Hacia la construcción de un corpus etiquetado gramaticalmente para Procesamiento del Lenguaje Natural en el Español Dominicano.

Towards the construction of a Part-of-Speech Tagged corpus for Natural Language Processing Tasks in Dominican Spanish.

Lisibonny Beato-Castro(1), Máximo Pérez Medrano(2), César Méndez Vargas(3), Eva Concepción Abreu(4)

https://doi.org/10.29197/pqs.n3.2017.02

1. Profesora a Tiempo Completo del Departamento de Ingeniería de Sistemas y Computación, Campus Santiago

Directora del Departamento de Ingeniería de Sistemas y Computación, Campus Santiago

Ingeniera de Sistemas y Computación, Magna Cum Laude, por la Pontificia Universidad Católica Madre y Maestra

Maestría en Tecnologías de la Información concentración Ingeniería de Software por la Universidad Politécnica de Madrid

Candidata a Doctora en Desarrollo de Sistemas de Software Complejos por la Universidad Politécnica de Madrid

Email: le.beato@ce.pucmm.edu.do

2. Profesor por Asignatura del Departamento de Ingeniería de Sistemas y Computación, Campus Santiago

Ingeniero de Sistemas y Computación por la Pontificia Universidad Católica Madre y Maestra

Magíster en Tecnología Educativa por la Pontificia Universidad Católica Madre y Maestra

Email: me.perez@ce.pucmm.edu.do

3. Estudiante de término de la carrera de Ingeniería de Sistemas y Computación, Campus Santiago

Email: 20130262@ce.pucmm.edu.do

4. Estudiante de término de la carrera de Ingeniería de Sistemas y Computación, Campus Santiago

Email: 20131319@ce.pucmm.edu.do

ISSN 2519 5107

Resumen:

En este artículo presentamos los detalles del proceso de construcción de las primeras etapas de un corpus anotado gramaticalmente para el español dominicano. Nuestro acercamiento a esta tarea es la de emplear anotadores humanos con conocimientos en lingüística sobre un pequeño conjunto de textos provenientes de microblogs. Adicionalmente, analizamos los principales fenómenos encontrados en el proceso de desarrollo del corpus y discutimos sobre estrategias para su consolidación y posibles aplicaciones futuras.

Palabras clave:

Lingüística Computacional, Lingüística de Corpus, Procesamiento de Lenguaje Natural, Twitter

Abstract:

In this article, we present the details of the first stages of the construction of a Part-of-Speech Tagged corpus for Dominican Spanish. Our approach to this task is to employ human annotators with strong background in linguistics working on a small amount of texts from microblogs. Additionally, we analyze the main phenomena found in the corpus development process and discuss strategies for its consolidation and possible future applications.

Keywords:

Computational Linguistics, Corpus Linguistics, Natural Language Processing, Twitter

1. Introducción

La anotación de un corpus es la práctica de agregar información interpretativa, especialmente lingüística, a un corpus de texto, mediante una codificación añadida a la representación electrónica del propio texto (Garside, Leech, & McEnery, 1997).

Un caso común de anotación de un corpus es el de la anotación morfo-sintáctica, también llamada etiquetado gramatical, en la que una etiqueta se asocia con cada unidad del texto, para indicar su categoría gramatical. Dicha categoría gramatical es una variable lingüística que puede tomar diferentes valores que condicionan la forma morfológica concreta de una palabra (Bosque, 1990). El etiquetado gramatical es una de las tareas más comúnmente realizadas cuando se anotan recursos lingüísticos, debido al rol clave que tiene este tipo de etiquetado en la fase de pre-procesamiento de los sistemas de procesamiento de lenguaje natural (Brill, 2000).

"Hasta donde sabemos, en español existen tres conjuntos de etiquetas gramaticales ampliamente utilizadas en la literatura científica relativa a procesamiento de lenguaje natural en español."

Varios investigadores han llegado a la conclusión de que existe, de una forma u otra, un conjunto de categorías gramaticales comunes a todos los idiomas (Carnie, 2013; Newmeyer, 2005); sin embargo, distintos recursos lingüísticos existentes emplean distintos conjuntos de etiquetas, condicionados por las particularidades del idioma y los propósitos para los que inicialmente dicho recurso fue desarrollado. Por ejemplo, en el idioma inglés, el corpus Brown (Francis & Kucera, 1979) propone un conjunto de 87 etiquetas simples que proporcionan codificaciones distintas para todas las clases de palabras que tienen un comportamiento gramatical distinto y permite la formación de etiquetas compuestas que aumentan el número de etiquetas a 187. Existe también el conjunto de etiquetas del corpus Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993) el cual está basado en Brown, pero propone un conjunto de etiquetas más reducido eliminando redundancia mediante la utilización de información lexical y sintáctica.

Hasta donde sabemos, en español existen tres conjuntos de etiquetas gramaticales ampliamente utilizadas en la literatura científica relativa a procesamiento de lenguaje natural en español:

- El conjunto de etiquetas EAGLES, que se basa en las etiquetas propuestas por el Expert Advisory Group on Language Engineering Standards (EAGLES) para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas. En (Ide & Véronis, 1993) EAGLES establece etiquetas de longitud variable donde cada carácter corresponde a una característica morfológica. El primer carácter en la etiqueta es siempre la categoría gramatical y, dependiendo de la categoría, hay información adicional variable como, por ejemplo, tipo, género, número, entre otros. Este conjunto de etiquetas, al contener información inflexional y léxico-semántica de las unidades del texto, llega a un alto número de etiquetas, del orden de varios cientos.

- Treetagger (Schmid, 1995), la cual es una herramienta para anotar texto en varios idiomas con información de categorías gramaticales y lemas. Fue desarrollado por Helmut Schmid como parte del proyecto TC en el Instituto de Lingüística Computacional de la Universidad de Stu-

ISSN 2519 5107

ttgart. Para el idioma inglés utiliza un conjunto de etiquetas muy reconocido, el de Penn Treebank, pero para el español TreeTagger utiliza un subconjunto de EAGLES con categorías más amplias con poca información morfológica: no contiene información de género y número para sustantivos y adjetivos ni información de modo, tiempo y persona para los verbos. Esto hace que el número de etiquetas de Treetagger sean de alrededor de 75.

- El conjunto de etiquetas de Petrov, Das, & McDonald (2011), quienes proponen un conjunto de 12 etiquetas, a las cuales llaman universales. El propósito de los autores es facilitar la investigación futura relativa a inducción no supervisada de estructura sintáctica y estandarizar buenas prácticas. Adicionalmente al conjunto de etiquetas, proponen el mapeo de conjuntos de etiquetas ya existentes para el etiquetado gramatical en 22 idiomas, incluidos el idioma español.

Estos conjuntos de etiquetas son utilizadas por los diversos etiquetadores gramaticales automáticos en español (Parra Escartín & Martínez Alonso, 2015). Las medidas de desempeño reportadas para estos etiquetadores son dadas por su aplicación a corpus lingüísticos en el idioma español que no tienen las particularidades que el español dominicano presenta, y en muchos casos, no tienen representación suficiente de textos muy cortos, como los de los microblogs, ni la informalidad y el ruido con los que los textos se escriben en las Redes Sociales.

En este artículo mostramos las primeras etapas del proceso de construcción de un corpus etiquetado gramaticalmente cuyo propósito es el de ser usado como Gold Standard en distintas tareas de procesamiento de lenguaje natural en español dominicano.

La concepción de este corpus surge en el contexto de un proyecto de Opinion Mining (Liu, 2012) en Twitter para tweets escritos en español dominicano, en donde se busca la identificación correcta de palabras de ciertas categorías gramaticales, como adjetivos, verbos y adverbios. Es de importancia crucial en la determinación de opiniones, sentimientos y emociones.

En vez de seleccionar el etiquetador gramatical automático de acuerdo a su desempeño reportado sobre corpus existentes, buscamos sentar las bases para un corpus creado con las características propias del entorno sobre el cual va a trabajar nuestra aplicación particular.

Para el desarrollo de este corpus contamos con la colaboración de un grupo de voluntarios los cuales son profesores dominicanos de español a nivel universitario. Estos etiquetaron manualmente una cantidad reducida de tweets. Cada tweet fue etiquetado por dos profesores diferentes y se calculó el nivel de acuerdo en la etiqueta, para finalmente crear una versión definitiva que es utilizada para comparar el desempeño de uno de los etiquetadores automáticos que existen en español. Para los propósitos de nuestro proyecto en este punto, el conjunto de etiquetas universal nos es más conveniente, debido a que no necesitamos información inflexional y a que la granularidad gruesa de este conjunto de etiquetas nos provee de simplicidad a la hora de realizar comparaciones y cálculos de acuerdo entre etiquetadores.

El corpus desarrollado como producto de esta investigación tiene el potencial de ser ampliado en otros proyectos mediante colaboración abierta distribuida, o crowdsourcing, y de ser utilizado en otras aplicaciones en donde sea necesario el análisis de texto escrito en español dominicano, no

solo a nivel morfo-sintáctico sino también a nivel semántico.

El resto de este artículo está organizado de la siguiente manera: La Sección 2 trata en detalle el proceso de construcción de un corpus inicial etiquetado gramaticalmente para tareas de procesamiento de lenguaje natural en español dominicano. En la Sección 3 se discuten los resultados obtenidos. La Sección 4 muestra las conclusiones del trabajo y los pasos futuros.

2. Metodología para la construcción del corpus

2.1. Selección inicial de tweets

Se recopilaron 84 tweets geo-localizados desde la República Dominicana mediante el API de Twitter[1], entre junio del 2016 y octubre de 2016. Este grupo de tweets incluía tweets con fechas que iban desde abril de 2016 a mayo de 2016.

Para elegir estos tweets primero se tomó en cuenta que estuvieran escritos en español, además de estar bien redactados, o sea, que no tuviesen faltas o tuviesen muy pocas faltas ortográficas. Para esto se decidió tomar tweets escritos por personalidades dominicanas que escriben frecuentemente en esta Red Social y que suelen escribir contenido de opinión o con carga emocional. Se tomó en cuenta que el conjunto de datos tuviese una representación de tweets de distinta longitud: desde tweets muy cortos hasta tweets de varias oraciones.

Otro aspecto tomado en cuenta para la selección es que estos no presentasen las convenciones especiales de Twitter como: hashtags, menciones, direcciones URL, emojis y emoticonos.

1 https://dev.twitter.com/overview/api

2.2. Correcciones sobre los tweets

A pesar de que se obtuvo un conjunto de tweets con cierto grado de calidad en su escritura, se hizo necesario hacer correcciones a algunos de ellos, en su gran mayoría por situaciones del uso no adecuado de signos de puntuación, pequeños errores ortográficos y omisión de tildes. Esto, para tratar de disminuir la posibilidad de que los etiquetadores, manuales y automáticos, etiqueten de forma errónea las unidades de texto en esta etapa de nuestra investigación.

A continuación, algunos ejemplos de las correcciones realizadas:

Tu falta de verguenza y tu falta de pantalones han hecho que la corrupción y la impunidad reynen en nuestro país **(Original)** Tu falta de vergüenza y tu falta de pantalones han hecho que la corrupción y la impunidad reinen en nuestro país. **(Corregido)**
El debate lo está ganando el moderador! **(Original)** ¡El debate lo está ganando el moderador! **(Corregido)**
Este domingo estaré con amigos sembrando árboles. Si sigue la depredación, además de sembrar, tendremos q ir a quemar camiones. **(Original)** Este domingo estaré con amigos sembrando árboles. Si sigue la depredación, además de sembrar, tendremos que ir a quemar camiones. **(Corregido)**

ISSN 2519 5107

El objetivo de hacer una selección de tweets con esta condición de calidad en su ortografía se debe a que en la presente etapa de la investigación uno de los objetivos es probar los niveles exactitud y precisión de los etiquetadores gramaticales automáticos que manejen el idioma español, y esto garantiza que los mismos clasifiquen con la menor dificultad posible las palabras que conforman los tweets. Cabe aclarar que somos conscientes de que estas situaciones especiales que provocan las características intrínsecas de los tweets deben ser abordadas en futuros trabajos como parte de nuestra investigación, y que debemos buscar soluciones a las situaciones particulares que presentan los tweets escritos en español dominicano en este sentido.

2.3. Etiquetado manual de los tweets

Al principio de la investigación creíamos que la tarea de etiquetado gramatical para construir el corpus con el que trabajaremos era una tarea de una complejidad baja, debido a la brevedad del texto y al reducido conjunto de etiquetas gramaticales a utilizar, y que podía ser abordada por cuatro voluntarios con estudios universitarios, pero sin ninguna experiencia previa en la realización de esta tarea. Para comprobar el nivel de aptitud que los voluntarios tenían para esta tarea, se les entregó una pequeña cantidad de tweets a cada uno. La primera tarea que debían hacer sobre los tweets era un proceso de tokenización (Webster & Kit, 1992), que se refiere a la segmentación del texto en unidades lingüísticas antes de que cualquier tipo de análisis sea realizado sobre este. Para nuestros propósitos, estos tokens pueden adoptar la forma de unigramas o n-gramas (Manning & Schütze, 1999). Posteriormente debía colocar a cada token la etiqueta de la categoría gramatical que mejor describiera la función que ejercía dicho token en el tweet, ayudados, en caso de ser necesario, por el diccionario en línea de la Real Academia Española de la Lengua[2]. Este ejercicio inicial evidenció que, para ambas tareas, los sujetos tuvieron varias dudas en cada uno de los tweets que analizaron que no siempre pudieron resolver utilizando sus propios conocimientos ni el diccionario en línea. La tarea resultó larga, compleja e inconclusa en la mayoría de los casos.

Se decidió, por esta razón, en vez de contar con voluntarios no entrenados, contar con la colaboración de un grupo de trece académicos dominicanos del área de español, los cuales, por su formación, pudieron abordarla sin las complicaciones anteriormente mencionadas.

El proceso seguido por este nuevo grupo de voluntarios se describe a continuación:

- Se repartieron 84 tweets entre los profesores de forma equitativa. Nos referiremos a este conjunto de tweets de cada profesor como "version_1".

- Cada tweet de la "version_1" tenía que ser tokenizado por el profesor del modo que él considerara más apropiado para, posteriormente, etiquetar cada token con la categoría gramatical a la que él considera pertenece de acuerdo al contexto donde aparece en el tweet. Estas etiquetas deben ser seleccionadas del conjunto de etiquetas universales propuestas por Petrov. También se le pidió especificar el lema o palabra base que corresponde a dicho token. Se les indicó a los profesores que podían tomarse el tiempo que necesitaran para el etiquetado y se les reiteró que lo importante era la calidad del etiquetado, no la rapidez con que se realizaba.

2 http://dle.rae.es

- Una vez terminada esta tarea, el profesor debería realizar la misma labor con el primer grupo de tweets de otro profesor. En este caso, a este segundo grupo de tweets trabajados por el profesor le llamamos "version_2". El profesor recibe los tweets sin ninguna información de lo que pasó con ellos en la etapa previa con el profesor que los etiqueta originalmente como su primer grupo. La idea es que cada tweet sea etiquetado por dos profesores diferentes para posteriormente ser comparado el grado de acuerdo en el etiquetado de ambos grupos.

2.4. Versión Final de los Tweets Etiquetados

Después de las tareas anteriores se planificaron varias sesiones en donde nos reunimos con los profesores para verificar diferencias en la tokenización y etiquetado de un mismo tweet en ambas versiones, discutirlas y definir una versión final etiquetada para su posterior comparación con las versiones originales etiquetadas y librerías automáticas de etiquetado gramatical en español.

Con el liderazgo de uno de los lingüistas voluntarios y apoyados en el Diccionario panhispánico de dudas (de la Lengua, Española, & de la Lengua, 2005) y el manual de la Nueva gramática de la lengua española (Española, 2010), se llevaron a cabo varias sesiones para analizar y discutir el porqué de las diferencias encontradas en las dos versiones etiquetadas de los tweets y determinar cuáles eran las acciones más adecuadas que tomar para una versión definitiva, tanto en la segmentación de las unidades lexicales de los tweets como de la etiqueta gramatical que debía ser colocada.

2.4.1. Tokenización

En lo referente al proceso de tokenización, los lingüistas concuerdan en que los verbos auxiliares no deben ir separados de los verbos a los cuales proporcionan información gramatical adicional.

> Prefiero/VERB Vivir/NOUN Y/CONJ Perder/VERB Que/CONJ No/ADV Haber_vivido/VERB Nada/PRON ./.

En el caso del ejemplo, el token asumiría la forma del bigrama haber_vivido y se le colocaría una única etiqueta gramatical.

En el caso de lugares con nombre compuesto, así como también otras entidades, tales como programas de televisión o nombres de instituciones, la forma de proceder en la segmentación sería unir todos los unigramas que los conforman en un único token.

> Todos/PRON en/ADP Santo_Domingo_Este/NOUN a/ADP votar/VERB en/ADP la/DET boleta/NOUN 15/NUM por/ADP el/DET diputado/NOUN que/PRON te/PRON representa/VERB en/ADP el/DET congreso/NOUN ./.

En el ejemplo, el token sería el trigrama Santo_Domingo_Este, también con una única etiqueta gramatical.

Los voluntarios consideran que hay ciertos pronombres, conjunciones, adverbios y partículas que deben tener la forma de n-gramas, debido a que como unigramas no tiene sentido práctico atribuirles una etiqueta gramatical separada. Algunos ejemplos encontrados son los pronombres lo_que y el_que, la conjunción para_que, los adverbios a_diario y a_la_vez y las partículas a_veces, al_fin y a_favor.

Los siguientes tweets presentan algunas de estas situaciones:

> 4/NUM años/NOUN de/ADP mala/ADJ suerte/NOUN para/ADP el_que/PRON vote/VERB por/ADP el/DET PLD/NOUN ./

ISSN 2519 5107

Aquí/ADV ,/. a_veces/PRT ,/. valen/VERB más/ADV las/DET relaciones/NOUN que/CONJ el/DET talento/NOUN .../. ¡/. lo/PRON vemos/VERB a_diario/ADV !/. No/ADV hay/VERB congruencia/NOUN en/ADP algunas/DET cosas/NOUN ./

2.4.2. Etiquetado Gramatical

Fueron diversas las discrepancias encontradas entre los etiquetadores en ambas versiones del etiquetado. Las siguientes subsecciones detallan estos casos.

2.4.2.1. Adjetivos y determinantes

Un caso que se encontró con mucha frecuencia fue el del etiquetado de palabras que indican demostrativo, cantidad indefinida o posesivo como un adjetivo. Tales son las palabras mi, este y nuestro. La gramática moderna propone que estas palabras deben considerarse como determinantes y es así como fueron etiquetadas para esta versión final.

Por ejemplo, en el siguiente tweet la palabra este se etiquetó finalmente como determinante:

Este/DET espectacular/ADJ hair_look/X lo/PRON logramos/VERB gracias/PRT a/ADP las/DET extensiones/NOUN de/ADP afrolatinard/X ./.

2.4.2.2. Adverbios y adjetivos

Hubo, también, confusión entre el uso de varias palabras y su función como adverbio o como adjetivo dentro del texto. Tal es el caso de la palabra mejor, que en algunos tweets funciona como adverbio, por la condición de que no varía cuando acompaña a palabras que pueden cambiar en género y número, y en otras como adjetivo, en donde si cambia el género y número de las palabras, dicho adjetivo debe también modificar su forma.

En el siguiente caso es adverbio, porque permanece invariante si la oración pasara al plural:

¡/. Dios/NOUN Ya/ADV Sabe/VERB Lo_que/PRON Necesitas/VERB ,/. Mejor/ADV Agradécele/VERB Lo_que/PRON Tienes/VERB !/.

En el siguiente caso es adjetivo, porque debería pluralizarse a mejores si la palabra a la que acompaña fuera asfaltos:

Bien/ADV pudiste/VERB gastar/VERB este/DET dinero/NOUN en/ADP un/DET mejor/ADJ asfalto/NOUN y/CONJ no/ADV ese/DET disparate/NOUN que/PRON están/VERB haciendo/VERB y/CONJ que/PRON ya/ADV la/DET lluvia/NOUN dañó/VERB ./.

2.4.2.3. Adjetivos y verbos como sustantivos

El caso de palabras que tradicionalmente son adjetivos y verbos, pero que en ciertos contextos asumen la forma de sustantivos, también fue encontrado en estas revisiones. En los siguientes tweets, los adjetivos malo y peor se etiquetaron en la versión final como sustantivos y lo mismo se hizo con los verbos conformarse, dejar e insistir:

Lo/DET malo/NOUN de/ADP cuando/ADV a/ADP uno/PRON le/PRON importa/VERB tanto/ADV alguien/PRON es/VERB que/PRON la/DET más/ADV mínima/ADJ tontería/NOUN a/ADP uno/PRON le/PRON duele/VERB mucho/ADV ./.

Lo/DET peor/NOUN es/VERB que/CONJ no/ADV podré/VERB ver/VERB los/DET juegos/NOUN de/ADP la/DET final/NOUN ./.

¡/. Conformarse/NOUN Y/CONJ Dejar/NOUN De/ADP Insistir/NOUN Es/VERB Como/ADV Ver/VERB A/ADP Alguien/PRON Ahogándose/ADV Y/CONJ Dejarlo/VERB Morir/VERB !/.

2.4.2.4. Usos de las palabras que y qué

Un caso especial es el etiquetado de las palabras que y qué en ambas versiones etiquetadas de los tweets. Son cuatro las distintas categorías gramaticales que se les colocaron a las mismas en la versión final: Conjunción, Partícula, Pronombre y Adjetivo.

En el siguiente tweet el que funciona como conjunción:

> Los/DET electores/NOUN jóvenes/ADJ deben/VERB saber/VERB que/PRT el/DET candidato/NOUN del/ADP PRM/NOUN no/ADV cree/VERB en/ADP la/DET tecnología/NOUN ./.

El tweet siguiente muestra un uso del que como partícula:

> Los/DET electores/NOUN jóvenes/ADJ deben/VERB saber/VERB que/PRT el/DET candidato/NOUN del/ADP PRM/NOUN no/ADV cree/VERB en/ADP la/DET tecnología/NOUN ./.

Su uso como pronombre puede verse en el siguiente tweet:

> Las/DET personas/NOUN que/PRON hacen/VERB daño/NOUN a/ADP un/DET país/NOUN usan/VERB traje/NOUN y/CONJ corbata/NOUN ,/. no/ADV tatuajes/NOUN y/CONJ dreadlocks/X ./.

En el siguiente caso el qué ejerce de modificador del sustantivo manera, por lo que es un adjetivo:

> Qué/ADJ manera/NOUN de/ADP quererse/VERB ganar/VERB a/ADP una/DET ciudad/NOUN que/PRON te/PRON odia/VERB ./.

2.4.2.5. Siglas y números como sustantivos

Varios etiquetadores no etiquetaron las siglas en los tweets en donde aparecían, sin embargo, lo correcto es colocarles la etiqueta de sustantivo. Tal es el caso de las siglas PLD en el siguiente tweet:

> Mi/DET programa/NOUN tuvo/VERB que/CONJ salir/VERB del/ADP aire/NOUN por/ADP esta/DET semana/NOUN gracias/ADP al/ADP PLD/NOUN ./.

También hubo diferencias en el etiquetado de algunas cantidades numéricas cuando estas ejercían la función de sustantivo. En el tweet siguiente el número 15 se debe etiquetar como sustantivo:

> Encuestas/NOUN no/ADV son/VERB votos/NOUN ./. Nos/PRON vemos/VERB el/DET 15/NOUN de/ADP mayo/NOUN ./.

No así en el siguiente tweet en donde 7:00 debe ser etiquetado como número:

> Yo/PRON entiendo/VERB que/CONJ están/VERB perdidos/ADJ ,/. pero/CONJ llamar/VERB a/ADP las/DET 7:00/NUM am/X es/VERB desesperación/NOUN ./.

2.5. Medidas para el cálculo de acuerdo entre anotadores:

En nuestra investigación reportamos el acuerdo general en las anotaciones y no identificamos de forma separada aquella que hizo cada profesor con respecto a otro. Esto da como resultado dos versiones de las anotaciones, version_1 y version_2, en las cuales se comparan las dos versiones etiquetadas de cada tweet.

Aparte de comparar las dos versiones iniciales de tweets anotados (version_1 y version_2) también realizamos una comparación de cada una de ellas con la versión producida después de la etapa de discusiones y limpieza con los etiquetadores, a la que llamamos “version_final”.

ISSN 2519 5107

Debido al hecho de que los etiquetadores podían tokenizar los tweets de forma libre y, por tanto, las diferentes versiones podrían tener distinta cantidad de tokens, aplicamos la estrategia de comparar a nivel de unigramas y cuando el token era un n-grama lo separamos y colocamos en cada unigrama resultante la misma etiqueta gramatical del token original.

Para estas comparaciones utilizamos la exactitud (Witten, Frank, Hall, & Pal, 2011) entre las versiones X y Y, una métrica ampliamente utilizada en el desarrollo de corpus para tareas de procesamiento de lenguaje natural. Se calcula de la siguiente manera:

$$\text{Exactitud (X,Y)} = \frac{\text{cantidad de tokens del corpus etiquetados identicamente en X y Y}}{\text{cantidad total de tokens en el corpus}}$$

La versión final contra el etiquetado realizado por una herramienta automática también es objeto de comparación en nuestro trabajo. En este caso se busca verificar el desempeño del etiquetador automático con respecto a la versión final, que se considera como correcta. Aparte de utilizar la exactitud para esta comparación, se calcula también el desempeño por cada etiqueta gramatical T, utilizando las medidas de precisión y exhaustividad (Witten et al., 2016).

Siendo:

Verdaderos Positivos (VP): Cantidad de tokens etiquetados como T por ambos etiquetadores.
Falsos Positivos (FP): Cantidad de tokens etiquetados como no T en la versión final, pero etiquetado como T por el etiquetador automático.
Falsos Negativos (FN): Cantidad de tokens etiquetados como T en la versión final pero etiquetado como no T por el etiquetador automático.

Las medidas de precisión y exhaustividad se calculan de la siguiente manera:

$$\text{Precisión (T)} = \frac{VP}{VP+FP}$$

$$\text{Exhaustividad (T)} = \frac{VP}{VP+FN}$$

Se calcula también la medida F (Witten et al., 2016), o F-Measure, que es la media armónica entre la precisión y la exhaustividad:

$$\text{F- Measure (T)} = \frac{2 \text{ x Precisión x Exhaustividad}}{\text{Precisión x Exhaustividad}}$$

3. Resultados

Acuerdo general entre versiones

En la tabla 1 se muestra el porcentaje de acuerdo entre las distintas versiones etiquetadas, el cual indica la proporción de tokens etiquetados de la misma forma por ambos anotadores, con respecto al total de tokens del conjunto de tweets. El acuerdo entre los anotadores de las versiones 1 y 2, consideradas como iniciales, es de un 86.72%. El acuerdo de cada una de estas versiones con respecto a la versión final (después de las sesiones de discusión y revisión) es de 91.68% y 92.60%, respectivamente.

Cantidad Total de Tokens		1,310
	Cantidad Tokens Etiquetados Igual	**Acuerdo**
Version_1 vs Version_2	1,136	86.72%
Version_1 vs Version_Final	1,201	91.68%
Version_2 vs Version_Final	1,213	92.60%

Tabla 1. Acuerdo entre etiquetadores para las distintas versiones de los tweets etiquetados.

Aunque los niveles de acuerdo de cada versión inicial con la versión final son superiores a los acuerdos entre ellas mismas, ninguna de las 2 versiones supera a los niveles de acuerdo reportados por herramientas automáticas de etiquetado gramatical que alcanzan una exactitud en el etiquetado gramatical en español de entre 97% y 98% (Padró & Stanilovsky, 2012).

La tabla 2 muestra las diferencias en el etiquetado por cada categoría gramatical entre las versiones 1 y 2 del mismo. La columna "Coincidencias" muestra el total de tokens que fueron etiquetados con la misma etiqueta en ambas versiones y la columna "Diferencias" muestra la cantidad de tokens que, en alguna de las dos versiones, no fue etiquetada con dicha etiqueta, pero en la otra versión si lo fue. La columna "% total de diferencias" muestra el porcentaje que representa esa diferencia con respecto al total. Debe notarse que la suma de todas las diferencias es de un 200% debido a que cada diferencia se cuenta dos veces. Esto es así porque no se asume que ninguna de las dos versiones es la correcta y se cuenta una diferencia en ambas direcciones: para el token X la version_1 tiene una diferencia con respecto a la versión_2 y la version_2 tiene una diferencia con respecto a la versión_1.

	Etiqueta	**Coincidencias**	**Diferencias**	**% total de diferencias**	
1	PRON	84	57	32.76	
2	ADJ	48	53	30.46	
3	ADV	87	41	23.56	
4	NOUN	221	41	23.56	
5	DET	115	40	22.99	
6	PRT	1	38	21.84	
7	ADP	132	26	14.94	
8	CONJ	44	23	13.22	
9	VERB	221	17	9.77	
10	X	11	10	5.75	
11	NUM	3	2	1.15	
12	.	169	0	0.00	
Total			348	200.00	

Tabla 2. Coincidencias y diferencias por etiqueta gramatical entre la version_1 y version_2 del etiquetado.

ISSN 2519 5107

El acuerdo entre los etiquetadores es bajo en categorías gramaticales importantes para algunas aplicaciones de procesamiento de lenguaje natural, tales como la de opinion mining. Existe un alto número de diferencias relativas en las etiquetas PRON, ADJ y ADV que representan a los pronombres, adjetivos y adverbios, respectivamente. Estas últimas dos categorías son un ejemplo de categorías importantes en el contexto de nuestro proyecto.

Un caso a resaltar es el de la etiqueta PRT, correspondiente a la categoría gramatical partícula, en la cual solo se coincidió en una ocasión y está involucrada en 38 casos de diferencias absolutas entre las dos versiones.

En la tabla 3 se muestran las etiquetas donde se verifican la mayor cantidad de diferencias entre las dos versiones iniciales.

La columna "Frecuencia Etiqueta 1" representa la suma para ambas versiones de la cantidad de tokens etiquetados con la "Etiqueta 1", y la columna "Frecuencia Etiqueta 2" representa la suma para ambas versiones de la cantidad de tokens etiquetados con la "Etiqueta 2". La columna "Diferencias" representa la cantidad de ocasiones en las que un anotador, para un mismo token, colocó la "Etiqueta 1" y el otro la "Etiqueta 2". El "% total de diferencias" indica la proporción que representa este tipo particular de diferencia con respecto al total de diferencias observadas en ambas versiones.

Etiqueta 1	Etiqueta 2	Frecuencia etiqueta 1 (f1)	Frecuencia Etiqueta 2 (f2)	f1+f2	Diferencias	% total de Diferencias
DET	PRON	270	225	495	17	9.29
PRON	CONJ	225	111	336	14	7.65
DET	ADJ	270	149	419	13	7.10
NOUN	ADJ	483	149	632	12	6.56
ADV	ADJ	215	149	364	11	6.01
ADP	PRT	290	40	330	10	5.46
NOUN	X	483	32	515	9	4.92
…	…	…	…	…	…	…
Total					183	100.00

Tabla 3. Pares de etiquetas gramaticales con los porcentajes de diferencias más altos entre la version_1 y version_2 del etiquetado.

En esta tabla se puede ver que las etiquetas gramaticales DET y ADJ, correspondientes a determinantes y adjetivos, están involucradas en varios casos de diferencias. Es importante resaltar el caso que las involucra a las dos como parte de un par, ya que en las sesiones de discusión para construir la versión final se resaltó el hecho de que actualmente algunas palabras, que antes se denominaban adjetivos, las nuevas reglas de la gramática española las consideran determinantes. Se ve, también, que el porcentaje de confusión más alto entre las dos versiones sucede con las etiquetas DET y PRON, que representan a los determinantes y los pronombres respectivamente. Evaluación

de la versión final etiquetada versus herramienta de etiquetado gramatical automática

Como forma de verificar qué tan bien la versión final—que asumimos como correcta para el español dominicano—es etiquetada por una herramienta automática, decidimos utilizar Freeling (Padró & Stanilovsky, 2012) en su versión 4.0, el cual es un conjunto de analizadores del lenguaje de código abierto para varios idiomas, incluido el español. Entre estos analizadores se encuentran tokenizadores, analizadores morfológicos, etiquetadores gramaticales, entre otros. La elección de esta herramienta para la evaluación se debe a que, de las consultadas, es la que emplea mayor variedad de estilos de tokenización y, dado que los voluntarios tenían total libertad para la realización de esta tarea, es la que podría producir tokens más similares a los resultantes de la anotación de dichos voluntarios.

El etiquetador gramatical de Freeling emplea dos acercamientos distintos: un modelo híbrido llamado relax, que combina reglas gramaticales estadísticas y reglas manuales, y otro basado en un modelo oculto de Markov (HMM) (Blunsom, 2004) . En nuestro caso, utilizamos este último modelo al que accedemos mediante una librería wrapper para el lenguaje de programación Python sobre una librería existente de Freeling en el lenguaje de programación C++.

Aunque por defecto Freeling utiliza el conjunto de etiquetas EAGLES para español, mapeamos dichas etiquetas al conjunto de etiquetas universal, mediante un recurso desarrollado en el trabajo de (Parra Escartín & Martínez Alonso, 2015).

Existen dos fenómenos que Freeling maneja de forma especial y que en la comparativa con la versión final pueden provocar significativas diferencias, ya que la versión final los trata de forma distinta:

- Las contracciones al y del, las cuales convierte a a el y de el, y cada palabra es etiquetada por Freeling de forma separada.

- Los verbos con enclíticos, los cuales son palabras compuestas que se forman añadiendo al verbo pronombres como me, te, se, lo, entre otros, inmediatamente después del verbo. La herramienta en este caso separa el verbo del enclítico y, por ejemplo, la palabra irse, se convertiría en dos palabras, ir y se, las cuales se etiquetarán de forma separada.

Para poder realizar la comparativa, lo que hacemos con los tokens etiquetados devueltos por Freeling es, en el caso de las contracciones, volver a unirlas y colocar la etiqueta de la primera palabra, en este caso la de a o la de de, según corresponda. Con los verbos con enclíticos lo que hacemos es unir las partes componentes y colocarle la etiqueta correspondiente a verbo.

Una vez realizadas estas modificaciones a lo devuelto por Freeling, se calculó la exactitud de la misma sobre la versión final y el resultado obtenido fue de un 85.34%, el cual es un número significativamente inferior al reportado por esta herramienta sobre conocidos corpus en el idioma español.

En la tabla 4 reportamos las medidas de precisión, exhaustividad y F-Measure por etiqueta:

ISSN 2519 5107

Etiqueta	Precisión	Exhaustividad	F-Measure
NUM	1.0000	1.0000	1.0000
.	0.9941	1.0000	0.9970
DET	0.9079	0.9388	0.9231
ADP	0.8980	0.9103	0.9041
VERB	0.9082	0.8319	0.8684
ADV	0.9529	0.7570	0.8438
NOUN	0.6935	0.9395	0.7979
PRON	0.9204	0.7232	0.8100
ADJ	0.8780	0.5902	0.7059
CONJ	0.6986	0.7846	0.7391
PRT	0.0000	0.0000	0.0000
X	0.0000	0.0000	0.0000

Tabla 4. Precisión, exhaustividad y F-measure de cada etiqueta gramatical dada por la herramienta Freeling sobre la versión final etiquetada.

Puede observarse que etiquetas gramaticales que tienen mucha carga semántica en diversos contextos y aplicaciones de procesamiento de lenguaje natural, tales como adjetivos, verbos, adverbios y sustantivos, tienen unos valores de F-measure por debajo del 90%. En el caso de los sustantivos, de forma especial, la precisión que se obtiene es significativamente baja aunque la exhaustividad de esta categoría ayuda a elevar el valor del F-measure.

Las etiquetas DET y ADP, referentes a determinantes y adposiciones, respectivamente, tienen valores de precisión y exhaustividad cercanos o superiores al 90%.

Es significativo el hecho de que las etiquetas PRT y X, relativas a partículas y Otras, respectivamente, no fueron acertadas en ningún momento por Freeling de forma correcta.

1. Conclusiones y trabajos futuros

En nuestra investigación hemos presentado los primeros pasos hacia la construcción de un corpus etiquetado gramaticalmente que puede ser utilizado en tareas de procesamiento de lenguaje natural para español dominicano.

En nuestro trabajo se evidenció que hay ciertas categorías gramaticales que presentan importantes diferencias en su etiquetado cuando se comparan con los resultados devueltos por una herramienta automática. El caso de las partículas, verbos, adjetivos, sustantivos y adverbios merece un análisis más exhaustivo para determinar las causas de estas diferencias.

En su estado actual, este corpus contiene textos cortos provenientes de microblogs y ha sido etiquetado manualmente por voluntarios entre-

nados para la tarea. En el futuro este corpus puede ser ampliado con textos dominicanos de otros contextos, por voluntarios no entrenados utilizando una herramienta que les guíe y que puede estar basada en reglas que se deriven de los resultados de la presente investigación.

Además, un corpus ampliado en el futuro podría contener texto con expresiones propias del español dominicano que puedan ser correctamente etiquetadas gramaticalmente e incluso contener texto informal con errores ortográficos, que también pueda ser etiquetado correctamente de acuerdo al contexto en donde aparece.

Nuestra investigación, también, deja abierta la posibilidad de comparar este corpus con los resultados devueltos por varios etiquetadores automáticos en español y determinar cuál o cuáles de estas herramientas se adaptan mejor a los propósitos particulares de tareas de procesamiento de lenguaje natural específicas, en español dominicano.

Agradecimientos

Agradecemos al V Fondo Concursable de Investigación PUCMM por financiar la presente investigación.

También damos las gracias al Departamento de Español de la Pontificia Universidad Católica Madre y Maestra Campus Santiago, en la figura de su director, el Prof. Francisco Cruz, por su colaboración y la de sus profesores en este trabajo de investigación.

Bibliografía

Blunsom, P. (2004). Hidden markov models. Lecture notes, August, 15, 18-19.

Bosque, I. (1990). Las categorías gramaticales: relaciones y diferencias: Síntesis.

Brill, E. (2000). Part-of-speech tagging. Handbook of natural language processing, 403-414.
Carnie, A. (2013). Syntax: A generative introduction: John Wiley & Sons.

de la Lengua, A. d. A., Española, E. A., & de la Lengua, A. d. A. (2005). Diccionario panhispánico de dudas: Real Academia Española.

Española, R. R. A. (2010). Nueva gramática de la lengua española manual: Espasa.

Francis, W. N., & Kucera, H. (1979). Brown corpus manual. Brown University, 2.

Garside, R., Leech, G. N., & McEnery, T. (1997). Corpus annotation: linguistic information from computer text corpora: Taylor & Francis.

Ide, N., & Véronis, J. (1993). Background and context for the development of a Corpus Encoding Standard. Retrieved from http://www.cs.vassar.edu/CES/CES3.ps.gz

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing (Vol. 999): MIT Press.

ISSN 2519 5107

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational linguistics, 19(2), 313-330.

Newmeyer, F. J. (2005). Possible and probable languages: A generative perspective on linguistic typology: Oxford University Press on Demand.

Padró, L., & Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. Paper presented at the LREC2012.

Parra Escartín, C., & Martínez Alonso, H. (2015). Choosing a Spanish Part-of-Speech tagger for a lexically sensitive task.

Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. arXiv preprint arXiv:1104.2086.

Schmid, H. (1995). Treetagger| a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 43, 28.

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. Paper presented at the Proceedings of the 14th conference on Computational linguistics-Volume 4.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2011). Output: Knowledge Representation. Data Mining: Practical machine learning tools and techniques (pp. 73). Kaufmann, Burlington.