjunto de textos del siglo XVI del CHEM

(Buelna y DLNE) y no sólo sobre el corpus de estudio. Por otro, sólo trabajé con los sufijos y no tomé en cuenta los prefijos.

Con respecto a haberme enfocado solamente a los sufijos, hay dos puntos que comentar. El primero es que la tradición gramatical española no cree que los sufijos y prefijos sean de la misma naturaleza. Así, trata a los primeros como parte de la derivación y flexión, mientras que a los segundos, por su gran parecido a las preposiciones, como composición (cf. RAE 1973, y Val 1999: 4775-4776). En segundo lugar, la prefijación no cambia ni el significado ni la categoría gramatical de la palabra (cf. Varela y Martín 1999). Por tanto, dejé de lado la prefijación y sólo experimenté con la sufijación.

En la Tabla 5.16 pongo la lista de los primeros 50 sufijos, ordenados por el índice de afijalidad, de los 831 obtenidos (en el anexo **¡Error! No se encuentra el origen de la referencia.** se puede ver la lista completa). Como anotaba en su momento, el método de Brill genera una regla por cada grupo de caracteres finales de cada palabra del corpus, muchas de estas son innecesarias. Al respecto, contar con los afijos descubiertos evitará esta situación, lo que contribuirá a generar sólo las reglas asociadas a éstos.

Ya que el método de descubrimiento de afijos no divide los resultados en morfemas mínimos, en el catálogo se pueden encontrar segmentos formados por varios afijos, algunas veces incluyendo enclíticos. También debo decir que el método fue realizado con base en la representación ortográfica del corpus. Por lo tanto, los sufijos descubiertos dieron muestra clara de la variación ortográfica de la época. En ellos se pueden observar dos tipos de variaciones: la falta de acento, y las alternancias de ciertas grafías (~iendo, ~yendo, ~jendo).

Las entradas de la Tabla 5.16 son parte del inventario de sufijos del español. Al respecto, no entraré en una discusión amplia sobre sus significados ni sobre la relación entre los sufijos y su forma. Por ejemplo, varios de ellos podrían ser clasificados como nominales o verbales, como el sufijo ~s, que puede ser marca de plural o de 2ª persona de singular.

Sin embargo, para dar muestra de la efectividad del método de descubrimiento de afijos, enseguida haré una comparación entre los sufijos flexivos obtenidos y los que consigna la gramática académica. A pesar de la distancia con el siglo XVI, tomaré como base la gramática descriptiva actual, apegado a la idea de que los cambios en la morfología flexiva del español medieval al español actual no han sido considerables[8].

[8] Los cambios en el verbo del latín al español, según Penny (2005/2202), exhibieron algunas tendencias analíticas (AMĀTUR > *es amado*) y sintéticas (*contar telo é* > (*te lo*) *contaré*); cambios fonológicos (DĪCERE > *dezir*, DĪCŌ

Tabla 5.16: Los primeros 50 sufijos del siglo XVI del CHEM ordenados por afijalidad

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
1	~a	1010	0.5634	0.9736	0.9659	0.3749	0.4331	0.8343
2	~s	1399	1.001	1	0.4611	0.4026	0.5001	0.8207
3	~o	997	0.5959	0.977	0.808	0.3273	0.579	0.7937
4	~ó	347	0.5298	0.9649	0.8552	0.8032	0.9059	0.7833
5	~as	439	0.3201	0.9649	0.9461	0.454	0.5752	0.7437
6	~os	521	0.3525	0.9715	0.8554	0.3583	0.4981	0.7265
7	~ar	280	0.2502	0.9601	0.9577	0.5907	0.7419	0.7227
8	~ado	268	0.2234	0.9681	0.9538	0.5826	0.7705	0.7151
9	~e	577	0.3638	0.9675	0.7987	0.3002	0.1974	0.71
10	~ase	101	0.1506	0.9627	0.9752	0.7266	0.8182	0.6962
11	~aron	132	0.1927	0.9671	0.9244	0.7543	0.9273	0.6947
12	~ados	121	0.1365	0.9736	0.9613	0.5817	0.6294	0.6904
13	~ando	121	0.138	0.9648	0.9503	0.6722	0.8472	0.6844
14	~an	311	0.1864	0.9532	0.906	0.4677	0.4456	0.6819
15	~en	272	0.2629	0.9581	0.8186	0.5282	0.207	0.6798
16	~ava	96	0.1306	0.9656	0.9294	0.7111	0.795	0.6752
17	~asen	66	0.1083	0.9627	0.9508	0.7253	0.8052	0.674
18	~er	134	0.122	0.9036	0.9834	0.6734	0.734	0.6696
19	~ará	37	0.07745	0.943	0.9787	0.7115	0.5321	0.6664
20	~ada	91	0.1035	0.9491	0.9459	0.5617	0.4784	0.6662
21	~aba	98	0.1421	0.9596	0.8868	0.7206	0.8513	0.6628
22	~ido	128	0.1388	0.9444	0.8835	0.6337	0.6995	0.6556
23	~n	731	0.4292	0.9878	0.5419	0.364	0.2257	0.653
24	~avan	66	0.09089	0.9543	0.8997	0.8049	0.8708	0.6483
25	~ía	170	0.2001	0.9471	0.7897	0.8252	0.8442	0.6456
26	~iendo	71	0.1124	0.9446	0.8768	0.7474	0.712	0.6446
27	~amente	34	0.05325	0.9526	0.92	0.5574	0.7227	0.6419
28	~amos	74	0.0802	0.897	0.9449	0.7475	0.8327	0.6407
29	~i	45	0.05045	0.8711	1	0.4054	0.1514	0.6405
30	~ieron	85	0.09915	0.9381	0.88	0.7456	0.8053	0.6391

> *digo*) y analógicos (*cozer* > *cuego*/*cuezes*); el cambio de acento verbal (TIMĒRE > /temére/ = VÉNDĔRE > /βendére/); el apócope de ~e en el español medieval, las vacilaciones entre vocales átonas (*escriviendo*/*escreviendo*) y la reducción de cuatro a tres conjugaciones.

En lo que respecta a los morfemas de número y persona, este autor menciona que "estos elementos no han sufrido más modificaciones que las que obedecen al cambio fonológico" (Penny 2005/2002: 188), en cuanto al aspecto: "Los cambios que condujeron del español medieval al moderno son escasos. El sistema continúa inalterado en su estructura general, aunque algunas formas verbales ocupan una posición diferente en el esquema moderno" (Penny 2005/2002: 196). Sobre el tiempo y modo, los cambios también son escasos. De manera concreta se podrían resaltar, por ejemplo, el cambio de *cantara* de indicativo a subjuntivo y el uso de *cantare,* en el sistema moderno, como futuro de subjuntivo.

Tabla 5.16: Los primeros 50 sufijos del siglo XVI del CHEM ordenados por afijalidad (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
31	~adas	52	0.07178	0.9408	0.9034	0.5778	0.7356	0.6387
32	~ían	69	0.1565	0.9284	0.8279	0.8734	0.8044	0.6376
33	~ara	28	0.06654	0.9249	0.9212	0.6512	0.02531	0.6375
34	~arse	34	0.04685	0.9254	0.9372	0.7391	0.8052	0.6365
35	~es	329	0.2091	0.9611	0.7283	0.3786	0.5117	0.6328
36	~aré	20	0.04872	0.8924	0.9284	0.7143	0.7167	0.6231
37	~are	19	0.05331	0.9155	0.8976	0.5938	0.422	0.6221
38	~ir	64	0.06744	0.9043	0.8931	0.6095	0.9156	0.6216
39	~iese	76	0.07738	0.9336	0.8487	0.7677	0.8111	0.6199
40	~ió	93	0.09501	0.932	0.8125	0.6596	0.8422	0.6132
41	~iesen	45	0.05509	0.8979	0.8839	0.7895	0.7117	0.6123
42	~iere	31	0.04486	0.8973	0.8805	0.7381	0.8065	0.6075
43	~aban	45	0.06638	0.9317	0.8206	0.7759	0.8504	0.6062
44	~se	377	0.198	0.9583	0.6544	0.706	0.251	0.6036
45	~ja	57	0.06182	0.873	0.8729	0.6951	0.7787	0.6026
46	~ida	40	0.06005	0.8757	0.8617	0.6897	0.5878	0.5991
47	~lo	102	0.08785	0.9346	0.7721	0.3806	0.268	0.5982
48	~ia	154	0.1184	0.9255	0.7479	0.4118	0.5513	0.5972
49	~ian	66	0.09176	0.9151	0.7824	0.7674	0.905	0.5964
50	~arán	17	0.03389	0.888	0.8672	0.68	0.619	0.5964

Todos los sufijos de flexión nominal fueron descubiertos, como puede verse en la Tabla 5.17. Para evaluar los sufijos de flexión verbal, es necesario recordar que el método no separa morfemas al interior de los sufijos, por lo que no da cuenta por separado de las tres vocales temáticas del español (~a~, ~e~, ~i~) ni de los morfemas de persona-número y tiempo-modo-aspecto. Así, la evaluación la hago con base en las formas aglutinadas.

Tabla 5.17: Sufijos de flexión nominal descubiertos

Flexión	Afijos	Porcentaje descubierto
De género	~a, ~e, ~o.	100%
De número	~es, ~s	
De género y número	~as, ~es, ~os	

En la Tabla 5.18 se muestra el resultado del descubrimiento de formas impersonales. Como puede apreciarse, se descubrieron todos los afijos. En esta tabla, incluyo los segmentos descubiertos como muestra de su variedad.

Tabla 5.18: Sufijos de flexión verbal impersonal descubiertos

Modo	**Afijos esperados**	**Porcentaje descubierto**	**Afijos descubiertos**
Infinitivo	~ar, ~er, ~ir	100%	(todos)
Gerundio	~ndo		~ando, ~iendo, ~yendo, ~endo, ~jendo, ~ndo Con enclíticos: ~ándole, ~ándose, ~iéndole, ~andose, ~andole, ~ndose, ~iendoles, ~iéndose, ~iendole, ~iendose, ~ndole, ~ándola, ~iéndola, ~ndoles, ~yendose, ~ndolos, ~ándoles, ~éndole, ~andoles, ~iéndolo, ~endole, ~ándolas, ~éndolo, ~ándolo, ~éndoselo, ~andoselo, ~iéndoselo, ~endolos
Participio	~do		~ado, ~ados, ~ido, ~idos, ~do, ~ydo, ~jdo, ~dos, ~ydos, ~ida, ~adas, ~ada

La Tabla 5.19 consigna los afijos descubiertos asociados a los paradigmas del modo indicativo. Marco los que no fueron encontrados en el catálogo y pongo un porcentaje para dar idea del grado de efectividad del método.

Tabla 5.19: Sufijos de flexión verbal del modo indicativo descubiertos

Indicativo	**1ª conjugación**	**2ª/3ª conjugación**	**Porcentaje descubierto**
Presente			
1ª	~o		100%
2ª	~as	~es	
3ª	~a	~e	
1ª	~ámos	~émos, ~ímos	
2ª	~áis	~éis, ~íis	
3ª	~an	~en	
Pasado simple			
1ª	~é	~i	83%
2ª	~áste	~íste	
3ª	~ó	~ió	
1ª	~ámos	~ímos	
2ª	**~ásteis**	**~ísteis**	
3ª	~áron	~iéron	

Tabla 5.19: Sufijos de flexión verbal del modo indicativo descubiertos (continuación)

Indicativo	**1ª conjugación**	**2ª/3ª conjugación**	**Porcentaje descubierto**
Pasado imperfecto			
1ª	~ába	~ía	83%
2ª	~ábas	~ías	
3ª	~ába	~ía	
1ª	~ábamos	~íamos	
2ª	**~ábais**	**~íais**	
3ª	~ában	~ían	
Futuro			
1ª	~aré	~eré, ~iré	83%
2ª	~arás	~erás, ~irás	
3ª	~ará	~erá, ~irá	
1ª	~arémos	~erémos, ~irémos	
2ª	**~aréis**	**~eréis, ~iréis**	
3ª	~arán	~erán, ~irán	
Condicional			
1ª	~aría	~ería, ~iría	56%
2ª	**~arías**	~erías, **~irías**	
3ª	~aría	~ería, ~iría	
1ª	**~aríamos**	**~eríamos, ~iríamos**	
2ª	**~aríais**	**~eríais, ~iríais**	
3ª	~arían	~erían, ~irían	

Por último, se pueden ver en la Tabla 5.20 los sufijos descubiertos que pertenecen al modo subjuntivo. En el caso del pasado imperfecto se descubrió al menos una forma de las alternantes, así que se tomó como descubierto el 100%.[9]

[9] Llaman la atención los afijos del pasado imperfecto de subjuntivo ya que parecen dar muestra de la competencia entre las formas basadas en ~se contra las basadas en ~ra, pero sólo un estudio a profundidad daría validez a esto.

Tabla 5.20: Sufijos de flexión verbal del modo subjuntivo descubiertos

Subjuntivo	**1ª conjugación**	**2ª/3ª conjugación**	**Porcentaje descubierto**
Presente			
1ª	~e	~a	100%
2ª	~es	~as	
3ª	~e	~a	
1ª	~émos	~ámos	
2ª	~éis	~áis	
3ª	~en	~an	
Pasado imperfecto			
1ª	~ára/áse	~iéra/iése	100%
2ª	~áras/**áses**	~iéras/**iéses**	
3ª	~ára/áse	~iéra/iése	
1ª	~áramos/**ásemos**	~iéramos/**iésemos**	
2ª	**~árais**/áseis	**~iérais**/iéseis	
3ª	~áran/ásen	~iéran/iésen	
Futuro			
1ª	~áre	~iére	75%
2ª	~áres	~iéres	
3ª	~áre	~iére	
1ª	**~áremos**	~iéremos	
2ª	**~áreis**	**~iéreis**	
3ª	~áren	~iéren	

Creo que la revisión anterior deja clara la efectividad de método y al mismo tiempo consigna algunos segmentos faltantes que bien pudieron ser agregados manualmente a la lista final de afijos descubiertos. Sin embargo, esto contravendría la intención de contar con un método de descubrimiento de reglas morfológicas lo menos supervisado posible. Dejo pendiente una revisión de los afijos derivativos, enclíticos y segmentos dudosos descubiertos.

Medina propone dos probabilidades asociadas a cada segmento. La primera (prob. 1) se refiere a la posibilidad de que el segmento sea un sufijo en un tipo de palabra. Esto es, la probabilidad de que al tomar cualquier entrada de la lista de tipos (*types*) del corpus, si el segmento aparece, éste sea un sufijo. La segunda probabilidad (prob. 2) es a nivel de ocurrencia de palabra (*token*) y se refiere a la probabilidad de que si en una cadena de ocurrencias de palabras aparece el segmento, éste sea un sufijo. En la Tabla 5.21 muestro los primeros 50 sufijos ordenados por la segunda probabilidad. Es interesante que los sufijos se vean tan diferentes a los de la Tabla 5.16. Por ejemplo, nótese que, como lo consignó Medina (2000, 2003), las altas probabilidades no están asociadas a altas medidas de afijalidad. De hecho, sólo 19 de estos cincuenta segmentos superan el 0.5. Además, aparecen sufijos más largos en comparación con los ordenados por afijalidad. Resalto esta situación y muestro estos sufijos, ya que realicé experimentos con este ordenamiento.

Tabla 5.21: Los primeros 50 sufijo del siglo XVI del CHEM ordenados por prob. 2

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
1	~erle	10	0.05096	0.8155	0.706	1	1	0.5241
2	~amyento	3	0.008536	0.4554	0.3829	1	1	0.2823
3	~jestad	1	0.008963	0.4315	0.2219	0.3333	0.9953	0.2208
4	~jdo	22	0.04679	0.8305	0.7082	0.9565	0.9833	0.5285
5	~guiente	3	0.0111	0.4441	0.4644	0.6	0.9783	0.3065
6	~dre	2	0.006402	0.6041	0.5142	0.2857	0.9769	0.3749
7	~cino	1	0.01536	0.5034	0.2416	0.5	0.9677	0.2535
8	~and	1	0.002561	0.5034	0.3829	0.3333	0.9672	0.2963
9	~és	3	0.01366	0.4409	0.2383	0.3333	0.9376	0.231
10	~jr	21	0.03567	0.8125	0.6384	0.84	0.9365	0.4955
11	~óse	10	0.04584	0.7142	0.5996	0.9091	0.9286	0.4532
12	~aron	132	0.1927	0.9671	0.9244	0.7543	0.9273	0.6947
13	~én	7	0.0128	0.6601	0.4952	0.6364	0.9269	0.3894
14	~liçia	1	0.008963	0.4315	0.2416	0.5	0.9231	0.2274
15	~ll	4	0.01536	0.5898	0.5996	0.3636	0.9216	0.4016
16	~ze	13	0.02049	0.6616	0.6882	0.4483	0.9203	0.4568
17	~ent	3	0.006402	0.2269	0.4644	0.6	0.92	0.2326
18	~hiz	1	0.008963	0.4315	0.2416	0.5	0.9167	0.2274
19	~ir	64	0.06744	0.9043	0.8931	0.6095	0.9156	0.6216
20	~sa	46	0.02516	0.793	0.7747	0.5287	0.9143	0.531
21	~erlo	11	0.04959	0.7974	0.6942	0.8462	0.913	0.5137
22	~ó	347	0.5298	0.9649	0.8552	0.8032	0.9059	0.7833
23	~ian	66	0.09176	0.9151	0.7824	0.7674	0.905	0.5964
24	~namentos	1	0.005122	0.5034	0.2416	0.5	0.8889	0.25
25	~ándolo	5	0.0146	0.355	0.3526	0.8333	0.8889	0.2407
26	~ándose	11	0.0802	0.8313	0.7314	0.8462	0.8824	0.5476
27	~vo	26	0.0197	0.7519	0.7449	0.7879	0.881	0.5055
28	~iéndose	7	0.01719	0.6358	0.6585	0.7778	0.875	0.4372
29	~verençia	1	0.003841	0.6712	0.2416	0.5	0.875	0.3055
30	~avan	66	0.09089	0.9543	0.8997	0.8049	0.8708	0.6483
31	~té	3	0.006402	0.505	0.6093	0.4286	0.8667	0.3736
32	~mos	162	0.1341	0.9557	0.4615	0.7465	0.8638	0.5171
33	~jendo	10	0.03303	0.7219	0.6339	0.7692	0.8636	0.4629
34	~die	1	0.005122	0.5034	0.3624	0.25	0.8571	0.2903
35	~aba	98	0.1421	0.9596	0.8868	0.7206	0.8513	0.6628
36	~aban	45	0.06638	0.9317	0.8206	0.7759	0.8504	0.6062
37	~ando	121	0.138	0.9648	0.9503	0.6722	0.8472	0.6844
38	~marca	1	0.003841	0.6712	0.2416	0.5	0.8462	0.3055
39	~bo	23	0.01692	0.7636	0.7037	0.575	0.8457	0.4947
40	~ron	170	0.1399	0.9465	0.246	0.5397	0.8443	0.4442

Tabla 5.21: Los primeros 50 sufijo del siglo XVI del CHEM ordenados por prob. 2 (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
41	~ía	170	0.2001	0.9471	0.7897	0.8252	0.8442	0.6456
42	~ió	93	0.09501	0.932	0.8125	0.6596	0.8422	0.6132
43	~cho	11	0.01781	0.6979	0.6848	0.2973	0.8408	0.4668
44	~iente	16	0.007682	0.4952	0.7628	0.4848	0.8408	0.4219
45	~erse	20	0.05256	0.8046	0.7894	0.7143	0.8393	0.5489
46	~jan	10	0.04597	0.7491	0.6342	0.625	0.8333	0.4764
47	~eçesidad	1	0.01152	0.8949	0.2416	0.5	0.8333	0.3827
48	~ándola	5	0.04942	0.6201	0.4635	0.8333	0.8333	0.3777
49	~amos	74	0.0802	0.897	0.9449	0.7475	0.8327	0.6407
50	~só	13	0.02039	0.6394	0.745	0.5652	0.8291	0.4683

Aunque con una evaluación sencilla, ha quedado demostrado que una porción considerable de los afijos de flexión se descubrió mediante el método de Medina. Es importante resaltar que sin echar mano de modelos como los ampliamente difundidos en el campo de la inteligencia artificial, o de mucha formalización como la usada en los basados en reglas simbólicas construidas a priori, el método logra descubrir exitosamente unidades morfológicas. Además, es crucial que sea un método no supervisado, esto es, que no necesite información explícita para realizar el descubrimiento. Así, los afijos descubiertos serán agregados a las plantillas con el fin de generar nuevas reglas de carácter morfológico. En la siguiente sección explico este proceso.

5.4. Generación de reglas

En esta sección, reporto los resultados de varios experimentos para la generación de reglas para identificar categorías gramaticales. El primer apartado describe las reglas obtenidas con el método original. Los siguientes, las obtenidas con el uso de sufijos y otras modificaciones que se detallarán adelante. El único cambio que hice, tanto para el método original, como para todos los demás experimentos, fue la eliminación de las plantillas que producían reglas con caracteres iniciales. Esto se debe a que, como ya dije, no involucré a los prefijos en la experimentación.

Para generar las reglas es necesario ejecutar un programa de cómputo que utiliza varios archivos previamente creados. Por tratarse de una cuestión completamente computacional, los pasos necesarios para crear los archivos y generar las reglas se detallan en el anexo **¡Error! No se encuentra el origen de la referencia.**. En los siguientes apartados no hago referencia a ningún aspecto relacionado con estos pasos.

5.4.1. Método original de Brill (experimento 1)

Para realizar el proceso de entrenamiento con el método original, se eliminaron las plantillas que obtenían reglas con caracteres iniciales, todas las demás plantillas se mantuvieron intactas. El umbral para calificar reglas, según la mejoría en el etiquetado, fue establecido en 2. Esto significa que el programa generó reglas válidas hasta que una de ellas obtuvo una calificación menor al umbral.

El número de reglas generadas fue de **139**. En la Tabla 5.22 muestro las primeras 35 (25%); el conjunto completo puede verse en el anexo **¡Error! No se encuentra el origen de la referencia.**. Cada regla, compuesta de varias partes, se interpreta de acuerdo con su plantilla de origen. En general, en todas aparece al final una cantidad que equivale a la calificación de la regla; la otra cantidad, que en ocasiones aparece, es la longitud de la cadena de caracteres que debe aparecer al final de la palabra (uno, dos, tres o cuatro). Algunas reglas incluyen una F al inicio de su nombre, por ejemplo FHASSUF, que indica un etiquetado sólo si la palabra ya tiene una determinada etiqueta; si el nombre no incluye la F, por ejemplo HASSUF, se cambia la etiqueta sin importar cuál sea.

Tabla 5.22: Primeras 35 reglas generadas por el método original de Brill

Núm.	Regla
1.	NCMS a fhassuf 1 NCFS 109
2.	NCMS s fhassuf 1 NCMP 88.8761904761905
3.	NCMS do fhassuf 2 VMP00SM 56.5
4.	NCMS r fhassuf 1 VMN0000 50.95
5.	ó hassuf 1 VMIS3S0 39
6.	NCMP as fhassuf 2 NCFP 36
7.	VMP00SM ndo fhassuf 3 VMG0000 31.3333333333333
8.	NCMS la fgoodright NCFS 27
9.	NCMS an fhassuf 2 VMII3P0 24
10.	ron hassuf 3 VMIS3P0 18.1818181818182
11.	se hassuf 2 VMSI3S0 17
12.	NCMS on fhassuf 2 NCFS 16
13.	VMN0000 or fhassuf 2 NCMS 14
14.	r addsuf 1 VMIP3S0 13.0714285714286
15.	d hassuf 1 NCFS 13
16.	NCFS que fgoodright VMII3S0 12.3666666666667
17.	años goodleft DN0CP0 11.6666666666667
18.	dos hassuf 3 VMP00PM 10.5
19.	ava hassuf 3 VMII3S0 9
20.	sen hassuf 3 VMSI3P0 9
21.	NCMS que fgoodright VMIP3S0 8.48809523809524
22.	1 char W 8
23.	ren hassuf 3 VMSF3P0 8
24.	NCMP las fgoodright NCFP 7

Tabla 5.22: Primeras 35 reglas generadas por el método original de Brill (continuación)

Núm.	Regla
25.	vuestra goodleft APS000 6.92553191489362
26.	NCFS ca fhassuf 2 AQ0FS0 6
27.	gan hassuf 3 VMSP3P0 6
28.	mos hassuf 3 VMIP1P0 6
29.	iado hassuf 4 NCMS 5.33333333333333
30.	muy goodright AQ0MS0 5.02777777777778
31.	ía hassuf 2 VMII3S0 5
32.	NCFS n faddsuf 1 VMSP3S0 5
33.	NCMS en fhassuf 2 VMSP3P0 5
34.	are hassuf 3 VMSF3S0 5
35.	VMIP3S0 o fhassuf 1 VMIS3S0 4.66666666666667

En seguida, explicaré algunas reglas con el fin de que se puedan interpretar y se siga la discusión sobre las mismas. Se hará evidente que muchas de ellas no parecen tener sentido (por ejemplo, la 43 presentada abajo), pero se presentan tal cual al ser las generadas por el método. Para lograr una mejor comprensión, dividiré las reglas generadas en tres tipos (véanse Tabla 5.23, Tabla 5.24 y Tabla 5.25).

Tabla 5.23: Reglas basadas en caracteres al final de la palabra (FHASSUF, HASSUF, FADDSUF, ADDSUF, FDELETESUF, DELETESUF)

No.	Regla	Interpretación
1.	NCMS a FHASSUF 1 NCFS 109	Cambia la etiqueta NCMS (nombre común masculino singular), por la etiqueta NCFS (nombre común femenino singular), si la palabra contiene la cadena de caracteres “a” al final (por ejemplo: *yglesia*/NCFS).
5.	ó HASSUF 1 VMIS3S0 39	Cambia la etiqueta (cualquiera que sea) por la etiqueta VMIS3S0 (verbo principal indicativo pasado simple tercera persona singular), si la palabra contiene la cadena de caracteres “ó” al final (por ejemplo: *partió*/VMIS3S0).
32.	NCFS n FADDSUF 1 VMSP3S0 5	Cambia la etiqueta NCFS (nombre común femenino singular), por la etiqueta VMSP3S0 (verbo principal subjuntivo presente tercera persona singular), si al agregar la cadena de caracteres “n” al final de la palabra el resultado es una palabra del corpus (por ejemplo: *dexe*/VMSP3S0).
14.	r ADDSUF 1 VMIP3S0 13.0714	Cambia la etiqueta (cualquiera que sea) por la etiqueta VMIP3S0 (verbo principal indicativo presente tercera persona singular), si al agregar la cadena de caracteres “r” al final de la palabra el resultado es una palabra del corpus. (por ejemplo: *llama*/VMIP3S0)

Tabla 5.23: Reglas basadas en caracteres al final de la palabra (FHASSUF, HASSUF, FADDSUF, ADDSUF, FDELETESUF, DELETESUF) (continuación)

No.	Regla	Interpretación
39.	VMIP3S0 n FDELETESUF 1 VMIP3P0 4	Cambia la etiqueta VMIP3S0 (verbo principal indicativo presente tercera persona singular), por la etiqueta VMIP3P0 (verbo principal indicativo presente tercera persona plural), si al quitar la cadena de caracteres "n" del final de la palabra el resultado es una palabra del corpus (por ejemplo: *hazen*/VMIP3P0).
53.	ll DELETESUF 2 DN0CP0 3	Cambia la etiqueta (cualquiera que sea) por la etiqueta DN0CP0 (determinante numeral común plural), si al quitar la cadena de caracteres "ll" del final de la palabra el resultado es una palabra del corpus (por ejemplo: *mjll*/DN0CP0).

Tabla 5.24: Reglas basadas en una letra al interior de la palabra (FCHAR, CHAR)

No.	Regla	Interpretación
43.	VMIC3S0 o FCHAR NCFS 3	Cambia la etiqueta VMIC3S0 (verbo principal indicativo condicional tercera persona singular), por la etiqueta NCFS (nombre común femenino singular), si la palabra contiene la cadena de caracteres "o".
22.	1 CHAR W 8	Cambia la etiqueta (cualquiera que sea) por la etiqueta W (fecha-hora), si la palabra contiene la cadena de caracteres "1" (uno) (por ejemplo: *13*/W).

Tabla 5.25: Reglas basadas en bigramas de palabras (FGOODRIGHT, GOODRIGHT, FGOODLEFT, GOODLEFT)

No.	Regla	Interpretación
8.	NCMS la FGOODRIGHT NCFS 27	Cambia la etiqueta NCMS (nombre común masculino singular), por la etiqueta NCFS (nombre común femenino singular), si la palabra aparece después de la palabra "la" (por ejemplo: *la*/DA0FS0 *carçel*/NCFS).
30.	muy GOODRIGHT AQ0MS0 5.02777	Cambia la etiqueta (cualquiera que sea) por la etiqueta AQ0MS0 (adjetivo calificativo masculino singular) si la palabra aparece después de la palabra "muy" (por ejemplo: *muy*/RG *grand*/AQ0MS0).
56.	VMIP3S0 lo FGOODLEFT RM 3	Cambia la etiqueta VMIP3S0 (verbo principal indicativo presente tercera persona singular), por la etiqueta RM (adverbio de modo), si la palabra aparece antes de la palabra "lo" (por ejemplo: *bien*/RM *lo*/DA0NS0 *que*/PR0CN00 *haze*/VMIP3S0).
17.	años GOODLEFT DN0CP0 11.66666	Cambia la etiqueta (cualquiera que sea) por la etiqueta DN0CP0 (determinante numeral común plural), si la palabra aparece antes de la palabra "años" (por ejemplo: *tres*/DN0CP0 *años*/NCMP).

Ahora, presento una discusión sobre el conjunto de reglas. Lo primero que se observa es que las siete iniciales son de tipo HASSUF (FHASSUS y HASSUF), es decir, miran el final de la palabra. De hecho, este tipo de reglas representaron el 50% de todas las reglas (70/139), las de ADDSUF

(FADDSUF y ADDSUF) el 6% (8/139) y las de DELETESUF (FDELETESUF y DELETESUF) el 8% (11/139); todas sumaron un total de 64%. Las que usan un carácter al interior de la palabra fueron el 9% (13/139). El 18% (25/139) fue para las que toman en cuenta la palabra anterior (GOODRIGHT) y el 9% (12/139) la palabra siguiente (GOODLEFT).

Lo anterior da muestra de la importancia de cada tipo de regla para el etiquetado. Como era de esperarse, para este corpus las que toman en cuenta el final de la palabra son las más generalizadoras y enseguida las que se basan en la palabra anterior. De hecho, hablando de las 35 reglas de la Tabla 5.22, el 77% (27/35) se refiere a caracteres finales.

Sobre el significado de las reglas, la primera utiliza el segmento final "a" como indicador de nombres femeninos (ver primera entrada en Tabla 5.22). La segunda toma la "s" como marca de plural en sustantivos. La tercera, cuarta y quinta identifican verbos: la cadena final "do" para identificar participios, "r" para infinitivos y la "ó" asociada a verbo indicativo pasado simple tercera persona singular. La quinta, a diferencia de las anteriores, no se fija en la etiqueta que tiene actualmente la palabra y, sin importar cuál sea, la cambia.

A medida que se revisan más reglas, es claro que las siguientes comienzan a ser más específicas y en algunos casos corrigen a las anteriores; esto se nota porque exigen una etiqueta para la palabra a etiquetar. Ejemplo de esto es la sexta, que mira la terminación "as" para marcar nombres como femeninos plurales (*navajas*/NCFP), ajustando posibles generalizaciones de la regla 2, que los había etiquetado como masculinos (NCMP). Por su parte, la siete asigna gerundio si la palabra termina con "ndo" (*echándo*/VMG0000) y corrige errores de la regla 3, que etiquetó con participio (VMP00SM) todas las palabras terminadas en "do".

De hecho, lo anterior muestra el espíritu del método, ya que desde la primera regla se está corrigiendo, o mejorando, el etiquetado inicial. Éste consistió en marcar todas las palabras con minúscula inicial como nombres comunes masculinos singulares (NCMS); de aquí que las primeras reglas se apliquen cuando la palabra tiene esta etiqueta. También de allí se entiende que la primera y más productiva sea la de pasar muchos de esos sustantivos masculinos a femeninos.

La regla 8 es de carácter contextual, ya que revisa si la palabra anterior es "la", en ese caso, la palabra actual, a la derecha de "la" (FGOODRIGHT), es etiquetada como nombre femenino singular. Las reglas 9, 10 y 11 usan finales de palabra para marcar verbos: "an" para indicativo pasado imperfecto tercera persona plural (*baylavan*/VMII3P0), "ron" para indicativo pasado simple tercera persona plural (*hizieron*/VMIS3P0) y "se" para subjuntivo imperfecto tercera singular (*ayunase*/VMSI3S0). La siguiente regla (12), es interesante porque indica que en nuestro corpus la terminación "on" está más asociada a sustantivos femeninos que masculinos (*conserbaçion*/NCFS).

La 13 corrige generalizaciones de la regla 4, que marcó infinitivos a palabras terminadas con "r", ésta cambia a sustantivos masculinos singulares los que terminan en "or", una clara marca que no pertenece al infinitivo (*jmperador*/NCMS).

La regla 14 marca una palabra como verbo indicativo presente tercera persona singular si al agregar una "r" al final se convierte en una palabra del lexicón (ADDSUF). La 15 asocia la terminación "d" con nombres femeninos singulares (*majestad*/NCFS). La 16 cambia éstos a verbos indicativos en pasado imperfecto tercera persona singular si están precedidos por "que" (*que*/CS *estava*/VMII3S0). La regla 17 es la primera de tipo GOODLEFT y resultó productiva por la frecuencia de combinaciones de determinante numeral seguido de la palabra "años".

Las siguientes reglas (18 - 35) se interpretan de la misma forma que las ya explicadas y para no extenderme, resaltaré sólo algunas de las más interesantes. La regla 19 refleja la variación ortográfica de la época, ya que asocia la terminación "ava" con indicativo pasado imperfecto tercera persona de singular (*tornava*/VMII3S0). La regla 22 es del tipo que revisa una letra dentro de la palabra, en este caso es el '1' (*uno*) y fue productiva porque en lo textos del corpus encontré fechas de manera frecuente.

La regla 23 es para subjuntivos futuros detectados por la terminación "ren" (*vinjeren*/VMSF3P0). La regla 24 es prácticamente la misma que la 8, ya que se basa en la palabra anterior "las" como marca de femenino plural (*las*/DA0FP0 *rayzes*/NCFP). La 25 es un ejemplo de que las reglas se basan en construcciones frecuentes, en este caso resultó abundante encontrar una preposición (APS000) antes de la palabra *vuestra* (*de*/APS000 *vuestra*/DP2FSS). Un ejemplo parecido es la 30, que marca adjetivo (AQ0MS0) después de *muy*. Las reglas 32, 33 y 34 pueden dar idea del uso frecuente de subjuntivo en español del siglo XVI, razón por la cual estas reglas aparecen entre las más generalizadoras; las tres etiquetan subjuntivos con las terminaciones "n" (presente tercera singular), "en" (presente tercera plural) y "are" (futuro tercera singular).

No dudo que dentro del conjunto de reglas que faltan por revisar existan algunas interesantes, pero prefiero seguir adelante con los experimentos basados en las modificaciones propuestas en esta tesis. Lo que sí conviene anotar aquí es que las reglas obtenidas reflejan la importancia de mirar el interior de la palabra para la asignación de categorías gramaticales en el español. Además, nos dicen que el método de Brill produce reglas útiles con poco corpus.

He expuesto los resultados de aplicar el método original de Brill al corpus de estudio. También, apunté que las basadas en partes finales de palabras son las más abundantes y en segundo lugar las que toman en cuenta la palabra anterior. Finalmente, expliqué cómo se interpretarían las

reglas obtenidas y discutí algunas de las más generalizadoras, es decir, las que aparecen al inicio de la lista ordenada.

Ahora bien, una vez obtenidos los resultados anteriores, ¿sería posible esperar que, con los sufijos previamente descubiertos incluidos en las plantillas de transformación, las reglas obtenidas de éstas sean más productivas que las reglas del método original? A continuación expondré los resultados del método con los sufijos descubiertos y otras modificaciones.

5.4.2. Método con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF (experimento 2)

En este apartado reporto los resultados del experimento 2 con modificaciones al método original de Brill (experimento 1). Primero detallo los cambios realizados y después comparo las reglas obtenidas. El primer cambio fue la integración de los sufijos descubiertos en las plantillas de transformación. Éstas se modificaron para generar reglas que tomen los sufijos en lugar de sólo probar con caracteres finales (uno, dos, tres o cuatro). Como expliqué en la sección **¡Error! No se encuentra el origen de la referencia.**, el método de Brill toma cada letra final de cada palabra del corpus y genera con ella una regla, el conjunto generado se evalúa y se obtiene la mejor. Con la inclusión de los afijos, únicamente se generaron las reglas con los afijos probables de cada palabra.

El segundo cambio fue la creación de otra plantilla para generar un nuevo tipo de regla. Ésta se basó tanto en la morfología de la palabra como en su contexto y produjo varias reglas útiles que mostraré más abajo. Su formato es el siguiente:

- Cambia la etiqueta X por la etiqueta Y si la palabra tiene el sufijo S y la palabra anterior tiene la etiqueta Z.

Esta plantilla genera un nuevo tipo de regla (FTAGANTSUFMP) que se explicará abajo. Es importante decir que en este caso se asoció a cada palabra un solo sufijo: el primero que coincidiera en lista ordenada por la segunda probabilidad. Se decidió este ordenamiento ya que sería la probabilidad de que, si el segmento aparece en la palabra, éste sea su sufijo.

Como tercera modificación, se eliminaron dos plantillas más. El método original incluye plantillas que verifican si los caracteres finales (uno, dos, tres o cuatro) al ser agregados (ADDSUF) o eliminados (DELETESUF) dan como resultado una palabra del corpus, en caso afirmativo se genera una regla. Este procedimiento intenta buscar segmentos finales más útiles para las reglas, pero está muy alejado de ser un método de descubrimiento de morfología. Por tanto, ya que se propone utilizar unidades afijales previamente descubiertas, hice experimentos sin estas plantillas.

Una vez terminado el proceso de entrenamiento, fueron obtenidas **126** reglas, 13 menos que con el método original. La última regla fue la misma que la del conjunto generado con el método de Brill (*señores GOODLEFT DD0MP0*) y obtuvo la misma calificación: 1.95, lo que indica que ambos conjuntos de reglas llegaron al mismo nivel de etiquetado. El conjunto total de reglas pueden verse en el anexo **¡Error! No se encuentra el origen de la referencia.**. Enseguida pongo las primeras 35 (Tabla 5.26).

Tabla 5.26: Primeras 35 reglas generadas por el método con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF

Num.	Regla
1	NCMS a fhassuf 1 NCFS 109
2	NCMS s fhassuf 1 NCMP 88.8761904761905
3	NCMS do fhassuf 2 VMP00SM 56.5
4	NCMS r fhassuf 1 VMN0000 51.95
5	ó hassuf 1 VMIS3S0 39
6	NCMP as fhassuf 2 NCFP 36
7	VMP00SM ndo fhassuf 3 VMG0000 31.3333333333333
8	NCMS la fgoodright NCFS 27
9	NCMS an fhassuf 2 VMII3P0 24
10	ron hassuf 3 VMIS3P0 18.1818181818182
11	se hassuf 2 VMSI3S0 17
12	NCMS ion fhassuf 3 NCFS 16
13	VMN0000 or fhassuf 2 NCMS 14
14	d hassuf 1 NCFS 13
15	NCFS que fgoodright VMII3S0 12.8666666666667
16	años goodleft DN0CP0 11.6666666666667
17	NCMS que fgoodright VMIP3S0 11.4880952380952
18	VMII3P0 RL an ftagantysufmp VMIP3P0 11
19	dos hassuf 3 VMP00PM 10.5
20	NCFS NCFS ava ftagantysufmp VMII3S0 9
21	sen hassuf 3 VMSI3P0 9
22	1 char W 8
23	ren hassuf 3 VMSF3P0 8
24	NCMP las fgoodright NCFP 7
25	mente hassuf 5 RM 7
26	vuestra goodleft APS000 6.92553191489362
27	NCFS ca fhassuf 2 AQ0FS0 6
28	gan hassuf 3 VMSP3P0 6
29	NCMP Fc mos ftagantysufmp VMIP1P0 6
30	iado hassuf 4 NCMS 5.33333333333333
31	muy goodright AQ0MS0 5.02777777777778
32	NCMS F n ftagantysufmp APS000 5
33	APS000 en fhassuf 2 VMSP3P0 5
34	NCMS PP3CS00 are ftagantysufmp VMSF3S0 5
35	NP00 , fgoodright NCMS 4.14285714285714

En primer lugar, haré un resumen general del tipo de reglas obtenidas. Las de tipo HASSUF disminuyeron en comparación con el método original, sin dejar de ser las más frecuentes, representando ahora el 45% (57/126). El nuevo tipo de regla (FTAGANTSUFMP), que también toma en cuenta los sufijos, representó el 10% (12/126). Las que usan una letra al interior de la palabra resultaron con el 12% (15/126). Las de tipo GOODRIGHT fueron nuevamente las segundas más frecuentes con el 21% (26/126), mientras que las GOODLEFT consiguieron el 13% (16/126).

Si se mira solamente las reglas más generalizadoras, las primeras 35, la proporción de tipos de reglas se parece bastante a la proporción del método original. Así, las que involucran sufijos corresponden al 74% (26/35), contra el 77% del original. Las que revisan una letra al interior de la palabra el 3% (1/35), igual que el original. Finalmente, las de tipo GOODRIGHT el 17% (6/35), contra el 14%; y las de GOODLEFT el 6% (2/35), también idéntico al original.

En otros aspectos, encontré que 75 de ellas se repiten en los dos; 49 del método original no se encuentran en el nuevo método, y 36 del nuevo no fueron generadas en el original; finalmente 15 reglas fueron equivalentes aunque representadas de distinta manera en el nuevo conjunto. Enseguida hablo con mayor detalle de algunas de estas diferencias.

Las reglas uno a la once son las mismas en los dos métodos. La regla 12 resultó equivalente en ambos, con la diferencia de que el nuevo método utilizó un sufijo más grande, es decir, más específico (véase Tabla 5.27). El carácter de equivalente tiene que ver con la etiqueta que se pone como resultado de la regla y del parecido gráfico de los segmentos involucrados.

Tabla 5.27: Comparación de regla 12 del método original y regla 12 del método con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF

Método original	Método con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF
12. NCMS on FHASSUF 2 NCFS 16	12. NCMS ion FHASSUF 3 NCFS 16

Otras reglas iguales fueron las siguientes: 13-17, 19, 21-24, 26-28, 30, 31 y 35. La regla 14 del método original no apareció en el nuevo porque en éste no usé la plantilla para reglas de tipo ADDSUF. A su vez, la regla 18 del nuevo conjunto (*VMII3P0 RL an FTAGANTYSUFMP VMIP3P0*), no apareció en el original porque es producto de la nueva plantilla. Aunque intuitivamente es una regla extraña, con ella se hace un ajuste a la aplicación de la regla 9: se cambia la etiqueta de verbos de indicativo pasado imperfecto tercera persona plural (VMII3P0), precedidos de adverbios de lugar

(RL), por la etiqueta de verbo presente tercera persona plural, si y sólo si la palabra tiene el sufijo ~an[10] (*donde*/RL *posan*/VMIP3P0).

Otras reglas equivalentes fueron la 19 del método original y la 20 de este experimento (Tabla 5.28). La nueva regla etiqueta palabras como verbos de indicativo pasado imperfecto tercera persona singular si incluyen el sufijo ~ava, tienen la etiqueta nombre común femenino singular y antes de ellas está una palabra etiquetada como nombre común femenino singular. La regla del método original sólo revisa que la palabra tenga la terminación "ava" y obtiene la misma calificación.

Tabla 5.28: Comparación de regla 19 del método original y regla 20 del método con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF

Método original	Método con sufijos descubiertos, nueva plantilla de regla y sin plantillas addsuf y deletesuf
19. ava HASSUF 3 VMII3S0 9	20. NCFS NCFS ava FTAGANTYSUFMP VMII3S0 9

También la regla 25 del nuevo conjunto y la 36 del original fueron equivalentes. Lo que llama la atención es que la regla del nuevo experimento aparece antes, además, tiene mejor calificación (Tabla 5.29).

Tabla 5.29: Comparación de regla 36 del método original y regla 25 del método con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF

Método original	Método con sufijos descubiertos, nueva plantilla de regla y sin plantillas addsuf y deletesuf
36. ente HASSUF 4 RM 4.5	25. mente HASSUF 5 RM 7

Esta es una muestra clara de la pertinencia de utilizar el método de descubrimiento de afijos. Por un lado, el hecho de que la regla esté mejor calificada es relevante ya que apoya la idea de que el uso de sufijos mejora el método original. Por otro lado, el método de Brill no podría haber determinado un segmento de palabra de longitud cinco, ya que sólo tomó en cuenta uno, dos, tres o cuatro. Por supuesto que esto podría solucionarse modificando las plantillas para que tomen en cuenta cinco caracteres, el problema es que se tendría que cambiar este parámetro cada vez que nos demos cuenta de que son necesarios segmentos más largos.

En este sentido, el descubrimiento de sufijos con el método de Medina nos brinda la posibilidad de contar con segmentos de longitud no predefinida. En el caso de utilizar el método en

[10] No deja de ser interesante que un adverbio de lugar pueda correlacionarse con el tiempo presente del verbo de la oración donde aparece.

otras lenguas, donde la morfología cambie en cuanto al tamaño de los afijos y secuencias afijales, el método mantendría su efectividad.

Sin embargo, una de las reglas equivalentes entre los dos métodos, que pertenece al método original, resultó con mejor calificación que la del nuevo experimento. Como puede verse en la Tabla 5.30, la diferencia entre las reglas es que la del experimento nuevo apareció cinco posiciones antes que la otra.

Tabla 5.30: Comparación de regla 70 del método original y regla 65 del método con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF

Método original	Método con sufijos descubiertos, nueva plantilla de regla y sin plantillas addsuf y deletesuf
70. ral HASSUF 3 AQ0MS0 2.8	65. NCMS al FHASSUF 2 AQ0MS0 2.56666666666667

Ya que incluí el nuevo tipo de plantilla, consigno en la Tabla 5.31 las reglas obtenidas en el nuevo experimento; cuando es posible, las comparo con las equivalentes en el experimento original.

Tabla 5.31: Reglas generadas con la nueva plantilla FTAGANTYSUFMP en el método con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF

Regla del método con sufijos descubiertos, nueva plantilla de regla y sin plantillas addsuf y deletesuf	Regla equivalente en el método original
18. VMII3P0 RL an ftagantysufmp VMIP3P0 11	(ninguna)
20. NCFS NCFS ava ftagantysufmp VMII3S0 9	19. ava hassuf 3 VMII3S0 9
29. NCMP Fc mos ftagantysufmp VMIP1P0 6	28. mos hassuf 3 VMIP1P0 6
32. NCMS F n ftagantysufmp APS000 5	(ninguna)
34. NCMS PP3CS00 are ftagantysufmp VMSF3S0 5	34. are hassuf 3 VMSF3S0 5
51. NCMS APS000 ll ftagantysufmp DN0CP0 3	53. ll deletesuf 2 DN0CP0 3
68. NP00 DP1CSS so ftagantysufmp NCMS 2	(ninguna)
69. NP00 F j ftagantysufmp NCMS 2	73. j hassuf 1 NCMS 2
74. NCMS APS000 ción ftagantysufmp NCFS 2	(ninguna)
95. NP00 APS000 re ftagantysufmp W 2	103. NP00 bre fhassuf 3 W 2
100. NCMS APS000 ante ftagantysufmp RL 2	(ninguna)
109. NCFS CS aria ftagantysufmp VMIC3S0 2	(ninguna)

Estas reglas se interpretan de la siguiente manera: la primera etiqueta es la de la palabra en cuestión; la segunda, la de la anterior; luego sigue el sufijo asociado a la palabra a etiquetar y finalmente la nueva etiqueta. Para ejemplificar, tómese la número 34. Esta regla establece que si la palabra en cuestión está etiquetada como NCMS, su palabra anterior tiene PP3CS00 (pronombre personal) y aparece el sufijo ~are, entonces se cambia su etiqueta por VMSF3S0 (verbo subjuntivo futuro). Una construcción en donde aplicaría esta regla sería: su/DP3CS0 Señoría/NCFS le/**PP3CS00** mand**are**/**VMSF3S0**.

Para cerrar esta sección, explico el ejemplo de los verbos de indicativo pasado imperfecto tercera persona plural. En ambos conjuntos, la regla nueve (9. *NCMS an fhassuf 2 VMII3P0*) cambia la etiqueta del etiquetado inicial por la de verbo (*estaban*/VMII3P0). Después, en el conjunto generado por el nuevo método surge una regla (18. *VMII3P0 RL an ftagantysufmp VMIP3P0*) que ajusta posibles errores de la anterior (*allí*/RL *hechaban*/VMII3P0), esto permitiría imaginar que una regla similar deberá surgir en el método original para resolver el mismo problema. Esto no sucede sino hasta la regla 58 (58. *VMII3P0 tan fhassuf 3 VMIP3P0*) y luego en la 117 se vuelven a corregir posibles fallas (117. *VMII3P0 e fgoodright VMSP3P0*). Lo que llama la atención es que en el conjunto del nuevo experimento, ya no aparece ninguna regla que corrija el etiquetado de la 9, porque aparentemente ya no es necesaria.

En conclusión, las modificaciones hechas al método original de Brill han producido un conjunto menor de reglas, pero no es claro por qué. Será necesario averiguar a qué se debe esta mejora. Por tal razón, en los siguientes apartados reporto los resultados de experimentos adicionales en los que cambio distintos aspectos del método propuesto en esta sección y comparo resultados.

5.4.3. Método con sufijos previamente descubiertos, nueva plantilla de regla, sin plantillas ADDSUF Y DELETESUF, y cambio de orden de sufijos (experimento 3)

Como había mencionado, el método de descubrimiento de afijos genera un catálogo de sufijos con dos probabilidades y un índice de afijalidad. El experimento anterior se realizó con base en el orden de la probabilidad de que un sufijo lo fuera en una palabra (prob. 2). Así, presento a continuación el resultado del cambio de orden de sufijos.

Volví a generar las reglas con los sufijos ordenados por el índice de afijalidad, todos los demás aspectos del método se mantuvieron inmóviles. El resultado obtenido fue de **125** reglas, una menos que el experimento anterior (experimento 2). De éstas, 110 fueron exactamente las mismas y siete equivalentes por cambio del tipo HASSUF al tipo FTAGANTYSUFMP. Nueve reglas del experimento anterior no aparecieron en el conjunto del experimento actual. A su vez, ocho del actual no están en el anterior.

El impacto del cambio de orden de los sufijos afecta directamente a las reglas de tipo FTAGANTYSUFMP, como puede verse en la Tabla 5.32. Ambas reglas identifican nombres femeninos singulares, pero una utiliza el sufijo ~ion y la otra ~n. Esto se debe a que más palabras resultaron asociadas al sufijo ~n ya que su lugar en el catálogo de sufijos ordenados por afijalidad es 23 y por probabilidad 2 es el 399. Es decir, entre mas arriba esté, más palabras están asociadas a él. En cambio, si está en la posición 399, otros sufijos (~aron, ~ian, ~avan) estarán asociados antes que ~n.

Tabla 5.32: Comparación de regla 12 con sufijos ordenados por probabilidad 2 y regla 12 con sufijos ordenados por afijalidad

Método con sufijos descubiertos ordenados por probabilidad 2, nueva plantilla de regla y sin plantillas addsuf y deletesuf (experimento 1)	Método con sufijos descubiertos ordenados por afijalidad, nueva plantilla de regla y sin plantillas addsuf y deletesuf (experimento 2)
12. NCMS ion fhassuf 3 NCFS 16	12. NCMS F n ftagantysufmp NCFS 16

Ya que el método de Brill consiste en corregir lo anteriormente generalizado, los cambios en las primeras reglas impactan en las siguientes. Además, la calificación de cada regla se calcula de nuevo cada vez que se selecciona una. Ya que se trata de un método que toma en cuenta todas las posibilidades, nada garantiza una secuencia lógica de reglas, es decir, nada asegura que ciertas reglas aparecerán para corregir todos los errores de las anteriores. Es más, podrían aparecer nuevas que reajusten las posibilidades de aparición de otras.

Ahora bien, parece que el orden de los sufijos sólo afectó al nuevo tipo de plantilla, por lo que será necesario realizar experimentos en los que elimine ésta y verifique el efecto del ordenamiento.

5.4.4. Método con sufijos previamente descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF Y DELETESUF (experimentos 4 y 5)

En esta sección reporto los resultados de la generación de reglas sin el uso de la nueva plantilla. Además, experimento primero con los sufijos ordenados por prob. 2 (experimento 4) y luego por el índice de afijalidad (experimento 5).

Una vez realizado el entrenamiento se obtuvieron **128** reglas (experimento 4), 11 reglas menos que el método original y 2 más que el experimento 2, que incluía la nueva plantilla de regla. Al respecto, ocho de las reglas de tipo FTAGANTYSUFMP fueron representadas por reglas HASSUF o FHASSUF y cuatro no aparecieron en el nuevo conjunto (véase Tabla 5.33).

Tabla 5.33: Equivalencia de reglas de la plantilla FTAGANTYSUFMP en el método con sufijos descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF

Regla del método con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF (experimento 1)	Regla del método con sufijos descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF (experimento 3)
18. VMII3P0 RL an ftagantysufmp VMIP3P0 11	(ninguna)
20. NCFS NCFS ava ftagantysufmp VMII3S0 9	19. ava hassuf 3 VMII3S0 9
29. NCMP Fc mos ftagantysufmp VMIP1P0 6	28. mos hassuf 3 VMIP1P0 6
32. NCMS F n ftagantysufmp APS000 5	(ninguna)
34. NCMS PP3CS00 are ftagantysufmp VMSF3S0 5	32. are hassuf 3 VMSF3S0 5
51. NCMS APS000 ll ftagantysufmp DN0CP0 3	49. ll hassuf 2 DN0CP0 3
68. NP00 DP1CSS so ftagantysufmp NCMS 2	(ninguna)
69. NP00 F j ftagantysufmp NCMS 2	68. j hassuf 1 NCMS 2
74. NCMS APS000 ción ftagantysufmp NCFS 2	73. NCMS ión fhassuf 3 NCFS 2
95. NP00 APS000 re ftagantysufmp W 2	95. NP00 bre fhassuf 3 W 2
100. NCMS APS000 ante ftagantysufmp RL 2	(ninguna)
109. NCFS CS aria ftagantysufmp VMIC3S0 2	111. aria hassuf 4 VMIC3S0 2

El nuevo conjunto de reglas, generado con este experimento, comparte 121 reglas con el conjunto del experimento 2. El método sin plantilla FTAGANTYSUFMP generó siete nuevas reglas que no aparecieron en el que sí la incluye. A su vez, cinco reglas del método que incluye la plantilla no aparecieron en el método sin ella. De las reglas nuevas que generó el experimento 4, en cinco de las siete interviene la etiqueta de verbo de indicativo presente tercera persona plural (VMIP3P0), como puede apreciarse en la Tabla 5.34.

Tabla 5.34: Nuevas reglas del método con sufijos descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF, asociadas a la etiqueta VMIP3P0

Num.	Regla
53.	san hassuf 3 VMIP3P0 3
54.	VMII3P0 tan fhassuf 3 VMIP3P0 3
71.	VMIP3P0 de fgoodright NCMS 2
102.	ndan hassuf 4 VMIP3P0 2
103.	VMSP3P0 , fgoodright VMIP3P0 2

Esta etiqueta es precisamente la misma que participa en la primera regla que no aparece en el nuevo experimento y que sí se encuentra en el experimento 2 (*VMII3P0 RL an ftagantysufmp VMIP3P0*), de hecho, ocupa el lugar 18 de la lista. Esto me parece interesante ya que, como decía antes, un cambio en las primeras reglas impacta en el acomodo de las demás. Para el caso concreto

del que hablo, la falta de esa regla permitió que otras entraran en escena. Como puede verse en la Tabla 5.34, surgieron reglas asociadas a sufijos como ~san, ~tan y ~ndan para identificar esta categoría gramatical.

Algunas reglas que muestran el uso de sufijos de longitud mayor a cuatro caracteres se pueden ver en la Tabla 5.35.

Tabla 5.35: Reglas del método con sufijos descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF Y DELETESUF, que usan sufijos mayores de cuatro caracteres

Num.	Regla
24.	mente hassuf 5 RM 7
81.	iones hassuf 5 NCFP 2
115.	arian hassuf 5 VMIC3P0 2
116.	arían hassuf 5 VMIC3P0 2
123.	uestro hassuf 6 DP2MSP 2

El experimento 5, que usa la lista de sufijos ordenados por afijalidad, generó el mismo conjunto de reglas que con el ordenamiento por prob. 2. Por lo anterior, el orden de los sufijos no fue pertinente en la producción de reglas sin nueva plantilla. Realicé dos experimentos más en los que modifico otro aspecto del método. A continuación expongo los resultados.

5.4.5. Método con sufijos previamente descubiertos, sin nueva plantilla de regla, sin plantillas ADDSUF Y DELETESUF y asociando un solo sufijo a cada tipo de palabra (experimentos 6 y 7)

El método original de Brill prueba con todos los caracteres finales (uno, dos, tres o cuatro), de todos los tipos de palabras, para generar reglas que compitan. Siguiendo esta idea, al incluir los sufijos en las plantillas, decidí ajustar los programas de cómputo para que se generaran plantillas usando todos los sufijos asociados a cada tipo de palabra. Esta situación redujo la cantidad de reglas generadas para competir. Sin embargo, tuve la curiosidad de conocer qué pasaría si en lugar de permitir que se generen reglas con todos los sufijos posibles, se generen con un solo sufijo por cada tipo de palabra.

En los presentes experimentos (6 y 7), generé reglas implementando la idea anterior. En el experimento 6 los sufijos fueron ordenados por la segunda probabilidad y en el 7 por el índice de afijalidad. Presento a continuación los resultados.

El experimento 6 generó **147** reglas (8 más que el de Brill) y el 7 generó **114** (25 menos que el de Brill). En todos los experimentos anteriores, incluyendo el de Brill, las primeras trece reglas habían resultado prácticamente iguales. En cambio, en estos experimentos, se realizó un reacomodo

de reglas y surgieron nuevas. En el experimento 6, sólo coincidió la primera regla y en el 7 las dos primeras. En la Tabla 5.36 puede verse la comparación de las 20 primeras reglas entre estos experimentos y el cuarto.

Tabla 5.36: Comparación de 20 primeras reglas de experimentos 4, 6 y 7

Experimento 4	Experimento 6	Experimento 7
NCMS a fhassuf 1 NCFS 109	NCMS a fchar NCFS 83.7916	NCMS a fhassuf 1 NCFS 109
NCMS s fhassuf 1 NCMP 88.876	ado hassuf 3 VMP00SM 46.333	NCMS s fhassuf 1 NCMP 88.876
NCMS do fhassuf 2 VMP00SM 56.5	los goodright NCMP 39.5125	ó hassuf 1 VMIS3S0 39
NCMS r fhassuf 1 VMN0000 51.95	ó hassuf 1 VMIS3S0 39	NCMS ar fhassuf 2 VMN0000 34.75
ó hassuf 1 VMIS3S0 39	NCFS ar fhassuf 2 VMN0000 37.75	NCMS la fgoodright NCFS 27
NCMP as fhassuf 2 NCFP 36	NCFS as fhassuf 2 NCFP 29	NCMP las fgoodright NCFP 24
VMP00SM ndo fhassuf 3 VMG0000 31.333	do hassuf 2 VMP00SM 21.5	NCMS an fhassuf 2 VMII3P0 24
NCMS la fgoodright NCFS 27	er hassuf 2 VMN0000 19.2	NCMS d fchar VMP00SM 22.755
NCMS an fhassuf 2 VMII3P0 24	NCFS el fgoodright NCMS 18.1666	er hassuf 2 VMN0000 19.2
ron hassuf 3 VMIS3P0 18.1818	VMP00SM , fgoodright VMG0000 13.5	VMP00SM el fgoodright NCMS 16.880
se hassuf 2 VMSI3S0 17	NCFS ados fhassuf 4 VMP00PM 13	VMP00SM n fchar VMG0000 13.666
NCMS ion fhassuf 3 NCFS 16	NCFS que fgoodright VMII3S0 12.366	NCFS que fgoodright VMII3S0 12.866
VMN0000 or fhassuf 2 NCMS 14	es hassuf 2 NCMP 12	aron hassuf 4 VMIS3P0 12
d hassuf 1 NCFS 13	aron hassuf 4 VMIS3P0 12	años goodleft DN0CP0 11.666
NCFS que fgoodright VMII3S0 12.866	años goodleft DN0CP0 11.666	NCMS que fgoodright VMIP3S0 11.488
años goodleft DN0CP0 11.666	NCMS la fgoodright NCFS 11	ir hassuf 2 VMN0000 11
NCMS que fgoodright VMIP3S0 11.488	ir hassuf 2 VMN0000 11	NCMS n fhassuf 1 NCFS 9
dos hassuf 3 VMP00PM 10.5	NCFS an fhassuf 2 VMIP3P0 11	en hassuf 2 VMSI3P0 9
ava hassuf 3 VMII3S0 9	NCMS s fhassuf 1 AQ0MP0 9.166	1 char W 8
sen hassuf 3 VMSI3P0 9	NCFS o fhassuf 1 NCMS 9	vuestra goodleft APS000 7.925

Los resultados de estos experimentos parecen interesantes por el cambio de reglas en comparación con los anteriores, pero el etiquetado que lograron no varió significativamente (un punto porcentual peor), lo que indica que los cambios realizados no contribuyen a mejorar la precisión del método de Brill.

Para terminar la sección de experimentación, puedo comentar que el uso de afijos sí ayuda a disminuir el número de reglas generadas por el método de Brill. Lo mismo ocurrió con el uso de la nueva plantilla. Además, a ésta sí le afecta el orden de los sufijos ya que se basa en la asignación de sólo uno por cada tipo de palabra.

Resta conocer cuál es el resultado de aplicar los conjuntos de reglas a un corpus nuevo (corpus de evaluación). Es necesario recordar que todos han sido generados con un umbral para calificar reglas de 2, esto es, la generación de reglas de etiquetado se ha detenido cuando la última mejor regla obtiene una calificación menor a dicho umbral. De esta manera, se espera que el etiquetado de todos los conjuntos de reglas alcance niveles parecidos de precisión. Con esta idea en

mente, en la siguiente sección reporto los niveles de etiquetado del corpus de evaluación con todos los experimentos realizados.

5.5. Resultados y evaluación

En esta sección enfatizo dos aspectos principales. El primero es el resultado de aplicar los distintos conjuntos de reglas al corpus de evaluación. Como ya lo había dicho, se reservó una porción del corpus para este procedimiento, la cual no fue involucrada para el proceso de generación de reglas. Una vez etiquetada, se comparó con el etiquetado manual y se obtuvo una medida de precisión basada en el total de palabras bien etiquetadas entre el total de palabras del corpus. Los resultados obtenidos pueden verse en la Tabla 5.37.

Tabla 5.37: Resultados del etiquetado con los experimentos realizados

Palabras etiquetadas	Experim. 1	Experim. 2	Experim. 3	**Experim. 4**	Experim. 6	Experim. 7
Correctas	1126	1128	1127	**1130**	1119	1112
Incorrectas	262	260	261	**258**	269	276
Total	1388	1388	1388	**1388**	1388	1388
Precisión (correctas/total)	81.12%	81.27%	81.20%	**81.41%**	80.62%	80.12%

Experim. 1: Método original.
Experim. 2: Método con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF.
Experim. 3: Método con sufijos previamente descubiertos, nueva plantilla de regla, sin plantillas ADDSUF y DELETESUF, y sufijos ordenados por afijalidad.
Experim. 4: Método con sufijos previamente descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF.
Experim. 6: Método con sufijos previamente descubiertos, sin nueva plantilla de regla, sin plantillas ADDSUF y DELETESUF y asociando un solo sufijo a cada tipo de palabra, sufijos ordenados por prob. 2.
Experim. 7: Método con sufijos previamente descubiertos, sin nueva plantilla de regla, sin plantillas ADDSUF y DELETESUF y asociando un solo sufijo a cada tipo de palabra, sufijos ordenados por afijalidad.

De acuerdo con la tabla anterior, se puede ver que el método con sufijos previamente descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF (experimento 4), logró escasamente superar a los demás. En realidad, con una diferencia tan mínima no se puede decir que no tengan la misma precisión en el etiquetado. De hecho, si se compara con el método original de Brill, la inclusión de afijos no mejora significativamente la precisión de los resultados; sin embargo, el nuevo método logra el mismo nivel de etiquetado utilizando **menos reglas y más lingüísticas**.

La Tabla 5.38 es una concentración de los resultados y variables involucradas en los distintos experimentos. Ésta brinda un panorama general de las modificaciones hechas al método original y sus repercusiones en los resultados.

Tabla 5.38: Comparación de experimentos (variables y resultados)

Variable	Experim. 1	Experim. 2	Experim. 3	Experim. 4	Experim. 6	Experim. 7
Sufijos descubiertos	No	Sí	Sí	Sí	Sí	Sí
Nueva plantilla	No	Sí	Sí	No	No	No
Plantillas ADDSUF y DELETESUF	Sí	No	No	No	No	No
Orden de sufijos	---	Prob. 2	Afijalidad	Prob. 2	Prob. 2	Afijalidad
Tipo de palabra asociado con todos sus sufijos posibles	---	Sí	Sí	Sí	No	No
Número de reglas	139	126	125	128	147	**114**
Palabras correctas	1126	1128	1127	1130	1119	1112
Total de palabras	1388	1388	1388	1388	1388	1388
Precisión (correctas/total)	81.12%	81.27%	81.20%	**81.41%**	80.62%	80.12%

Experim. 1: Método original.
Experim. 2: Método con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF.
Experim. 3: Método con sufijos previamente descubiertos, nueva plantilla de regla, sin plantillas ADDSUF y DELETESUF, y sufijos ordenados por afijalidad.
Experim. 4: Método con sufijos previamente descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF.
Experim. 6: Método con sufijos previamente descubiertos, sin nueva plantilla de regla, sin plantillas ADDSUF y DELETESUF y asociando un solo sufijo a cada tipo de palabra, sufijos ordenados por prob. 2.
Experim. 7: Método con sufijos previamente descubiertos, sin nueva plantilla de regla, sin plantillas ADDSUF y DELETESUF y asociando un solo sufijo a cada tipo de palabra, sufijos ordenados por afijalidad.

De la tabla anterior se infiere que la inclusión de sufijos beneficia al método de Brill, reduciendo sus reglas de etiquetado. Además, ordenar los sufijos por afijalidad tiende a reducir las reglas, pero no ayuda a mejorar la precisión (aunque en realidad no la empeora significativamente). También, generar reglas con todos los sufijos posibles asociados a cada tipo de palabra del corpus produce mejores resultados que asociar uno solo; creo que esto se debe a que va mejor con la naturaleza del método de Brill. Finalmente, la nueva plantilla, que mira el sufijo de la palabra y la etiqueta de la palabra anterior, también ayuda a reducir el número de reglas, pero su implementación resulta difícil en un método que divide las reglas léxico-morfológicas de las contextuales; una regla de este tipo combina ambos niveles de etiquetado.

6. Conclusiones

A continuación, expongo una síntesis de los capítulos presentados y hago un resumen de la experimentación efectuada. Después, muestro las desventajas del método propuesto, sus problemas y las cuestiones que quedan pendientes por resolver. Luego, expongo los logros obtenidos y las ventajas de la propuesta realizada. Además, reformulo las hipótesis con el fin de dejar una propuesta para futuras investigaciones. Para terminar, brindo unas conclusiones generales.

El primer capítulo fue dedicado a la morfología. En él expuse la complejidad de los fenómenos asociados a este nivel del lenguaje. Comencé presentando algunos aspectos relacionados al concepto de morfema y palabra. Después, abordé el concepto de morfología y la manera como ésta se expresa en distintas lenguas. Finalmente, revisé los cuatro principales fenómenos de formación de palabras: flexión, derivación, composición e incorporación.

En el segundo capítulo, se estudió el concepto de categoría gramatical. Primeramente, abordé su definición desde distintas perspectivas lingüísticas y puntualicé algunos problemas terminológicos asociados. En seguida, discutí los principales criterios para identificar categorías. Para terminar el capítulo, presenté algunas cuestiones sobre la posibilidad de que algunas de ellas sean universales.

El siguiente capítulo me permitió profundizar en los aspectos computacionales de la identificación de categorías gramaticales. Presenté la definición de este proceso y las características de los conjuntos de etiquetas utilizados para marcar las categorías. Después, hice un recuento de la historia de los programas de cómputo para etiquetado y caracterícé los dos principales métodos utilizados: estadísticos y basados en reglas. Como parte de los segundos, resalté el analizador gramatical elaborado en El Colegio de México en los años setenta y expliqué a profundidad el método propuesto por Eric Brill, sobre el cuál basé mi experimentación.

En el cuarto capítulo, expuse algunas propuestas de segmentación de palabras basadas en métodos supervisados y no supervisados. Después, resumí la propuesta de Medina para determinar unidades morfológicas a partir de un índice de afijalidad. Éste se obtiene de medir las características de un afijo: capacidad combinatoria, poco contenido de información en contextos con mucho contenido de información y aportación a la economía del sistema. Este método de segmentación de afijos fue la clave para las modificaciones realizadas al método original de Brill.

En el capítulo cinco, reporté los experimentos llevados a cabo para generar las reglas de etiquetado en un corpus de siglo XVI. Además, este capítulo incluyó la descripción del corpus de estudio, obtenido a partir del Corpus Histórico del Español en México y de las etiquetas utilizadas.

Se indicaron las modificaciones hechas al método original de Brill: adición y eliminación de algunas plantillas y la inclusión del catálogo de afijos descubiertos. También, consigné el resultado y evaluación del proceso de descubrimiento de afijos y del etiquetado del corpus de evaluación.

Ahora, presento un resumen de los experimentos realizados. El primero estuvo dedicado a la generación de reglas de identificación de categorías con el método original de Brill. La única modificación que hice fue la eliminación de plantillas para generar reglas con partes iniciales de palabras, lo demás quedó intacto. El resultado fue de 139 reglas. Una vez aplicadas, se obtuvo una precisión en el etiquetado de 81.12% en el corpus de evaluación.

El segundo experimento fue realizado con el uso del catálogo de afijos descubiertos con el método de Medina. El catálogo estuvo ordenado por prob. 2 (probabilidad de que un sufijo lo sea en una ocurrencia de palabra del corpus). La inclusión de los afijos al método de Brill permitió mejorar dos aspectos del mismo. El primero fue el ahorro en la generación de reglas que compiten: en lugar de generar reglas con cada terminación de palabra (uno, dos, tres o cuatro caracteres) de todas las palabras del corpus, se generaron sólo las reglas asociadas a los sufijos posibles de cada palabra. El segundo fue que no se dependió de un número arbitrario de caracteres al final de la palabra, ya que el método descubrió sufijos sin importar su longitud.

Además, eliminé dos plantillas del método original. Éstas sólo generaban reglas cuando al agregar (ADDSUF) o eliminar (DELETESUF) el segmento final, la palabra aparecía atestiguada en el corpus. Estas plantillas le permitían al método original obtener partes finales más útiles, pero como los sufijos se descubrieron previamente, ya no fueron necesarias. La última modificación fue la adición de una plantilla de carácter morfosintáctico. Ésta (FTAGANTYSUFMP) generó reglas cuando la palabra tenía un sufijo y etiqueta determinados, y la palabra anterior una etiqueta específica.

El resultado de este experimento fue un conjunto de 126 reglas, trece menos que el método original, con una precisión de 81.27%. Debido a la escasa diferencia con respecto al primer experimento, puedo decir que la precisión es similar al método de Brill. Después, con el fin de verificar si el ordenamiento del catálogo de afijos tenía alguna pertinencia en el conjunto de reglas, se realizó un experimento similar, pero con los sufijos ordenados por afijalidad. El resultado fue un conjunto de 125 reglas, una menos que con el ordenamiento por prob. 2, y la precisión obtenida se mantuvo prácticamente inmóvil (81.20%). Hasta este momento, el uso de afijos y la nueva plantilla habían demostrado que brindaban mejoras al método de Brill. Sin embargo, surgió la interrogante de cuál sería el efecto sin la nueva regla, es decir, únicamente con el uso de afijos.

De esta manera, los siguientes dos experimentos se realizaron sin esta plantilla, uno con sufijos ordenados por prob. 2 y el otro por afijalidad. Los resultados en ambos casos fueron 128

reglas producidas y una precisión de 81.41%. Ésta medida, si bien la mayor, no fue significativamente mejor a las obtenidas en los experimentos anteriores, pero el número de reglas siguió siendo menor que en el experimento de Brill (once menos). Así, el cambio de orden de afijos no tuvo ningún efecto en el resultado.

Después de los experimentos anteriores, había comprobado la pertinencia de usar afijos en lugar de meras partes finales de palabras. Sin embargo, realicé dos experimentos más para confirmar los descubrimientos. En estos, se generaron reglas tomando sólo un sufijo por cada palabra y no todos los posibles. Esta situación impactó en el número de reglas que compitieron para ser seleccionadas, contraviniendo la naturaleza del método de Brill: producir una gran cantidad de reglas.

El resultado con ordenamiento por prob. 2 mostró un incremento en reglas, alcanzando 147, ocho más que el método de Brill. El ordenamiento por afijalidad, por otra parte, redujo más el número de reglas, dejando el conjunto en 114, veinticinco menos que el método original. Ambos conjuntos de reglas tuvieron una disminución en la precisión de un punto porcentual con respecto al método de Brill. Éste había obtenido 81.12% y estos experimentos obtuvieron 80.62% (prob. 2) y 80.12% (afijalidad).

El hecho de que se obtuviera aproximadamente la misma precisión en casi todos los experimentos, no me permite aceptar la primera hipótesis de esta tesis, ya que la identificación automática de categorías gramaticales que toma en cuenta unidades morfológicas no obtuvo mejores resultados en términos de precisión, que la que usa caracteres finales. Sin embargo, la reducción en número de reglas es a mi juicio un indicador robusto de que el método de identificación de categorías de Brill con los sufijos descubiertos exhibe **mayor economía en reglas** que el método de Brill solo; esta situación no me permite rechazar la segunda hipótesis.

Una vez revisadas las hipótesis con estos resultados, bien vale la pena reformularlas con el fin de proponer el inicio de una nueva investigación. Las hipótesis reformuladas pueden parafrasearse de la siguiente manera:

i. Las reglas de identificación de categorías gramaticales que toman en cuenta unidades morfológicas son más productivas que aquellas que utilizan caracteres iniciales o finales de palabras[1].

[1] Una regla productiva sería aquella que etiquete correctamente un conjunto de palabras del corpus de tal manera que no surgan otras reglas que corrijan su etiquetado. La productividad no se trata solamente de etiquetar muchas palabras, sino también de que su etiquetado sea correcto.

ii. Por lo anterior, el conjunto de reglas de identificación de categorías gramaticales que toman en cuenta unidades morfológicas es más económico, en el sentido de que involucra menos reglas, que aquel que se limita a utilizar caracteres iniciales o finales de palabras.

Ahora, atenderé los problemas, desventajas y asuntos pendientes del método propuesto. El etiquetado manual del corpus representó un factor crucial en el desarrollo de la tesis. Por un lado, el trabajo de etiquetar los textos fue tardado y tedioso, lo que implicó un tiempo considerable para contar con los primeros textos para experimentar. Por ello, el corpus obtenido fue escaso en comparación con los comúnmente usados para procesamiento automático.

En otros temas, la separación tajante entre reglas léxicas y contextuales del método de Brill original no permite la inserción adecuada de plantillas de reglas que podría llamar morfosintácticas. La parte de entrenamiento léxico aplica las reglas generadas a tipos de palabra, lo que cancela la posibilidad de revisar el contexto. Por eso, para incluir la plantilla que utilicé (FTAGANTYSUFMP), tuve que generar un conjunto de bigramas de etiquetas e incorporarlo en el proceso de generación de reglas. Restaría, como trabajo futuro, modificar el método y los programas de cómputo para incluir más aspectos morfosintácticos.

Una de las desventajas del método utilizado es la misma que la de todos los análisis basados en corpus: los resultados obtenidos pueden cambiar con nuevos ejemplos. Por lo anterior, es posible que en otros corpus los resultados que consiga cambien. De esta manera, queda pendiente probar el método propuesto en corpus más grandes, de diversas épocas del español, y en otras lenguas de morfología concatenativa similar. Sólo así se conocerá mejor su efectividad.

Como en toda investigación, no fue posible abarcar todo, así que quedaron algunos pendientes que pueden dar paso a nuevas indagaciones. Por ejemplo, sería conveniente probar con los sufijos ordenados por la prob. 1 (probabilidad a nivel de tipo de palabra) y ver los efectos que esto tiene. También, es posible cambiar el umbral de calificación de reglas; esto traería tal vez como consecuencia un conjunto mayor y mejor precisión en el etiquetado.

Además, quedó pendiente incluir los prefijos en la experimentación. Al respecto, serían necesarias varias acciones: verificar la pertinencia de los prefijos descubiertos, agregar las plantillas de reglas y comparar nuevamente resultados. Otra variable a modificar en el proceso de generación de reglas es el etiquetado de estado inicial. En los experimentos utilicé sólo uno y sería interesante probar con otros: asignar etiqueta de nombre común a todas las palabras o asignar etiquetas a partir de un diccionario, por dar algunos ejemplos.

Un trabajo pendiente es investigar a mayor profundidad la naturaleza de las reglas obtenidas en los experimentos. Algunas de ellas no son lo que un lingüista esperaría como *regla lingüística*, pero no se puede negar que reflejan el etiquetado del corpus de estudio y, por tanto, encierran algún tipo de regularidad lingüística o no. Lo cierto es que, cuando menos en mi caso, despiertan curiosidad.

En otros aspectos, aunque el descubrimiento de afijos no fue el objetivo central de la tesis, quedó pendiente una mejor evaluación de los mismos y una discusión más amplia sobre su representatividad como un catálogo de afijos de la época.

Un pendiente más es aplicar las reglas de carácter contextual que se generan como parte del método original de Brill y ver la mejora que aporta al etiquetado del corpus.

Como dije, el método de Brill parte de un corpus etiquetado del cual infiere reglas. Esto implica que el conocimiento del lingüista quedó plasmado en el etiquetado, por lo que el programa de cómputo no descubre las categorías. Esta situación deja un reto pendiente: descubrirlas sin etiquetado previo y sin información lingüística a priori; esto es, que el método descubra las categorías que emergen naturalmente del corpus, no las impuestas por la tradición y el pensamiento introspectivo de alguien. Entrar en este problema sería un desafío muy interesante y en ese sentido, esta tesis puede ser un camino hacia él.

Para terminar la exposición de los pendientes que deja esta tesis, quedaría uno de naturaleza totalmente computacional: el programa que realiza el etiquetado es ejecutado en línea de comando, por lo que a futuro será conveniente el desarrollo de una interfaz de usuario.

Ahora voy a presentar las ventajas del método y los logros obtenidos. El primero es que a pesar del tamaño del corpus, se lograron niveles alentadores de precisión. Así, con el resultado de esta tesis, se pueden etiquetar más documentos, corregirlos y volver a generar reglas que mejoren la precisión. Hay más ventajas al corregir un etiquetado que al hacerlo de cero completamente a mano.

Por otra parte, haber obtenido menos reglas se traduce en un proceso de etiquetado más económico, de hecho, niveles similares de precisión con un conjunto de reglas más pequeño es signo de que éstas son más productivas ya que hacen lo mismo, pero con menos.

Al respecto, las reglas obtenidas son una invitación a estudiar la lengua. Tienen la bondad de ser obtenidas a partir de datos duros, lo que las hace un conocimiento empírico valioso; aunque, como decía, dentro de ellas puede haber algo que como lingüistas no aceptemos o nos sorprenda. Además, piénsese en las posibilidades de generarlas a partir de corpus de lenguas menos estudiadas que el español, para observar las reglas más productivas generadas e inferir cosas acerca de dichas lenguas.

El uso del método de descubrimiento de afijos fue una buena decisión. Éste tiene la ventaja de que se enfrenta a una lengua con una mirada totalmente neutral, es decir, sin conocimiento lingüístico a priori. Los afijos se descubren por la cuantificación de sus características y no por un aparato descriptivo hecho por el lingüista o el computólogo.

A propósito de lo anterior, un logro asociado al método de descubrimiento de afijos fue que se encontraron la gran mayoría de sufijos flexivos del español del siglo XVI, situación valiosa ya que el corpus procesado fue pequeño. Con la inclusión de éstos, se logró eliminar la dependencia a un número fijo de caracteres y se mejoró el proceso de entrenamiento reduciendo el número de reglas a evaluar. Además, se modificó el método de Brill para incluir unidades morfológicas en lugar de segmentos de palabras. Si se dice, en la literatura lingüística para la identificación de categorías, que se debe tomar en cuenta la morfología, es relevante que ahora ya se cuente con un método para abundar en asuntos de descubrimiento morfológico. Esto brinda un método independiente de la lengua y aplicable a lenguas de morfología concatenativa.

Tomar los lineamientos de EAGLES dio otra ventaja, ya que son esfuerzos de estandarización de recursos lingüísticos y permitirán contar con un corpus intercambiable y reusable.

Sobre los objetivos planteados para la tesis, puedo decir que se profundizó en el conocimiento del método de Brill, lo que permitió modificar las plantillas para incluir los afijos y agregar una más de tipo morfosintáctico. Al respecto, fue necesario poner esfuerzo en la interpretación tanto del algoritmo como del programa de cómputo.

También, como se había planteado, se obtuvo el catálogo de afijos del español del siglo XVI de manera automática a partir de medidas de entropía, número de cuadros e índice de economía, es decir, de un índice de afijalidad. Si bien al principio sólo se había contemplado la entropía, al conocer la propuesta de Medina decidí usar el índice de afijalidad y los índices de probabilidad del catálogo.

Un objetivo que también se cumplió, fue el etiquetado de un corpus de estudio de forma manual. Luego, a partir de él, se generaron las reglas para etiquetar automáticamente otros con una precisión del 81%. La comparación del método de Brill con el método que usa los afijos, como ya mencioné, dio como resultado una reducción del conjunto de reglas generadas.

Computacionalmente, el objetivo de contar con un etiquetador para el siglo XVI, del Corpus Histórico del Español en México, quedó cumplido. Restaría su adaptación para trabajar con documentos XML, ya que actualmente sólo se etiquetan archivos de texto plano.

Para terminar este capítulo, enseguida expongo unas conclusiones finales. Quedó demostrado que la inclusión de afijos descubiertos a partir de criterios lingüísticos mejoró el método de Brill haciéndolo más económico: **la misma precisión pero con menos reglas**. Tal vez por la naturaleza del método era difícil que la precisión mejorara. De esta manera, esta tesis aporta: un conjunto de reglas de etiquetado aplicables a español del siglo XVI y un método más económico para generar reglas de etiquetado, que puede utilizarse en distintas lenguas de morfología concatenativa.

Brill resaltó que su método no requiere de información morfológica explícita, como una lista de afijos, porque automáticamente encuentra segmentos de palabra útiles para el etiquetado. Sin embargo, creo que su método está muy alejado del verdadero descubrimiento de morfología. Por un lado, hay una definición arbitraria de letras finales o iniciales, no obstante se sabe que la morfología concatenativa de las lenguas es variada. Por otro, la manera como revisa que las partes finales tienen propiedades combinatorias no llega ni a formar un *cuadro*, en el sentido de Greenberg.

Además, como mostré en el capítulo de morfología, las lenguas no se ajustan a divisiones tajantes entre niveles. De esta manera, los fenómenos de flexión, composición e incorporación tienen implicaciones en los niveles morfológico, sintáctico o en ambos, según el marco teórico. En este sentido, obtener un método computacional que intente etiquetar las categorías de una lengua deberá tomar en cuenta todos estos hechos lingüísticos.

Pero es justo decir que el método de Brill nunca estuvo pensado así y que su naturaleza es otra. Recuérdese que es una propuesta desde la perspectiva de la inteligencia artificial y el aprendizaje automático. Aunque no toma verdaderamente en cuenta la morfología, las reglas que obtiene reflejan conocimiento empírico y heurístico que no debe despreciarse. En ellas quedan reflejas regularidades del lenguaje de diversos niveles.

Como se vio, el método es una suerte de explosión de posibilidades, desde las menos afortunadas hasta las que alcanzan a tener sentido, que compiten para quedar fijas en un conjunto final. Ya que sus resultados terminan siendo válidos y útiles, cabe la posibilidad de preguntarnos ¿Será posible que cada hablante genere un conjunto de posibilidades inmenso que termine reducido en su propia construcción gramatical de su lengua a partir de restricciones, condiciones o intuiciones estadísticas?

A. Catálogo de sufijos del siglo XVI del CHEM

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
1	~a	1010	0.5634	0.9736	0.9659	0.3749	0.4331	0.8343
2	~s	1399	1.001	1	0.4611	0.4026	0.5001	0.8207
3	~o	997	0.5959	0.977	0.808	0.3273	0.579	0.7937
4	~ó	347	0.5298	0.9649	0.8552	0.8032	0.9059	0.7833
5	~as	439	0.3201	0.9649	0.9461	0.454	0.5752	0.7437
6	~os	521	0.3525	0.9715	0.8554	0.3583	0.4981	0.7265
7	~ar	280	0.2502	0.9601	0.9577	0.5907	0.7419	0.7227
8	~ado	268	0.2234	0.9681	0.9538	0.5826	0.7705	0.7151
9	~e	577	0.3638	0.9675	0.7987	0.3002	0.1974	0.71
10	~ase	101	0.1506	0.9627	0.9752	0.7266	0.8182	0.6962
11	~aron	132	0.1927	0.9671	0.9244	0.7543	0.9273	0.6947
12	~ados	121	0.1365	0.9736	0.9613	0.5817	0.6294	0.6904
13	~ando	121	0.138	0.9648	0.9503	0.6722	0.8472	0.6844
14	~an	311	0.1864	0.9532	0.906	0.4677	0.4456	0.6819
15	~en	272	0.2629	0.9581	0.8186	0.5282	0.207	0.6798
16	~ava	96	0.1306	0.9656	0.9294	0.7111	0.795	0.6752
17	~asen	66	0.1083	0.9627	0.9508	0.7253	0.8052	0.674
18	~er	134	0.122	0.9036	0.9834	0.6734	0.734	0.6696
19	~ará	37	0.07745	0.943	0.9787	0.7115	0.5321	0.6664
20	~ada	91	0.1035	0.9491	0.9459	0.5617	0.4784	0.6662
21	~aba	98	0.1421	0.9596	0.8868	0.7206	0.8513	0.6628
22	~ido	128	0.1388	0.9444	0.8835	0.6337	0.6995	0.6556
23	~n	731	0.4292	0.9878	0.5419	0.364	0.2257	0.653
24	~avan	66	0.09089	0.9543	0.8997	0.8049	0.8708	0.6483
25	~ía	170	0.2001	0.9471	0.7897	0.8252	0.8442	0.6456
26	~iendo	71	0.1124	0.9446	0.8768	0.7474	0.712	0.6446
27	~amente	34	0.05325	0.9526	0.92	0.5574	0.7227	0.6419
28	~amos	74	0.0802	0.897	0.9449	0.7475	0.8327	0.6407
29	~i	45	0.05045	0.8711	1	0.4054	0.1514	0.6405
30	~ieron	85	0.09915	0.9381	0.88	0.7456	0.8053	0.6391
31	~adas	52	0.07178	0.9408	0.9034	0.5778	0.7356	0.6387
32	~ían	69	0.1565	0.9284	0.8279	0.8734	0.8044	0.6376
33	~ara	28	0.06654	0.9249	0.9212	0.6512	0.02531	0.6375
34	~arse	34	0.04685	0.9254	0.9372	0.7391	0.8052	0.6365
35	~es	329	0.2091	0.9611	0.7283	0.3786	0.5117	0.6328
36	~aré	20	0.04872	0.8924	0.9284	0.7143	0.7167	0.6231
37	~are	19	0.05331	0.9155	0.8976	0.5938	0.422	0.6221
38	~ir	64	0.06744	0.9043	0.8931	0.6095	0.9156	0.6216
39	~iese	76	0.07738	0.9336	0.8487	0.7677	0.8111	0.6199
40	~ió	93	0.09501	0.932	0.8125	0.6596	0.8422	0.6132

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
41	~iesen	45	0.05509	0.8979	0.8839	0.7895	0.7117	0.6123
42	~iere	31	0.04486	0.8973	0.8805	0.7381	0.8065	0.6075
43	~aban	45	0.06638	0.9317	0.8206	0.7759	0.8504	0.6062
44	~se	377	0.198	0.9583	0.6544	0.706	0.251	0.6036
45	~ja	57	0.06182	0.873	0.8729	0.6951	0.7787	0.6026
46	~ida	40	0.06005	0.8757	0.8617	0.6897	0.5878	0.5991
47	~lo	102	0.08785	0.9346	0.7721	0.3806	0.268	0.5982
48	~ia	154	0.1184	0.9255	0.7479	0.4118	0.5513	0.5972
49	~ian	66	0.09176	0.9151	0.7824	0.7674	0.905	0.5964
50	~arán	17	0.03389	0.888	0.8672	0.68	0.619	0.5964
51	~idos	38	0.03777	0.8698	0.8673	0.5938	0.6	0.5916
52	~ero	36	0.01604	0.8039	0.954	0.4235	0.4775	0.5913
53	~io	123	0.05795	0.896	0.8102	0.4767	0.6301	0.5881
54	~iera	26	0.03413	0.8615	0.8643	0.6842	0.7345	0.5866
55	~le	106	0.08422	0.9385	0.7351	0.4344	0.1012	0.586
56	~jo	40	0.04181	0.8512	0.8561	0.6154	0.4702	0.583
57	~ándole	15	0.08297	0.8574	0.8063	0.8824	0.6774	0.5822
58	~aría	31	0.05568	0.9342	0.7547	0.8611	0.6957	0.5815
59	~arle	17	0.04173	0.8782	0.8225	0.8095	0.6486	0.5808
60	~ra	167	0.09409	0.913	0.7318	0.5154	0.8021	0.5796
62	~ya	41	0.04238	0.8124	0.8714	0.6721	0.5862	0.5754
61	~era	64	0.03827	0.8287	0.8593	0.5378	0.5	0.5754
63	~é	74	0.1349	0.8899	0.6996	0.7957	0.2669	0.5748
64	~arme	17	0.0247	0.8861	0.8016	0.7727	0.6765	0.5708
65	~r	548	0.2358	0.964	0.5118	0.5563	0.4569	0.5705
66	~emos	55	0.0992	0.877	0.7351	0.7971	0.7086	0.5704
67	~ante	22	0.02805	0.8545	0.8123	0.449	0.465	0.565
68	~açion	23	0.02171	0.8396	0.8325	0.3538	0.329	0.5646
69	~ays	14	0.02469	0.8426	0.825	0.6087	0.7547	0.5641
70	~ieren	19	0.02763	0.828	0.8305	0.6786	0.7838	0.562
71	~alle	13	0.02649	0.8635	0.7941	0.6842	0.4194	0.5614
72	~les	88	0.05968	0.9009	0.7184	0.3451	0.3856	0.5596
73	~eis	21	0.06164	0.7712	0.8422	0.7778	0.5	0.5583
74	~aria	19	0.02635	0.8585	0.7895	0.5588	0.625	0.5581
75	~re	97	0.05929	0.8836	0.7278	0.4491	0.4061	0.5569
76	~eys	25	0.07718	0.8372	0.7511	0.7353	0.3182	0.5551
77	~ta	82	0.05051	0.8841	0.7305	0.3779	0.7643	0.555
78	~yo	31	0.02503	0.7662	0.8692	0.7209	0.1974	0.5535
79	~la	58	0.04442	0.8916	0.7224	0.3353	0.114	0.5528
80	~arlo	12	0.02433	0.832	0.7993	0.6667	0.7143	0.5519
81	~los	69	0.05769	0.9204	0.6766	0.4059	0.2347	0.5516
82	~aran	16	0.02233	0.8342	0.796	0.6957	0.5429	0.5508
83	~erse	20	0.05256	0.8046	0.7894	0.7143	0.8393	0.5489
84	~eros	28	0.01459	0.7654	0.8663	0.5714	0.6825	0.5488
85	~ación	23	0.0206	0.855	0.7698	0.4035	0.7351	0.5485
86	~imos	19	0.0308	0.7916	0.8217	0.5758	0.6774	0.548
87	~al	27	0.01076	0.7306	0.9017	0.2571	0.3792	0.5477
88	~ándose	11	0.0802	0.8313	0.7314	0.8462	0.8824	0.5476
89	~aremos	11	0.02095	0.7825	0.8352	0.7333	0.75	0.5462
90	~aciones	12	0.01419	0.7867	0.8375	0.4138	0.5147	0.5461

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
91	~ales	25	0.008809	0.7536	0.8758	0.3165	0.332	0.5461
92	~do	555	0.1871	0.9491	0.4993	0.5078	0.7509	0.5451
93	~idas	18	0.02739	0.79	0.8094	0.5	0.4304	0.5423
94	~ydo	23	0.05278	0.7968	0.7747	0.7667	0.3535	0.5414
95	~eras	8	0.01296	0.7391	0.8715	0.3636	0.2254	0.5412
96	~imiento	18	0.0165	0.7866	0.8186	0.4737	0.6162	0.5406
97	~nos	44	0.03946	0.8732	0.7005	0.4583	0.5209	0.5377
98	~ran	59	0.03644	0.8525	0.7167	0.6277	0.5241	0.5352
99	~osa	11	0.01502	0.7157	0.8749	0.4783	0.05322	0.5352
100	~osas	6	0.008963	0.7141	0.8739	0.3529	0.01474	0.5323
101	~to	116	0.05172	0.8567	0.6862	0.2843	0.6556	0.5315
102	~sa	46	0.02516	0.793	0.7747	0.5287	0.9143	0.531
103	~iéndole	17	0.02275	0.8215	0.7452	0.85	0.6735	0.5298
104	~allos	13	0.01862	0.7643	0.8056	0.5909	0.4286	0.5295
105	~jdo	22	0.04679	0.8305	0.7082	0.9565	0.9833	0.5285
106	~arian	10	0.03227	0.8305	0.7185	0.6667	0.6	0.5271
107	~ria	48	0.04079	0.8225	0.7151	0.4898	0.4337	0.5261
108	~tas	29	0.02534	0.8651	0.6836	0.29	0.5065	0.5247
109	~iesse	14	0.02387	0.8519	0.6972	0.7778	0.7419	0.5243
110	~ira	9	0.01238	0.7646	0.7956	0.5294	0.2545	0.5242
111	~erle	10	0.05096	0.8155	0.706	1	1	0.5241
112	~so	62	0.03641	0.7971	0.738	0.5536	0.5812	0.5238
113	~illa	18	0.01892	0.7793	0.7717	0.6	0.5	0.5233
114	~y	22	0.0227	0.757	0.7875	0.3729	0.01272	0.5224
115	~go	51	0.02947	0.7379	0.7992	0.51	0.321	0.5222
116	~allo	12	0.02273	0.815	0.7261	0.6316	0.6182	0.5213
117	~anos	6	0.005975	0.7579	0.7977	0.1818	0.06667	0.5205
118	~andose	13	0.0193	0.8332	0.7069	0.6842	0.6522	0.5198
119	~mas	5	0.008451	0.8141	0.7351	0.2083	0.3274	0.5192
120	~ios	24	0.01627	0.7816	0.7564	0.2609	0.5403	0.5181
121	~rado	11	0.01036	0.8017	0.7417	0.1774	0.07068	0.5179
122	~arles	15	0.02373	0.8205	0.7089	0.8824	0.7857	0.5177
123	~í	18	0.04382	0.7794	0.7294	0.72	0.3959	0.5175
124	~mos	162	0.1341	0.9557	0.4615	0.7465	0.8638	0.5171
125	~ie	13	0.01468	0.636	0.8976	0.4062	0.5796	0.5161
127	~arlos	11	0.01676	0.7947	0.7322	0.6875	0.7059	0.5146
126	~ieran	10	0.02279	0.7838	0.7373	0.5882	0.6923	0.5146
128	~erlo	11	0.04959	0.7974	0.6942	0.8462	0.913	0.5137
129	~tos	48	0.02227	0.801	0.7176	0.2365	0.397	0.5136
130	~rar	11	0.01094	0.7296	0.797	0.193	0.1756	0.5125
131	~erme	9	0.03101	0.6894	0.8164	0.6429	0.7143	0.5123
132	~ras	24	0.01878	0.7769	0.7358	0.2609	0.6296	0.5105
133	~ino	13	0.01014	0.6873	0.8327	0.3939	0.3248	0.51
134	~na	31	0.02602	0.7781	0.7255	0.3039	0.1715	0.5099
135	~tar	31	0.02453	0.7696	0.7339	0.3444	0.6085	0.5093
136	~da	153	0.09376	0.9013	0.5319	0.4951	0.6754	0.509
137	~jos	11	0.01699	0.7368	0.7709	0.3548	0.1206	0.5082
139	~arnos	9	0.02162	0.8215	0.6802	0.6923	0.6818	0.5078
138	~is	40	0.02945	0.8131	0.681	0.6452	0.3278	0.5078
140	~ación	10	0.01472	0.7635	0.7432	0.5263	0.7609	0.5071

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
141	~oso	10	0.01383	0.6622	0.8421	0.3571	0.2258	0.5061
142	~vo	26	0.0197	0.7519	0.7449	0.7879	0.881	0.5055
143	~ed	8	0.01729	0.6505	0.8473	0.5333	0.1891	0.505
144	~illas	10	0.01024	0.7238	0.7802	0.4348	0.5946	0.5047
145	~idad	39	0.01665	0.7815	0.7102	0.5493	0.6693	0.5028
146	~yan	15	0.04285	0.7303	0.7341	0.625	0.6146	0.5024
147	~j	13	0.02374	0.6157	0.8671	0.4062	0.2083	0.5022
148	~das	53	0.0603	0.8649	0.5799	0.2896	0.412	0.5017
149	~ría	59	0.05645	0.799	0.6493	0.7024	0.703	0.5016
150	~ros	28	0.03174	0.8389	0.6338	0.2745	0.6252	0.5015
151	~sar	2	0.01536	0.783	0.7025	0.07407	0.05	0.5003
152	~vido	1	0.006402	0.8054	0.6874	0.05882	0.004808	0.4997
153	~me	76	0.04544	0.8713	0.5794	0.6667	0.1586	0.4987
154	~sas	11	0.008032	0.774	0.7111	0.2245	0.591	0.4977
155	~de	32	0.01917	0.7485	0.7248	0.3951	0.06542	0.4975
156	~erlos	6	0.03222	0.772	0.6877	0.6	0.6	0.4973
157	~nas	14	0.01847	0.7776	0.6938	0.2295	0.375	0.4966
158	~no	37	0.02865	0.8152	0.6454	0.3058	0.1493	0.4964
159	~iar	2	0.005122	0.7551	0.7279	0.05882	0.02885	0.496
160	~jr	21	0.03567	0.8125	0.6384	0.84	0.9365	0.4955
161	~bo	23	0.01692	0.7636	0.7037	0.575	0.8457	0.4947
162	~tó	15	0.01545	0.7733	0.686	0.2459	0.352	0.4916
163	~res	52	0.03231	0.8983	0.5411	0.3636	0.264	0.4906
164	~antes	9	0.01067	0.755	0.7016	0.45	0.1031	0.4891
165	~mo	12	0.01291	0.8218	0.6324	0.1519	0.6477	0.489
166	~ma	16	0.01352	0.7552	0.6968	0.254	0.1186	0.4885
167	~bía	3	0.01024	0.7644	0.6906	0.1579	0.02241	0.4884
168	~ella	2	0.02049	0.6415	0.8029	0.1333	0.00638	0.4883
169	~ençia	16	0.01585	0.6367	0.8078	0.2712	0.4628	0.4868
170	~yendo	13	0.02925	0.7168	0.7122	0.5909	0.3542	0.4861
171	~l	55	0.02621	0.8471	0.5832	0.2619	0.1712	0.4855
172	~ellos	9	0.01494	0.6263	0.8133	0.4737	0.04065	0.4849
173	~ito	13	0.01487	0.6393	0.794	0.3514	0.3374	0.4827
174	~tado	6	0.01387	0.7947	0.6394	0.08824	0.06714	0.4826
175	~adores	6	0.007256	0.6932	0.7466	0.25	0.4493	0.4823
176	~andole	8	0.01601	0.7688	0.6601	0.7273	0.3913	0.4817
177	~asse	7	0.01134	0.7414	0.6802	0.5385	0.3889	0.4776
178	~san	1	0.01024	0.7551	0.6653	0.07143	0.09375	0.4769
179	~lar	2	0.004481	0.7131	0.7124	0.06452	0.01802	0.4767
180	~jan	10	0.04597	0.7491	0.6342	0.625	0.8333	0.4764
181	~ias	17	0.006854	0.6808	0.7416	0.1771	0.1486	0.4764
182	~eza	8	0.01617	0.6564	0.7558	0.4444	0.5	0.4761
183	~ten	12	0.02006	0.6647	0.7412	0.4138	0.5106	0.4753
184	~cio	17	0.01431	0.7842	0.6241	0.7727	0.7905	0.4742
185	~iamos	10	0.02446	0.7032	0.6941	0.6667	0.5429	0.4739
186	~mente	68	0.05321	0.9427	0.4254	0.5714	0.702	0.4738
187	~çio	20	0.01082	0.7019	0.7084	0.4545	0.7511	0.4737
188	~viendo	1	0.005122	0.7551	0.6585	0.1111	0.007812	0.4729
189	~las	24	0.02107	0.8125	0.5845	0.2162	0.07987	0.4727
190	~ares	6	0.0128	0.6592	0.7432	0.3158	0.5696	0.4717

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
191	~el	9	0.00939	0.501	0.9047	0.2903	0.07799	0.4717
192	~yas	7	0.007865	0.7078	0.6982	0.4118	0.2	0.4713
193	~sen	123	0.09823	0.9282	0.3865	0.6474	0.7619	0.471
194	~amjento	12	0.01846	0.7696	0.6244	0.6	0.7	0.4708
195	~aren	5	0.007426	0.7709	0.6339	0.3846	0.2632	0.4707
196	~ro	21	0.02335	0.8171	0.571	0.1312	0.2647	0.4705
197	~este	1	0.005122	0.7551	0.6496	0.08333	0.0006146	0.4699
198	~ga	21	0.01902	0.6521	0.7378	0.3818	0.303	0.4697
199	~or	41	0.01702	0.6989	0.6922	0.277	0.1194	0.4694
200	~tan	8	0.01104	0.7057	0.6912	0.186	0.1994	0.4693
201	~dos	58	0.109	0.918	0.3806	0.1871	0.2806	0.4692
202	~endo	78	0.08175	0.9152	0.4086	0.5493	0.6644	0.4685
203	~só	13	0.02039	0.6394	0.745	0.5652	0.8291	0.4683
204	~biese	1	0.01536	0.839	0.5495	0.0625	0.04082	0.468
205	~uestro	1	0.0128	0.9061	0.4832	0.25	0.5625	0.4674
206	~cho	11	0.01781	0.6979	0.6848	0.2973	0.8408	0.4668
207	~ame	1	0.005122	0.7551	0.6394	0.1	0.05882	0.4665
208	~sava	3	0.01024	0.6991	0.6899	0.25	0.375	0.4664
209	~co	17	0.0116	0.6588	0.7244	0.1977	0.6512	0.4649
210	~alla	8	0.01264	0.5917	0.7856	0.6667	0.08989	0.4633
211	~jendo	10	0.03303	0.7219	0.6339	0.7692	0.8636	0.4629
212	~car	5	0.00717	0.688	0.6934	0.122	0.1543	0.4629
213	~des	36	0.03013	0.8196	0.5385	0.3913	0.5698	0.4627
214	~gos	1	0.003841	0.6712	0.7121	0.03333	0.03111	0.4624
215	~ren	30	0.03683	0.8328	0.5148	0.3704	0.2873	0.4615
216	~ienda	4	0.0144	0.6789	0.6899	0.3333	0.06452	0.4611
217	~gar	3	0.005548	0.6113	0.7661	0.08824	0.05479	0.461
218	~ança	12	0.007256	0.6544	0.7208	0.75	0.8269	0.4608
219	~yr	9	0.03827	0.7545	0.5896	0.5294	0.1579	0.4608
220	~eres	10	0.01472	0.5891	0.7725	0.5556	0.2061	0.4588
221	~rando	3	0.01536	0.5808	0.7776	0.1154	0.1515	0.4579
222	~dores	20	0.02279	0.8392	0.5113	0.4082	0.3226	0.4578
223	~ca	21	0.0175	0.7013	0.6534	0.1981	0.1566	0.4574
224	~t	20	0.01229	0.6976	0.661	0.4255	0.6038	0.457
225	~ze	13	0.02049	0.6616	0.6882	0.4483	0.9203	0.4568
226	~rse	56	0.0736	0.885	0.4116	0.6437	0.7152	0.4567
227	~aronse	2	0.01472	0.7511	0.604	0.25	0.2	0.4566
228	~que	21	0.01341	0.5941	0.7621	0.5676	0.08994	0.4565
229	~sos	6	0.00939	0.6956	0.662	0.12	0.03274	0.4557
231	~dor	30	0.02953	0.8503	0.4852	0.4348	0.4431	0.455
230	~te	77	0.04924	0.8743	0.4415	0.2048	0.3727	0.455
232	~ello	9	0.02177	0.6243	0.7188	0.5	0.02838	0.455
233	~pas	1	0.005122	0.7551	0.604	0.125	0.06667	0.4547
235	~edad	8	0.01937	0.6027	0.7408	0.4	0.246	0.4543
234	~ora	2	0.003201	0.5873	0.7725	0.1111	0.01149	0.4543
236	~ndose	3	0.0128	0.8763	0.473	0.05556	0.05263	0.454
237	~óse	10	0.04584	0.7142	0.5996	0.9091	0.9286	0.4532
238	~remos	17	0.01589	0.7545	0.5887	0.5667	0.6053	0.453
239	~erias	3	0.007682	0.6712	0.6783	0.4286	0.3529	0.4524
240	~rme	35	0.02908	0.8217	0.5016	0.6604	0.4609	0.4508

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
241	~dios	1	0.01536	0.839	0.4974	0.1111	0.001634	0.4506
242	~be	16	0.01464	0.6783	0.658	0.5161	0.1081	0.4503
243	~arla	5	0.008195	0.6343	0.706	0.5	0.5	0.4495
244	~llo	28	0.04024	0.8154	0.4921	0.4912	0.2847	0.4493
245	~illo	8	0.005442	0.6225	0.7185	0.5333	0.5556	0.4488
246	~nado	3	0.01024	0.6059	0.7252	0.08108	0.06329	0.4471
247	~yó	7	0.02341	0.6295	0.6818	0.3889	0.04192	0.4449
248	~ano	11	0.01071	0.5793	0.7434	0.2558	0.06832	0.4444
250	~ron	170	0.1399	0.9465	0.246	0.5397	0.8443	0.4442
249	~rió	1	0.003841	0.6712	0.6577	0.1	0.08772	0.4442
251	~ario	1	0.007682	0.6712	0.6469	0.0303	0.003115	0.4419
252	~ndo	161	0.1138	0.9059	0.305	0.482	0.3384	0.4416
253	~taba	1	0.003841	0.6712	0.6475	0.03448	0.005025	0.4409
254	~tase	1	0.003841	0.6712	0.6472	0.03704	0.02703	0.4407
255	~sto	10	0.0178	0.7348	0.5689	0.2632	0.1505	0.4405
257	~ve	17	0.01529	0.6345	0.67	0.6296	0.5806	0.4399
256	~ré	6	0.02945	0.8424	0.448	0.1429	0.09756	0.4399
258	~tor	2	0.002561	0.5034	0.8138	0.09524	0.02899	0.4399
259	~ez	6	0.007042	0.5648	0.7443	0.3333	0.4385	0.4387
260	~dad	31	0.01995	0.7209	0.5747	0.2263	0.6277	0.4385
261	~nbre	10	0.00717	0.8102	0.4973	0.4762	0.31	0.4382
262	~iendoles	3	0.0239	0.685	0.604	0.375	0.2727	0.4376
263	~eria	3	0.01579	0.4721	0.825	0.1304	0.08475	0.4376
264	~elo	9	0.01053	0.6033	0.6982	0.1429	0.259	0.4373
265	~iéndose	7	0.01719	0.6358	0.6585	0.7778	0.875	0.4372
266	~iendole	9	0.0286	0.7035	0.579	0.75	0.8	0.437
267	~gan	3	0.004695	0.6152	0.6905	0.1034	0.05042	0.4368
269	~çion	58	0.0251	0.7741	0.5099	0.5179	0.5523	0.4364
268	~oles	10	0.02151	0.6751	0.6126	0.1724	0.3727	0.4364
270	~ones	23	0.01715	0.8381	0.4536	0.2054	0.3783	0.4363
271	~alo	5	0.005634	0.6494	0.6514	0.2941	0.08621	0.4355
273	~ojos	1	0.008963	0.8629	0.4331	0.1667	0.07692	0.435
272	~ería	1	0.006402	0.6041	0.6946	0.0625	0.005376	0.435
274	~irle	2	0.005122	0.7551	0.544	0.3333	0.2857	0.4347
275	~ese	38	0.04616	0.8475	0.4101	0.2815	0.1495	0.4346
276	~ate	1	0.003841	0.6712	0.6281	0.06667	0.03333	0.4344
278	~rá	50	0.04159	0.8048	0.4559	0.5814	0.7211	0.4341
277	~on	67	0.03784	0.9053	0.3593	0.1309	0.04634	0.4341
279	~iendose	3	0.01963	0.6994	0.5812	0.375	0.3077	0.4334
280	~itos	2	0.003201	0.5873	0.709	0.08333	0.07317	0.4332
281	~tre	1	0.01024	0.8809	0.4082	0.07143	0.01183	0.4331
282	~ere	2	0.03457	0.811	0.4537	0.03279	0.008523	0.4331
283	~rían	8	0.02081	0.6836	0.5941	0.3478	0.4808	0.4328
284	~ana	2	0.002561	0.5034	0.7907	0.07692	0.03004	0.4322
285	~sado	2	0.0128	0.5978	0.6856	0.06061	0.00905	0.432
286	~ador	7	0.005305	0.6616	0.6265	0.2258	0.1693	0.4311
287	~eso	5	0.01076	0.4782	0.8029	0.2174	0.01458	0.4306
288	~ente	18	0.02077	0.7626	0.508	0.09045	0.1799	0.4305
289	~dado	5	0.01844	0.5778	0.6944	0.1923	0.1188	0.4302
290	~llas	11	0.02479	0.7118	0.5513	0.22	0.231	0.4293

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
291	~dio	4	0.01344	0.6355	0.6358	0.1111	0.01127	0.4282
292	~arían	6	0.008536	0.683	0.5911	0.6	0.5882	0.4275
293	~cto	9	0.01252	0.752	0.518	0.4286	0.5072	0.4275
294	~llos	20	0.0219	0.8105	0.4494	0.3333	0.03851	0.4273
295	~pa	6	0.008963	0.5077	0.7586	0.4	0.6289	0.4251
296	~ne	4	0.0144	0.6152	0.6433	0.09524	0.02185	0.4243
297	~mjento	25	0.02407	0.8287	0.4199	0.6098	0.7273	0.4242
298	~amiento	7	0.01244	0.7091	0.5453	0.4118	0.7692	0.4223
299	~iente	16	0.007682	0.4952	0.7628	0.4848	0.8408	0.4219
300	~encia	9	0.006829	0.53	0.7268	0.2571	0.1231	0.4212
301	~vas	3	0.006402	0.6712	0.5832	0.2143	0.09524	0.4203
302	~ura	10	0.007939	0.7515	0.5012	0.2941	0.1637	0.4202
303	~allas	4	0.007042	0.5873	0.6644	0.5	0.6667	0.4196
304	~ydos	1	0.003841	0.6712	0.583	0.09091	0.04545	0.4194
305	~ce	1	0.003841	0.6712	0.5827	0.07692	0.03226	0.4192
306	~ón	71	0.02851	0.9063	0.3226	0.3498	0.6759	0.4191
307	~tras	1	0.003841	0.6712	0.581	0.05263	0.03106	0.4187
308	~nta	5	0.01101	0.726	0.5184	0.07463	0.08254	0.4185
309	~ita	1	0.005122	0.5034	0.7456	0.05556	0.07692	0.418
310	~idor	6	0.004695	0.618	0.6302	0.375	0.2656	0.4177
311	~xo	4	0.008643	0.6987	0.5447	0.1905	0.01254	0.4174
312	~rian	11	0.01839	0.7135	0.5167	0.3793	0.3393	0.4162
313	~eren	10	0.01946	0.7915	0.4371	0.2174	0.1145	0.416
314	~ala	1	0.003841	0.3356	0.9061	0.0625	0.01075	0.4152
315	~id	3	0.006402	0.5593	0.6783	0.4286	0.4	0.4147
316	~esta	3	0.008109	0.5849	0.6495	0.12	0.003394	0.4142
317	~bido	3	0.02603	0.6883	0.5267	0.09091	0.02427	0.4137
318	~acion	4	0.005122	0.5537	0.682	0.2667	0.3333	0.4136
319	~lla	16	0.03017	0.7401	0.4689	0.2581	0.05625	0.4131
320	~po	10	0.009091	0.5051	0.7226	0.4167	0.2951	0.4122
321	~lana	1	0.008963	0.8629	0.3624	0.25	0.5	0.4114
322	~cha	2	0.01472	0.6208	0.5979	0.07692	0.001761	0.4111
323	~bieron	1	0.02433	0.6888	0.518	0.05263	0.01724	0.4104
324	~enta	9	0.004553	0.4437	0.7824	0.2432	0.5725	0.4102
325	~ys	46	0.04996	0.8749	0.3036	0.7667	0.5023	0.4095
326	~çe	2	0.01088	0.5537	0.6616	0.05556	0.02113	0.4087
327	~esen	19	0.04353	0.8257	0.3556	0.2468	0.125	0.4083
329	~baxo	1	0.005122	0.7551	0.4644	0.2	0.6667	0.4082
328	~u	6	0.01259	0.3496	0.8625	0.25	0.003286	0.4082
330	~sta	8	0.01232	0.64	0.5708	0.1633	0.01183	0.4077
331	~rla	3	0.01963	0.7239	0.4785	0.15	0.08889	0.4074
332	~mjentos	3	0.01024	0.7809	0.4305	0.2	0.1034	0.4072
334	~selo	12	0.01227	0.8158	0.3924	0.4286	0.4688	0.4068
333	~ben	3	0.0111	0.6482	0.5612	0.1765	0.08403	0.4068
335	~erá	7	0.01482	0.5139	0.6906	0.4375	0.25	0.4065
338	~ete	4	0.005762	0.4405	0.7727	0.1481	0.396	0.4063
337	~çía	1	0.003841	0.6712	0.544	0.08333	0.01818	0.4063
336	~reo	1	0.003841	0.6712	0.544	0.1667	0.01667	0.4063
339	~un	4	0.008323	0.4986	0.7102	0.2	0.02632	0.4057
340	~ejas	2	0.003841	0.6712	0.5402	0.2857	0.2083	0.4051

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
341	~tes	3	0.007682	0.7195	0.4852	0.02804	0.004178	0.4041
342	~eles	2	0.003201	0.4195	0.789	0.09091	0.03175	0.4039
343	~bas	3	0.002988	0.5593	0.6485	0.1429	0.03478	0.4036
344	~ça	25	0.01188	0.6629	0.5359	0.5556	0.5236	0.4035
345	~rle	12	0.03767	0.8315	0.3381	0.3243	0.2459	0.4024
346	~vieron	9	0.009816	0.5468	0.6498	0.4737	0.4235	0.4021
347	~imientos	4	0.007682	0.5888	0.6093	0.5714	0.2857	0.4019
348	~ores	5	0.006146	0.7025	0.4963	0.06757	0.25	0.4017
350	~ll	4	0.01536	0.5898	0.5996	0.3636	0.9216	0.4016
349	~reys	2	0.01536	0.6376	0.5519	0.1667	0.3077	0.4016
351	~vio	1	0.01152	0.5593	0.6339	0.07692	0.0101	0.4016
352	~rles	4	0.02561	0.8645	0.3077	0.1538	0.08163	0.3993
353	~bos	2	0.009603	0.5377	0.6496	0.1667	0.05797	0.399
354	~nó	1	0.002561	0.5034	0.6906	0.0625	0.01316	0.3989
355	~illos	5	0.008195	0.4256	0.7624	0.3125	0.3667	0.3987
356	~escrivo	1	0.006402	0.8054	0.3829	0.3333	0.07692	0.3983
357	~escrevir	1	0.006402	0.8054	0.3829	0.3333	0.07143	0.3983
358	~d	49	0.0428	0.8454	0.305	0.2237	0.5042	0.3977
359	~iga	3	0.01024	0.449	0.7314	0.2308	0.05825	0.3969
361	~ais	6	0.005548	0.4925	0.6899	0.5	0.5882	0.396
360	~z	10	0.00781	0.5905	0.5898	0.2174	0.4212	0.396
362	~m	8	0.006722	0.5044	0.6755	0.2963	0.04938	0.3956
364	~ldo	5	0.007682	0.5683	0.6093	0.7143	0.5	0.3951
363	~us	6	0.01131	0.3834	0.7908	0.4	0.009091	0.3951
365	~lle	21	0.0375	0.6962	0.4513	0.6176	0.6078	0.395
366	~ndole	4	0.01761	0.6967	0.4705	0.06154	0.03876	0.3949
367	~osos	3	0.004695	0.5817	0.5947	0.1667	0.1	0.3937
368	~edor	5	0.01357	0.4671	0.6925	0.3333	0.1176	0.3911
369	~ina	3	0.007256	0.4475	0.7178	0.15	0.4494	0.3908
371	~én	7	0.0128	0.6601	0.4952	0.6364	0.9269	0.3894
370	~p	2	0.006402	0.5034	0.6585	0.2222	0.08333	0.3894
372	~cas	2	0.01152	0.4614	0.6945	0.06061	0.01852	0.3891
373	~able	4	0.003521	0.5453	0.6177	0.2857	0.5	0.3889
374	~mismo	7	0.006036	0.5554	0.6048	0.7778	0.6473	0.3888
375	~sse	31	0.03432	0.7921	0.339	0.7045	0.7432	0.3885
376	~miento	36	0.03439	0.7664	0.3644	0.5806	0.795	0.3884
377	~dieron	1	0.006402	0.6041	0.5537	0.05882	0.008	0.3881
378	~eran	5	0.009731	0.5761	0.5776	0.1667	0.03352	0.3878
379	~vida	1	0.0128	0.6041	0.544	0.1667	0.01064	0.387
381	~venya	1	0.0128	0.9061	0.2416	0.5	0.5	0.3868
380	~pare	1	0.0128	0.9061	0.2416	0.5	0.3333	0.3868
384	~raçon	1	0.003841	0.6712	0.4832	0.25	0.6667	0.3861
383	~poniéndole	1	0.003841	0.6712	0.4832	0.25	0.1111	0.3861
382	~ramento	1	0.003841	0.6712	0.4832	0.25	0.003236	0.3861
386	~je	7	0.00567	0.5174	0.6342	0.3182	0.3488	0.3858
385	~esse	3	0.0128	0.7458	0.3988	0.1111	0.07692	0.3858
387	~ramos	2	0.02305	0.5034	0.6302	0.125	0.09091	0.3856
388	~eron	28	0.05254	0.8313	0.2704	0.2014	0.3227	0.3848
389	~erla	1	0.01793	0.5753	0.561	0.2	0.1667	0.3847
391	~diese	2	0.008963	0.5339	0.6109	0.1	0.02062	0.3846

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
390	~rra	2	0.007042	0.6376	0.5093	0.08696	0.00335	0.3846
392	~ydas	2	0.01985	0.3756	0.7543	0.2	0.1667	0.3833
393	~rlos	14	0.02021	0.7821	0.3471	0.4828	0.4286	0.3832
394	~eçesidad	1	0.01152	0.8949	0.2416	0.5	0.8333	0.3827
395	~si	4	0.01056	0.5453	0.5911	0.4	0.04706	0.3823
396	~vos	1	0.003841	0.6712	0.4713	0.08333	0.009804	0.3821
397	~ven	4	0.007682	0.5732	0.5643	0.2222	0.3023	0.3817
398	~neros	1	0.002561	0.5034	0.6382	0.1111	0.02564	0.3814
399	~tenellos	1	0.005122	0.7551	0.3829	0.3333	0.3333	0.3811
400	~imyento	4	0.005762	0.4782	0.6577	0.4	0.4167	0.3805
401	~ción	15	0.01895	0.7809	0.3415	0.5556	0.7857	0.3804
402	~imjento	6	0.008963	0.5397	0.5911	0.6	0.303	0.3799
403	~ejo	2	0.01024	0.4674	0.6585	0.2222	0.03448	0.3787
404	~uelas	4	0.004481	0.5873	0.5436	0.5	0.4286	0.3785
405	~ción	46	0.01525	0.64	0.4796	0.46	0.7513	0.3783
407	~ándola	5	0.04942	0.6201	0.4635	0.8333	0.8333	0.3777
406	~yón	12	0.008323	0.7338	0.391	0.7059	0.7273	0.3777
408	~chas	2	0.003841	0.5034	0.6255	0.09524	0.003026	0.3776
409	~ydad	4	0.006082	0.5831	0.5428	0.4	0.5	0.3773
410	~lado	3	0.005975	0.358	0.7657	0.1071	0.1887	0.3765
411	~iéndola	1	0.008963	0.5753	0.544	0.1667	0.25	0.3761
412	~tales	1	0.008963	0.4315	0.6875	0.09091	0.04	0.376
413	~erra	4	0.008003	0.4754	0.6435	0.3636	0.1552	0.3757
414	~ver	1	0.01793	0.5753	0.5331	0.06667	0.005602	0.3754
415	~ento	4	0.006722	0.5693	0.5495	0.0241	0.01118	0.3752
416	~dre	2	0.006402	0.6041	0.5142	0.2857	0.9769	0.3749
417	~jmjento	5	0.01255	0.5306	0.5812	0.625	0.6	0.3748
418	~mbre	10	0.007298	0.7312	0.3852	0.4167	0.2808	0.3746
419	~ba	47	0.0553	0.7515	0.3164	0.2717	0.1798	0.3744
420	~té	3	0.006402	0.505	0.6093	0.4286	0.8667	0.3736
421	~olo	3	0.02091	0.6596	0.4398	0.07692	0.03543	0.3734
422	~mento	11	0.006635	0.5193	0.5913	0.5	0.05319	0.3724
423	~men	3	0.008536	0.5034	0.6047	0.2143	0.1	0.3722
424	~edades	1	0.002561	0.5034	0.6093	0.1429	0.08333	0.3717
425	~yese	3	0.01195	0.4937	0.6093	0.4286	0.3636	0.3716
426	~ientes	2	0.005122	0.5034	0.6061	0.1	0.02564	0.3715
431	~león	1	0.008963	0.8629	0.2416	0.5	0.25	0.3712
428	~ljcenciado	1	0.008963	0.8629	0.2416	0.5	0.2	0.3712
429	~liçençiado	1	0.008963	0.8629	0.2416	0.5	0.1429	0.3712
427	~ocote	1	0.008963	0.8629	0.2416	0.5	0.08333	0.3712
430	~licenciado	1	0.008963	0.8629	0.2416	0.5	0.04082	0.3712
432	~llama	1	0.008963	0.8629	0.2416	0.5	0.005435	0.3712
433	~çia	30	0.01174	0.7598	0.3416	0.297	0.4591	0.371
434	~ientos	3	0.006402	0.5649	0.5393	0.09375	0.08571	0.3702
435	~án	5	0.01793	0.8106	0.2814	0.1042	0.1014	0.37
436	~echo	2	0.008323	0.4674	0.6339	0.1538	0.03733	0.3699
437	~tad	4	0.02369	0.4978	0.5817	0.1538	0.0124	0.3677
438	~cados	1	0.002561	0.5034	0.5938	0.05	0.01205	0.3666
439	~tido	2	0.008323	0.4894	0.6013	0.1111	0.02353	0.3663
440	~tra	5	0.004866	0.5034	0.5899	0.1923	0.01799	0.3661

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
441	~poner	2	0.003841	0.6292	0.4644	0.4	0.09091	0.3658
442	~bierdes	1	0.01536	0.839	0.2416	0.5	0.5	0.3653
443	~ndoles	3	0.003414	0.6152	0.4772	0.1111	0.06522	0.3653
444	~nte	40	0.0331	0.8085	0.2515	0.1544	0.4101	0.3644
445	~dia	3	0.01152	0.4576	0.623	0.08571	0.009934	0.364
446	~mientos	4	0.008003	0.7055	0.3776	0.2667	0.08333	0.3637
447	~ño	4	0.005122	0.4602	0.625	0.2353	0.03694	0.3635
448	~qué	1	0.01536	0.6712	0.4022	0.1429	0.002427	0.3629
449	~elle	4	0.009603	0.4743	0.604	0.5	0.4444	0.3627
450	~dió	1	0.02689	0.5753	0.4819	0.03571	0.005917	0.3614
451	~eo	2	0.009603	0.3636	0.7103	0.1176	0.01724	0.3612
452	~ña	8	0.01408	0.501	0.5662	0.4706	0.7597	0.3604
453	~via	1	0.005122	0.5034	0.5723	0.09091	0.007874	0.3603
454	~arselo	1	0.0128	0.6041	0.4635	0.1667	0.1667	0.3601
455	~cia	26	0.0134	0.782	0.2841	0.5	0.6609	0.3599
456	~tura	9	0.01067	0.4198	0.6485	0.45	0.3119	0.3597
457	~ás	5	0.01024	0.6768	0.39	0.3846	0.08655	0.359
458	~otro	1	0.006402	0.6041	0.4644	0.2	0.003115	0.3583
459	~lor	3	0.004695	0.5593	0.5106	0.3333	0.3889	0.3582
460	~ntas	1	0.003841	0.6712	0.399	0.03448	0.003906	0.358
461	~va	48	0.04665	0.7524	0.2732	0.2927	0.1828	0.3574
462	~çeso	2	0.005762	0.5034	0.561	0.4	0.01042	0.3567
463	~yría	1	0.005122	0.5034	0.561	0.2	0.2	0.3565
464	~alas	1	0.005122	0.5034	0.561	0.2	0.07143	0.3565
465	~vió	1	0.01152	0.4475	0.6094	0.08333	0.02083	0.3561
466	~aja	1	0.006402	0.4027	0.6585	0.1111	0.09091	0.3559
467	~rio	3	0.008536	0.5444	0.5126	0.04478	0.01222	0.3552
468	~iva	1	0.01024	0.5034	0.5511	0.1111	0.05263	0.3549
470	~sión	5	0.006146	0.3898	0.6661	0.2083	0.5126	0.354
469	~sy	4	0.009283	0.5087	0.544	0.6667	0.06731	0.354
471	~ición	1	0.002561	0.5034	0.556	0.04762	0.02564	0.354
472	~amento	4	0.005442	0.5705	0.4832	0.5	0.03135	0.3531
473	~dan	2	0.009603	0.5263	0.523	0.06897	0.01639	0.3529
481	~mina	1	0.003841	0.6712	0.3829	0.3333	0.3333	0.3527
474	~buelto	1	0.003841	0.6712	0.3829	0.3333	0.2	0.3527
475	~quito	1	0.003841	0.6712	0.3829	0.3333	0.2	0.3527
482	~chillos	1	0.003841	0.6712	0.3829	0.3333	0.1667	0.3527
477	~pongan	1	0.003841	0.6712	0.3829	0.3333	0.125	0.3527
479	~sillas	1	0.003841	0.6712	0.3829	0.3333	0.08333	0.3527
480	~tuvieron	1	0.003841	0.6712	0.3829	0.3333	0.03571	0.3527
483	~mí	1	0.003841	0.6712	0.3829	0.3333	0.01075	0.3527
476	~pués	1	0.003841	0.6712	0.3829	0.3333	0.004739	0.3527
478	~otras	1	0.003841	0.6712	0.3829	0.3333	0.002457	0.3527
484	~tio	3	0.0128	0.4414	0.6036	0.1765	0.3704	0.3526
485	~eva	1	0.002561	0.5034	0.5511	0.1111	0.01562	0.3524
486	~issimo	2	0.004481	0.5873	0.4644	0.4	0.4	0.352
487	~más	2	0.004481	0.5873	0.4644	0.4	0.075	0.352
488	~iessen	2	0.003201	0.5873	0.4644	0.4	0.3636	0.3516
490	~rnos	5	0.02484	0.7421	0.2875	0.2632	0.2069	0.3515

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
489	~jas	1	0.003841	0.3356	0.7151	0.03846	0.01064	0.3515
491	~zen	2	0.005122	0.5369	0.5118	0.1818	0.02674	0.3513
493	~enpeño	1	0.006402	0.8054	0.2416	0.5	0.5	0.3511
492	~escribo	1	0.006402	0.8054	0.2416	0.5	0.07692	0.3511
494	~enbío	1	0.006402	0.8054	0.2416	0.5	0.05556	0.3511
495	~ierdes	3	0.005122	0.5034	0.544	0.5	0.8	0.3508
496	~yda	1	0.03585	0.4315	0.5846	0.1111	0.05556	0.3506
497	~edo	1	0.003841	0.3356	0.7122	0.1111	0.02439	0.3505
498	~rán	12	0.0271	0.6317	0.3921	0.3158	0.3582	0.3503
499	~ebo	1	0.002561	0.5034	0.544	0.1667	0.03571	0.35
500	~brar	1	0.002561	0.5034	0.544	0.08333	0.02703	0.35
501	~aje	1	0.002561	0.5034	0.5436	0.125	0.09524	0.3499
502	~den	2	0.007682	0.3775	0.6633	0.07407	0.4011	0.3495
503	~ío	3	0.007256	0.4881	0.5528	0.2308	0.04412	0.3494
504	~stad	5	0.005378	0.6208	0.42	0.3846	0.034	0.3487
505	~cado	5	0.005634	0.4463	0.5939	0.1429	0.04918	0.3486
506	~sion	3	0.005122	0.3356	0.7016	0.15	0.09615	0.3474
507	~gas	1	0.003841	0.3356	0.702	0.0625	0.03571	0.3472
508	~iones	2	0.005122	0.6376	0.3976	0.02469	0.01124	0.3468
509	~chos	2	0.005122	0.5034	0.5309	0.08	0.002509	0.3465
510	~ad	14	0.01591	0.7675	0.2553	0.08092	0.21	0.3462
511	~tubo	1	0.003841	0.6712	0.3624	0.25	0.0102	0.3458
512	~ela	2	0.007682	0.3524	0.6749	0.1176	0.1154	0.345
513	~ardes	2	0.009603	0.5393	0.4832	0.5	0.4	0.3441
514	~rada	1	0.003841	0.3356	0.6899	0.04348	0.04348	0.3431
515	~idores	1	0.003841	0.3356	0.6899	0.08333	0.03448	0.3431
516	~ciones	5	0.004866	0.641	0.3829	0.4167	0.3846	0.3429
517	~nunçió	1	0.003841	0.6712	0.3526	0.1667	0.07692	0.3425
518	~bia	3	0.007256	0.3835	0.6339	0.2308	0.07143	0.3416
519	~nçia	12	0.01291	0.7024	0.3081	0.1558	0.1497	0.3411
520	~sario	3	0.002561	0.5034	0.5156	0.2308	0.2973	0.3405
521	~nero	3	0.006402	0.4139	0.5996	0.2727	0.4483	0.34
522	~brado	1	0.002561	0.5034	0.5138	0.0625	0.02439	0.3399
523	~engo	2	0.007042	0.3859	0.6246	0.3333	0.03356	0.3392
524	~eer	2	0.01152	0.4772	0.5208	0.25	0.1017	0.3365
525	~za	10	0.008067	0.5561	0.4425	0.2857	0.1197	0.3356
526	~lando	1	0.003841	0.3356	0.6648	0.08333	0.02564	0.3348
527	~anes	1	0.002561	0.5034	0.4974	0.1111	0.09524	0.3345
528	~anto	1	0.003841	0.3356	0.6601	0.09091	0.003636	0.3332
529	~yendose	1	0.01024	0.5034	0.4832	0.25	0.4	0.3323
530	~ponga	2	0.003841	0.6292	0.3624	0.5	0.05556	0.3318
532	~zo	3	0.008963	0.4602	0.5252	0.2	0.008639	0.3314
531	~cargo	1	0.007682	0.5034	0.4832	0.25	0.004854	0.3314
534	~tienda	1	0.006402	0.6041	0.3829	0.3333	0.125	0.3311
533	~çiençia	1	0.006402	0.6041	0.3829	0.3333	0.04255	0.3311
535	~can	1	0.006402	0.4027	0.5839	0.05882	0.01923	0.331
536	~entos	1	0.01536	0.4195	0.5579	0.0137	0.007547	0.3309
538	~bamos	1	0.005122	0.5034	0.4832	0.25	0.2857	0.3306
537	~ierades	1	0.005122	0.5034	0.4832	0.25	0.25	0.3306
539	~ísimo	2	0.007682	0.5034	0.4786	0.1818	0.1111	0.3299

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
540	~eado	2	0.008323	0.516	0.4644	0.4	0.1538	0.3296
541	~iembre	4	0.004161	0.4614	0.5208	0.5	0.4068	0.3288
542	~igo	2	0.005122	0.3692	0.6112	0.09524	0.01238	0.3285
543	~jó	1	0.006402	0.4027	0.5729	0.1	0.05556	0.3273
544	~bio	2	0.008963	0.481	0.4917	0.1538	0.275	0.3272
545	~b	3	0.009816	0.3661	0.6048	0.3333	0.1364	0.3269
546	~ge	5	0.007682	0.4128	0.5595	0.3846	0.25	0.3267
547	~ime	2	0.003841	0.3356	0.6394	0.2	0.1667	0.3263
548	~ión	12	0.01462	0.5848	0.3793	0.075	0.1051	0.3263
549	~iado	2	0.004481	0.2936	0.6792	0.04651	0.008163	0.3258
550	~azer	1	0.002561	0.5034	0.4713	0.08333	0.001972	0.3258
551	~lles	7	0.009877	0.6335	0.3333	0.4118	0.2903	0.3255
552	~vedad	1	0.005122	0.5034	0.4644	0.2	0.05263	0.3243
553	~cen	1	0.005122	0.5034	0.4635	0.1667	0.06667	0.324
554	~varon	1	0.002561	0.5034	0.4644	0.2	0.02564	0.3234
555	~tener	1	0.002561	0.5034	0.4635	0.1667	0.007042	0.3231
556	~á	26	0.06003	0.7264	0.1813	0.268	0.756	0.3226
557	~cos	1	0.006402	0.4027	0.5565	0.03333	0.03614	0.3219
558	~quen	2	0.006402	0.3775	0.5812	0.25	0.1875	0.3217
559	~ndolos	1	0.007682	0.5034	0.4537	0.05556	0.05263	0.3216
560	~otra	1	0.003841	0.3356	0.6246	0.1667	0.002326	0.3213
561	~ronse	1	0.003841	0.6712	0.2872	0.08333	0.07143	0.3207
562	~bre	4	0.004481	0.5034	0.4522	0.07273	0.02809	0.32
563	~ste	2	0.01088	0.4424	0.506	0.08333	0.001132	0.3198
565	~myento	8	0.009443	0.6268	0.3212	0.5	0.5909	0.3192
564	~ves	3	0.005122	0.4419	0.5106	0.3333	0.1351	0.3192
566	~ntes	5	0.0169	0.697	0.2425	0.07812	0.02041	0.3188
567	~mal	1	0.01152	0.5593	0.3829	0.3333	0.006579	0.3179
568	~jeron	4	0.0144	0.4751	0.4635	0.6667	0.4444	0.3177
569	~ij	2	0.003841	0.5034	0.4451	0.2857	0.2857	0.3175
570	~yon	4	0.003201	0.5873	0.3589	0.4	0.3846	0.3165
571	~tal	2	0.005122	0.3356	0.6048	0.2222	0.01156	0.3152
572	~cion	13	0.009652	0.5757	0.3597	0.5909	0.6829	0.315
573	~niendo	1	0.002561	0.5034	0.4376	0.125	0.02083	0.3145
574	~ntar	3	0.003841	0.4475	0.4884	0.1	0.03488	0.3132
576	~biado	1	0.002561	0.5034	0.4331	0.1667	0.05	0.313
575	~monio	1	0.002561	0.5034	0.4331	0.1667	0.0137	0.313
577	~ose	2	0.02497	0.619	0.2944	0.02778	0.01093	0.3128
578	~zas	3	0.003841	0.4698	0.4642	0.2143	0.15	0.3126
579	~nde	1	0.01024	0.5034	0.4179	0.03333	0.5172	0.3105
580	~çi	2	0.005122	0.5034	0.4228	0.25	0.1429	0.3104
581	~nto	4	0.02241	0.7047	0.2041	0.02094	0.007395	0.3104
582	~erán	4	0.01344	0.3762	0.5402	0.5714	0.3333	0.31
583	~jdad	1	0.002561	0.5034	0.4235	0.1111	0.25	0.3098
584	~nes	4	0.008963	0.5593	0.3603	0.02632	0.02367	0.3095
586	~jentos	1	0.007682	0.839	0.08149	0.0625	0.09375	0.3094
585	~dolos	1	0.003841	0.6712	0.2531	0.04545	0.00551	0.3094
587	~yos	1	0.007682	0.3356	0.5846	0.1111	0.05128	0.3093
588	~seria	2	0.01857	0.5227	0.3829	0.6667	0.6667	0.3081
589	~ua	2	0.008323	0.2894	0.6262	0.1111	0.02251	0.308

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
590	~elles	1	0.008963	0.2876	0.6246	0.1667	0.1429	0.3071
591	~yeron	2	0.003841	0.3356	0.5812	0.25	0.07143	0.3069
593	~guiente	3	0.0111	0.4441	0.4644	0.6	0.9783	0.3065
592	~zes	1	0.006402	0.2014	0.7119	0.04762	0.002439	0.3065
596	~verençia	1	0.003841	0.6712	0.2416	0.5	0.875	0.3055
607	~marca	1	0.003841	0.6712	0.2416	0.5	0.8462	0.3055
606	~mun	1	0.003841	0.6712	0.2416	0.5	0.6	0.3055
597	~trate	1	0.003841	0.6712	0.2416	0.5	0.5	0.3055
598	~tocar	1	0.003841	0.6712	0.2416	0.5	0.5	0.3055
599	~rrededor	1	0.003841	0.6712	0.2416	0.5	0.5	0.3055
602	~çepa	1	0.003841	0.6712	0.2416	0.5	0.5	0.3055
603	~perlado	1	0.003841	0.6712	0.2416	0.5	0.5	0.3055
604	~pendençia	1	0.003841	0.6712	0.2416	0.5	0.5	0.3055
605	~movidos	1	0.003841	0.6712	0.2416	0.5	0.5	0.3055
609	~conozco	1	0.003841	0.6712	0.2416	0.5	0.5	0.3055
611	~conoçer	1	0.003841	0.6712	0.2416	0.5	0.5	0.3055
595	~com	1	0.003841	0.6712	0.2416	0.5	0.3333	0.3055
600	~curso	1	0.003841	0.6712	0.2416	0.5	0.3333	0.3055
608	~cresçio	1	0.003841	0.6712	0.2416	0.5	0.3333	0.3055
601	~pende	1	0.003841	0.6712	0.2416	0.5	0.2727	0.3055
610	~fundir	1	0.003841	0.6712	0.2416	0.5	0.25	0.3055
594	~qujeren	1	0.003841	0.6712	0.2416	0.5	0.2	0.3055
613	~rón	1	0.003841	0.6712	0.2395	0.1111	0.25	0.3048
612	~man	2	0.01088	0.3006	0.603	0.1429	0.05263	0.3048
614	~ngo	1	0.002561	0.5034	0.4073	0.09091	0.005208	0.3044
615	~uras	4	0.004802	0.3985	0.5011	0.2857	0.07463	0.3015
616	~ende	2	0.01024	0.1464	0.7456	0.1111	0.1071	0.3008
617	~ritos	1	0.008963	0.2876	0.604	0.125	0.07407	0.3002
618	~pos	1	0.003841	0.3356	0.561	0.1	0.02326	0.3001
619	~ntos	1	0.003841	0.6712	0.2239	0.01163	0.002907	0.2996
620	~melo	2	0.004481	0.3524	0.5402	0.2857	0.3	0.299
621	~quyere	1	0.003841	0.6712	0.2219	0.3333	0.1111	0.299
622	~creto	1	0.003841	0.6712	0.2219	0.3333	0.05	0.299
623	~pública	1	0.003841	0.6712	0.2219	0.3333	0.04651	0.299
624	~ándoles	4	0.01729	0.4145	0.4635	0.6667	0.4167	0.2984
625	~xico	1	0.007682	0.5034	0.3829	0.3333	0.25	0.298
627	~ljdad	1	0.005122	0.5034	0.3829	0.3333	0.5	0.2972
626	~cya	1	0.005122	0.7551	0.1313	0.125	0.1667	0.2972
628	~ion	5	0.008195	0.5614	0.3207	0.03145	0.03216	0.2967
630	~and	1	0.002561	0.5034	0.3829	0.3333	0.9672	0.2963
632	~calles	2	0.002561	0.5034	0.3829	0.6667	0.1818	0.2963
633	~roba	1	0.002561	0.5034	0.3829	0.3333	0.1429	0.2963
629	~mones	1	0.002561	0.5034	0.3829	0.3333	0.0625	0.2963
631	~servjcio	1	0.002561	0.5034	0.3829	0.3333	0.01351	0.2963
635	~yase	2	0.007042	0.3955	0.4832	0.5	0.5	0.2953
634	~dra	1	0.01152	0.3356	0.5387	0.08333	0.02941	0.2953
636	~g	1	0.003841	0.3356	0.544	0.1667	0.125	0.2945
637	~içion	1	0.008963	0.2876	0.5853	0.06667	0.06667	0.294
638	~sí	1	0.0128	0.4027	0.4644	0.2	0.004301	0.2933
639	~vamos	1	0.003841	0.3356	0.5402	0.1429	0.1429	0.2932

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
641	~éndole	4	0.01921	0.694	0.1639	0.1739	0.3269	0.2924
640	~ellas	1	0.01793	0.2157	0.6435	0.09091	0.004808	0.2924
642	~mana	1	0.005122	0.5034	0.3677	0.2	0.0625	0.2921
643	~hera	1	0.005122	0.5034	0.3677	0.2	0.003509	0.2921
644	~ñas	1	0.006402	0.4027	0.4649	0.07143	0.03846	0.2913
645	~çientos	1	0.002561	0.5034	0.3677	0.2	0.3333	0.2912
646	~dandoles	1	0.007682	0.5034	0.3624	0.25	0.2	0.2912
647	~mjente	2	0.01344	0.4759	0.3829	0.6667	0.6667	0.2908
648	~die	1	0.005122	0.5034	0.3624	0.25	0.8571	0.2903
649	~nço	1	0.005122	0.5034	0.3624	0.25	0.09091	0.2903
650	~sclavos	1	0.002561	0.5034	0.3624	0.25	0.01587	0.2895
652	~ebi	1	0.006402	0.4027	0.4569	0.1111	0.2143	0.2887
651	~çera	1	0.005122	0.2517	0.6094	0.08333	0.03226	0.2887
653	~jgo	4	0.004481	0.4363	0.4228	0.5	0.5714	0.2879
654	~guna	2	0.005122	0.3356	0.5221	0.2	0.04437	0.2876
655	~estas	2	0.003841	0.3775	0.4808	0.1333	0.00905	0.2874
656	~manos	2	0.007682	0.3883	0.4635	0.3333	0.03614	0.2865
657	~ssen	4	0.008323	0.6622	0.1883	0.3077	0.2727	0.2863
658	~cta	1	0.002561	0.5034	0.3526	0.1667	0.04545	0.2862
659	~mesmo	1	0.002561	0.5034	0.3526	0.1667	0.02899	0.2862
662	~jente	4	0.01184	0.4195	0.4237	0.2857	0.1739	0.285
661	~mando	1	0.007682	0.3356	0.5118	0.09091	0.04348	0.285
660	~sino	2	0.01216	0.4805	0.3624	0.5	0.009901	0.285
665	~tuvjera	1	0.006402	0.6041	0.2416	0.5	0.5	0.284
666	~tobiese	1	0.006402	0.6041	0.2416	0.5	0.3333	0.284
664	~dexa	1	0.006402	0.6041	0.2416	0.5	0.2857	0.284
663	~dicho	1	0.01024	0.3775	0.4644	0.2	0.0001893	0.284
667	~dido	1	0.007682	0.3356	0.5072	0.02857	0.00813	0.2835
668	~vano	1	0.003841	0.6712	0.1744	0.2	0.1429	0.2832
669	~ysmo	1	0.003841	0.6712	0.1744	0.2	0.04167	0.2832
670	~iene	4	0.004161	0.4824	0.3612	0.3636	0.08489	0.2826
671	~ubo	3	0.02518	0.4204	0.4022	0.4286	0.05344	0.2826
672	~amyento	3	0.008536	0.4554	0.3829	1	1	0.2823
674	~ovo	3	0.02006	0.2658	0.561	0.6	0.1304	0.2823
673	~idençia	2	0.008323	0.4055	0.4331	0.3333	0.08696	0.2823
675	~nda	1	0.008963	0.4315	0.4042	0.02941	0.005291	0.2816
676	~rdes	6	0.008749	0.5543	0.28	0.4	0.3438	0.281
677	~icaron	1	0.003841	0.3356	0.5016	0.07692	0.05882	0.2803
678	~stra	1	0.002561	0.5034	0.3278	0.1	0.0009479	0.2779
679	~andoles	1	0.006402	0.6041	0.2219	0.1667	0.1429	0.2774
680	~iéndolo	2	0.01088	0.3051	0.5142	0.2857	0.25	0.2767
681	~jese	3	0.00939	0.3803	0.4376	0.375	0.3	0.2758
682	~tro	3	0.008536	0.3319	0.4839	0.12	0.2224	0.2748
683	~ço	1	0.01024	0.1258	0.6877	0.05	0.009709	0.2746
685	~ceso	1	0.003841	0.3356	0.4832	0.25	0.375	0.2742
684	~sigo	1	0.003841	0.3356	0.4832	0.25	0.03226	0.2742
686	~él	3	0.008536	0.4111	0.4022	0.4286	0.01754	0.2739
687	~jn	2	0.01921	0.2397	0.561	0.4	0.1667	0.2733
688	~est	2	0.005762	0.2517	0.561	0.4	0.375	0.2728
689	~endole	1	0.003841	0.6712	0.143	0.07143	0.04348	0.2727

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
690	~zer	1	0.005122	0.5034	0.3086	0.05882	0.001919	0.2724
691	~eça	1	0.006402	0.2014	0.6048	0.1111	0.209	0.2709
692	~fetan	1	0.01152	0.5593	0.2416	0.5	0.75	0.2708
693	~zco	2	0.003201	0.5873	0.2219	0.2222	0.1429	0.2708
694	~dir	1	0.005122	0.2517	0.5528	0.07692	0.02326	0.2699
695	~quiera	2	0.006402	0.4405	0.3624	0.5	0.1111	0.2698
696	~ías	2	0.005122	0.3692	0.4301	0.1111	0.006711	0.2681
697	~ctos	1	0.003841	0.3356	0.4644	0.2	0.09091	0.2679
698	~zio	2	0.003841	0.3356	0.4635	0.3333	0.2857	0.2676
699	~sorrey	1	0.008963	0.4315	0.3624	0.25	0.025	0.2676
700	~ándolas	1	0.0128	0.4027	0.3829	0.3333	0.25	0.2662
701	~siento	1	0.0128	0.4027	0.3829	0.3333	0.1667	0.2662
702	~yera	2	0.009603	0.4045	0.3829	0.6667	0.6667	0.2657
704	~nojos	1	0.006402	0.4027	0.3829	0.3333	0.3333	0.264
703	~tus	1	0.006402	0.4027	0.3829	0.3333	0.1667	0.264
705	~yentos	4	0.004161	0.5915	0.196	0.5	0.6	0.2639
706	~umbre	2	0.007682	0.4171	0.3624	0.25	0.6296	0.2624
707	~ban	4	0.01569	0.4924	0.2717	0.05556	0.0223	0.2599
708	~quél	1	0.01408	0.2746	0.4832	0.25	0.04348	0.2573
709	~cura	1	0.006402	0.4027	0.3624	0.25	0.03333	0.2572
710	~fe	1	0.006402	0.4027	0.3624	0.25	0.02564	0.2572
711	~miente	1	0.01024	0.3775	0.3829	0.3333	0.4	0.2569
712	~éndolo	1	0.01921	0.4698	0.2795	0.1	0.09091	0.2562
713	~nça	4	0.007042	0.4992	0.2577	0.1739	0.2027	0.2546
714	~cino	1	0.01536	0.5034	0.2416	0.5	0.9677	0.2535
715	~yesen	1	0.008963	0.1438	0.604	0.125	0.06667	0.2523
716	~pecjalmente	1	0.007682	0.5034	0.2416	0.5	0.5	0.2509
717	~forçar	1	0.007682	0.5034	0.2416	0.5	0.5	0.2509
720	~namentos	1	0.005122	0.5034	0.2416	0.5	0.8889	0.25
718	~mysa	1	0.005122	0.5034	0.2416	0.5	0.07143	0.25
719	~mjnas	1	0.005122	0.5034	0.2416	0.5	0.0625	0.25
721	~negavan	1	0.002561	0.5034	0.2416	0.5	0.5	0.2492
725	~opuesto	1	0.002561	0.5034	0.2416	0.5	0.5	0.2492
726	~nun	1	0.002561	0.5034	0.2416	0.5	0.5	0.2492
727	~nego	1	0.002561	0.5034	0.2416	0.5	0.5	0.2492
724	~sierto	1	0.002561	0.5034	0.2416	0.5	0.3333	0.2492
723	~sirven	1	0.002561	0.5034	0.2416	0.5	0.08333	0.2492
722	~suyo	1	0.002561	0.5034	0.2416	0.5	0.01923	0.2492
728	~jentes	1	0.003841	0.3356	0.4073	0.09091	0.04762	0.2489
729	~bi	2	0.006402	0.4405	0.2974	0.1333	0.08696	0.2481
730	~sçio	1	0.002561	0.5034	0.2383	0.1111	0.04545	0.2481
731	~rlo	1	0.002561	0.5034	0.2371	0.03226	0.02273	0.2477
732	~uxo	1	0.002561	0.5034	0.2346	0.2	0.02778	0.2468
733	~van	6	0.03542	0.502	0.2028	0.06452	0.02128	0.2468
734	~ntura	2	0.004481	0.4195	0.3138	0.25	0.1529	0.2459
735	~ades	2	0.004481	0.2936	0.4351	0.04651	0.02817	0.2444
736	~entras	2	0.006402	0.2397	0.4832	0.5	0.25	0.2431
737	~scandalo	1	0.002561	0.5034	0.2219	0.3333	0.1429	0.2426
738	~spañoles	1	0.002561	0.5034	0.2219	0.3333	0.01667	0.2426
739	~beer	1	0.007682	0.3356	0.3829	0.3333	0.1176	0.2421

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
740	~cabo	1	0.007682	0.3356	0.3829	0.3333	0.04	0.2421
741	~stro	3	0.01024	0.3542	0.3603	0.2727	0.04605	0.2416
742	~ándolo	5	0.0146	0.355	0.3526	0.8333	0.8889	0.2407
743	~esçió	1	0.007682	0.1678	0.544	0.1667	0.01205	0.2398
744	~sita	2	0.005762	0.2265	0.4832	0.5	0.375	0.2385
745	~vja	1	0.0128	0.302	0.4005	0.1111	0.02041	0.2385
746	~rededor	1	0.007682	0.3356	0.3624	0.25	0.4444	0.2352
747	~éndoselo	1	0.002561	0.5034	0.196	0.25	0.25	0.234
748	~issima	1	0.003841	0.3356	0.3624	0.25	0.25	0.234
749	~pedido	1	0.003841	0.3356	0.3624	0.25	0.04545	0.234
750	~sençia	2	0.01024	0.1278	0.561	0.4	0.07317	0.233
751	~ent	3	0.006402	0.2269	0.4644	0.6	0.92	0.2326
752	~ti	1	0.007682	0.1678	0.5208	0.125	0.05556	0.2321
753	~ques	1	0.007682	0.1678	0.5208	0.125	0.01	0.2321
754	~eçido	1	0.007682	0.1678	0.5186	0.06667	0.05263	0.2314
755	~és	3	0.01366	0.4409	0.2383	0.3333	0.9376	0.231
756	~jeren	1	0.003841	0.3356	0.3526	0.1667	0.09091	0.2307
757	~nze	2	0.01024	0.1342	0.544	0.3333	0.1746	0.2295
758	~xa	1	0.003841	0.3356	0.3468	0.1111	0.3	0.2288
759	~çido	4	0.01088	0.2413	0.4315	0.1429	0.08475	0.2279
760	~liçia	1	0.008963	0.4315	0.2416	0.5	0.9231	0.2274
761	~hiz	1	0.008963	0.4315	0.2416	0.5	0.9167	0.2274
762	~çon	1	0.002561	0.5034	0.1744	0.2	0.07692	0.2268
763	~andoselo	1	0.008963	0.2876	0.3829	0.3333	0.2	0.2265
764	~içia	1	0.006402	0.2014	0.4712	0.1429	0.1071	0.2263
765	~tasen	1	0.01024	0.1258	0.5315	0.07692	0.05556	0.2225
766	~cyo	2	0.004481	0.2936	0.3677	0.4	0.4	0.2219
767	~rey	1	0.01793	0.2157	0.4302	0.08333	0.01852	0.2213
768	~jestad	1	0.008963	0.4315	0.2219	0.3333	0.9953	0.2208
769	~rtos	1	0.01024	0.3775	0.2742	0.05556	0.00641	0.2207
770	~ue	1	0.005122	0.2517	0.4037	0.01724	6.34E-05	0.2202
771	~rda	1	0.003841	0.3356	0.3201	0.08333	0.01	0.2199
772	~signo	1	0.0128	0.4027	0.2416	0.5	0.6667	0.219
773	~altos	1	0.0128	0.4027	0.2416	0.5	0.5	0.219
774	~ç	1	0.01024	0.1258	0.5208	0.125	0.125	0.219
775	~çiones	7	0.01116	0.2737	0.3697	0.1591	0.1222	0.2182
776	~nu	1	0.006402	0.4027	0.2416	0.5	0.5	0.2169
777	~llegarse	1	0.006402	0.4027	0.2416	0.5	0.5	0.2169
778	~éronle	1	0.01665	0.3872	0.2416	0.25	0.1	0.2152
779	~viso	1	0.01024	0.2517	0.3829	0.3333	0.09091	0.215
780	~nçias	1	0.003841	0.3356	0.3037	0.05	0.01754	0.2144
781	~çientas	1	0.007682	0.1678	0.4644	0.2	0.1429	0.2133
782	~dones	1	0.005122	0.2517	0.3829	0.3333	0.1429	0.2133
783	~stes	1	0.005122	0.2517	0.3761	0.1429	0.1111	0.211
784	~visiones	1	0.006402	0.4027	0.2219	0.3333	0.06667	0.2103
785	~chada	3	0.003414	0.3915	0.2306	0.375	0.3333	0.2085
786	~myentos	2	0.009603	0.2607	0.3526	0.3333	0.3333	0.2076
787	~lança	1	0.01152	0.2237	0.3829	0.3333	0.3333	0.2061
788	~screvir	1	0.006402	0.4027	0.196	0.25	0.05882	0.2017
789	~dençia	1	0.01024	0.1258	0.4644	0.2	0.09091	0.2002

Tabla A.1: Catálogo de sufijos del siglo XVI del CHEM (continuación)

No.	Sufijo	Frec.	Cuadros	Economía	Entropía	Prob. 1	Prob. 2	Afijalidad
790	~umbres	1	0.006402	0.2014	0.3829	0.3333	0.3333	0.1969
791	~çin	1	0.006402	0.2014	0.3829	0.3333	0.25	0.1969
792	~unbres	1	0.006402	0.2014	0.3829	0.3333	0.125	0.1969
793	~moneda	1	0.007682	0.3356	0.2416	0.5	0.5882	0.195
795	~calce	1	0.007682	0.3356	0.2416	0.5	0.3333	0.195
794	~derredor	1	0.007682	0.3356	0.2416	0.5	0.2857	0.195
797	~escriba	1	0.003841	0.3356	0.2416	0.5	0.3333	0.1937
796	~sello	1	0.003841	0.3356	0.2416	0.5	0.25	0.1937
798	~patos	1	0.01408	0.183	0.3829	0.3333	0.2	0.1934
799	~çiplinas	1	0.006402	0.2014	0.3677	0.2	0.07143	0.1918
800	~rrey	2	0.01729	0.224	0.3331	0.2857	0.2179	0.1915
801	~ho	2	0.006402	0.3835	0.1825	0.04762	0.0006331	0.1908
802	~jesse	1	0.007682	0.3356	0.2219	0.3333	0.2	0.1884
803	~cias	1	0.01152	0.2237	0.3278	0.1	0.05882	0.1877
805	~ssor	1	0.003841	0.3356	0.2219	0.3333	0.7143	0.1871
804	~rzida	1	0.003841	0.3356	0.2219	0.3333	0.25	0.1871
806	~entre	1	0.0128	0.302	0.2416	0.5	0.006667	0.1855
807	~ndan	1	0.003841	0.3356	0.2021	0.06667	0.02174	0.1805
808	~artos	1	0.006402	0.2014	0.3312	0.2	0.09091	0.1797
809	~iéndoselo	1	0.008963	0.1438	0.3829	0.3333	0.3333	0.1786
810	~nientos	2	0.01024	0.1342	0.3829	0.6667	0.1474	0.1758
811	~endolos	1	0.007682	0.1678	0.3501	0.1429	0.1429	0.1752
812	~dizen	1	0.01024	0.1258	0.3829	0.3333	0.008929	0.173
813	~riesgo	1	0.01536	0.2517	0.2416	0.5	0.3333	0.1696
814	~ajo	2	0.01344	0.1419	0.3526	0.3333	0.09524	0.1693
816	~suelto	1	0.01152	0.1119	0.3829	0.3333	0.25	0.1688
815	~seis	1	0.01152	0.1119	0.3829	0.3333	0.03704	0.1688
817	~has	1	0.01665	0.3098	0.1792	0.04167	0.001486	0.1685
818	~llego	1	0.005122	0.2517	0.2416	0.5	0.3333	0.1661
819	~axo	1	0.01665	0.1549	0.3266	0.1111	0.04878	0.166
820	~bol	1	0.01536	0.2517	0.2219	0.3333	0.2857	0.163
821	~ole	1	0.01536	0.2517	0.2114	0.01235	0.006369	0.1595
822	~scrivo	1	0.005122	0.2517	0.196	0.25	0.07143	0.1509
823	~aparte	1	0.006402	0.2014	0.2416	0.5	0.1111	0.1498
824	~quj	1	0.02689	0.1918	0.2219	0.3333	0.03175	0.1468
825	~garra	1	0.01408	0.183	0.2416	0.5	0.6667	0.1462
826	~scrivjo	1	0.006402	0.2014	0.2219	0.3333	0.4	0.1432
827	~color	1	0.007682	0.1678	0.2416	0.5	0.1429	0.139
828	~dula	1	0.008963	0.1438	0.2219	0.3333	0.8125	0.1249
829	~hos	1	0.006402	0.2014	0.1597	0.03571	0.001242	0.1225
830	~suelle	1	0.01152	0.1119	0.2416	0.5	0.5	0.1217
831	~soluçión	1	0.01152	0.1119	0.2416	0.5	0.5	0.1217

B. Procedimiento de creación de archivos y generación de reglas

```
# GENERACIÓN DE REGLAS LÉXICAS-MORFOLÓGICAS
# Dividir corpus de entrenamiento en corpus de entrenamiento léxico y corpus de entrenamiento
# contextual etiquetados
cat CorpusEntEti.txt | Utilities/divide-in-two-rand.prl CorpusEntLexEti.txt CorpusEntConEti.txt

# Crear corpus de entrenamiento etiquetado sin etiquetas quitando etiquetas a corpus de
# entrenamiento etiquetado
cat CorpusEntEti.txt | Utilities/tagged-to-untagged.prl > CorpusEntEtiSinEti.txt

# Agregar a corpus de entrenamiento etiquetado sin etiquetas el corpus adicional no etiquetado
cat CorpusEntEtiSinEti.txt CorpusAdicional.txt > CorpusEntNoEti.txt

# Crear lista de tipos ordenados de mayor a menor frecuencia BIGWORDLIST.
cat CorpusEntNoEti.txt | Utilities/wordlist-make.prl | sort +1 -rn | awk '{print $1}' >
BIGWORDLIST

# Crear lista de {tipo, etiqueta, conteo}. Es decir, el número de veces que una palabra está etiquetada
# con cierta etiqueta SMALLWORDTAGLIST.
cat CorpusEntLexEti.txt | Utilities/word-tag-count.prl | sort +2 -rn > SMALLWORDTAGLIST

# Crear lista de bigramas BIGBIGRAMLIST
cat CorpusEntNoEti.txt | Utilities/bigram-generate.prl | awk '{print $1, $2}' > BIGBIGRAMLIST

# Crear archivo con {sufijo, etiqueta_anterior, etiqueta_actual}
cat CorpusEntEti.txt | Utilities/tag-bigram-suf-generate-2.prl sufsmasprob_2_s16.txt
BIGWORDLIST 300 | sort +3 -rn > BIGRAMASDETAGSLISTCONSUFMP

# Generación de reglas con método modificado (v. 13)
Learner_Code/unknown-lexical-learn-sufs-13.prl BIGWORDLIST SMALLWORDTAGLIST
BIGBIGRAMLIST 300 LEXRULEOUTFILE-MEDINA-13 sufsmasprob_2_s16.txt
BIGRAMASDETAGSLISTCONSUFMP

# Generación de reglas con método Brill (v. 1)
Learner_Code/unknown-lexical-learn-brill-1.prl BIGWORDLIST SMALLWORDTAGLIST
BIGBIGRAMLIST 300 LEXRULEOUTFILE-BRILL-1
```

C. Reglas generadas por el método original de Brill

Tabla C-1: Reglas generadas por el método de Brill original

Num.	Regla
1.	NCMS a fhassuf 1 NCFS 109
2.	NCMS s fhassuf 1 NCMP 88.8761904761905
3.	NCMS do fhassuf 2 VMP00SM 56.5
4.	NCMS r fhassuf 1 VMN0000 50.95
5.	ó hassuf 1 VMIS3S0 39
6.	NCMP as fhassuf 2 NCFP 36
7.	VMP00SM ndo fhassuf 3 VMG0000 31.3333333333333
8.	NCMS la fgoodright NCFS 27
9.	NCMS an fhassuf 2 VMII3P0 24
10.	ron hassuf 3 VMIS3P0 18.1818181818182
11.	se hassuf 2 VMSI3S0 17
12.	NCMS on fhassuf 2 NCFS 16
13.	VMN0000 or fhassuf 2 NCMS 14
14.	r addsuf 1 VMIP3S0 13.0714285714286
15.	d hassuf 1 NCFS 13
16.	NCFS que fgoodright VMII3S0 12.3666666666667
17.	años goodleft DN0CP0 11.6666666666667
18.	dos hassuf 3 VMP00PM 10.5
19.	ava hassuf 3 VMII3S0 9
20.	sen hassuf 3 VMSI3P0 9
21.	NCMS que fgoodright VMIP3S0 8.48809523809524
22.	1 char W 8
23.	ren hassuf 3 VMSF3P0 8
24.	NCMP las fgoodright NCFP 7
25.	vuestra goodleft APS000 6.92553191489362
26.	NCFS ca fhassuf 2 AQ0FS0 6
27.	gan hassuf 3 VMSP3P0 6
28.	mos hassuf 3 VMIP1P0 6
29.	iado hassuf 4 NCMS 5.33333333333333
30.	muy goodright AQ0MS0 5.02777777777778
31.	ía hassuf 2 VMII3S0 5
32.	NCFS n faddsuf 1 VMSP3S0 5
33.	NCMS en fhassuf 2 VMSP3P0 5
34.	are hassuf 3 VMSF3S0 5
35.	VMIP3S0 o fhassuf 1 VMIS3S0 4.66666666666667
36.	ente hassuf 4 RM 4.5
37.	NP00 , fgoodright NCMS 4.14285714285714
38.	lico hassuf 4 AQ0MS0 4
39.	VMIP3S0 n fdeletesuf 1 VMIP3P0 4
40.	después goodleft CC 4
41.	ría hassuf 3 VMIC3S0 4
42.	NCMP es faddsuf 2 NCMS 3
43.	VMIC3S0 o fchar NCFS 3
44.	NCMS z fhassuf 1 NCFS 3
45.	NCMS le fgoodright VMIS3S0 3

Tabla C-1: Reglas generadas por el método de Brill original (continuación)

Num.	Regla
46.	ea hassuf 2 AQ0FS0 3
47.	NCFS ba fhassuf 2 VMII3S0 3
48.	o addsuf 1 APS000 3
49.	Señor goodright NCMS 3
50.	de addsuf 2 APS000 3
51.	enos hassuf 4 AQ0MP0 3
52.	_ char W 3
53.	ll deletesuf 2 DN0CP0 3
54.	ués hassuf 3 RT 3
55.	VMIP3S0 de fgoodright RL 3
56.	VMIP3S0 lo fgoodleft RM 3
57.	san hassuf 3 VMIP3P0 3
58.	VMII3P0 tan fhassuf 3 VMIP3P0 3
59.	que deletesuf 3 CC 3
60.	pues goodleft CC 3
61.	has hassuf 3 AQ0FP0 3
62.	é hassuf 1 VMIS1S0 3
63.	VMIP3S0 i fhassuf 1 VMIS1S0 3
64.	ia deletesuf 2 VMIC3S0 3
65.	ere hassuf 3 VMSF3S0 3
66.	rá hassuf 2 VMIF3S0 3
67.	\ char F 3
68.	VMIS1S0 a fchar VMIF1S0 3
69.	han goodright VMP00SM 2.9491341991342
70.	ral hassuf 3 AQ0MS0 2.8
71.	smo hassuf 3 RM 2.6
72.	días goodleft DN0CP0 2.52380952380952
73.	j hassuf 1 NCMS 2
74.	VMN0000 del fgoodright NCMS 2
75.	NCMS ión fhassuf 3 NCFS 2
76.	santa goodright NCFS 2
77.	Ç char NP00 2
78.	diez goodright NCMP 2
79.	VMP00PM sus fgoodright NCMP 2
80.	NCMS sto fhassuf 3 VMP00SM 2
81.	vo hassuf 2 VMIS3S0 2
82.	vio hassuf 3 VMIS3S0 2
83.	ndio hassuf 4 VMIS3S0 2
84.	ones hassuf 4 NCFP 2
85.	muchas goodright NCFP 2
86.	VMII3S0 e fhassuf 1 VMIP3S0 2
87.	be hassuf 2 VMIP3S0 2
88.	va addsuf 2 VMIP3S0 2
89.	NCMS uno fhassuf 3 AQ0MS0 2
90.	hermano goodleft AQ0MS0 2
91.	NCMS es fgoodright AQ0MS0 2
92.	cra hassuf 3 AQ0FS0 2
93.	VMII3S0 s faddsuf 1 AQ0FS0 2

Tabla C-1: Reglas generadas por el método de Brill original (continuación)

Num.	Regla
94.	NCFS cosa fgoodright AQ0FS0 2
95.	ntra hassuf 4 APS000 2
96.	NP00 otras fgoodleft APS000 2
97.	juramento goodright APS000 2
98.	VMII3S0 ga fhassuf 2 VMSP3S0 2
99.	NCFS se fgoodright VMSP3S0 2
100.	los deletesuf 3 AQ0MP0 2
101.	tos hassuf 3 AQ0MP0 2
102.	NCMP ntes fhassuf 4 AQ0MP0 2
103.	NP00 bre fhassuf 3 W 2
104.	W z fchar DN0CP0 2
105.	ebe hassuf 3 DN0CP0 2
106.	VMG0000 q fchar RT 2
107.	pues deletesuf 4 RT 2
108.	hasta goodright RT 2
109.	nan hassuf 3 VMIP3P0 2
110.	nen hassuf 3 VMIP3P0 2
111.	dan hassuf 3 VMIP3P0 2
112.	no deletesuf 2 CC 2
113.	después goodleft CC 2
114.	NCMS ay fgoodleft CC 2
115.	tos addsuf 3 RG 2
116.	NCMP no fgoodright RG 2
117.	VMII3P0 e fgoodright VMSP3P0 2
118.	NCMS me fgoodright VMIP1S0 2
119.	preguntas goodleft AQ0FP0 2
120.	NCFP cosas fgoodright AQ0FP0 2
121.	ada hassuf 3 VMP00SF 2
122.	VMP00SF la fgoodright NCFS 2
123.	NCFS dra fhassuf 3 VMIF3S0 2
124.	{ char F 2
125.	VMP00PM p fchar AQ0MPP 2
126.	rían hassuf 4 VMIC3P0 2
127.	ran hassuf 3 VMIF3P0 2
128.	NCMS l fdeletesuf 1 APCMS0 2
129.	as deletesuf 2 DI0FP0 2
130.	NCFS ez fhassuf 2 NCCS 2
131.	endo deletesuf 4 VSG0000 2
132.	én hassuf 2 I 2
133.	sta deletesuf 3 APDFS0 2
134.	NP00 í fchar AQSFS0 2
135.	imo hassuf 3 AQSMS0 2
136.	NCMS - fchar Fg 2
137.	stra hassuf 4 DP1FSP 2
138.	VMSI3S0 f fchar VSSI3S0 2
139.	señores goodleft DD0MP0 1.95238095238095

D. Reglas generadas por el método de Brill con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF

Tabla D-1: Reglas generadas por el método de Brill con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF

Num.	Regla
1.	NCMS a fhassuf 1 NCFS 109
2.	NCMS s fhassuf 1 NCMP 88.8761904761905
3.	NCMS do fhassuf 2 VMP00SM 56.5
4.	NCMS r fhassuf 1 VMN0000 51.95
5.	ó hassuf 1 VMIS3S0 39
6.	NCMP as fhassuf 2 NCFP 36
7.	VMP00SM ndo fhassuf 3 VMG0000 31.3333333333333
8.	NCMS la fgoodright NCFS 27
9.	NCMS an fhassuf 2 VMII3P0 24
10.	ron hassuf 3 VMIS3P0 18.1818181818182
11.	se hassuf 2 VMSI3S0 17
12.	NCMS ion fhassuf 3 NCFS 16
13.	VMN0000 or fhassuf 2 NCMS 14
14.	d hassuf 1 NCFS 13
15.	NCFS que fgoodright VMII3S0 12.8666666666667
16.	años goodleft DN0CP0 11.6666666666667
17.	NCMS que fgoodright VMIP3S0 11.4880952380952
18.	VMII3P0 RL an ftagantysufmp VMIP3P0 11
19.	dos hassuf 3 VMP00PM 10.5
20.	NCFS NCFS ava ftagantysufmp VMII3S0 9
21.	sen hassuf 3 VMSI3P0 9
22.	1 char W 8
23.	ren hassuf 3 VMSF3P0 8
24.	NCMP las fgoodright NCFP 7
25.	mente hassuf 5 RM 7
26.	vuestra goodleft APS000 6.92553191489362
27.	NCFS ca fhassuf 2 AQ0FS0 6
28.	gan hassuf 3 VMSP3P0 6
29.	NCMP Fc mos ftagantysufmp VMIP1P0 6
30.	iado hassuf 4 NCMS 5.33333333333333
31.	muy goodright AQ0MS0 5.02777777777778
32.	NCMS F n ftagantysufmp APS000 5
33.	APS000 en fhassuf 2 VMSP3P0 5
34.	NCMS PP3CS00 are ftagantysufmp VMSF3S0 5
35.	NP00 , fgoodright NCMS 4.14285714285714
36.	VMIP3S0 o fhassuf 1 VMIS3S0 4
37.	NCFS í fchar VMII3S0 4
38.	NCFS ba fhassuf 2 VMII3S0 4
39.	NCFS se fgoodright VMSP3S0 4
40.	VMIP3S0 n fhassuf 1 VMIP3P0 4
41.	después goodleft CC 4
42.	ría hassuf 3 VMIC3S0 4

Tabla D-1: Reglas generadas por el método de Brill con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF (continuación)

Num.	Regla
43.	NCMS es fgoodright AQ0MS0 3.8
44.	VMIC3S0 o fchar NCFS 3
45.	NCMS z fhassuf 1 NCFS 3
46.	NCMS le fgoodright VMIS3S0 3
47.	NCFS le fgoodright VMIP3S0 3
48.	ello goodleft APS000 3
49.	NCFS no fgoodright VMSP3S0 3
50.	_ char W 3
51.	NCMS APS000 ll ftagantysufmp DN0CP0 3
52.	pués hassuf 4 RT 3
53.	VMIP3S0 de fgoodright RL 3
54.	VMIP3S0 lo fgoodleft RM 3
55.	has hassuf 3 AQ0FP0 3
56.	ebi hassuf 3 VMIS1S0 3
57.	é hassuf 1 VMIS1S0 3
58.	ere hassuf 3 VMSF3S0 3
59.	rá hassuf 2 VMIF3S0 3
60.	\ char F 3
61.	ran hassuf 3 VMIF3P0 3
62.	VMIS1S0 a fchar VMIF1S0 3
63.	mo hassuf 2 RM 2.76891891891892
64.	VMP00SM su fgoodright NCMS 2.66666666666667
65.	NCMS al fhassuf 2 AQ0MS0 2.56666666666667
66.	días goodleft DN0CP0 2.52380952380952
67.	VMII3S0 me fgoodright VMIP3S0 2.14375
68.	NP00 DP1CSS so ftagantysufmp NCMS 2
69.	NP00 F j ftagantysufmp NCMS 2
70.	NP00 y fhassuf 1 NCMS 2
71.	VMP00SM edo fhassuf 3 NCMS 2
72.	VMIP3P0 y fgoodleft NCMS 2
73.	VMN0000 del fgoodright NCMS 2
74.	NCMS APS000 ción ftagantysufmp NCFS 2
75.	santa goodright NCFS 2
76.	Ç char NP00 2
77.	diez goodright NCMP 2
78.	VMP00PM sus fgoodright NCMP 2
79.	NCMS sto fhassuf 3 VMP00SM 2
80.	vio hassuf 3 VMIS3S0 2
81.	vo hassuf 2 VMIS3S0 2
82.	iones hassuf 5 NCFP 2
83.	muchas goodright NCFP 2
84.	çe hassuf 2 VMIP3S0 2
85.	be hassuf 2 VMIP3S0 2
86.	VMSP3S0 . fgoodleft VMIP3S0 2
87.	hermano goodleft AQ0MS0 2
88.	magestad goodleft AQ0FS0 2
89.	NCFS cosa fgoodright AQ0FS0 2

Tabla D-1: Reglas generadas por el método de Brill con sufijos descubiertos, nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF (continuación)

Num.	Regla
90.	pena goodleft APS000 2
91.	VMII3S0 ga fhassuf 2 VMSP3S0 2
92.	VMIS1S0 y fgoodright VMSP3S0 2
93.	tos hassuf 3 AQ0MP0 2
94.	NCMP ntes fhassuf 4 AQ0MP0 2
95.	NP00 APS000 re ftagantysufmp W 2
96.	W z fchar DN0CP0 2
97.	VMG0000 q fchar RT 2
98.	VMII3S0 más fgoodleft RT 2
99.	hasta goodright RT 2
100.	NCMS APS000 ante ftagantysufmp RL 2
101.	VMII3S0 ca fhassuf 2 RL 2
102.	iera hassuf 4 VMSI3S0 2
103.	pues goodleft CC 2
104.	NCMS ay fgoodleft CC 2
105.	NCMP no fgoodright RG 2
106.	NCMS me fgoodright VMIP1S0 2
107.	preguntas goodleft AQ0FP0 2
108.	NCFP cosas fgoodright AQ0FP0 2
109.	NCFS CS aria ftagantysufmp VMIC3S0 2
110.	rada hassuf 4 VMP00SF 2
111.	{ char F 2
112.	VMP00PM p fchar AQ0MPP 2
113.	arian hassuf 5 VMIC3P0 2
114.	arían hassuf 5 VMIC3P0 2
115.	VMIP3S0 l fhassuf 1 APCMS0 2
116.	én hassuf 2 I 2
117.	NP00 í fchar AQSFS0 2
118.	VMII3S0 ra fhassuf 2 VSII3S0 2
119.	RM d fchar AQSMS0 2
120.	NCMS - fchar Fg 2
121.	uestro hassuf 6 DP2MSP 2
122.	stra hassuf 4 DP1FSP 2
123.	VMIF3P0 que fgoodright VSII3P0 2
124.	u hassuf 1 DP3CS0 2
125.	VMSI3S0 f fchar VSSI3S0 2
126.	señores goodleft DD0MP0 1.95238095238095

E. Reglas generadas por el método de Brill con sufijos descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF

Tabla E-1: Reglas generadas por el método de Brill con sufijos descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF

Num.	Regla
1.	NCMS a fhassuf 1 NCFS 109
2.	NCMS s fhassuf 1 NCMP 88.8761904761905
3.	NCMS do fhassuf 2 VMP00SM 56.5
4.	NCMS r fhassuf 1 VMN0000 51.95
5.	ó hassuf 1 VMIS3S0 39
6.	NCMP as fhassuf 2 NCFP 36
7.	VMP00SM ndo fhassuf 3 VMG0000 31.3333333333333
8.	NCMS la fgoodright NCFS 27
9.	NCMS an fhassuf 2 VMII3P0 24
10.	ron hassuf 3 VMIS3P0 18.1818181818182
11.	se hassuf 2 VMSI3S0 17
12.	NCMS ion fhassuf 3 NCFS 16
13.	VMN0000 or fhassuf 2 NCMS 14
14.	d hassuf 1 NCFS 13
15.	NCFS que fgoodright VMII3S0 12.8666666666667
16.	años goodleft DN0CP0 11.6666666666667
17.	NCMS que fgoodright VMIP3S0 11.4880952380952
18.	dos hassuf 3 VMP00PM 10.5
19.	ava hassuf 3 VMII3S0 9
20.	sen hassuf 3 VMSI3P0 9
21.	1 char W 8
22.	ren hassuf 3 VMSF3P0 8
23.	NCMP las fgoodright NCFP 7
24.	mente hassuf 5 RM 7
25.	vuestra goodleft APS000 6.92553191489362
26.	NCFS ca fhassuf 2 AQ0FS0 6
27.	gan hassuf 3 VMSP3P0 6
28.	mos hassuf 3 VMIP1P0 6
29.	iado hassuf 4 NCMS 5.33333333333333
30.	muy goodright AQ0MS0 5.02777777777778
31.	NCMS en fhassuf 2 VMSP3P0 5
32.	are hassuf 3 VMSF3S0 5
33.	NP00 , fgoodright NCMS 4.14285714285714
34.	VMIP3S0 o fhassuf 1 VMIS3S0 4
35.	NCFS í fchar VMII3S0 4
36.	NCFS ba fhassuf 2 VMII3S0 4
37.	NCFS se fgoodright VMSP3S0 4
38.	VMIP3S0 n fhassuf 1 VMIP3P0 4
39.	después goodleft CC 4
40.	ría hassuf 3 VMIC3S0 4
41.	NCMS es fgoodright AQ0MS0 3.8
42.	VMIC3S0 o fchar NCFS 3

Tabla E-1: Reglas generadas por el método de Brill con sufijos descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF (continuación)

Num.	Regla
43.	NCMS z fhassuf 1 NCFS 3
44.	NCMS le fgoodright VMIS3S0 3
45.	NCFS le fgoodright VMIP3S0 3
46.	ello goodleft APS000 3
47.	NCFS no fgoodright VMSP3S0 3
48.	_ char W 3
49.	ll hassuf 2 DN0CP0 3
50.	pués hassuf 4 RT 3
51.	VMIP3S0 de fgoodright RL 3
52.	VMIP3S0 lo fgoodleft RM 3
53.	san hassuf 3 VMIP3P0 3
54.	VMII3P0 tan fhassuf 3 VMIP3P0 3
55.	has hassuf 3 AQ0FP0 3
56.	ebi hassuf 3 VMIS1S0 3
57.	é hassuf 1 VMIS1S0 3
58.	ere hassuf 3 VMSF3S0 3
59.	rá hassuf 2 VMIF3S0 3
60.	\ char F 3
61.	ran hassuf 3 VMIF3P0 3
62.	VMIS1S0 a fchar VMIF1S0 3
63.	mo hassuf 2 RM 2.76891891891892
64.	VMP00SM su fgoodright NCMS 2.66666666666667
65.	NCMS al fhassuf 2 AQ0MS0 2.56666666666667
66.	días goodleft DN0CP0 2.52380952380952
67.	VMII3S0 me fgoodright VMIP3S0 2.14375
68.	j hassuf 1 NCMS 2
69.	NP00 y fhassuf 1 NCMS 2
70.	VMP00SM edo fhassuf 3 NCMS 2
71.	VMIP3P0 de fgoodright NCMS 2
72.	VMN0000 del fgoodright NCMS 2
73.	NCMS ión fhassuf 3 NCFS 2
74.	santa goodright NCFS 2
75.	Ç char NP00 2
76.	diez goodright NCMP 2
77.	VMP00PM sus fgoodright NCMP 2
78.	NCMS sto fhassuf 3 VMP00SM 2
79.	vio hassuf 3 VMIS3S0 2
80.	vo hassuf 2 VMIS3S0 2
81.	iones hassuf 5 NCFP 2
82.	muchas goodright NCFP 2
83.	çe hassuf 2 VMIP3S0 2
84.	be hassuf 2 VMIP3S0 2
85.	VMSP3S0 . fgoodleft VMIP3S0 2
86.	hermano goodleft AQ0MS0 2
87.	magestad goodleft AQ0FS0 2
88.	NCFS cosa fgoodright AQ0FS0 2
89.	pena goodleft APS000 2

Tabla E-1: Reglas generadas por el método de Brill con sufijos descubiertos, sin nueva plantilla de regla y sin plantillas ADDSUF y DELETESUF (continuación)

Num.	Regla
90.	juramento goodright APS000 2
91.	VMII3S0 ga fhassuf 2 VMSP3S0 2
92.	VMIS1S0 y fgoodright VMSP3S0 2
93.	tos hassuf 3 AQ0MP0 2
94.	NCMP ntes fhassuf 4 AQ0MP0 2
95.	NP00 bre fhassuf 3 W 2
96.	W z fchar DN0CP0 2
97.	VMG0000 q fchar RT 2
98.	VMII3S0 más fgoodleft RT 2
99.	hasta goodright RT 2
100.	VMII3S0 ca fhassuf 2 RL 2
101.	iera hassuf 4 VMSI3S0 2
102.	ndan hassuf 4 VMIP3P0 2
103.	VMSP3P0 , fgoodright VMIP3P0 2
104.	pues goodleft CC 2
105.	NCMS ay fgoodleft CC 2
106.	NCMP no fgoodright RG 2
107.	VMII3P0 e fgoodright VMSP3P0 2
108.	NCMS me fgoodright VMIP1S0 2
109.	preguntas goodleft AQ0FP0 2
110.	NCFP cosas fgoodright AQ0FP0 2
111.	aria hassuf 4 VMIC3S0 2
112.	rada hassuf 4 VMP00SF 2
113.	{ char F 2
114.	VMP00PM p fchar AQ0MPP 2
115.	arian hassuf 5 VMIC3P0 2
116.	arían hassuf 5 VMIC3P0 2
117.	VMIP3S0 l fhassuf 1 APCMS0 2
118.	én hassuf 2 I 2
119.	NP00 í fchar AQSFS0 2
120.	VMII3S0 ra fhassuf 2 VSII3S0 2
121.	RM d fchar AQSMS0 2
122.	NCMS - fchar Fg 2
123.	uestro hassuf 6 DP2MSP 2
124.	stra hassuf 4 DP1FSP 2
125.	VMIF3P0 que fgoodright VSII3P0 2
126.	u hassuf 1 DP3CS0 2
127.	VMSI3S0 f fchar VSSI3S0 2
128.	señores goodleft DD0MP0 1.95238095238095

7. Referencias

ALCOBA, SANTIAGO. 1999. "La flexión verbal", en I. Bosque y V. Demonte (dirs.), *Gramática descriptiva de la lengua española*, v. 3, Madrid: Espasa-Calpe y RAE, 4915-4991.

ALLEN, JAMES. 1995. *Natural language understanding*. 2ª. Redwood City, California: Benjamin/Cummings.

AMBADIANG, THÉOPHILE. 1993. *La morfología flexiva*. Madrid: Taurus.

_____. 1999. "La flexión nominal. Género y número", en I. Bosque y V. Demonte (dirs.), *Gramática descriptiva de la lengua española*, v. 3, Madrid: Espasa-Calpe y RAE, 4843-4913.

ANANIADOU, SOPHIA Y J. MCNAUGHT (eds.)**.** 2006. *Text Mining for Biology and Biomedicine*. Norwood, MA: Artech House.

ANDERSON, STEPHEN R. 1985a. "Inflectional morphology", en T. Shopen (ed.), *Language typology and syntactic description. Grammatical categories and the lexicon*, v. III, Cambridge: Cambridge University Press, 150-201.

_____. 1985b. "Typological distinction in word formation", en T. Shopen (ed.),*Language typology and syntactic description. Grammatical categories and the lexicon*, v. III, Cambridge: Cambridge University Press, 3-56.

ANGUIANO PEÑA, GILBERTO. 2007. *Indización semiautomática para almacenar y recuperar información del léxico del español usado en México*. Tesis Maestría (Maestría Bibliotecología y Estudios de la Información), México, UNAM.

ANWARD, JAN. 2000. "A dynamic model of part-of-speech differentiation", en P. M. Vogel y B. Comrie (eds.), *Approaches to the Typology of Word Classes*, Berlín: Mouton de Gruyter, 3-45.

ARONOFF, MARK HOWARD. 1976. *Word formation in generative grammar*, Cambridge, Mass: The MIT press.

BEARD, ROBERT. 1998. "Derivation", en Andrew Spencer y Arnold M. Zwicky (eds.), *The handbook of morphology*, Oxford y Malden, Mass.: Blackwell, 44-65.

BELLO, ANDRÉS. 1984/1848. *Gramática de la lengua castellana*. Madrid: EDAF.

BENIERS, ELISABETH (trad. y ed.). 2000. *Lecturas de morfología*. México: UNAM.

BHAT, D. N. S. 2000. "Word classes and sentential functions", en P. M. Vogel y B. Comrie (eds.), *Approaches to the Typology of Word Classes*Berlín: Mouton de Gruyter, 47-63.

BLOOMFIELD, LEONARD. 1961/1933. *Language*. London: George Allen.

BOSQUE, IGNACIO. 1999. "El sintagma adjetival. Modificadores y complementos del adjetivo. Adjetivo y participio", en I. Bosque y V. Demonte (dirs.), *Gramática descriptiva de la lengua española,* v. 1, Madrid: Espasa-Calpe y RAE, 217-312.

BOSQUE, IGNACIO Y V. DEMONTE (dirs.). 1999. *Gramática descriptiva de la lengua española*. 3 v. Madrid: Espasa-Calpe y RAE.

BRANTS, THORSTEN. 2000. "TnT: a statistical part-of-speech tagger", en *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP 2000),* Seattle, WA, Morgan Kaufmann, pp.224-231.

BRILL, ERIC. 1992. "A simple rule_based part of speech tagger", en *Proceedings of the Third Conference on Applied Natural Language Processing,* Trento, Italy: ACL, 112-116.

_____. 1993. *A Corpus-Based Approach to Language Learning*. Ph.D. Dissertation, Department of Computer and Information Science, University of Pennsylvania.

_____. 1994. "Some Advances in Transformation-Based Part of Speech tagging", en *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI´94)*, Seattle, Washington, 722-727.

_____. 1995. "Transformation-Based Error-Driven Learning and Natural language Processing: A Case Study in Part-of-Speech Tagging", en *Computational Linguistics,* 21:4, pp. 543-565.

BROWN, K Y J. MILLER. 1999. *Concise encyclopedia of grammatical categories*. Oxford, UK: Elsevier Science.

CARSTAIRS-MCCARTHY, ANDREW. 1998. "Paradigmatic Structure: Inflectional Paradigms and Morphological Classes", en Andrew Spencer y Arnold M. Zwicky (eds.), *The handbook of morphology,* Oxford y Malden, Mass.: Blackwell, 323-334.

CHARNIAK, EUGENE. 1996/1993. *Statistical language learning*. Cambridge, Massachusetts: The MIT Press.

CHURCH, KENNETH. 1988. "A stochastic parts program and noun phrase parser for unrestricted text", en *Second Conference on Applied Natural Language Processing (proceedings)*, Austin, Texas, 136-143.

CLOEREN, JAN. 1999. "Tagsets", en H. van Halteren (ed.), *Syntactic Wordclass Tagging*, Dordrecht, Netherlands: Kluwer Academic, 37-54.

COMPANY COMPANY, CONCEPCIÓN. 1994. *Documentos lingüísticos de la Nueva España. Altiplano Central*. México: UNAM.

CREUTZ, M. Y K. LAGUS. 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)*. Finland: Espoo.

CROFT, WILLIAM. 1991. *Syntactic categories and grammatical relations: the cognitive organization of information*. Chicago: University of Chicago.

_____. 2000. "Part of speech as language universals and as language-particular categories", en P. M. Vogel y B. Comrie (eds.), *Approaches to the Typology of Word Classes*, Berlín: Mouton de Gruyter, 65-102.

CROMM, OLIVER. 1996. *Affixerkennung in deutschen Wortformen. Eine Untersuchung zum nicht-lexikalischen Segmentierungsverfahren von N. D. Andreev*. Francfort del Meno: Abschluß des Ergänzungsstudiums Linguistische Datenverarbeitung.

DAVIES, MARK. 2003a. "Annotation without lexicons: an alternative to the standard bootstrapping approach", en P. Rayson, et al. (eds.), *Proceedings from Corpus Linguistics 2003*, Lancaster, 174-83**.**

_____. 2003b. "Relational n-gram databases as a basis for unlimited annotation on very large corpora", en K. Simov (ed.), *Workshop on the Shallow Processing of Large Corpora*, Lancaster University, Lancaster, 23-33.

DELBECQUE, NICOLE. 2008. "Semántica cognitiva y categorización lingüística", en M. J. Rodríguez Espiñeira y Jesús Pena Seijas (coords.), *Categorización lingüística y límites intercategoriales*, Santiago de Compostela: Universidad de Santiago de Compostela, 20-56.

DEMONTE, VIOLETA. 1999. "El adjetivo: clases y usos. La posición del adjetivo en el sintagma nominal", en I. Bosque y V. Demonte (dirs.), *Gramática descriptiva de la lengua española*, v. 1, Madrid: Espasa-Calpe y RAE, 129-215.

DEROSE, S. J. 1988. "Grammatical Category Disambiguation by Statistical Optimization", en *Computational Linguistics*, 14: 1, 31-39.

DIXON, R. M. W. 2000/1982. "¿Dónde quedaron todos los adjetivos?", en Elisabeth Beniers (trad. y ed.), *Lecturas de morfología*, México: UNAM, 87-171

_____. 1999. "Adjectives", en K. Brown y J. Miller (eds.), *Concise encyclopedia of grammatical categories*, Oxford, UK: Elsevier Science, 1-7.

FABB, NIGEL. 1998. "Compounding", en Andrew Spencer y Arnold M. Zwicky (eds.), *The handbook of morphology*, Oxford y Malden, Mass.: Blackwell, 66-83.

FERNÁNDEZ LEBORANS, MARÍA JESÚS. 1999. "El nombre propio", en I. Bosque y V. Demonte (dirs.), *Gramática descriptiva de la lengua española*, v. 1, Madrid: Espasa-Calpe y RAE, 77-128.

GARCÍA HIDALGO, M. I. 1979. "La formalización del Analizador Gramatical del DEM", en Luis Fernando Lara, R. Ham Chande y M. I. García Hidalgo, *Investigaciones lingüísticas en lexicografía*, Jornadas, 89, México: El Colegio de México, 85-155.

GERDTS, DONNA B. 1998. "Incorporation", en Andrew Spencer y Arnold M. Zwicky (eds.), *The handbook of morphology*, Oxford y Malden, Mass.: Blackwell, 84-99.

GOLDSMITH, J. 2001. "Unsupervised Learning of the Morphology of a Natural Language", en *Computational Linguistics*, 27, 2, 153 – 198.

GONZÁLEZ CALVO, JOSÉ MANUEL. 1998. *Estudios de morfología española*. Cáceres: Universidad de Extremadura.

HAMMOND, MICHAEL. 2003. *Programming for linguistics: Perl for language researchers*. Oxford: Blackwell.

HAM CHANDE, ROBERTO. 1979. "Del 1 al 100 en Lexicografía", en Luis Fernando Lara, R. Ham Chande y M. I. García Hidalgo, *Investigaciones lingüísticas en lexicografía*, Jornadas, 89, México: El Colegio de México, 41-83.

HARRIS, ZELLIG S. 1955. "From Phoneme to Morpheme", en *Language*, 31, 2, 190–222.

HINDLE, DONALD. 1989. "Acquiring disambiguation rules from text", en *Proceedings of the 27th Meeting of the Association for Computational Linguistics (ACL-89*), Vancouver, 118-125.

HOCKETT, CHARLES F. 1953. "Review of The mathematical theory of communication", en *Language*, 29, 1.

_____. 1971/1958. *Curso de lingüística moderna*. Traducido por Emma Gregores y Jorge Alberto Suárez. Buenos Aires: EUDEBA.

JIMÉNEZ, HÉCTOR Y GUILLERMO MORALES. 2002. **"**Sepe: A POS Tagger for Spanish", en Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002*, Berlin: Springer, 250-259.

KAGEURA, KYO. 1999. "Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences", en *Journal of Quantitative Linguistics*, 6, 149–166.

KAY, MARTIN. 2003. "Introduction", en Ruslan Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, XVII-XX.

KLENK, URSULA Y HAGEN LANGER. 1989. "Morphological Segmentation Without a Lexicon". *Literary and Linguistic Computing*, 4, 4, 247–253.

LARA, LUIS FERNANDO. 2004. "¿Es posible una teoría de la palabra?", en *Lexis*, XXVII, 1-2, 401-427.

LARA, LUIS FERNANDO. 2006. *Curso de lexicología.* México: El Colegio de México.

LARA, LUIS FERNANDO, R. HAM CHANDE Y M. I. GARCÍA HIDALGO. 1979. *Investigaciones lingüísticas en lexicografía*, Jornadas, 89, México: El Colegio de México.

LÁZARO MORA, FERNANDO A. 1999. "La derivación apreciativa", en I. Bosque y V. Demonte (dirs.), *Gramática descriptiva de la lengua española*, v. 3, Madrid: Espasa-Calpe y RAE, 4645-4682.

LEECH, GEOFFREY Y NICHOLAS SMITH. 1999. "The use of tagging", en Hans Van Halteren (ed.). *Syntactic Wordclass Tagging*, Dordrecht, Netherlands: Kluwer Academic, 23-36.

LEECH, GEOFFREY Y ANDREW WILSON. 1999. "Standards for Tagsets", en Hans Van Halteren (ed.). *Syntactic Wordclass Tagging*, Dordrecht, Netherlands: Kluwer Academic, 55-80.

LOPE BLANCH, J. M. 1985. *El habla de Diego de Ordaz. Contribución a la historia del español americano*. Publicaciones del Centro de Lingüística Hispánica, 20. México: UNAM, IIF.

MANNING, CHRISTOPHER D. Y HINRICH SCHÜTZE. 1999. *Foundations of Statistical Natural Language Processing*, Cambridge, Mass.: The MIT Press.

MEDINA URREA, ALFONSO. 2000. "Automatic Discovery of Affixes by means of a Corpus: A Catalog of Spanish Affixes", en *Journal of Quantitative Linguistics*, 7: 2, 97–114.

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

MEDINA URREA, ALFONSO Y M. ALVARADO GARCÍA. 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 URREA, ALFONSO Y E. CRISTINA BUENROSTRO. 2003. "Características cuantitativas de la flexión verbal del chuj", en *Estudios de Lingüística Aplicada*, 38, 15-31.

MEDINA URREA, ALFONSO, J. A. HERRERA CAMACHO Y M. ALVARADO GARCÍA. 2009. "Towards the Speech Synthesis of Raramuri: A Unit Selection Approach based on Unsupervised Extraction of Suffix Sequences", en Alexander Gelbukh (ed.), *Advances in Computational Linguistics*, Research in Computing Science, 41, Berlín: Springer, 243-256.

MEDINA URREA, ALFONSO Y JAROSLAVA HLAVÁČOVÁ. 2005. "Automatic Recognition of Czech Derivational Prefixes", en Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, CICLing 2005, Berlin: Springer, 189-197.

MEDINA URREA, ALFONSO Y C. MÉNDEZ CRUZ. 2006. "Arquitectura del corpus histórico del Español de México (CHEM)", en A. Hernández y J. L. Zechinelli (eds.), *Avances en la ciencia de la computación (VII Encuentro Internacional de Computación ENC'06)*, México: Sociedad Mexicana de Ciencia de la Computación, 248-253.

MEYA, MONTSERRAT. 1986. "Morphologische Analyse des Spanischen", en Christoph Schwarz y Gregor Thurmair (eds.), *Informationslinguistische Texterschließung*, v. 4, Zürich: Georg Olms Verlag, 134–156.

MITKOV, RUSLAN (ed.). 2003. *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

MONACHINI, MONICA Y NICOLETTA CALZOLARI. 1999. "Standardization in the Lexicon", en H. van Halteren (ed.), *Syntactic Wordclass Tagging*, Dordrecht, Netherlands: Kluwer Academic, 149-174.

MORALES CARRASCO, RAÚL Y ALEXANDER GELBUKH. 2003. "Evaluation of TnT Tagger for Spanish", en *4th Mexican International Conference on Computer Science (ENC-2003)*, IEEE Computer Society Press, 18-25.

MORENO DE ALBA, JOSÉ G. 1986. *Morfología derivativa nominal en el español de México*. México: UNAM.

NIDA, EUGENE A. 1949/1946. *Morphology. The Descriptive Analysis of Words*. 2a. Ann Arbor: The University of Michigan.

OFLAZER, KEMAL. 1999. "Morphological Analysis", en H. van Halteren (ed.), *Syntactic Wordclass Tagging*, Dordrecht, Netherlands: Kluwer Academic, 175-205.

PENA, JESÚS. 1999. "Partes de la morfología. Las unidades del análisis morfológico", en I. Bosque y V. Demonte (dirs.), *Gramática descriptiva de la lengua española*, v. 3, Madrid: Espasa-Calpe y RAE, 4305-4366.

PENNY, RALPH. 2005/2002. *Gramática histórica del español.*, Traducido por J. I. Pérez Pascual y M. E. Pérez Pascual, 2ª ed., Barcelona: Ariel.

PLA, F., A. MOLINA Y N. PRIETO. 2001. "Evaluación de un etiquetador morfosintáctico basado en bigramas especializados para el castellano", en *Procesamiento del Lenguaje Natural*, 27, SEPLN, 215-221.

PORTER, M.F. 1980. "An Algorithm for Suffix Stripping", en *Program*, 14, 3, 130–137.

RAE. 1973. *Esbozo de una nueva gramática de la lengua española*. Madrid: Espasa Calpe.

SANTIAGO L., RAMÓN Y EUGENIO BUSTOS G. 1999. "La derivación nominal", en I. Bosque y V. Demonte (dirs.), *Gramática descriptiva de la lengua española*, v. 3, Madrid: Espasa-Calpe y RAE, 4505-4594.

SAPIR, EDWARD. 1954/1921. *El lenguaje. Introducción al estudio del habla*. Traducido por Margit y Antonio Alatorre. Breviarios, 96. México: FCE.

SHANNON CLAUDE L. Y WARREN WEAVER. 1964. *The mathematical theory of communication*. Chicago: The University of Illinois.

SHOPEN, TIMOTHY. 1985. *Language typology and syntactic description. Grammatical categories and the lexicon*. v. III. Cambridge: Cambridge University Press.

SPENCER, ANDREW. 1991. *Morphological theory: an introduction to word structure in generative grammar*. Cambridge: Cambridge University Press.

_____. 1998. “Morphophonological Operations”, en Andrew Spencer y Arnold M. Zwicky (eds.), *The handbook of morphology*, Oxford y Malden, Mass.: Blackwell, 123-143.

SPENCER, ANDREW Y ARNOLD M. ZWICKY (eds.). 1998. *The handbook of morphology*. Oxford y Malden, Mass.: Blackwell.

STUMP, GREGORY T. 1998. “Inflection”, en Andrew Spencer y Arnold M. Zwicky (eds.), *The handbook of morphology*, Oxford y Malden, Mass.: Blackwell, 13-43.

TRASK, R. L. 1999. “Part of Speech”, en K. Brown y J. Miller (eds.), *Concise encyclopedia of grammatical categories*, Oxford, UK: Elsevier Science, 278-284.

THURMAIR, GREGOR. 1986. “Ein Morphologisches Prozessorsegment zur Erzeugung von Grundformen mithilfe von Lernverfahren”, en Christoph Schwarz y Gregor Thurmair (eds.), *Informationslinguistische Texterschließung*, v. 4, Zürich: Georg Olms Verlag, 8–31.

TORRUELLA, J. Y LLISTERRI, J. 1999. "Diseño de corpus textuales y orales", en J.M. Blecua et al (eds.), *Filología e informática: Nuevas tecnologías en los estudios filológicos.* Barcelona: Editorial Milenio y Universidad Autónoma de Barcelona.

VAL ÁLVARO, JOSÉ F. 1999. “La composición”, en I. Bosque y V. Demonte (dirs.), *Gramática descriptiva de la lengua española*, v. 3, Madrid: Espasa-Calpe y RAE, 4759-44841.

VAN DER AUWERA. 1999. “Adverbs and Adverbials”, en K. Brown y J. Miller (eds.), *Concise encyclopedia of grammatical categories*, Oxford, UK: Elsevier Science, 8-11.

VAN HALTEREN, HANS (ed.). 1999. *Syntactic Wordclass Tagging*. Dordrecht, Netherlands: Kluwer Academic.

VAN HALTEREN, HANS Y ATRO VOUTILAINEN. 1999. “Automatic Taggers: An Introduction”, en H. van Halteren (ed.), *Syntactic Wordclass Tagging*, Dordrecht, Netherlands: Kluwer Academic, 109-115.

VARELA, SOLEDAD Y J. MARTÍN GARCÍA. 1999. “La prefijación”, en I. Bosque y V. Demonte (dirs.), *Gramática descriptiva de la lengua española*, v. 3, Madrid: Espasa-Calpe y RAE, 4993-5040.

VOGEL, PETRA M. Y BERNARD COMRIE. 2000. *Approaches to the Typology of Word Classes.* Berlín: Mouton de Gruyter.

VOUTILAINEN, ATRO. 1999a. “Orientation”, en Hans Van Halteren (ed.), *Syntactic Wordclass Tagging*, Dordrecht, Netherlands: Kluwer Academic, 3-7.

_____**.** 1999b. “A short history of tagging”, en Hans Van Halteren (ed.), *Syntactic Wordclass Tagging*, Dordrecht, Netherlands: Kluwer Academic, 9-21.

_____. 2003. “Part-of-Speech Tagging”, en Ruslan Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 219-232.

WEAVER, WARREN. 1964. “Recent Contributions to the Mathematical Theory of Communication”, en Claude L. Shannon y Warren Weaver, *The mathematical theory of communication*, Chicago: The University of Illinois, 3-28.

WEISS, SHOLOM, N. INDURKHYA, T. ZHANG Y F. DAMERAU. 2005. *Text mining. Predictive Methods for Analyzing Unstructured Information.* New York: Springer.

WIERZBICKA, ANNA. 2000. “Lexical prototypes as a universal basis for cross-linguistic identification of ‘part of speech’”, en P. M. Vogel y B. Comrie (eds.), *Approaches to the Typology of Word Classes*, Berlín: Mouton de Gruyter, 285-317.