

UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Nuevas fuentes de información para entrenamiento de etiquetadores gramaticales

Tesista: Fernando Jorge Rodriguez
Director: Dr. José Castaño

Buenos Aires, Marzo de 2012.



Índice

1. Introducción	3
1.1. Motivación	3
1.2. Trabajo realizado	4
1.3. Etiquetado gramatical	5
2. Definiciones y marco teórico	5
2.1. Etiquetas	6
2.2. Conjuntos de etiquetas	9
2.2.1. Especificidad de etiquetas: Treebank vs C5 y Brown	12
2.3. Corpus	13
2.3.1. Un poco de historia	15
2.3.2. Métodos	16
2.4. Etiquetadores gramaticales automáticos	16
2.4.1. Etiquetadores gramaticales basados en reglas	17
2.4.2. Etiquetadores gramaticales estocásticos	18
2.4.3. Etiquetadores gramaticales basado en HMM	18
2.5. Corpora de entrenamiento y corpora de verificación	21
2.6. Evaluación de etiquetadores gramaticales	22
2.7. Análisis de error	22
2.8. Palabras desconocidas	23
2.9. Etiquetador Gramatical TnT	24
2.9.1. Modelo teórico	25
2.9.2. Suavizado	25
2.9.3. Manejo de palabras desconocidas	26
3. Desarrollo	27
3.1. Diccionario COBUILD	27
3.1.1. Método de construcción	31
3.1.2. Evidencia	31
3.1.3. Un corpus	31
3.1.4. The Bank of English	31
3.1.5. La lista de palabras principales	32
3.1.6. Frecuencia	32
3.1.7. Ejemplos	33
3.1.8. Información gramatical	33
3.1.9. Pragmatismo	33
3.1.10. Definición del estilo	33
3.2. Traducción de etiquetas	34
3.3. Extracción de la información	36
3.4. Nuevo Corpus generado	38
3.5. Experimentación	41
3.6. Conclusiones	41

1. Introducción

1.1. Motivación

El etiquetado o anotado gramatical, también conocido como Part-of-speech tagging, POS tagging o simplemente POST, es el proceso de asignar una etiqueta gramatical a cada una de las palabras de un texto según su categoría léxica. Como se menciona más adelante, el etiquetado gramatical juega un papel importante en áreas de la lingüística computacional como por ejemplo síntesis del habla, reconocimiento del habla y recuperación de la información.

Este proceso es realizado manualmente por lingüistas o automáticamente mediante programas conocidos como etiquetadores gramaticales. Estos programas son entrenados a partir de un corpus anotado correctamente (denominado corpus de entrenamiento) del cual aprenden. De esta manera el etiquetador obtiene, procesa y retiene información sobre como está anotado el corpus de entrenamiento, información que utiliza posteriormente durante el proceso de etiquetado gramatical.

Una vez finalizado el entrenamiento, se procede a etiquetar el texto requerido utilizando la información adquirida.

Basicamente un corpus etiquetado o anotado es una lista de palabras con su correspondiente etiqueta gramatical. Por ejemplo:

A DT
form NN
of IN
asbestos NN
once RB
used VBN
to TO
make VB
Kent NNP
cigarette NN
filters NNS
has VBZ
caused VBN
a DT
high JJ
percentage NN
of IN
cancer NN
deaths NNS
among IN
a DT
group NN
of IN
workers NNS
exposed VBN
to TO
it PRP
more RBR

than IN
 30 CD
 years NNS
 ago RB
 , ,
 researchers NNS
 reported VBD
 . .

La complejidad del etiquetado gramatical reside en que la etiqueta no depende solo de la palabra o tipo de palabra que se está etiquetando, por el contrario, la ubicación y el contexto de la palabra en donde ésta aparece es un factor determinante al asignar la etiqueta correspondiente.

Como fué mencionado anteriormente, la etapa de entrenamiento de un etiquetador gramatical obtiene, procesa y retiene información sobre el contexto y la ubicación de una palabra y la etiqueta asignada. Luego se utiliza esta información en el proceso de etiquetación, donde el etiquetador analiza para cada palabra su ubicación y contexto, y en base a ello y al conocimiento adquirido previamente con el corpus de entrenamiento determina una etiqueta gramatical.

Uno de los grandes problemas del etiquetado gramatical reside en la falta de corpus anotados para utilizar durante el entrenamiento.

Los corpus de entrenamiento son etiquetados manualmente por lingüistas especializados. Es un trabajo profundamente meticuloso y tedioso ya que el lingüista debe dar una etiqueta gramatical palabra por palabra en corpus del orden del MILLON de palabras. Además de la complejidad del trabajo, el tiempo empleado para etiquetar un corpus es sumamente extenso y como consecuencia el valor económico es significativo, ya que intervienen grupos de trabajo altamente capacitados durante un tiempo prolongado. El resultado de este complejo proceso artesanal es una tabla de palabras con su correspondiente etiqueta gramatical como se mostró anteriormente.

Ante la importancia que adquieren los corpus etiquetados es inevitable pensar en algún otro tipo de texto que posea información de etiquetas. Por ejemplo algunos diccionarios contienen una palabra, su definición y algunos ejemplos en donde ésta aparece con cada uno de sus sentidos. Es decir que de alguna manera un diccionario contiene por cada palabra uno o más contextos en donde ésta aparece etiquetada. Entonces si tomamos todos los ejemplos de cada palabra de un diccionario y su etiqueta podemos construir un corpus parcialmente anotado.

1.2. Trabajo realizado

La idea de este trabajo es suplir la falta de corpus de entrenamiento utilizando la información de etiquetado que posee un diccionario generando una nueva fuente de información para entrenar etiquetadores automáticos. Se utiliza esta nueva fuente de información para etiquetar, medir y comparar resultados en el rendimiento conseguido. Se combina e integra esta información obtenida con corpus de entrenamiento clásicos e inspeccionan resultados en el etiquetado. Este trabajo menciona detalladamente la forma de extraer la información relevante sobre etiquetas gramaticales a partir de un diccionario y las decisiones que fueron aplicadas. Por último se realizan mediciones sobre el rendimiento de

los etiquetadores gramaticales entrenados con esta nueva fuente de información y los corpus clásicos de entrenamiento y se presentan las conclusiones.

1.3. Etiquetado gramatical

Como se mencionó anteriormente, el etiquetado gramatical, también conocido como Part-of-speech tagging, POS tagging o simplemente POST, es el proceso de asignar una etiqueta a cada una de las palabras de un texto según su categoría léxica. Este proceso se realiza en base a la definición de la palabra y la de sus palabras vecinas, es decir, el contexto en que ésta aparece. Por ejemplo en *Does that flight serve dinner*, *dinner* es un sustantivo y por lo tanto recibe la etiqueta para sustantivos NN.

El etiquetado gramatical brinda una gran cantidad de información sobre una palabra y sus vecinas. Por ejemplo, las etiquetas distinguen entre pronombres posesivos (mi, tu, su, etc.) y pronombres personales (Yo, Tú, Él, etc.). Saber si una palabra es un pronombre posesivo o personal nos brinda información sobre las palabras que pueden ocurrir a continuación: los pronombres posesivos generalmente son sucedidos por un sustantivo (como en *Mi comida*) mientras que los personales son sucedidos por un verbo (como en *Yo duermo*).

Utilizando esta deducción podemos aseverar que si una palabra fué etiquetada como pronombre personal, es muy probable que la próxima palabra sea un verbo. Este conocimiento puede ser de útil aplicación en modelos lingüísticos para reconocimiento del habla (voz a texto). Pero esta no es la única información que una etiqueta gramatical nos puede ofrecer.

Una etiqueta gramatical también nos puede acercar información relacionada con la pronunciación de la palabra. En inglés la palabra *content* puede ser un sustantivo o un adjetivo y su pronunciación varía dependiendo de este hecho. Utilizando estas ideas podemos producir pronunciaciones más naturales en un sistema de síntesis del habla (texto a voz) o también podemos obtener más exactitud en un sistema de reconocimiento del habla (voz a texto).

Otra aplicación importante del etiquetado gramatical en sistemas de recuperación de la información es el reconocimiento de sustantivos u otro tipo de palabras importantes dentro de un documento, para guardar y utilizar esta información en búsquedas posteriores.

Por último, la asignación automática de etiquetas gramaticales juega un papel importante en algoritmos de desambiguación del sentido de la palabra y en modelos lingüísticos basados en n-gramas utilizados en sistemas de reconocimiento del habla.

2. Definiciones y marco teórico

A continuación se presentan definiciones y teorías que ayudan a comprender el trabajo realizado. Se presenta el concepto de etiqueta gramatical, es decir, una etiqueta que identifica el rol que cumple una palabra dentro de cierto contexto. Se muestran los tipos de etiquetas que han sido utilizados intentando abarcar los distintos significados que pueden tener las palabras. Hasta el día de hoy no se ha llegado a un consenso sobre un conjunto de etiquetas adecuado y se siguen explorando distintas alternativas. Se explica el concepto de etiquetado gramatical, es decir, la tarea de asignar a cada palabra una etiqueta gramatical

adecuada según el contexto en donde ésta aparece. Se muestran ejemplos de que esta tarea está muy lejos de ser trivial, introduciendo el concepto de ambigüedad gramatical. Esto ocurre cuando una palabra puede tener muchos significados (y por lo tanto distintas etiquetas gramaticales) dependiendo del contexto en dónde aparece.

Se exhibe la importancia del etiquetado gramatical dentro de distintas áreas como la computación lingüística, reconocimiento y síntesis del habla. Se muestra como se maneja este proceso utilizando programas que lo realizan automáticamente, es decir, etiquetadores gramaticales automáticos. Se explica en profundidad como funcionan estos etiquetadores gramaticales automáticos, mostrando como las implementaciones actuales utilizan un proceso de entrenamiento. Este proceso ocurre a partir de un corpus previamente anotado que el etiquetador automático toma como ejemplo para reproducir el etiquetado.

Se presenta el concepto de corpus y corpus anotados gramaticalmente como conjuntos de información extremadamente valiosos para todas estas tareas. Se muestra la forma de medir, evaluar y comparar el rendimiento de los etiquetadores gramaticales, introduciendo los conceptos de corpus de entrenamiento y corpus de verificación. Se muestran técnicas de análisis de error para el proceso de etiquetación automática. Se exhibe también el manejo de ciertos casos especiales dentro del proceso de etiquetación automática; las palabras desconocidas. Y por último se explican en detalle los etiquetadores automáticos utilizados en el presente trabajo.

2.1. Etiquetas

Tradicionalmente la definición de POS o etiqueta gramatical se ha basado en funciones sintácticas y morfológicas, es decir que se agrupan en clases las palabras que funcionan similarmente con respecto a lo que puede ocurrir a su alrededor (sus propiedades de distribución sintáctica) o con respecto a los afijos que poseen (sus propiedades morfológicas). Mientras que las clases de palabras tienen tendencia hacia la coherencia semántica (por ejemplo los sustantivos generalmente describen gente, lugares o cosas y los adjetivos generalmente describen propiedades), este no es necesariamente el caso y en general no se utiliza coherencia semántica como criterio para la definición de POS o etiqueta gramatical.

Las etiquetas gramaticales pueden ser divididas en dos grandes categorías: clases cerradas y clases abiertas. Las clases cerradas son aquellas que tienen miembros relativamente fijos. Por ejemplo, las preposiciones son una clase cerrada porque hay un conjunto cerrado de ellas, es decir que son un grupo de palabras que raramente varía ya que raramente se agregan nuevas preposiciones. En contraste, los sustantivos y los verbos son clases abiertas ya que continuamente se introducen y eliminan nuevos verbos y sustantivos al lenguaje. Es probable que cualquier hablante o corpus tenga una clase abierta de palabras diferente, pero todos los hablantes de un lenguaje y corpora suficientemente grandes, seguramente van a compartir el conjunto de clases de palabras cerradas. Las clases de palabras cerradas también son generalmente palabras funcionales como *de*, *y* o *tu*, que tienden a ser muy cortas, ocurrir frecuentemente y generalmente tienen usos estructurales en gramática.

Hay cuatro clases abiertas principales:

- **Sustantivos** Es el nombre dado a la clase sintáctica que denota personas, lugares o cosas. Pero desde que las clases sintácticas como sustantivos son definidas sintáctica y morfológicamente en vez que semánticamente, algunas palabras para personas, lugares y cosas pueden no ser sustantivos y a la inversa, algunos sustantivos pueden no ser palabras para personas, lugares o cosas. Por lo tanto los sustantivos incluyen términos concretos como *barco* y *silla*, abstracciones como *banda ancha* y *relación*. Se puede definir a una palabra como sustantivo basándose en características como la capacidad de ocurrir con determinantes (una *cabra*, su *banda ancha*), tomar posesivos (los ingresos anuales de *IBM*) y para la mayoría pero no todos los sustantivos, ocurrir en la forma plural (*cabras*, *teléfonos*). Los sustantivos tradicionalmente son agrupados en sustantivos propios y sustantivos comunes.

- **Sustantivos propios:** Son nombres de personas específicas o entidades y usualmente son escritos en mayúscula.
- **Sustantivos comunes:** En algunos lenguajes se dividen en sustantivos contables e incontables.
 - **Sustantivos contables:** Son aquellos que permiten establecer su número en unidades. En general esta clase posee forma singular y plural (*silla/s*, *dedo/s*).
 - **Sustantivos incontables:** Se refieren a sustantivos que no se puede determinar su número en unidades (*harina*, *nieve*, *azúcar*).

- **Verbos:** Los verbos son una clase de palabras que incluye a la mayoría de las palabras referidas a acciones y procesos. Tienen ciertas formas morfológicas como tiempo, modo, persona, regularidad, etc. Además, el verbo puede concordar en género, persona y número con algunos de sus argumentos o complementos (a los que normalmente se conoce como sujeto, objeto, etc.). En español concuerda con el sujeto siempre en número y casi siempre en persona (la excepción es el caso del llamado sujeto inclusivo: *Los españoles somos así*).

Algunos ejemplos:

Marisol *canta* una ópera.

La comida *está* caliente.

- **Adjetivos:** Las palabras pertenecientes a esta clase expresan propiedades o cualidades. Por ejemplo *Ese hombre es **alto***. Los adjetivos tienen género y número al igual que los sustantivos. El género y el número de los adjetivos depende del sustantivo al que acompañan. Hay adjetivos que presentan una sola forma para el masculino y para el femenino. Son adjetivos de una sola terminación (*verde*, *especial*, *amable*, *grande*, etc.). Por el otro lado, los adjetivos de dos terminaciones presentan distintas formas para el masculino y el femenino (*feo-fea*, *pequeño-pequeña*, *blanco-blanca*, etc.) Se clasifican en:

- **Determinativos:** Preceden al sustantivo, lo concretan y lo presentan
 - **Demostrativos:** *Esta* niña

- **Posesivos:** *Mi* niña
- **Numerales:** *Tres* niñas
- **Indefinidos:** *Algunas* niñas
- **Exclamativos:** ¡*Qué* niña!
- **Interrogativos:** ¿*Qué* niña?
- **Calificativos:** Califican al sustantivo, es decir, añaden cualidades al sustantivo. Los adjetivos calificativos se dividen en especificativos y explicativos o epítetos.
 - **Adjetivos calificativos especificativos:** Son aquellos que concretan el significado del sustantivo. Suelen aparecer detrás del sustantivo.
Ej: Quiero una corbata *azul*.
 - **Adjetivos calificativos explicativos o epítetos:** indican cualidades que ya de por sí lleva el sustantivo. Suelen ir delante del sustantivo.
Ej: *Blanca* nieve, *Verde* hierba.
- **Adverbios:** Los adverbios son otro ejemplo de clase abierta de palabras: se definen como modificadores del verbo, adjetivo o de otro adverbio. Tradicionalmente se dividen en:
 - **Adverbios de lugar:** aquí, allí, ahí, allá, acá, arriba, abajo, cerca, lejos, delante, detrás, encima, debajo, enfrente, atrás, alrededor, etc.
 - **Adverbios de tiempo absoluto:** pronto, tarde, temprano, todavía, aún, ya, ayer, hoy, mañana, siempre, nunca, jamás, próximamente, prontamente, anoche, enseguida, ahora, mientras.
 - **Adverbios de modo:** bien, mal, regular, despacio, deprisa, así, tal, como, aprisa, adrede, peor, mejor, fielmente, estupendamente, fácilmente - todas las que se formen con las terminaciones "mente".
 - **Adverbios de cantidad o grado:** muy, poco, muy poco, mucho, bastante, más, menos, algo, demasiado, casi, sólo, solamente, tan, tanto, todo, nada, aproximadamente.

Por otro lado tenemos las clases cerradas de palabras que detallamos a continuación:

- **Preposiciones:** Las preposiciones son enlaces que relacionan los componentes de una oración para brindarles sentido. La unión se lleva a cabo con una o varias palabras. La significación que dan las preposiciones responde a circunstancias de movimiento, lugar, tiempo, modo, causa, posesión, pertenencia, materia y procedencia.
Algunos ejemplos:

Me levanté de la cama a las ocho de la mañana.

*Dejé mis cuadernos **sobre** el sillón.*

*Corrí apresurado **hacia** la calle pero no logré divisarte.*

*Lucía se divierte **con** sus muñecas.*

- **Determinantes:** Los determinantes son clases cerradas de palabras que ocurren con sustantivos, generalmente marcando el principio de una frase sustantiva. Un pequeño subtipo de determinantes es el artículo: *a, el*. Otros determinantes incluyen *ese* (como en *el libro ese*).
- **Pronombres:** Los pronombres son formas que generalmente actúan como una clase de atajo para referirse a alguna frase sustantiva, entidad o evento. Se dividen en:
 - **Pronombres personales:** Hacen referencia a personas o entidades (Yo, tú, él, ella, nosotros, ellos, etc.)
 - **Pronombres posesivos:** Son formas de pronombres personales que indican una posesión actual o mas generalmente solo una relacion abstracta entre la persona y algun objeto (mío, tuyo, suyo, mi, nuestro, etc.)
- **Conjunciones:** Las conjunciones son utilizadas para unir dos frases, cláusulas o sentencias. Las conjunciones coordinantes como *y, o* unen dos elementos de igual estado. Las conjunciones subordinativas son utilizadas cuando uno de los elementos es de algún tipo de estado integrado. Por ejemplo *Me molestó **que** no me lo dijeras*.
- **Verbos auxiliares:** Los verbos auxiliares son verbos que proporcionan información gramatical y semántica adicional a un verbo de significado completo. Dichos verbos auxiliares brindan la información gramatical de modo, tiempo, persona y número y las formas no personales. Por ejemplo *¿por qué no **has** llegado a la hora prevista?* o también *La avenida principal de la ciudad **fue** clausurada por obras de refacción*.
- **Numerales:** Los determinantes numerales o simplemente numerales son los que expresan de modo preciso y exacto la cantidad de objetos designados por el sustantivo al que acompañan, delimitan o designan. Limitan el significado general del sustantivo, precisando con exactitud la cantidad de objetos que aquel designa o el lugar de orden que ocupan. Los numerales pueden ser de varias clases. Los más importantes son:
 - **Cardinales:** informan una cantidad exacta:
Quiero *cuatro* libros.
 - **Ordinales:** informan del orden de colocación:
Quiero el *cuarto* libro.
 - **Fraccionarios:** informan de particiones de la unidad:
Quiero la *cuarta* parte.
 - **Multiplicativos:** informan de múltiplos:
Quiero *dobles* raciones.

2.2. Conjuntos de etiquetas

La sección anterior dió una descripción general de los tipos de clases sintácticas a las que pertenecen las palabras. Esta sección presenta los conjuntos de etiquetas actuales utilizados en la etiquetación gramatical. Es decir, las etiquetas que se corresponden con cada una de estas clases sintácticas. Todavía no

existe un consenso sobre el conjunto de etiquetas o tagset más adecuado. Generalmente los conjuntos de etiquetas grandes ofrecen una descripción sintáctica más específica mientras que los conjuntos de etiquetas más pequeños usualmente brindan una información lingüística más acotada. Una de las características clave para decidir que conjunto de etiquetas es el más adecuado justamente depende del nivel de detalle lingüístico que se esté buscando. Otro hecho notable referido a los conjuntos de etiquetas es que los más pequeños generalmente están contenidos en los conjuntos mayores. Ya que las etiquetas más específicas que se encuentran en los conjuntos mayores pueden ser convertidas en etiquetas de menor especificidad con la consecuente pérdida de detalle lingüístico. Por el otro lado, también se pueden convertir las etiquetas pertenecientes a un conjunto pequeños a conjuntos mayores que poseen etiquetas de mayor especificidad, ya que generalmente existen etiquetas equivalentes en los conjuntos de mayor tamaño.

Etiqueta	Descripción	Ejemplo
CC	Coordinating conjunction	<i>and</i>
CD	Cardinal number	<i>1, third</i>
DT	Determiner	<i>the</i>
EX	Existential	<i>there there is</i>
FW	Foreign word	<i>d'hoevre</i>
IN	Preposition/subordinating conjunction	<i>in, of, like</i>
JJ	Adjective	<i>green</i>
JJR	Adjective, comparative	<i>greener</i>
JJS	Adjective, superlative	<i>greenest</i>
LS	List marker	<i>1)</i>
MD	Modal	<i>could, will</i>
NN	Noun, singular or mass	<i>table</i>
NNS	Noun plural	<i>tables</i>
NNP	Proper noun, singular	<i>John</i>
NNPS	Proper noun, plural	<i>Vikings</i>
PDT	Predeterminer both	<i>the boys</i>
POS	Possessive ending	<i>friend's</i>
PRP	Personal pronoun	<i>I, he, it</i>
PRP\$	Possessive pronoun	<i>my, his</i>
RB	Adverb	<i>however, usually, naturally, here, good</i>
RBR	Adverb, comparative	<i>better</i>
RBS	Adverb, superlative	<i>best</i>
RP	Particle	<i>give up</i>
SYM	Symbol	<i>+, %, &</i>
TO	To	<i>to go, to him</i>
UH	Interjection	<i>uhhuhhuhh</i>
VB	Verb, base form	<i>take</i>
VBD	Verb, past tense	<i>took</i>
VBG	Verb, gerund/present participle	<i>taking</i>
VCN	Verb, past participle	<i>taken</i>
VBP	Verb, sing. present, non-3d	<i>take</i>
VBZ	Verb, 3rd person sing. present	<i>takes</i>
WDT	Wh-determiner	<i>which</i>
WP	Wh-pronoun	<i>who, what</i>
WP\$	Possessive wh-pronoun	<i>whose</i>
WRB	Wh-abverb	<i>where, when</i>
\$	Dollar sign	<i>\$</i>
#	Pound sign	<i>#</i>
"	Left quote	<i>(' or ")</i>
"	Right quote	<i>(' or ")</i>
(Left parenthesis	<i>([, (, {, i)</i>
)	Right parenthesis	<i>(],), }, i)</i>
,	Comma	<i>,</i>
.	Sentence-final punc	<i>(. ! ?)</i>
:	Mid-sentence punc	<i>(: ; ... -)</i>

Más allá de que no exista aún un consenso sobre cuál conjunto de etiquetas utilizar, hay un pequeño número de conjuntos de etiquetas o tagsets populares

para el idioma inglés, muchos de los cuales evolucionaron a partir del conjunto de etiquetas utilizado para etiquetar el corpus Brown. Este conjunto de etiquetas se conoció como el Brown Corpus Tag-set, un conjunto de 87 etiquetas que se utilizó para etiquetar el corpus Brown: un corpus de 1 millón de palabras construido a partir de ejemplos provenientes de 500 textos de diferentes géneros (diarios, novelas, no ficción, académico, etc.) que fué ensamblado en la Universidad Brown entre 1963 y 1964. Este corpus fué etiquetado gramaticalmente aplicando en primera instancia un etiquetador automático, el programa TAGGIT, y luego corregido manualmente.

Al lado del conjunto de etiquetas Brown se encuentran dos de los conjuntos de etiquetas más utilizados: el conjunto de etiquetas reducido Pen Treebank de 45 etiquetas y el conjunto de etiquetas CLAWS C5 de tamaño medio con 62 etiquetas que fué utilizado para etiquetar el corpus British National Corpus (BNC).

El conjunto de etiquetas Penn Treebank mostrado anteriormente también fué utilizado para etiquetar el corpus Brown, el corpus Wall Street Journal y el corpus Switchboard entre otros. En realidad, quizás en parte por su pequeño tamaño es uno de los conjuntos de etiquetas más utilizado. A continuación se exhiben algunos ejemplos de oraciones del corpus Brown etiquetadas con el conjunto de etiquetas Penn Treebank. Representaremos una palabra etiquetada mediante la colocación de una barra oblicua seguida de su etiqueta:

1. The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.
2. **There/EX** are/VBP 70/CD children/NNS **there/RB**
3. Although/IN preliminary/JJ findings/NNS were/VBD **reported/VBN** more/RBR than/IN a/DT year/NN ago/IN ./, the/DT latest/JJS results/NNS appear/VBP in/IN today/NN 's/**POS** New/NNP England/NNP Journal/NNP of/IN Medicine/NNP ./.

El primer ejemplo exhibe los determinantes *the* y *a*, los adjetivos *grand* y *other*, los sustantivos comunes *jury*, *number* y *topics* y el verbo en tiempo pasado *commented*. El segundo ejemplo muestra el uso de la etiqueta ET para marcar la construcción existencial *there* y otro uso de *there* que es etiquetado como un adverbio (RB). El tercer ejemplo muestra la segmentación del morfema posesivo 's y un ejemplo de la construcción pasiva 'were reported', en la cual el verbo *reported* está marcado como un pasado participio (VBN) en vez de como un pasado simple (VBD). También es interesante notar que el sustantivo propio *New England* está etiquetado como NNP. Finalmente, se puede observar que como *New England Journal of Medicine* es un sustantivo propio, el etiquetado de Treebank elige marcar cada sustantivo separado como NNP, incluyendo *journal* y *medicine*, que en otros serían etiquetados como sustantivos comunes (NN).

2.2.1. Especificidad de etiquetas: Treebank vs C5 y Brown

El conjunto de etiquetas del Penn Treebank fué seleccionado a partir del conjunto de etiquetas original del corpus Brown (de 87 etiquetas). Este conjunto reducido deja afuera información que puede ser recuperada desde la identidad del ítem léxico. Por ejemplo los conjuntos de etiquetas Brown original y C5

incluyen una etiqueta separada para cada una de las diferentes formas de los verbos *do* (C5 propone la etiqueta VDD para *did* y VDG para *doing*), *be* y *have*. Estas etiquetas fueron omitidas en el conjunto de Treebank.

Ciertas distinciones sintácticas no fueron marcadas en el conjunto de etiquetas Penn Treebank porque las oraciones de Treebank fueron parseadas, no meramente etiquetadas, y por lo tanto alguna información sintáctica es representada en la estructura de la frase. Por ejemplo, la etiqueta simple IN es utilizada para preposiciones como para conjunciones subordinadas ya que la estructura de árbol de la oración las desambigua (las conjunciones subordinadas siempre preceden cláusulas, las preposiciones preceden frases nominales o frases preposicionales). La mayoría de las situaciones de etiquetado, sin embargo, no involucran corpora parseado. Por esta razón el conjunto del Penn Treebank no es suficientemente específico en ciertos casos. Los conjuntos de etiquetas de Brown original y C5, por ejemplo, distinguen preposiciones (IN) de conjunciones subordinadas (CS), como en los siguiente ejemplos:

1. **after/CS** spending/VBG a/AT few/AP days/NNS at/IN the/AT Brown/NP Palace/NN Hotel/NN
2. **after/IN** a/AT wedding/NN trip/NN to/IN Corpus/NP Christi/NP ./.

También tienen dos etiquetas para la palabra *to*; en Brown el uso del infinitivo es etiquetado como TO, mientras que las preposiciones utilizan IN:

1. **to/TO** give/VB priority/NN **to/IN** teacher/NN pay/NN raises/NNS

Brown también tiene la etiqueta NR para sustantivos adverbiales como *home*, *west*, *Monday* y *tomorrow*. Como el Treebank carece de esta etiqueta, hay una política mucho menos consciente para sustantivos adverbiales; *Monday*, *Tuesday* y otros días de la semana son marcados como NNP, *tomorrow*, *west* y *home* son marcados algunas veces como NN y algunas veces como RB. Esto hace el conjunto de etiquetas del Treebank menos útil para tareas de alto nivel lingüístico como la detección del tiempo de frases. Sin embargo, el conjunto de etiquetas de Treebank ha sido el más utilizado para la evaluación de algoritmos de etiquetación automática. Esta es la razón por la cual elegimos este conjunto de etiquetas para utilizar en el desarrollo del presente trabajo.

2.3. Corpus

Un corpus es una colección de textos escritos o transcripciones del lenguaje oral para cierto idioma. Para construir un corpus se reúne una cantidad considerable de textos escritos y transcripciones orales para luego ser preservado en algún formato (generalmente electrónico). Generalmente el tamaño de un corpus es superior al millón de palabras. La palabra corpus viene del latín..., su plural es corpora. El corpus se utiliza para el estudio del lenguaje genuino, como se expresa cotidianamente.

Diferentes tipos de corpora permiten el análisis de distintas clases de discursos para hallar evidencia cuantitativa sobre la existencia de patrones en el lenguaje o para verificar teorías. Los primeros estudios sobre un corpus se enfocaron en palabras; su frecuencia y ocurrencia. Con el desarrollo de la tecnología y de motores de búsqueda más precisos y eficientes, las posibilidades crecieron ampliamente. Hoy en día es posible realizar búsquedas para una palabra

perteneciente a cierta clase sintáctica o patrones completos como por ejemplo: preposición + sustantivo o determinante + sustantivo, o una palabra particular + clase de palabra específica sucediéndola. Estas búsquedas facilitan, por ejemplo, a los publicadores de diccionarios para hallar colocaciones.

La lingüística también es aplicada sobre el corpus para realizar estudios de traducción donde utilizando corpus de las 2 lenguas se hace evidente el significado de las palabras y las supuestas equivalencias que podrían diferir en el uso específico. Por otra parte algunos aspectos gramaticales fuertemente conectados a la léxica permiten a los lingüistas mostrar diferencias en el uso de ciertas estructuras gramaticales en las traducciones. En el caso de inglés, también las diferencias entre sus variedades americano y británico pueden ser fácilmente analizadas gracias al corpora.

El cambio histórico de los significados de las palabras y la gramática es analizado como consecuencia del desarrollo del corpora. Y aunque la cantidad de viejos textos disponibles en formato electrónico es mucho más pequeña que la cantidad de textos contemporáneos, el trabajo es factible. Por lo tanto, las diferencias en los aspectos gramaticales concernientes a la voz pasiva fueron trazados y resulta que con el siglo 19 la voz pasiva en el idioma inglés comienza a ser utilizada más y más frecuentemente.

Cuando corpora escrito y hablado se hizo disponible, los lingüistas comenzaron a analizarlo para verificar si había patrones o diferencias entre lo hablado y lo escrito. Parece que aparte de algunas características obvias como salidas en falso y vacilaciones que se producen en el habla pero no en la escritura, la utilización de un gran número de expresiones deícticas es más frecuente en los discursos orales. Probablemente esto es debido a los signos lingüísticos extra en los que el lenguaje hablado es más vago. Adicionalmente ciertas características gramaticales manifestadas en el habla deben ser consideradas agramaticales en la escritura.

Los lingüistas utilizan una metodología que intenta describir naturalmente las ocurrencias del lenguaje apoyadas en sus observaciones de gran cantidad de evidencia encontrada en el corpora. Más aún, las operaciones estadísticas están generalmente involucradas en el trabajo sobre el corpora, especialmente cuando las frecuencias de uso de algún aspecto léxico son medidas. Grandes bases de datos de ocurrencia natural del lenguaje han ayudado a realizar progresos en el estudio de fraseología, especialmente cuando se ha descubierto que ciertos significados de las palabras se correlacionan con las estructuras gramaticales en donde son utilizadas. La lingüística de corpus ha encontrado aplicaciones en muchos campos como: análisis crítico del discurso, estilismos, lingüística forense así como traducciones y enseñanza del lenguaje.

En las traducciones es útil utilizar corpora paralelo que permite una mejor elección de equivalencias y estructuras gramaticales que podrían reflejar el significado deseado. Estudios adicionales sobre corpora revelaron que los traductores no traducen palabra por palabra sino unidades más grandes (cláusulas o sentencias).

Los estudios de corpora probablemente han tenido una gran influencia en la enseñanza del lenguaje. Primero que nada, han influido en la forma en que se hacen los diccionarios. Segundo los aprendices del lenguaje han sido estudiados para mejorar el conocimiento de los maestros. Y hoy en día los aprendices son alentados a hacer uso de corpora por sí mismos para incrementar su conocimiento. Inclusive la información de los resultados reunidos de corpora han

influenciado en el diseño y contenido de los libros de aprendizaje.

Los lingüistas de corpus creen que un análisis confiable del lenguaje ocurre mejor en ejemplos recolectados de campo; contextos naturales y con interferencia experimental mínima. Dentro del corpus lingüístico existen visiones divergentes en torno al nivel de las anotaciones. Desde John Sinclair abogando anotaciones mínimas y permitiendo a los textos «hablar por ellos mismos» a otros como el equipo de Survey of English Usage (University College, London) abogando anotaciones como un camino hacia un riguroso entendimiento lingüístico.

2.3.1. Un poco de historia

El punto de inflexión en corpus lingüístico moderno fué la publicación de Henry Kucera y W. Nelson Francis: *Computational Analysis of Present-Day American English* en 1967. Un trabajo basado en el análisis del corpus Brown, una compilación cuidadosamente seleccionada de inglés americano actual totalizando alrededor de 1 millón de palabras obtenidas de una amplia variedad de fuentes. Kucera y Francis sometieron este corpus a una gran variedad de análisis computacional desde el cual compilaron un rico y nutrido corpus combinando elementos de lingüística, enseñanza de lenguaje, psicología, estadística y sociología. Una publicación clave adicional fué «Towards a description of English Usage» de Randolph Quirk (1960) en la que introdujo Survey of English Usage.

Poco después el publicador de Boston Houghton-Mifflin se acercó a Kucera para suministrarle la citación base de 1 millón de palabras para su nuevo diccionario *American Heritage Dictionary* (AHD), el primer diccionario que fué compilado utilizando corpus lingüístico. El AHD dió el paso innovador de combinar elementos prescriptivos (como debe utilizarse el lenguaje) con información descriptiva (como se usa actualmente).

Otros publicadores siguieron el ejemplo. El publicador inglés Collins creó y compiló el diccionario *Cobuild* utilizando el corpus *Bank of English*. Fué diseñado para usuarios que están aprendiendo inglés como lengua extranjera.

El corpus Brown también dió lugar a un número de corpora similarmente estructurada: el corpus LOB (1960, inglés británico), Kolhapur (inglés indio), Wellington (inglés de Nueva Zelanda), *Australian Corpus of English* (inglés australiano) y el *Flob corpus* (1990, inglés británico).

Otro corpora representa muchos lenguajes, variedades y modos e incluye el *International Corpus of English*, el *British National Corpus*: una colección de 100 millones de palabras provenientes de textos escritos e inglés hablado, creado en los 1990s por un consorcio de publicadores, universidades (Oxford y Lancaster) y la *British Library*. Para inglés americano contemporáneo, el trabajo se ha centrado en el *American National Corpus* (más de 400 millones de palabras de inglés americano contemporáneo).

El primer corpus computarizado de lenguaje hablado transcripto fué construido en 1971 por el *Montreal French Project*, conteniendo 1 millón de palabras que inspiró a Shana Poplack a crear un corpus mucho más grande de Francés hablado.

Al lado de estos corpora de lenguaje vivo se encuentra corpora computarizado que también fué construido a partir de colecciones de textos en lenguajes antiguos. Como ejemplo tenemos la base de datos *Andersen-Forbes* de la biblia hebrea, desarrollada desde los años 1970, en donde cada cláusula es parseada utilizando grados que representan más de 7 niveles de sintaxis y cada segmen-

to es etiquetado con 7 campos de información. El Quarc Arabic Corpus es un corpus anotado para el lenguaje árabe clásico del corán. Este es un proyecto reciente con múltiples capas de anotación incluyendo segmentación morfológica, etiquetado gramatical y análisis sintáctico utilizando dependencia gramatical.

2.3.2. Métodos

Corpus lingüísticos han generado una cantidad de métodos de investigación intentando trazar un camino desde los datos hacia la teoría. Wallib y Nelson (2001) introdujeron lo que ellos llamaron la perspectiva 3A: anotación, abstracción y análisis.

La anotación consiste en la aplicación de un esquema a los textos. Las anotaciones incluyen marcado estructural, etiquetado gramatical, parsing y varias representaciones más. La abstracción consiste en la traducción (mapeo) de términos del esquema a términos en el modelo teórico. Típicamente incluye búsqueda lingüística directa y también puede incluir aprendizaje por reglas para parsers. El análisis consiste de exploración estadística, manipulación y generalización desde los datos. También podría incluir evaluaciones estadísticas, optimización basada en reglas o métodos de descubrimiento del conocimiento.

La mayoría del corpora léxico de hoy en día está anotado gramaticalmente. Sin embargo, incluso los corpus lingüísticos que trabajan con texto plano no anotado inevitablemente aplican algún método para aislar términos que pueden ser interesantes en las palabras circundantes.

2.4. Etiquetadores gramaticales automáticos

Como se mencionó anteriormente, el etiquetado gramatical es el proceso de asignar una etiqueta gramatical a cada palabra dentro de un texto. Generalmente las etiquetas gramaticales también son aplicadas a los signos de puntuación, por lo tanto el etiquetado requiere que los signos de puntuación sean separados de las palabras. Este proceso se realiza previamente o como parte del etiquetado y es conocido como *tokenización*; es el proceso encargado de separar puntos, comas, paréntesis y otros caracteres de las palabras así como también desambiguar el fin de oración (por ejemplo un punto o signo de pregunta) de un signo de puntuación (como en una abreviación por ejemplo *étc.*)

La entrada para un algoritmo de etiquetación automática es una cadena de palabras y un conjunto de etiquetas. La salida es la mejor etiqueta encontrada para cada palabra. Consideremos las siguientes oraciones etiquetadas gramaticalmente:

Book/VB that/DT flight/NN ./.

Does/VBZ that/DT flight/NN serve/VB dinner/NN ?/.

Asignar una etiqueta gramatical a una palabra no es una tarea trivial incluso en estos sencillos ejemplos. Por ejemplo, la palabra *book* es ambigua. Es decir que tiene más de un uso posible y por lo tanto más de una etiqueta gramatical posible. Puede ser un verbo (como en **book** *that flight* o *to book the suspect*) o un sustantivo (como en *hand me that book* o *a book of matches*). Análogamente *that* puede ser un determinante (como en *Does that flight serve dinner*) o un complementador (como en *I thought that your flight was earlier*).

El problema del etiquetado gramatical reside en resolver estas ambigüedades, eligiendo la etiqueta adecuada según el contexto. ¿Pero qué magnitud tiene el problema de la ambigüedad de las palabras? Podemos apreciar que la mayoría de las palabras en inglés no son ambiguas, o lo que es lo mismo, tienen una única etiqueta posible. Pero sin embargo muchas de las palabras más comunes del inglés son ambiguas, es decir que las palabras más utilizadas, las que se emplean con mayor frecuencia, pueden tener más de una etiqueta. Por ejemplo *can* puede ser un auxiliar (puede), un sustantivo (lata o contenedor de metal) o un verbo (poner algo en la lata).

Afortunadamente muchas de las palabras ambiguas son fácilmente desambiguables. Esto sucede porque las etiquetas asociadas a una palabra no suelen ocurrir con la misma frecuencia. Por ejemplo *a* puede ser un determinante o la letra *a* (quizás como parte de un acrónimo o una inicial), pero es preciso notar que el sentido de *a* es mucho más frecuente como determinante que como letra. Es decir que es mucho más frecuente encontrar *a* en oraciones como *My father bought a new car* o *There is a hair in my soup* que en oraciones como *Written by A. Kamio* o *The letter a is the first letter of the alphabet*.

Existen distintos métodos computacionales para asignar una etiqueta gramatical a una palabra. La mayoría de los algoritmos de etiquetado automático pertenecen a una de dos clases: etiquetadores basados en reglas o etiquetadores estocásticos.

Los etiquetadores basados en reglas generalmente incluyen una gran cantidad de reglas de desambiguación escritas a mano que especifican, por ejemplo, que una palabra ambigua es un sustantivo antes que un verbo si es seguida por un determinante.

Los etiquetadores estocásticos generalmente resuelven la ambigüedad de etiquetas utilizando un corpus de entrenamiento del cual “aprenden” como etiquetar. Este aprendizaje se realiza extrayendo información sobre la probabilidad de que una palabra dada tenga cierta etiqueta en cierto contexto.

Adicionalmente existe una tercera clase de etiquetadores que es una mezcla de estos dos: etiquetadores basados en la transformación. Como los etiquetadores basados en reglas, están basados en reglas que determinan cuando una palabra ambigua debe tener cierta etiqueta. Y como los etiquetadores estocásticos tienen un componente de aprendizaje automático: las reglas son inducidas automáticamente de un corpus de entrenamiento previamente etiquetado.

2.4.1. Etiquetadores gramaticales basados en reglas

Los primeros algoritmos de asignación de etiquetas gramaticales estaban basados en un proceso de dos partes. En la primer etapa utilizaban un diccionario para asignar a cada palabra una lista de potenciales etiquetas gramaticales. En la segunda etapa utilizaban grandes listas de reglas de desambiguación escritas a mano para reducir la lista de etiquetas hasta llegar a una para cada palabra. De esta manera eliminaban las etiquetas inconsistentes con el contexto.

Las aproximaciones modernas de etiquetado gramatical basado en reglas mantienen los principios originales teniendo en cuenta que los diccionarios y el conjunto de reglas han adquirido un tamaño considerablemente mayor. Los etiquetadores actuales manejan alrededor de 3800 reglas y un diccionario de etiquetas del orden de las 56.000 entradas para el idioma inglés.

2.4.2. Etiquetadores gramaticales estocásticos

La inclusión de probabilidades en el proceso de etiquetación gramatical no es una idea nueva. Surge como una consecuencia natural a partir del hecho de que una palabra es empleada con un sentido gramatical mucho más frecuentemente que con otro. Como mencionamos anteriormente, *a* es mucho más frecuentemente utilizada como determinante que como letra. Y también a partir de la construcción gramatical; cierta etiqueta es más frecuentemente precedida por ciertas otra/s. Por ejemplo, como mencionamos antes, los pronombres posesivos generalmente son sucedidos por verbos. Es decir que es más probable encontrar oraciones cuyas palabras estén etiquetadas como PP sucedido por NN que PP sucedido por otra etiqueta.

A continuación vamos a presentar 2 tipos de etiquetadores gramaticales estocásticos: etiquetadores estocásticos basados en el modelo oculto de Markov o simplemente etiquetadores HMM ¹ y etiquetadores estocásticos basados en el modelo de máxima entropía.

2.4.3. Etiquetadores gramaticales basado en HMM

El uso del modelo oculto de Markov para realizar etiquetado gramatical es un caso especial de la inferencia bayesiana, un paradigma que fué conocido a partir del trabajo de Bayes (1763). La inferencia Bayesiana o clasificación Bayesiana fue aplicada exitosamente a problemas del lenguaje a partir de 1950. La clasificación bayesiana puede apreciarse como una tarea para la cual contamos con un conjunto de observaciones y el trabajo consiste en determinar a que conjunto de clases pertenece. En lo que respecta al etiquetado gramatical, se puede utilizar este mismo concepto para tratarlo como una tarea de clasificación de secuencia. En ese caso, la observación será una secuencia de palabras (digamos una oración) para la cual el trabajo consiste en asignar una secuencia de etiquetas gramaticales. Como ejemplo tomemos la oración que aparece a continuación:

Secretariat is expected to race tomorrow

En este caso las observaciones son la secuencia de palabras (es decir la oración misma) y nuestro objetivo es asignarles las etiquetas correspondientes. Ya que una palabra puede ser ambigua y tener más de una etiqueta posible, hay una pregunta clave que debemos hacernos: ¿Cuál es la mejor secuencia de etiquetas que corresponden a esta secuencia de palabras? La interpretación bayesiana comienza considerando todas las posibles secuencias de clases –en nuestro caso, todas las posibles secuencias de etiquetas gramaticales. El objetivo aquí es elegir la secuencia de etiquetas que es más probable dada la secuencia de observaciones de n palabras w_1^n . En otras palabras, queremos obtener, de todas las secuencias de n etiquetas t_1^n la secuencia de etiquetas tal que $P(t_1^n | w_1^n)$ sea mayor. Utilizaremos la notación $\hat{}$ para decir “nuestra estimación de la secuencia de etiquetas correcta”.

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} fP(t_1^n | w_1^n) \quad (1)$$

¹Por las siglas en inglés de Hidden Markov Model

La ecuación anterior se lee así: de todas las secuencias de etiquetas de longitud n , queremos la secuencia particular t_1^n que maximiza el lado derecho.

Mientras que esta ecuación nos garantiza obtener la secuencia de etiquetas óptima, todavía no queda del todo claro como utilizarla. Es decir, para una secuencia de etiquetas dada t_1^n y una secuencia de palabras w_1^n , no sabemos como computar directamente $P(t_1^n|w_1^n)$. Aquí entra en juego la clasificación Bayesiana, ofreciendo una forma de transformar la ecuación en un conjunto de otras probabilidades más sencillas de computar. Las reglas de Bayes reemplazan la probabilidad condicional $P(x|y)$ por otras tres probabilidades:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \quad (2)$$

Podemos sustituir (2) en (1) para obtener (3):

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} \frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)} \quad (3)$$

Convenientemente podemos simplificar (3) eliminando el denominador $P(w_1^n)$. Esto sucede ya que estamos eligiendo una de todas las secuencias de etiquetas, computando $\frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)}$ en cada una de ellas. Pero $P(w_1^n)$ no cambia en ninguna secuencia de etiquetas, entonces estamos preguntando siempre por la misma observación w_1^n , que tiene la misma probabilidad $P(w_1^n)$. Por lo tanto podemos quitar el denominador con la garantía de que el máximo sea el mismo:

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(w_1^n|t_1^n)P(t_1^n) \quad (4)$$

En resumen, la secuencia de etiquetas más probable \hat{t}_1^n dada alguna palabra w_1^n puede ser computada tomando el producto de dos probabilidades para cada secuencia de etiquetas y eligiendo la secuencia que lo maximiza.

Desafortunadamente todavía sigue siendo muy difícil computar esta ecuación directamente. Los etiquetadores gramaticales basados en HMM realizan dos suposiciones simplificadoras. La primera es que la probabilidad de aparición de una palabra depende solo de su etiqueta gramatical, es decir que es independiente de las palabras y etiquetas que tiene alrededor. Más técnicamente:

$$P(w_1^n|t_1^n) \approx \prod_{i=1}^n P(w_i|t_i) \quad (5)$$

La segunda suposición es que la probabilidad de aparición de una etiqueta gramatical depende solo de la etiqueta previa (sin tener en cuenta las etiquetas anteriores a la etiquetaa previa), esto es la suposición de bigrama.

$$P(t_1^n) \approx \prod_{i=1}^n P(t_i|t_{i-1}) \quad (6)$$

Utilizando estas suposiciones obtenemos esta nueva ecuación, la cual es utilizada por los etiquetadores gramaticales basados en bigramas para estimar la secuencia de etiquetas gramaticales más probable.

$$\hat{t}_1^n = \operatorname{argmax}_{t_1^n} P(t_1^n|w_1^n) \approx \operatorname{argmax}_{t_1^n} \prod_{i=1}^n P(w_i|t_i)P(t_i|t_{i-1}) \quad (7)$$

La ecuación anterior contiene dos clases de probabilidades, probabilidades de transición de etiquetas y probabilidades de palabras. Tomemos un momento para ver que es lo que representan estas probabilidades.

Las probabilidades de transición de etiquetas, $P(t_i|t_{i-1})$, representan la probabilidad de que ocurra una etiqueta dada la etiqueta previa. Por ejemplo, es muy probable que un determinantes preceda a un adjetivos o a un sustantivo, como *that/DD flight/NN* y *the/DT yellow/JJ hat/NN*. Por lo tanto esperamos que las probabilidades $P(NN|DT)$ y $P(JJ|DT)$ sean altas.

Por otro lado, es infrecuente que los adjetivos precedan a los determinantes, entonces la probabilidad $P(DT|JJ)$ será pequeña. Podemos computar la máxima probabilidad estimada o MLE ² de una probabilidad de transición de etiquetas $P(NN|DT)$ etiquetando y contando las etiquetas gramaticales en un corpus. Esto es: de todas las veces que vemos DT, cuántas de esas veces vemos NN después de DT. Lo expresamos más formalmente con el siguiente cociente:

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_i)} \quad (8)$$

Elijamos un corpus específico para examinar, por ejemplo el corpus Brown. Éste es un corpus de 1 millón de palabras de Inglés Americano. El corpus Brown ha sido etiquetado dos veces, la primera en los años sesenta con el conjunto de etiquetas 87-tag y de vuelta en los años noventa con el conjunto de etiquetas Treebank. En el corpus Brown etiquetado con el conjunto de etiquetas Treebank, la etiqueta DT ocurre 116.454 veces. De esas veces, DT es seguido por NN 56.509 veces. Por lo tanto el MLE de esta probabilidad de transición se calcula como sigue:

$$P(NN|DT) = \frac{C(DT, NN)}{C(DT)} = \frac{56509}{116454} = .49 \quad (9)$$

Claramente la probabilidad de obtener un sustantivo común después de un determinante es .49 y de hecho alta como sospechábamos.

Por otro lado las probabilidades de la palabra, $P(w_i|t_i)$, representan la probabilidad de que dada una etiqueta esta esté asociada con cierta palabra. Por ejemplo si tenemos la etiqueta VBZ (verbo singular de tiempo presente en tercera persona) y quisiéramos adivinar el verbo asociado a esa etiqueta, probablemente elegiríamos el verbo *is*³, debido a que el verbo *to be* es muy común en inglés. Podemos computar $P(is|VBZ)$ de nuevo contando de cuántas veces que vemos VBZ en un corpus cuántas de esas veces VBZ está etiquetando la palabra *is*. Esto es computar el siguiente cociente:

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)} \quad (10)$$

En el corpus Brown etiquetado con Treebank, la etiqueta VBZ ocurre 21.627 veces y VBZ es la etiqueta para *is* 10.073 veces. Entonces:

$$P(is|VBZ) = \frac{C(VBZ, is)}{C(VBZ)} = \frac{10,073}{21,627} = 0,47 \quad (11)$$

²Por sus siglas en inglés Maximum Likelihood Estimated

³*is* es el presente en tercera persona del verbo *to be*

En resumen, el etiquetado HMM es la tarea de elegir con la mayor probabilidad una secuencia de etiquetas para una secuencia de palabras dada. HMM incluye la suposición de ciertos hechos para simplificar las ecuaciones originales mejorando así la eficiencia de los cálculos.

2.5. Corpora de entrenamiento y corpora de verificación

Los etiquetadores gramaticales que se basan en modelos estocásticos poseen un proceso de entrenamiento sobre un corpus etiquetado previamente en el cual se generan las probabilidades que se utilizan para tomar decisiones frente a palabras ambiguas.

Este corpus de entrenamiento necesita ser cuidadosamente considerado. Si el corpus de entrenamiento es muy específico al dominio, es decir que el corpus de entrenamiento de alguna manera es similar al corpus que se desea etiquetar, las probabilidades van a ser muy ajustadas y no tendrá un buen rendimiento en oraciones de diferentes dominios. Pero si el corpus de entrenamiento es muy general, estas probabilidades no van a llegar a hacer el trabajo suficiente de reflejar el dominio.

Supongamos que estamos intentando etiquetar una oración particular. Si nuestra oración es parte del corpus de entrenamiento, las probabilidades de las etiquetas para esa oración van a ser extraordinariamente precisas y vamos a sobreestimar la precisión de nuestro etiquetador. Se desprende como conclusión que el corpus de entrenamiento no debe ser parcial incluyendo esa oración. Por lo tanto, al trabajar con etiquetadores basados en modelos estocásticos, dado un corpus de datos relevante, es una tarea habitual dividir los datos en un corpus de entrenamiento y un corpus de verificación.

Una vez realizada esta división se entrena el etiquetador con el corpus de entrenamiento, se ejecuta el proceso de etiquetación y luego se comparan los resultados con el corpus de verificación.

En general existen dos métodos para entrenar y verificar un etiquetador gramatical. En el primer método, se divide el corpus disponible en tres partes: un corpus de entrenamiento, un corpus de verificación y un corpus de test de desarrollo. Se entrena el etiquetador con el corpus de entrenamiento. Entonces se utiliza el corpus de test de desarrollo (también llamado *devtest*) para eventualmente afinar o ajustar algunos parámetros y en general decidir cual es el mejor modelo. Una vez que se elige el supuesto mejor modelo, se corre contra el corpus de verificación para ver su rendimiento.

En el segundo método de entrenamiento y verificación, se elige aleatoriamente una división de corpus de entrenamiento y verificación para nuestros datos. Se entrena el etiquetador y luego se calcula el error en el corpus de verificación. A continuación se repite con un corpus de entrenamiento y de verificación diferente seleccionado aleatoriamente.

La repetición de este proceso, llamado validación cruzada, generalmente es realizada 10 veces. Luego se promedian esas 10 corridas para obtener un promedio en la proporción del error.

Al comparar modelos es importante utilizar verificaciones estadísticas para determinar si la diferencia entre los modelos es significativa.

2.6. Evaluación de etiquetadores gramaticales

Los etiquetadores gramaticales generalmente son evaluados comparando su precisión contra un corpus de verificación Gold Standard etiquetado por humanos. Definimos precisión como el porcentaje de todas las etiquetas en el corpus de verificación donde el etiquetador y el Gold standard concuerdan. Los algoritmos más corrientes de etiquetado gramatical tienen una precisión alrededor del 96 %-97 % para corpus de etiquetas simples como el corpus del Penn Treebank. Estas precisiones son para palabras y puntuaciones, la precisión para palabras solas es menor.

Naturalmente uno tiende a preguntarse qué tan bueno es un 97 %. El rendimiento de un proceso de etiquetado puede ser comparado contra un límite inferior y un límite superior. Una manera de establecer un límite superior es ver que tan bien realizan la tarea los humanos.

Marcus, por ejemplo, encontró que los etiquetadores humanos concuerdan alrededor del 96 %-97 % de las etiquetas en la versión Penn Treebank del corpus Brown. Esto sugiere que el Gold Standard debe tener un 3 %-4 % de margen de error, y por lo tanto no tiene sentido obtener una precisión del 100 %. Ratnaparkhi mostró que en las palabras donde su etiquetador ha tenido problemas de ambigüedad de etiquetación fueron exactamente las mismas en donde los humanos han etiquetado inconsistentemente en el corpus de entrenamiento. Dos experimentos por Voutilainen encontraron que cuando a los humanos se les permitió discutir etiquetas, alcanzaron un consenso en el 100 % de las etiquetas.

Por otro lado el límite inferior sugerido por Gale es elegir la etiqueta más probable con unigrama para cada palabra ambigua. La etiqueta más probable para cada palabra puede ser computada desde un corpus etiquetado a mano (que puede ser el mismo que el corpus de entrenamiento para el etiquetador que está siendo evaluado).

2.7. Análisis de error

Para mejorar el rendimiento de un etiquetador gramatical necesitamos entender donde está funcionando mal. Por eso el análisis de error tiene un papel preponderante. Esta tarea se realiza construyendo una matriz de confusión o tabla de contingencia. Una matriz de confusión es una matriz de $n \times n$ donde la celda (x, y) contiene el número de veces que una palabra con correcta etiqueta x fué etiquetada por el modelo como y . Por ejemplo, la siguiente tabla muestra una porción de la matriz de confusión para los experimentos de etiquetado con HMM de Franz.

	IN	JJ	NN	NNP	RB	VBD	VCN
IN	-	.2			.7		
JJ	.2	-	3.3	2.1	1.7	.2	2.7
NN		8.7	-				.2
NNP	.2	3.3	4.1	-	.2		
RB	2.2	2.0	.5		-		
VBD		.3	.5			-	4.4
VCN		2.8				2.6	-

Las etiquetas de la fila indican las etiquetas correctas, las etiquetas de las columnas indican las etiquetas asignadas por el etiquetador, y cada celda indica el porcentaje del error de etiquetado general. Por lo tanto 4.4 % del total de errores fueron causados por fallida etiquetacion de VBD como VBN. La matriz anterior y el análisis de error relacionado en Franz, Kupiec y Ratnaparkhi sugieren que algunos de los mayores problemas que encaran los etiquetadores actuales son:

1. NN contra NNP contra JJ: Estas etiquetas son difíciles de distinguir. Es especialmente importante distinguir entre sustantivos propios para extracción de la información y traducción automática.
2. RP contra RB contra IN: Todas estas etiquetas pueden aparecer en secuencias de satélites inmediatamente después del verbo.
3. VBD contra VBN contra JJ: Distinguir estas etiquetas es importante para el *parsing* parcial (los participios son utilizados para encontrar pasivos), y para etiquetar correctamente los ejes de frases nominales.

El análisis de error es una parte crucial de cualquier aplicación lingüística computacional. Puede ayudar a encontrar *bugs*, encontrar problemas en los datos de entrenamiento y lo más importante, ayuda en el desarrollo de conocimiento y/o algoritmos para utilizar en la solución de problemas.

2.8. Palabras desconocidas

Todos los algoritmos de etiquetado gramatical presentados anteriormente requieren un diccionario que liste las posibles etiquetas de cada palabra para que posteriormente el proceso de etiquetado se encargue de identificar la etiqueta correcta. Pero claro, hay un problema, incluso el diccionario más grande no es capaz de contener cada palabra posible. Los sustantivos propios y los acrónimos son creados muy frecuentemente, incluso nuevos sustantivos comunes y verbos ingresan al lenguaje en una proporción sorprendente. Por lo tanto, para construir un etiquetador completo no podemos utilizar siempre un diccionario para obtener $P(w_i|t_i)$. Necesitamos algún método para adivinar la etiqueta de una palabra desconocida.

El algoritmo más básico para manejar palabras desconocidas es suponer que cada palabra desconocida es ambigua entre todas las posibles etiquetas, con igual probabilidad. Entonces el etiquetador debe confiar únicamente en etiquetas contextuales para sugerir la etiqueta adecuada. Un algoritmo ligeramente más complejo está basado en la idea de que la distribución de probabilidad de las etiquetas sobre las palabras desconocidas es muy similar a la distribución de las etiquetas sobre palabras que ocurren solo una vez en un corpus de entrenamiento, una idea sugerida por Baayen y Sproat (1996) y Dermatas y Kokkinakis (1995). Estas palabras que ocurren solo una vez son conocidas como *hapax legomena*.

Por ejemplo, palabras desconocidas y *hapax legomena* son similares en el hecho de que son más probables de ser sustantivos, seguidos por verbos, pero infrecuentemente suelen ser determinantes o intersecciones. Entonces la probabilidad $P(w_i|t_i)$ para una palabra desconocida es determinada por el promedio de la distribución sobre todos los conjuntos de palabras de un solo elemento en el corpus de entrenamiento. En resumen, la idea es utilizar “cosas que hemos visto una vez” como un estimador para “cosas que nunca hemos visto”.

La mayoría de los algoritmos para palabras desconocidas, de todas maneras, hace uso de una fuente de información mucho más poderosa: la morfología de las palabras. Para el inglés, por ejemplo, palabras que terminadas en *s* tienden a ser sustantivos plurales (NNS), palabras terminadas en *ed* tienden a ser pasado participio (VBN), palabras terminadas en *able* tienden a ser adjetivos (JJ), y así. Incluso si nunca vimos una palabra, podemos utilizar hechos sobre su forma morfológica para adivinar su etiqueta. Al lado del conocimiento morfológico, la información ortográfica puede ser de mucha ayuda. Por ejemplo, palabras que comienzan con letras mayúsculas generalmente son nombres propios (NP). La presencia de un guión es también una característica útil; las palabras con guión en la versión Brown del Treebank son más probables de ser adjetivos (JJ). Esta prevalencia de JJs es causada por las instrucciones de etiquetado para el Treebank, que especifican que modificadores prenominales deben ser etiquetados como JJ si contienen un guión.

¿Cómo son combinadas y utilizadas estas características en los etiquetadores gramaticales? Un método es entrenar por separado estimadores de probabilidad para cada característica, asumiendo independencia, y multiplicando las probabilidades. Weischedel (1993) construyó un modelo así, basado en cuatro clases específicas. Utilizaron 3 terminaciones inflexionales (*ed*, *s*, *ing*), 32 terminaciones derivacionales (como *ion*, *al*, *ive* y *ly*), 4 valores de mayúscula dependiendo si una palabra es inicio de oración (+/- mayúscula, +/- inicio) y donde una palabra fué guionada. Para cada característica, entrenaron estimadores de máxima verosimilitud de la probabilidad de la característica dada una etiqueta desde un corpus de entrenamiento etiquetado. Entonces combinaron las características para estimar la probabilidad de una palabra desconocida asumiendo independencia y multiplicando:

$$P(w_i|t_i) = p(\text{palabra desconocida}|t_i)p(\text{mayúscula}|t_i)p(\text{final/guion}|t_i) \quad (12)$$

Otro acercamiento basado en HMM, debido a Samuelsson (1993) y Brants (2000), generaliza el uso de morfología en una manera basada en datos. En este acercamiento, en vez de preseleccionar ciertos sufijos a mano, todas las secuencias finales de letras de todas las palabras son consideradas. Consideran sufijos menores a diez letras, computando para cada sufijo de longitud i la probabilidad de la etiqueta t_i :

$$P(t_i|l_{n-i+1}, \dots, l_n) \quad (13)$$

0m Estas probabilidades son suavizadas utilizando sucesivamente menores y menores sufijos. Esta información de sufijos es mantenida para palabras en mayúscula y minúscula.

En general, la mayoría de los modelos de palabras desconocidas intentan capturar el hecho de que las palabras desconocidas son improbable de ser clases cerradas de palabras como las preposiciones. Brants modela este hecho computando solamente las probabilidades de sufijos desde el corpus de entrenamiento para palabras cuya frecuencia en el corpus de entrenamiento es ≤ 10 .

2.9. Etiquetador Gramatical TnT

TnT(Trigrams' n'Tags) es un etiquetador gramatical estocástico basado en HMM. Según Brants este etiquetador tiene un rendimiento mejor o igual a otros etiquetadores actuales de diferentes bases teóricas, incluyendo etiquetadores basados en máxima entropía.

2.9.1. Modelo teórico

TnT utiliza modelos de Markov de segundo orden para la etiquetación gramatical. Técnicamente calcula, dada una secuencia de T palabras w_1, \dots, w_T

$$\operatorname{argmax}_{t_1, \dots, t_T} \left[\prod_{i=1}^T P(t_i | t_{i-1}, t_{i-2}) P(w_i | t_i) \right] P(t_{T+1} | t_T) \quad (14)$$

para hallar las etiquetas t_1, \dots, t_T . Las etiquetas adicionales t_{-1}, t_0 y t_T son delimitadores del principio y el final de la secuencia. Estas etiquetas adicionales mejoran levemente los resultados del etiquetado marcando una particularidad de TnT con respecto a otros etiquetadores. Las probabilidades son estimadas desde un corpus etiquetado previamente (el ya mencionado corpus de entrenamiento). Para ello TnT utiliza probabilidades de máxima verosimilitud \hat{P} obtenidas a partir de la frecuencia relativa y luego aplica una técnica de suavizado

$$\text{Unigramas: } \hat{P}(t_3) = \frac{f(t_3)}{N} \quad (15)$$

$$\text{Bigramas: } \hat{P}(t_3 | t_2) = \frac{f(t_2, t_3)}{f(t_2)} \quad (16)$$

$$\text{Trigramas: } \hat{P}(t_3 | t_1, t_2) = \frac{f(t_1, t_2, t_3)}{f(t_1, t_2)} \quad (17)$$

$$\text{Léxico: } \hat{P}(w_3 | t_3) = \frac{f(w_3, t_3)}{f(t_3)} \quad (18)$$

donde t_1, t_2 y t_3 pertenecen al conjunto de etiquetas y w_3 pertenece al lexicon. N es el número de *tokens* del corpus de entrenamiento. La probabilidad de máxima verosimilitud se calcula como cero si el denominador o el nominador son cero.

2.9.2. Suavizado

TnT aplica una técnica de suavizado sobre las frecuencias contextuales. Esto tiene lugar debido al problema de los datos esparsos en las probabilidades de los trigramas. Es decir, no hay suficientes instancias de cada trigramas para calcular confiablemente su probabilidad asociada. Incluso estableciendo a cero la probabilidad de un trigramas que no aparece en el corpus genera el efecto indeseado de convertir la probabilidad de una secuencia completa en cero. TnT utiliza interpolación lineal de unigramas, bigramas y trigramas para realizar este proceso de suavizado. Es decir que se estima la probabilidad de un trigramas como sigue

$$P(t_3 | t_1, t_2) = \lambda_1 \hat{P}(t_3) + \lambda_2 \hat{P}(t_3 | t_2) + \lambda_3 \hat{P}(t_3 | t_1, t_2) \quad (19)$$

donde \hat{P} son los estimadores de máxima verosimilitud presentados anteriormente y λ_1, λ_2 y λ_3 son los pesos asociados a estos estimadores, tales que $\lambda_1 + \lambda_2 + \lambda_3 = 1$. TnT utiliza interpolación lineal con independencia de contexto. Es decir que λ_1, λ_2 y λ_3 tienen el mismo valor para todos los trigramas, o lo que es lo mismo, λ_1, λ_2 y λ_3 son independientes del trigramas que se está calculando. Los valores λ_1, λ_2 y λ_3 son estimados por interpolación de borrado. La idea es que se dará mayor peso a la información de unigramas, bigramas o trigramas más abundante. A continuación se presenta el algoritmo utilizado para realizar esta tarea

Algoritmo 1 Cálculo de λ_1, λ_2 y $\lambda_3 = 0$

Establecer $\lambda_1 = \lambda_2 = \lambda_3 = 0$
por cada trigramas t_1, t_2, t_3 con $f(t_1, t_2, t_3) > 0$
 según el máximo de los tres valores siguientes:
 caso $\frac{f(t_1, t_2, t_3)-1}{f(t_1, t_2)-1}$: incrementar λ_1 en $f(t_1, t_2, t_3)$
 caso $\frac{f(t_2, t_3)-1}{f(t_2)-1}$: incrementar λ_2 en $f(t_1, t_2, t_3)$
 caso $\frac{f(t_3)-1}{N-1}$: incrementar λ_3 en $f(t_1, t_2, t_3)$
 fin
fin
normalizar λ_1, λ_2 y λ_3

2.9.3. Manejo de palabras desconocidas

TnT, al igual que muchos otros etiquetadores gramaticales, maneja las palabras desconocidas mediante análisis de sufijos. Los sufijos son fuertes predictores del tipo de palabra. Por ejemplo las palabras terminadas en *able* en el Wall Street Journal parte del Penn Treebank son adjetivos (JJ) en el 98 % de los casos (ej.: *fashionable, variable*) y sustantivos (NN) en el 2 % restante.

La distribución de probabilidades para un sufijo particular es generada a partir de todas las palabras en el corpus de entrenamiento que comparten el mismo sufijo (de alguna longitud máxima predefinida). El término sufijo se entiende en este contexto como la secuencia final de letras de una palabra, que no coincide necesariamente con el significado lingüístico de sufijo.

La fórmula utilizada para calcular la probabilidad de que una etiqueta pertenezca a cierto sufijo es $P(t|l_{n-m+1}, \dots, l_n)$, es decir, la probabilidad de una etiqueta t dadas las últimas letras l_i de una palabra de n letras. TnT aplica una técnica de suavizado utilizando sufijos cada vez más pequeños aplicando un peso θ_i a cada uno:

$$P(t|l_{n-m+1}, \dots, l_n) = \frac{\hat{P}(t|l_{n-i+1}, \dots, l_n) + \theta_i P(t|l_{n-i}, \dots, l_n)}{1 + \theta_i} \quad (20)$$

para $i = m, \dots, 0$, utilizando el estimador de máxima verosimilitud \hat{P} para las frecuencias en el lexicon, los pesos θ_i y el caso base

$$P(t) = \hat{P}(t) \quad (21)$$

El estimador de máxima verosimilitud para un sufijo de longitud i es

$$\hat{P}(t|l_{n-i+1}, \dots, l_n) = \frac{f(t, l_{n-i+1}, \dots, l_n)}{f(l_{n-i+1}, \dots, l_n)} \quad (22)$$

TnT utiliza desvío estándar del estimador de máxima verosimilitud para calcular los pesos θ_i .

Decisiones de diseño:

1. La primer decisión de diseño que afronta TnT es encontrar un buen valor para n , la longitud máxima de sufijo utilizada. TnT elige tomar la longitud del mayor sufijo encontrado en el corpus de entrenamiento, con la restricción de que sea menor o igual a 10.

2. Se utiliza independencia de contexto para calcular θ_i , la misma idea que se utilizó para calcular λ_i .
3. Se utilizan estimadores distintos para mayúsculas y minúsculas. Es decir, se mantienen dos árboles de sufijos distintos, uno para mayúsculas y otro para minúsculas.
4. La otra decisión relevante es: ¿Qué palabras del lexicon deben ser utilizadas para el manejo de sufijos? Basándose en el hecho de que las palabras desconocidas son más probablemente infrecuentes, TnT utiliza sufijos de palabras infrecuentes. Por lo tanto, restringe el procedimiento de cálculo de probabilidades de sufijos a palabras con una frecuencia menor o igual a 10.

Adicionalmente, TnT discrimina la información sobre mayúsculas y minúsculas. Esto es debido a que las probabilidades de las etiquetas de palabras con mayúsculas son distintas a las de las palabras con minúsculas. Para llevar esto a cabo se utilizan flags en las probabilidades contextuales. En vez de

$$P(t_3|t_1, t_2) \quad (23)$$

se utiliza

$$P(t_3, c_3|t_1, c_1, t_2, c_2) \quad (24)$$

donde c_1 , c_2 y c_3 son 1 si la palabra contiene mayúsculas y 0 en otro caso. Esto es equivalente a doblar el conjunto de etiquetas y utilizar etiquetas diferentes según si la palabra aparece en mayúscula o no.

3. Desarrollo

3.1. Diccionario COBUILD

Como se menciona anteriormente, para suplir la falta de corpus de entrenamiento sin caer en la tediosa y costosa tarea de anotar un nuevo corpus manualmente, se introduce una fuente de información existente y manualmente anotada. Estamos hablando de un diccionario, que no es ni más ni menos que un conjunto de palabras, donde cada palabra representa una entrada. Y a su vez cada entrada proporciona una explicación sobre el significado de esa palabra, algunas características como su pronunciación y clase gramatical y uno o más ejemplos que muestran su uso. Entonces si pegamos o concatenamos todas esas entradas podemos ver al diccionario como un grupo de palabras con información gramatical para aquellas palabras asociadas a una entrada. Es decir, estas entradas del diccionario:

```
siren
sirens
s*a*!i*%er%e0n
```

```
A woman is described as a siren when she is attractive and dangerous to men.
One of the women, another of those sirens, haughtily regarded us as we talked.
countable noun
noun
```

sirloin
 sirloins
 s*\$e*:l!o!in
 A sirloin is a piece of beef which is cut from the lower part of a cows back.
 ...a sirloin of Scotch beef.
 mass noun
 noun

sissy
 sissies
 s*!isi1
 A boy is described as a sissy, especially by other boys, if he does not like sport and is
 Youre a lot of cry-babies and sissies... ...Mummys little sissy boy.
 countable noun: also vocative
 noun

Pueden concatenarse y verse como:

A	talked
woman	.
is	A
described	sirloin NN
as	is
a	a
siren NN	piece
when	of
she	beef
is	which
attractive	is
and	cut
dangerous	from
to	the
men	lower
.	part
One	of
of	a
the	cows
women	back
,	.
another	...
of	a
those	sirloin NN
sirens NNS	of
,	Scotch
haughtily	beef
regarded	.
us	Long
as	fibres
we	are

picked	and
carefully	is
from	afraid
the	to
sisal NN	do
leaves	things
.	that
A	are
boy	slightly
is	dangerous
described	.
as	Youre
a	a
sissy NN	lot
,	of
especially	cry-babies
by	and
other	sissies NNS
boys	...
,	...
if	Mummys
he	little
does	sissy NN
not	boy
like	.
sport	

Esta última información conforma un corpus parcialmente anotado que se utilizará como base para construir un nuevo corpus completamente anotado. El cual servirá como una nueva fuente de información para entrenar etiquetadores gramaticales.

Claramente el primer paso para llevar a cabo esta tarea es elegir un diccionario y extraer esa información. El diccionario elegido fué Cobuild. A continuación se detallan las características que lo hicieron distintivo frente a otros diccionarios.

Cobuild es un diccionario basado en la información del corpus Bank of English y el corpus Collins. Su siglas significan: Collins Birmingham University International Language Database.

El corpus Collins es una base de datos con alrededor de 2.5 billones de palabras en Inglés. Contiene material escrito de websites, diarios, revistas y libros publicados en todo el mundo, y material hablado de radio, TV y conversaciones diarias.

A su vez, el Bank of English forma parte del corpus Collins. Contiene 650 millones de palabras cuidadosamente seleccionadas para dar un reflejo preciso y balanceado del Inglés que es usado día a día.

Gracias a la extensa amplitud del corpus se pueden apreciar una gran cantidad de ejemplos de como la gente utiliza realmente las palabras. Se puede observar como son utilizadas, que significan, que palabras ocurren juntas y que tan a menudo. Esta información sobre la frecuencia ha ayudado a decidir que palabras incluir en el diccionario Cobuild. Por ejemplo, alrededor del 90 % del

inglés hablado y escrito está constituido de aproximadamente 3.500 palabras.

The Bank of English contiene un amplio rango de tipos diferentes de lenguaje escrito y hablado proveniente de cientos de fuentes diferentes. Aunque la mayoría de las fuentes son británicas, aproximadamente el 25 % de la información proviene de fuentes de inglés americano y alrededor del 5 % de otras variedades nativas del inglés como Australia y Singapur. Los textos escritos provienen de diarios, revistas, libros de ficción y de no ficción, folletos, informes y cartas. Dos tercios del corpus están confeccionados a partir del lenguaje de los medios: diarios, revistas, radio y televisión. Esta es una categoría significativa en vista de que millones de personas leen y escuchan el lenguaje presente en los medios. También fueron incluidas publicaciones internacionales, nacionales y locales para capturar un rango general de temas importantes y estilos. Hay otros cientos de libros y revistas de especial interés que abordan temas desde aeróbicos a zoología. Sin embargo los libros de texto técnicos, científicos, manuales, etc., no fueron incluidos en el corpus.

El lenguaje hablado informal es representado por grabaciones de conversaciones diarias casuales, reuniones, entrevistas y discusiones. Alrededor de 15 millones de palabras de The Bank of English son transcripciones de lenguaje hablado de esta clase. Luego son seleccionados para incluir un amplio rango de temas y situaciones de habla.

El propósito de recolectar toda esta valiosa información en computadoras fue para permitir a los lingüistas (escritores de diccionarios) tener acceso a la mayor cantidad de información posible sobre cada una de las palabras que está definiendo. Desde luego, los lingüistas son elegidos a partir de su habilidad con el lenguaje, pero incluso el lingüista más experimentado no puede deducir utilizando solo su intuición todos los hechos relevantes sobre todas las palabras en un lenguaje. El corpus, y el software que se utiliza para analizarlo ayudó al equipo de Cobuild a ordenar la información y ganar valiosa percepción sobre la manera en que las palabras son utilizadas: sus significados, sus patrones gramaticales típicos y las maneras en que están relacionadas con otras palabras.

Muchas palabras tienen más de una clase de palabra gramatical y a menudo es de mucha ayuda para los lingüistas mirar solo a una clase de palabra por vez. Para ayudarlos a hacer esto, se ha desarrollado un software que muestra las clases de palabras en cada línea del corpus. De esta manera los lingüistas pueden mirar la información completa de la clase de palabra o pueden preguntar solo por verbos, sustantivos, etc.

Este tipo de software les permite a los lingüistas a tomar decisiones sobre los diferentes sentidos de las palabras, el lenguaje de las definiciones, la selección de ejemplos, y la información gramatical dada. El corpus permite realizar esta tarea con confianza y exactitud. Y cuanto más grande es el corpus mayor es la confianza y la exactitud.

El diccionario Cobuild fue concebido teniendo especial atención en los ejemplos expuestos. Como se mencionó antes, el proceso de agregado de palabras al diccionario es muy cuidadoso: cuando un editor quiere agregar una nueva palabra al diccionario, busca en el corpus cada ejemplo que contenga esa palabra. La palabra aparece en una larga lista de oraciones y el editor decide cuál de todos los ejemplos expresa mejor el sentido que está buscando en esa palabra. Todos los ejemplos del diccionario Cobuild muestran patrones gramaticales típicos, vocabulario típico y contextos típicos para cada palabra. En consecuencia, Cobuild presenta una cantidad exhaustiva del vocabulario inglés derivado de

observaciones directas del lenguaje.

3.1.1. Método de construcción

En 1987 se publicó el diccionario Cobuild basado en un corpus de 20 millones de palabras. A continuación se construyó un nuevo corpus, el Bank of English con alrededor de 650 millones de palabras. La nueva edición del diccionario Cobuild se basa en este nuevo corpus.

La construcción de este nuevo diccionario fue un proceso en donde se decidió que palabras y frases presentes en el corpus incluir. Luego se examinó el lenguaje palabra por palabra y frase por frase con el objetivo de dar clara cuenta de cada significado y uso. Entonces se escribió una definición, se eligieron ejemplos típicos, y se agregó información sobre la pronunciación, la gramática, semántica, pragmatismos y frecuencia para completar cada entrada.

3.1.2. Evidencia

Un diccionario debe comenzar por la evidencia, los hechos. Los hablantes de un lenguaje conocen mucho sobre éste, a partir de que cada día leen y hablan sin esfuerzo durante horas. Sin embargo no son capaces de explicar que es lo que hacen. La mayoría de las personas no son conscientes de la habilidad que poseen para utilizar un lenguaje; no pueden examinarlo en detalle, simplemente lo utilizan para comunicarse. Aquellos que aprenden a observar el lenguaje cuidadosamente pueden expresar y organizar algunos de los hechos sobre éste basados en la experiencia. De todas maneras hay muchos hechos sobre el lenguaje que no pueden ser descubiertos simplemente pensando y reflexionando sobre él, incluso leyendo y escuchando muy atentamente. Es por eso que Cobuild estableció la utilización de computadoras para identificar estos hechos.

3.1.3. Un corpus

El resultado fue que Cobuild estableció un nuevo tipo de evidencia, una colección de textos en inglés llamado corpus ubicado en una computadora de manera que pueda consultarse instantáneamente. Los creadores de Cobuild sabían que necesitaban millones de palabras de inglés, hablado y escrito, americano y británico, formal e informal, sobre hechos y sobre ficción, etc. Esta evidencia reunida durante varios años, permitió encontrar las palabras y expresiones más utilizadas. Cuando una palabra tiene varios significados existe la capacidad de ver cuales son los significados importantes, y cuales frases se deben incluir. Tomaron como filosofía y fueron conscientes de que todos los detalles de un uso natural de una palabra son esenciales y no pueden ser falsificados. Se dieron cuenta de que debían utilizar ejemplos reales siguiendo la tradición de los grandes lingüistas, en lugar de crearlos.

3.1.4. The Bank of English

Hace varios años que se ha hecho mucho más fácil reunir grandes cantidades de lenguaje hablado y escrito. Los publicadores de libros, revistas y diarios se hicieron conscientes de la gran cantidad de lenguaje que pasaba a través de sus manos y de que debe haber muchas buenas razones para conservarlo en formatos electrónicos. Un negocio apareció para el lenguaje electrónico entre la gente que

quería encontrar o verificar sentencias, particularmente en las noticias, revistas y lenguaje legal. Gradualmente millones de palabras comenzaron a estar disponibles para los estudiosos del lenguaje. Hoy en día el problema no es encontrar el lenguaje sino manejarlo y realizar selecciones sensibles y balanceadas para las tareas analíticas.

Diseñando el corpus The Bank of English se balancearon un número de factores (inglés hablado y escrito, americano y británico y otras características: hablantes de comunidades nativas, libros y revistas y más clasificaciones dentro de éstas)

Dentro del componente hablado, el tipo de lenguaje más difícil de recolectar fue como siempre la conversación informal grabada en la vida diaria de la gente común, sin pensar de que su lenguaje está siendo preservado en un corpus. Cada conversación tiene que ser grabada y transcrita por expertos para luego ser ingresada en una computadora. Esta clase de lenguaje improvisado es de un interés particular para los constructores de diccionarios. El Bank of English cuenta con un total de 15 millones de palabras de este tipo de grabaciones de lenguaje hablado.

3.1.5. La lista de palabras principales

Es mucho más fácil decidir que palabras y frases incluir y cuales omitir, cuando se tienen cifras exactas a partir de una cantidad tan grande de lenguaje. Las computadoras pueden verificar instantáneamente la actividad del lenguaje de miles de hablantes y escritores. Un diccionario (incluso un gran diccionario) es capaz de elegir solo los hechos a presentar más importantes del lenguaje y los compiladores necesitan buena evidencia para sus selecciones. Cobuild se especializa en presentar las palabras y frases que son frecuentes en el uso diario. Lejos de ser un registro histórico del lenguaje es más bien una muestra del lenguaje contemporáneo.

3.1.6. Frecuencia

Cobuild brinda información sobre la frecuencia de las palabras principales. Se establecieron 5 bandas de frecuencias. Comenzando con las palabras muy comunes (las de mayor frecuencia), oscila entre un vocabulario básico a uno intermedio y hasta cubrir el núcleo del vocabulario. Las palabras principales sin marca de frecuencia son las menos comunes, sin embargo vale la pena incluirlas en el diccionario. El punto es que el idioma inglés utiliza un número bastante pequeño de palabras para la mayoría de los propósitos pero también tiene disponible un rico y amplio vocabulario. Por ejemplo *be* y *because* son naturalmente pertenecientes a la banda de mayor frecuencia, por el otro lado, palabras como *barracuda*, *basalt* y *basrelief* no son tan frecuentes. Estas últimas son claramente utilizadas en ocasiones particulares. Cabe aclarar que incluso las palabras infrecuentes incluidas en el diccionario fueron seleccionadas por su utilidad relativa entre miles de palabras posibles.

Entonces Cobuild cuenta con un sistema de frecuencia que marca las palabras principales: una marca significa que la palabra tiene una alta frecuencia y por lo tanto es una palabra común dentro del lenguaje inglés. Dos o más marcas significan que la palabra es parte esencial del vocabulario, cuantas más marcas posee, menos frecuente es.

3.1.7. Ejemplos

Todos los ejemplos fueron seleccionados del corpus The Bank of English. Como antes dijimos, los ejemplos son seleccionados cuidadosamente para mostrar los patrones que frecuentemente son hallados junto a una palabra o frase. El compilador tiene docenas, centenas o miles de ejemplos disponibles y rápidamente escoge los *colocados* (palabras particulares ubicadas cerca de la palabra principal) y las estructuras típicas en donde la palabra o frase es encontrada más a menudo.

Esto significa que los ejemplos cumplen varias funciones. Desde luego ayudan a mostrar el significado de la palabra exhibiendo su uso. Las investigaciones incluso sugieren que un gran número de usuarios comienza con los ejemplos antes que con el significado. Las definiciones de Cobuild son bastante claras por sí mismas y los ejemplos muestran el fraseo característico alrededor de la palabra. Como los ejemplos son piezas de texto genuinas y han sido elegidas cuidadosamente en base al uso de la palabra, pueden ser de confianza para exhibir la palabra en un contexto natural.

3.1.8. Información gramatical

Casi cada sentido de cada entrada en el diccionario Cobuild tiene junto a esta una clasificación gramatical, usualmente una clase de palabra y a menudo también una nota estructural. Esta es la información sobre la que se sustenta este trabajo, ya que en base a ella se construirá el nuevo corpus de entrenamiento.

3.1.9. Pragmatismo

Muchos usos de una palabra necesitan más de una frase para explicar apropiadamente su significado. La gente utiliza palabras para realizar muchas cosas: hacer invitaciones, expresar sus sentimientos, enfatizar que es lo que está diciendo, etc. El corpus nos brinda evidencia para tales usos que son difíciles de tomar desde cualquier otra fuente, porque nosotros solo los advertimos cuando vemos reunidos muchos ejemplos de ellos.

El estudio y descripción de las formas en que la gente utiliza el lenguaje para realizar cosas es llamado pragmatismo. Este aspecto del lenguaje es muy importante y fácil de omitir. Esto sucede cuando el lenguaje está dando significado adicional. Cobuild posee mucha información sobre pragmatismo y la expone mediante un símbolo especial en cada entrada. Por ejemplo *and things like that* es definido como una expresión utilizada para ensanchar el rango de una lista.

3.1.10. Definición del estilo

La característica más distintiva de Cobuild en su primera versión fué el uso de frases completas en las definiciones. El significado fué establecido de la forma en que una persona ordinaria podría explicárselo a otra.

Generalmente los diccionarios ofrecen definiciones breves y tradicionales, mientras que Cobuild expone definiciones realmente amplias y ricas. Si se observan detenidamente las definiciones particulares se puede apreciar que cada palabra es elegida para ilustrar ciertos aspectos del significado. Y en la medida

en que es posible, las palabras utilizadas en una definición son más frecuentes que la palabra que está siendo definida.

Las definiciones cortas no pueden decir demasiado. Por ejemplo, el primer sentido de verbo de *mean* podría ser definido como solo *signify*, que es cierto, pero no es todo lo que se puede decir. Cobuild expone esto: *If you want to know...* es decir que ese sentido surge cuando alguien está en la búsqueda de información. La palabra *if* indica que esta es una opción, pero una perfectamente normal, y *you* nos dice que no es una característica de ningún grupo particular de gente (compararlo con *if a policeman arrests you...*). Entonces la definición dice lo que alguien puede querer saber sobre el significado de una *palabra, código, señal o gesto*, indicando que esas son las típicas clases de temas que serán encontradas con este sentido de *mean*. Solo después de toda esta información viene el equivalente de *signify*: *lo que se refiere a o a que mensaje transmite*. Entonces hay 12 palabras antes de la palabra principal en este sentido, pero cada una de ellas transmite información vital que sería muy difícil de incluir en una definición corta.

3.2. Traducción de etiquetas

Ya que el diccionario COBUILD contiene una clase de etiquetas gramaticales distintas a las etiquetas gramaticales utilizadas por los corpus de verificación, se tomó la decisión de traducir estas en etiquetas de PenTreeBank. Para reaalizar sin inconvenientes la tarea posterior de evaluación y medición del rendimiento de los etiquetadores gramaticales entrenados con la información adicional obtenida de Cobuild.

A continuación se muestra la tabla de traducción empleada:

Cuadro 1: Tabla de traducción de etiquetas

Etiqueta COBUILD	Etiqueta PenTreeBank
coordinating conjunction	CC
number	CD
determiner	DT
determiner + countable noun in singular	DT
preposition	IN
subordinating conjunction	IN
preposition, or adverb after verb	IN
preposition after noun	IN
adjective	JJ
classifying adjective	JJ
qualitative adjective	JJ
adjective colour	JJ
ordinal	JJ
adjective after noun	JJ
modal	MD
adverb	RB
noun	NN
uncountable noun	NN
noun singular	NN

Cuadro 1: Tabla de traducción de etiquetas

Etiqueta COBUILD	Etiqueta PenTreeBank
countable or uncountable noun	NN
countable noun with supporter	NN
uncountable or countable noun	NN
noun singular with determiner	NN
mass noun	NN
uncountable noun with supporter	NN
partitive noun	NN
noun singular with determiner with supporter	NN
countable noun + of	NN
countable noun, or by + noun	NN
countable noun or partitive noun	NN
count or uncountable noun	NN
countable noun or vocative	NN
partitive noun + uncountable noun	NN
noun singular with determiner + of	NN
noun in titles	NN
noun vocative	NN
uncountable noun + of	NN
indefinite pronoun	NN
uncountable noun, or noun singular	NN
countable noun, or in + noun	NN
partitive noun + noun in plural	NN
countable or uncountable noun with supporter	NN
uncountable noun, or noun before noun	NN
uncountable or countable noun with supporter	NN
noun before noun	NN
noun plural with supporter	NNP
noun in names	NNP
proper noun or vocative	NNP
proper noun	NNP
noun plural	NNS
predeterminer	PDT
pronoun	PP
possessive	PPS
adverb with verb	RB
adverb after verb	RB
sentence adverb	RB
adverb + adjective or adverb	RB
adverb + adjective	RB
preposition or adverb	RB
adverb after verb, or classifying adjective	RB
adverb or sentence adverb	RB
adverb with verb, or sentence adverb	RB
exclamation	UH
exclam	UH
verb	VB
verb + object	VB

Cuadro 1: Tabla de traducción de etiquetas

Etiqueta COBUILD	Etiqueta PenTreeBank
verb or verb + object	VB
ergative verb	VB
verb + adjunct	VB
verb + object + adjunct	VB
verb + object <i>noun group or reflexive</i>	VB
verb + object or reporting clause	VB
verb + object <i>reflexive</i>	VB
verb + object, or phrasal verb	VB
verb + to-infinitive	VB
ergative verb + adjunct	VB
verb + object + adjunct <i>to</i>	VB
verb + object, or verb + adjunct	VB
verb + object + adjunct <i>with</i>	VB
verb + adjunct <i>with</i>	VB
verb + complement	VB
verb + object, or verb	VB
verb + object + to-infinitive	VB
verb + reporting clause	VB
verb or ergative verb	VB
verb + adjunct <i>from</i>	VB
wh: used as determiner	WDT
wh: used as relative pronoun	WP
wh: used as pronoun	WP
wh: used as adverb	WRB
phrase + noun group	
convention	
combining form	
prefix	
phrasal verb	
other	
phrase	
suffix	
wh	
phrase after noun	
phrase + reporting clause	

3.3. Extracción de la información

El diccionario COBUILD guarda su información en un archivo de texto plano con un formato particular. El primer desafío de este trabajo fué comprender y extraer la información almacenada en ese archivo. A continuación se muestra un pequeño fragmento del mismo para ejemplificar

```

DICTIONARY_ENTRY
ace

```

aces
 *e!*is
 If you are or come within an ace of something, you very nearly do or experience it.
 He came within an ace of being run over.
 phrase: verb inflects
 phrase
 DI000183
 004

DICTIONARY_ENTRY

ace
 aces
 *e!*is
 A person who is ace at something is extremely good at it; an informal use.
 ...an ace marksman.
 classifying adjective
 adjective
 DI000183
 005
 expert

DICTIONARY_ENTRY

ace
 aces
 *e!*is
 If you say that something is ace, you mean that you think that it is very good;
 an informal use.
 Their new records really ace!
 qualitative adjective or exclamation
 adjective
 DI000183
 006
 great
 lousy

Cada entrada posee una cantidad variable de campos y no es posible identificarlos exactamente. Sin embargo, contienen generalmente la palabra, sus formas, la pronunciación y uno o más ejemplos donde se indica como se emplea (mediante una etiqueta gramatical). Estas entradas, que constituyen la fuente de información principal sobre la cual se basa este trabajo y que conforman el diccionario COBUILD, fueron cuidadosamente procesadas y refinadas intentando mantener toda la información disponible.

El primer desafío de esta etapa consistió en reconocer y registrar información relacionada a las formas flexionadas de la palabra (plurales, pasados, etc.). En muchas entradas ocurre la palabra, uno o más ejemplos en dónde esta aparece con su correspondiente uso (indicado por medio de etiquetas gramaticales) pero las apariciones dentro de los ejemplos ocurren con formas flexionadas. Tomemos como ejemplo esta entrada:

DICTIONARY_ENTRY

```

bite
bites, biting, bit, bitten
b*a*!it
If an object or surface bites, it grips another object or surface rather than slipping
on it or against it.
Let the clutch in slowly until it begins to bite.
verb
verb
DI002405
009
catch
grip

```

Aquí arriba podemos observar una entrada del diccionario para la palabra *bite*, conteniendo dos ejemplos de esta palabra con sus respectivas etiquetas gramaticales. Podemos notar que en el primer ejemplo la palabra aparece en la forma flexionada; *bites*. Es decir que la entrada del diccionario nos está ofreciendo más información que la que se observa a simple vista. Se puede observar que en el primer ejemplo la palabra aparece en la forma flexionada *bites* y la entrada asigna la etiqueta *verb* mientras que en el segundo ejemplo la palabra aparece en su forma regular *bite* con la misma etiqueta. Reconociendo la forma flexionada (*bites*) podemos adicionarle información extra a la etiqueta *verb*, es decir que aparte de guardar la etiqueta *verb* podemos guardar la información de que ocurre en una forma flexionada.

Las entradas de COBUILD exponen las formas derivadas de la palabra que pueden contener los ejemplos. En el ejemplo presentado anteriormente la palabra es *bite* y las formas derivadas de *bite* que muestra la entrada son *bites*, *biting*, *bit* y *bitten*. Con esta información y la etiqueta que fué anotada en COBUILD (*verb* en ambos casos) podemos inferir y generar información adicional para las etiquetas asignadas. Como ya mencionamos, en este caso la forma *bites* (derivada de la palabra *bite*) que aparece en el primer ejemplo posee la etiqueta *verb*. La tarea aquí será reconocer que *bites* es un verbo flexionado a partir de que *bites* está etiquetada como verbo y de que la palabra de la cual deriva es *bite*. Es decir, inferir el tipo de la forma derivada a partir de la palabra y la etiqueta asignada por Cobuild.

Con el objetivo de identificar las formas derivadas de una palabra se desarrollaron reglas y métodos para su reconocimiento, buscando preservar y aprovechar toda la información que ofrece COBUILD. Entonces, a partir de esta información: la palabra, la forma en que ocurre y la etiqueta asignada aplicamos las siguientes reglas para agregar información adicional a la etiqueta gramatical.

Aplicando el algoritmo de extracción y reconocimiento de formas derivadas explicado anteriormente se obtiene un nuevo corpus parcialmente anotado a partir del diccionario Cobuild. A continuación este corpus será procesado y utilizado como corpus de entrenamiento.

3.4. Nuevo Corpus generado

A partir del corpus parcialmente anotado obtenido en el proceso de extracción, se completarán las anotaciones automáticamente con un etiquetador gramatical. Manteniendo las etiquetas gramaticales obtenidas a partir de la in-

Algoritmo 2 Reconocimiento de formas derivadas

Traducir la etiqueta asignada por Cobuild a PenTreeBank

Si la etiqueta obtenida es

JJ:

Si la forma termina en *er* o empieza en *more* o *less* aplicar **JJR**

Si la forma termina en *est* o empieza en *most* o *least* aplicar **JJS**

RB:

Si la forma termina en *er* o empieza en *more* o *less* aplicar **RBR**

Si la forma termina en *est* o empieza en *most* o *least* aplicar **RBS**

NN:

Si la forma termina en *s* aplicar **NNS**

VB:

Si la forma termina en *ed* aplicar **VBD—VBN**

Si la forma termina en *ing* aplicar **VBG**

Si la forma es igual a la palabra y la palabra anterior es *to* aplicar **VBP**

Si la forma termina en *s* aplicar **VBZ**

formación procedente del diccionario Cobuild. Es decir, una vez finalizado el proceso de extracción de información desde el diccionario, se obtiene un corpus nuevo con las etiquetas gramaticales correspondientes a las palabras definidas en el diccionario. A continuación se exhibe un fragmento del mismo:

A
canary NN
is
a
small
yellow
bird
which
sings
beautifully
.
People
sometimes
keep
canaries NNS
in
cages
as
pets
.

Este es el resultado de extracción y reconocimiento de formas flexionadas correspondiente a la entrada de Cobuild:

DICTIONARY_ENTRY

canary

canaries

A canary is a small yellow bird which sings beautifully. People sometimes keep canaries in cages as pets.

countable noun

noun

Se puede apreciar que se ha reconocido *canaries* como el plural de *canary* (etiqueta NNS) y que se han reconocido y extraído los ejemplos de estas palabras asignando las etiquetas gramaticales traducidas a partir de las etiquetas del diccionario correspondientes a *canary* (countable noun/NN) y *canaries* (noun/NNS).

Una vez realizada esta tarea, se procederá a completar las anotaciones gramaticales para todas las palabras restantes. Este proceso se realiza anotando el corpus plano (sin las etiquetas obtenidas mediante Cobuild) con el etiquetador gramatical automático TnT. Luego se une este corpus anotado por TnT con el corpus anotado parcialmente procedente de Cobuild, preservando todas las etiquetas del diccionario. El resultado que se muestra a continuación es un nuevo corpus obtenido a partir de Cobuild, con las anotaciones que este provee y completado con anotaciones obtenidas mediante etiquetación automática utilizando TnT.

A DT

canary NN

is VBZ

a DT

small JJ

yellow JJ

bird NN

which WDT

sings VBZ

beautifully RB

..

People NNS

sometimes RB

keep VB

canaries NNS

in IN

cages NNS

as IN

pets NNS

..

3.5. Experimentación

3.6. Conclusiones