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. Por ejemplo tomemos la oración siguiente:

(1) *There is no asbestos in our products now.*

El resultado de etiquetarla gramaticalmente es:

(2) *There/EX is/VBZ no/DT asbestos/NN in/IN our/PRP products/NNS now/RB ./.*

donde cada palabra está sucedida por una barra oblicua seguida de la etiqueta gramataical asignada. Se puede apreciar por ejemplo que la palabra *is* fué etiquetada como VBZ (verbo de tiempo presente en tercera persona singular), *products* fué etiquetada como NNS (sustantivo plural), étc. Es decir que a cada palabra se le asignó un código que se corresponde con una función gramatical.

A simple vista el etiquetado gramatical parece una tarea trivial o al menos sencilla, sin embargo no es así. La complejidad de este proceso reside en la ambigüedad gramatical inherente al lenguaje. Por ejemplo, la palabra *premio* puede funcionar como sustantivo:

(1) *Gané un premio*

o como verbo

(2) *Por tu esfuerzo te premio*

En (1), *premio* tendría que recibir la etiqueta gramatical NN (sustantivo común) mientras que en (2) tendría que recibir la etiqueta gramatical VB (verbo). Ahora bien, sería interesante conocer que factor indica cual es la etiqueta correspondiente a una palabra ambigüa. En (1) se puede deducir que *premio* es sustantivo porque está precedido por la palabra *un* mientras que en (2) *premio* está precedido por la palabra *te* y a partir de este hecho se puede deducir que en esta caso *premio* funciona como verbo. En fin, las palabras circundantes brindan información vital para deducir el sentido gramatical en palabras ambigüas.

Como se menciona más adelante, el etiquetado gramatical juega un papel importante en áreas de la lingüistica computacional como síntesis del habla, reconocimiento del habla y recuperación de la información. El etiquetado gramatical es realizado manualmente por lingüistas; especialistas en el lenguaje que se ocupan de determinar una etiqueta gramatical para cada palabra. Desde luego que también es realizado automáticamente por computadoras, mediante programas conocidos como etiquetadores gramaticales. Algunas implementaciones actuales de estos programas están basadas en el aprendizaje; tomaun un corpus

[1] anotado correctamente con el cual se entrenan y luego emplean el conocimiento adquirido para etiquetar el corpus objetivo. En esa primer etapa conocida como entrenamiento, el etiquetador gramatical obtiene, procesa y retiene información sobre cada palabra, su etiqueta asignada y su contexto. Posteriormente el etiquetador determina una etiqueta para cada palabra del corpus objetivo, analizando su ubicación y contexto y utilizando el conocimiento adquirido en la etapa previa.

Uno de los grandes problemas del etiquetado gramatical reside en la falta de corpus anotados para utilizar como corpus de 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 millón de palabras. Además de la laboriosidad 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 períodos prolongados. El resultado de este complejo proceso es una tabla de palabras con su correspondiente etiqueta gramatical, como se muestra a continuación:

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

[1]Colección de textos escritos y/o transcripciones del lenguaje oral para cierto idioma

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

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. Esta es la idea central de este trabajo.

1.2. Trabajo realizado

Como se mencionó en la sección anterior, la idea de este trabajo es suplir la falta de corpus de enternamiento utilizando la información de etiquetado que posee un diccionario, generando una nueva fuente de información que servirá para entrenar etiquetadores automáticos. 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. Esta nueva fuente de información se utiliza para entrenar etiquetadores gramaticales automáticos. Una vez entrenados dichos etiquetadores se emplean para etiquetar un corpus objetivo y se analiza el resultado obtenido. Se realiza el mismo procedimiento, pero ahora combinando la nueva fuente de información generada con un corpus de entrenamiento clásico. Se vuelve a etiquetar un corpus objetivo y se analiza el resultado obtenido. Por último se realizan mediciones sobre el rendimiento de los etiquetadores gramaticales entrenados con esta nueva fuente de información y con los corpora 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, étc.) y pronombres personales (Yo, Tú, Él, étc.). 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 sucedios 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 linguísticos 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 desambigüación del sentido de la palabra y en modelos lingüísticos basados en n-gramas utilizados en sistemas de reconocimiento del habla.

1.4. Corpus

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

Los corpora son utilizados por lingüistas para describir naturalmente el lenguaje basados en la evidencia obtenida de sus observaciones. En su trabajo generalmente utilizan operaciones estadísticas sobre los corpora para medir la frecuencia de algún aspecto léxico. Los corpora, grandes cantidades de ocurrencia natural del lenguaje, han ayudado a realizar progresos en diferentes campos del lenguaje como el estudio de fraseología, análisis crítico del discurso, estilismos, lingüística forense, traducciones y enseñanza del lenguaje entre otros.

Diferentes tipos de corpora permiten el análisis de distintas clases de discuros 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 ocurencia. 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
- determinante + sustantivo

- una palabra particular + clase de palabra específica sucediéndola.

Cuando corpora escritos y hablados se hicieron disponibles, los lingüistas comenzaron a analizarlos para verificar patrones o diferencias entre el lenguaje hablado y el lenguaje escrito. Parece que aparte de algunas características obvias como salidas en falso y vacilaciones que se producen en el habla, la utilización de un gran número de expresiones deícticas es más frecuente en los discursos orales. Probablemente esto es debido a los signos linguísticos 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.

Otra área importante de estudio lingüístico de corpora es el cambio histórico de los significados de las palabras y la gramática. 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. En efecto fueron establecidas las diferencias en los aspectos gramaticales concernientes a la voz pasiva.

Por otro lado, en las traducciones es habitual utilizar corpora paralelos que permiten 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 influído en la forma en que se hacen los diccionarios. Segundo los aprendices del lenguaje han sido estudiados para mejorar el conocimiento de los maestros.

Los lingüistas 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 linguístico.

1.4.1. Un poco de historia

El punto de inflexión en corpus linguístico 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 linguística, 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 editor de Boston Houghton-Mifflin se acercó a Kucera para suministrarle el material base de 1 millón de palabras para su nuevo diccionario American Heritage Dictionary (AHD), el primer diccionario que fué compilado utilizando corpus linguístico. El AHD dió el paso innovador de combinar elementos prescriptivos (como debe utilizarse el lenguaje) con información descriptiva (como se usa actualmente).

Otros editores siguieron el ejemplo. El editor 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).

Otros corpora representan más lenguajes, variedades y modos: International Corpus of English, el British National Corpus es una colección de 100 millones de palabras provenientes de textos escritos e inglés hablado creado en los 1990s por un consorcio de editores, 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 construído 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 construído 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 segmento es etiquetado con 7 campos de información. El Quaric 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.

1.4.2. Métodos

Los corpora linguísticos 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 linguística 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 de los corpora de hoy en día está anotado gramaticalmente y aplican algún método para aislar términos que pueden ser interesantes en las palabras circundantes.

2. Desarrollo

2.1. 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 alamacenada 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 arriba presentada se corresponde con una entrada del diccionario. Tienen la característica de poseer una cantidad variable de campos y no es posible identificarlos exactamente. Sin embargo, contienen algunos rasgos comunes: la palabra, sus formas, la pronunciación y uno o más ejemplos donde se indica como se emplea (mediante una etiqueta gramatical). Estas entradas, que

conforman el diccionario COBUILD y que constituyen la fuente de información principal sobre la cual se basa este trabajo, fueron cuidadosamente procesadas y refinadas intentando mantener toda la información disponible. El primer desafío de esta etapa consistió en recuperar las entradas con toda la información gramatical disponible; explícita e implícita. Una primer tarea fué reconocer y registrar información relacionada a las formas flexionadas de la palabra (plurales, pasados, etc.), es decir, obtener información gramatical implícita.

2.1.1. Reconocimiento de formas flexionadas

En muchas entradas del diccionario COBUILD ocurre la palabra, uno o más ejemplos en donde ésta aparece con cierto sentido (indicado por medio de etiquetas gramaticales) pero dentro de los ejemplos hay apariciones de formas flexionadas. Tomemos la siguiente 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 se puede observar una entrada del diccionario para la palabra *bite*, que contiene dos ejemplos de esta palabra con sus respectivas etiquetas:

(1) *If an object or surface bites, it grips another object or surface rather than slipping on it or against it.*

(2) *Let the clutch in slowly until it begins to bite.*

En (2) aparece la palabra *bite* en su forma regular con la etiqueta *verb* mientras que en (1) aparece la forma flexionada *bites* con la etiqueta *verb*. En este caso el ejemplo está ofreciendo más información gramatical que la expuesta por medio de la etiqueta. Reconociendo la forma flexionada (*bites*) podemos adicionarle información extra a la etiqueta *verb*; en vez de guardar la etiqueta de Tree Bank correspondiente a *verb* (VB), en este caso guardaríamos la etiqueta VBZ que contiene más información gramatical que VB.

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) se puede inferir y generar etiquetas de Tree Bank con información adicional. Como ya se mencionó anteriormente, 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 reconocer información adicional a la etiqueta gramatical.

Algoritmo 1 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**

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

2.2. Traducción de etiquetas

Para cada una de sus definiciones, el diccionario COBUILD expone información gramatical expresada mediante etiquetas. Estas etiquetas gramaticales poseen un formato propio. Por ejemplo en la siguiente entrada de COBUILD para la palabra *canary*

```
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 expone la definición y un ejemplo, ambos con información gramatical sobre la palabra:

(1) *A canary is a small yellow bird which sings beautifully.*
(2) *People sometimes keep canaries in cages as pets.*

Se puede apreciar la etiqueta de COUBUILD *countable noun* para la aparición de *canary* en (1) y la etiqueta *noun* para la aparición de *canary* en (2).

Como la idea de este trabajo es producir un corpus anotado a partir de este diccionario para utilizar como fuente de entrenamiento de etiquetadores gramaticales es necesatio que el conjunto de etiquetas empleado sea el mismo que emplea el gold standard para poder medir posteriormente los resultados. Es por eso que se tomó la decisión de traducir estas etiquetas propias de COBUILD en etiquetas de Tree Bank, conjunto con el cual está anotado el gold standard.

A continuación se presenta 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
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

Cuadro 1: Tabla de traducción de etiquetas

Etiqueta COBUILD	**Etiqueta PenTreeBank**
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
verb or verb + object	VB
ergative verb	VB
verb + adjunct	VB
verb + object + adjunct	VB
verb + object $noungrouporreflexive$	VB
verb + object or reporting clause	VB
verb + object $reflexive$	VB
verb + object, or phrasal verb	VB
verb + to-infinitive	VB
ergative verb + adjunct	VB

Cuadro 1: Tabla de traducción de etiquetas

Etiqueta COBUILD	**Etiqueta PenTreeBank**
verb + object + adjunct *to*	VB
verb + object, or verb + adjunct	VB
verb + object + adjunct *with*	VB
verb + adjunct *with*	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 *from*	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	

2.3. 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 información 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 definiadas 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
. .