Creación automática de *treebanks* para el español americano

Minor Sandí Salazar
PCI
Universidad de Costa Rica
San Pedro, Costa Rica
minor.sandisalazar@ucr.ac.cr

Gabriela Marín Raventós
ECCI/CITIC
Universidad de Costa Rica
San Pedro, Costa Rica
gabriela.marin@ucr.ac.cr

***Abstract*— Para un lingüista o computólogo orientado al análisis de textos en español, analizar oraciones sintácticamente puede ser un esfuerzo de mucho tiempo cuando la cantidad de oraciones es elevada. Esta labor es más delicada si se toma en cuenta la variante del español que se emplee. Más aún, seleccionar cómo se etiquetan las palabras según la función puede complicar el proceso y el su impacto al compartir el conocimiento adquirido. Esta investigación propone realizar parte del esfuerzo en forma automática, utilizando reglas gramaticales, con el fin de analizar las oraciones sintácticamente y etiquetarlas con dependencias universales; un etiquetado estándar y nemónico, capaz de ser aplicado a diferentes idiomas.**

Keywords—treebank, análisis sintáctico, dependencias universales, reglas gramaticales

I. INTRODUCCIÓN

Desde épocas tan tempranas como la Edad Media, existía el interés por analizar oraciones gramaticalmente. Roger Bacon, en el siglo XIII ya afirmaba que ¨…en su sustancia, la gramática es una y la misma en todos los lenguajes…¨[1]

Esta pasión por el análisis de las oraciones se ha mantenido a lo largo de los siglos. Si se avanza a los años cincuenta del siglo pasado, puede descubrirse cómo las Ciencias de la Computación y la Lingüística comenzaron a cruzar sus caminos. Las emergentes teorías del lenguaje y los primeros lenguajes de programación que surgieron estrecharon sus pasos conforme avanzaron los años [2].

La década siguiente contempló la creación de los primeros textos con información morfológica y sintáctica para el inglés, llamados *treebanks*. Este proceso que culminó en los años noventa con la aparición de un texto analizado morfológica y sintácticamente, el Penn Treebank [3]. El Penn Treebank fue la base para el desarrollo de nuevos *treebanks* para otros idiomas, entre ellos el español.

Desde mediados de los años noventa el análisis sintáctico para el español comenzó a ver incrementadas sus investigaciones, enfatizándose el español de la Península Ibérica. Estos esfuerzos culminaron con la creación de textos analizados morfológica y sintácticamente como Áncora [4].

Todas estas investigaciones tienen como elemento común la búsqueda de la automatización parcial o total de los análisis morfosintácticos. Además, cada equipo de expertos ha generado su propio etiquetado de categorías gramaticales y funciones sintácticas, siguiendo una corriente de análisis diferente, lo que puede complicar la transmisión del conocimiento.

El avance del campo en inglés es mucho mayor que el del español. En nuestro caso, el español americano como un todo, existe un agravante adicional, el rezago con respecto al español peninsular [5]. Para conocer las diferencias puede remitirse al documento citado.

Nuestra investigación nace de la necesidad de analizar textos cada vez más voluminosos de las redes sociales. La anotación manual típicamente la realizan lingüistas. Pero si se tiene un millón de oraciones, ¿cuánto tiempo se duraría? ¿cuál etiquetado debería ser utilizado para compartir resultados y que sea fácilmente entendible? ¿cómo explicaría el proceso empleado de una forma que pueda ser reproducible en otros lugares, textos e incluso idiomas?

Este artículo relata el proceso por el cual se analizaron sintácticamente conjuntos de oraciones en español americano cuyas palabras poseen un sistema de etiquetado utilizado previamente para análisis de textos en diferentes lenguas. Como hipótesis se propuso la posibilidad de automatizar este análisis mediante el uso de reglas gramaticales, con la finalidad de reducir el involucramiento de anotadores humanos.

Para validar los resultados obtenidos en el proceso descrito se eligieron métricas conocidas en el área de la Lingüística Computacional. Estas métricas comparan los textos anotados creados por la aplicación implementada con sus originales, para validar su eficacia. Finalmente, se agregan las conclusiones obtenidas luego de la investigación, así como trabajos futuros a partir de las áreas de mejora encontradas.

II. CONSTRUCCION DEL PROCESO AUTOMATIZADO DE ETIQUETADO

Con el fin de garantizar el poder compartir los resultados de los etiquetados a la hora de automatizar el proceso de creación de *treebanks* se seleccionó el etiquetado conocido como dependencias universales, el cual es fácil de entender y utilizado en varios idiomas [6]. Estos etiquetados incluyen la identificación de categorías morfológicas y funciones

sintácticas para cada oración. A continuación, se describen las etapas llevadas a cabo.

A. *Creación de un treebank de prueba*

Se investigó mediante una búsqueda literaria, textos en español analizados morfológicamente. Se recolectaron sus características generales, con las que se creó un cuadro comparativo para seleccionar los textos que más se ajustaban a los objetivos de la investigación.

De este cuadro se seleccionó un *treebank* etiquetado con dependencias universales [6]. Este texto se constituyó en nuestro ¨*treebank* dorado¨, utilizado para evaluar el proceso de automatización.

Al *treebank* dorado se le eliminó el etiquetado sintáctico, y a este texto se le llamó corpus anotado morfológicamente, constituyéndose en el corpus de prueba.

B. *Implementación de un proceso automático para el análisis sintáctico*

Para el análisis de los textos anotados seleccionados se propuso e implementó la siguiente serie de etapas que se resumen en la Figura 1.

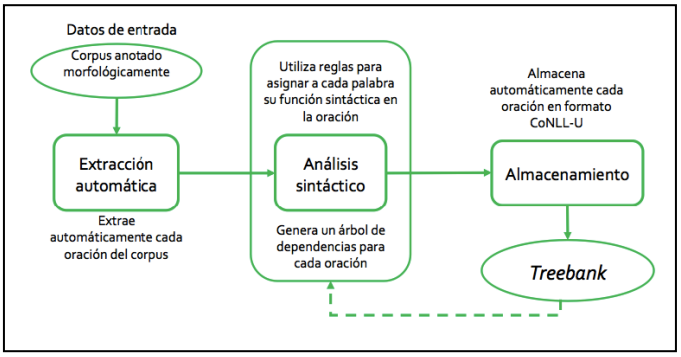

Figura 1: Análisis sintáctico y automático para oraciones

1) Extracción automática: En esta etapa se procesan en forma automática todas las oraciones simples del corpus anotado morfológicamente que fue seleccionado. Cada oración se procesa individualmente.

2) Análisis de oraciones mediante reglas gramaticales: En esta etapa se analizan en forma automática todas las oraciones extraídas previamente del corpus anotado.

Para analizar las estructuras oracionales mediante reglas gramaticales existen dos aproximaciones: la gramática generativa [7] y la gramática de dependencias [8]. En este caso se eligió la última para trabajar con las relaciones entre palabras, en las que cada una de ellas según su categoría gramatical, puede gobernar o ser gobernada por otros vocablos. De esta forma su función gramatical en la oración es determinada. Un ejemplo de este tipo de análisis se encuentra en la Figura 2, indicando cuál palabra es la raíz oracional.

El análisis sintáctico se realizó a través de la creación de reglas gramaticales que generan un árbol para representar cada oración, utilizando las dependencias universales.

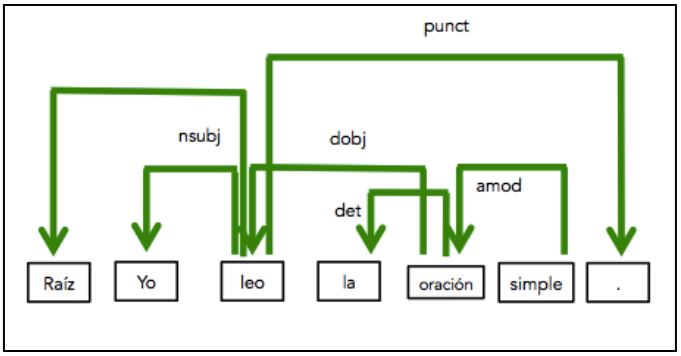

Figura 2: Oración analizada según relaciones de dependencia, empleando dependencias universales.

Para cada oración O de T

a. Se detecta la voz de O, analizando el verbo conjugado.

b. Se encuentra la raíz de O, de acuerdo a la voz definida para O y a las características del verbo conjugado (régimen, copulativo, semicopulativo, auxiliar).

c. Se asigna la función sintáctica para las demás palabras de O, iniciando con su primera palabra.

Figura 3: Algoritmo para analizar sintácticamente una oración

En total, se crearon 1030 reglas gramaticales para asignar a cada palabra una de las 28 funciones sintácticas analizadas. Estas reglas incluían particularidades del español americano, como el uso diferenciado de pronombres clíticos con respecto a la variante peninsular. Un ejemplo se presenta en la Figura 3.

3) Almacenamiento de oraciones: Luego del análisis sintáctico, las oraciones se almacenan en un archivo de texto. Este módulo utilizó un formato de almacenamiento empleado para distinguir en forma natural las dependencias universales. El texto fue codificado mediante UTF-8.

Para la implementación del modelo, se decidió utilizar el lenguaje de programación C++, entre otros detalles por la capacidad demostrada para manejar hileras de caracteres así como la posibilidad de administrar la memoria eficientemente.

III. Evaluación de la Calidad del Etiquetado

Luego de almacenar las oraciones, se emplearon varias métricas conocidas en el área de la Lingüística Computacional para evaluar la calidad del etiquetado. Con ellas se comparó el *treebank* dorado contra el *treebank* obtenido de aplicar el algoritmo de etiquetado de dependencias universales al corpus únicamente anotado morfológicamente, con fines de evaluación de la herramienta.

Las métricas utilizadas fueron propuestas por Buchholz y Marsi como parte de las conferencias CONLL [9], y son:

A. *Labeled attachement score (LAS)*

Esta métrica mide el porcentaje de palabras en el *treebank* generado a las que se asignó correctamente su etiquetado sintáctico, así como la palabra con la que se relaciona gramaticalmente, con respecto al *treebank* base. Para ello, se

divide el total de palabras a las que se les identificó correctamente ambos conceptos (su función gramatical y la palabra con la que se relaciona) entre el total de palabras en el *treebank*.

B. Labeled attachement score 2 (LAS2)

LAS2 mide el porcentaje de palabras en el *treebank* generado a las que se asignó correctamente su etiquetado sintáctico con respecto al *treebank* base. LAS2 se calcula dividiendo el total de palabras a las que se les identificó su función gramatical entre el total de palabras en el *treebank*.

C. Unlabeled attachement score (UAS)

UAS calcula el porcentaje de palabras en el *treebank* generado a las que se asignó correctamente la palabra con la que se relaciona, según el *treebank* base. UAS se calcula dividiendo el total de palabras a las que se les identificó la palabra con la que se relaciona entre el total de palabras en el *treebank*.

La Figura 4 resume las diferencias entre estas métricas.

Métricas	Acierto al asignar función	Acierto al asignar relación
LAS	✓	✓
LAS2	✓	✗
UAS	✗	✓

Figura 4 Diferencia semántica entre las métricas empleadas

IV. RESULTADOS

Con las reglas gramaticales obtenidas de la revisión de la literatura y el consejo de expertos, se ejecutó la prueba del algoritmo extrayendo 6.122 oraciones simples del corpus del *treebank* dorado, a las que se les eliminaron las anotaciones sintácticas. Se analizaron cerca de 106 mil palabras en estas 6.122 oraciones. La figura 5 presenta los resultados de esta corrida inicial.

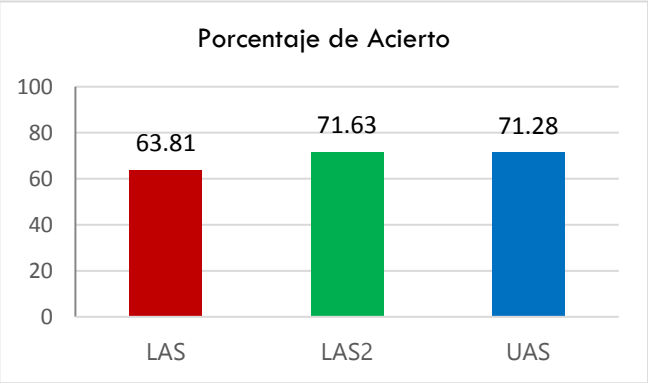

Figura 5 Resultados de la primera ejecución del algoritmo

Luego de estos resultados iniciales, se procedió a seguir incluyendo reglas, para determinar hasta dónde era posible mejorar los indicadores con la inclusión de reglas gramaticales.

Después de 609 iteraciones de mejora del algoritmo, se obtuvieron los resultados mostrados en la Figura 6. La evolución del desempeño del algoritmo desde la iteración 1, al incluir paulatinamente nuevas reglas se encuentran en la Figura 7.

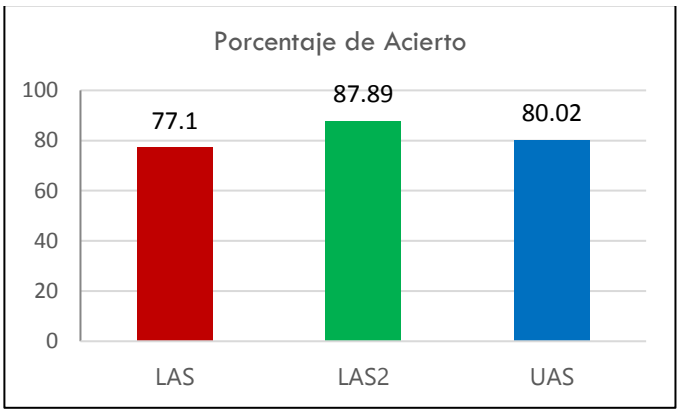

Figura 6 Resultados obtenidos luego de 609 iteraciones

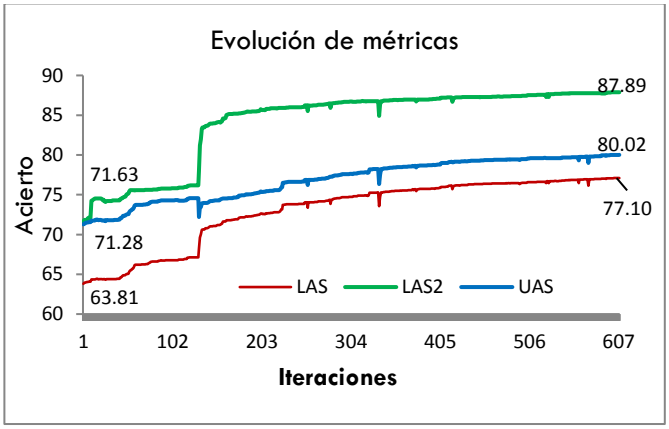

Figura 7 Evolución de las métricas al incluir nuevas reglas

Finalmente, aunque el enfoque de esta investigación no se orientaba hacia la eficiencia del proceso, pudo constatarse que el tiempo de procesamiento de 6.122 oraciones fue de aproximadamente dos minutos. Obviamente un tiempo mucho más corto que el requerido por un anotador humano.

En el Cuadro 1 se presentan las tasas de acierto del etiquetado automático por categoría gramatical.

Cuadro 1 Evaluación del *treebank* según categorías gramaticales

Categoría gramatical	LAS	LAS2	UAS
Determinante	98.143	99.540	98.206
Adposición	94.600	98.252	94.904
Adjetivo	84.231	90.939	85.191
Conjunción Subordinada	84.211	92.398	84.795
Auxiliar	81.766	88.868	85.988
Verbo auxiliar	78.472	86.283	81.259

Número	72.112	86.589	76.029
Pronombre	70.261	74.545	74.988
Nombre	63.908	82.076	68.350
Nombre Propio	59.699	72.072	67.663
Adverbio	59.444	85.159	63.413
Símbolo	50.000	57.025	64.256
Conjunción Coordinada	39.397	94.824	39.722

El cuadro 1 muestra que los resultados reflejan una fortaleza en cuanto al análisis de determinantes, adposiciones, conjunciones subordinadas, verbos, auxiliares, adjetivos, pronombres y números, puesto que el acierto para determinar sus funciones y relaciones sobrepasa el 70%. En un grado intermedio se encuentran los nombres, nombres propios, adverbios y símbolos, que no sobrepasan el 70% pero su acierto es mayor al 40%. Finalmente, conjunciones coordinadas y otros manifiestan una clara debilidad, al obtener aciertos menores al 40%. Para el caso de los nombres y nombres propios, es necesario destacar que muchas palabras de estas categorías actúan como aposiciones, las que debido a la dificultad para su identificación, no se abordaron en nuestra investigación.

Las funciones sintácticas presentan un comportamiento variado. El 91% de las raíces de la oración fueron identificadas exitosamente, en su función y relación. Determinantes y casos son otros ejemplos de eficacia, al rondar en 97.9% y 93.19% respectivamente. Los complementos indirectos superaron el 85% de eficacia en su detección. La identificación de oraciones activas y pasivas sobrepasaron el 80%.

Un caso un tanto distinto es el del complemento directo, que superó el 70% de eficacia. Para ciertos casos el analizador tendió a confundirlo con el complemento indirecto, pues su estructura sintáctica es similar. En dichos casos se requiere información semántica para diferenciarlos, hecho que se encuentra fuera de los límites de esta investigación.

En resumen, 14 de las veintiocho funciones sintácticas sobrepasaron el 70% de acierto, siete sobrepasaron el 50% y el resto presenta valores menores al 50%, siendo este comportamiento esperado para algunas de ellas, por su complejidad gramatical.

V. CONCLUSIONES

La aplicación creada para efectuar los análisis para cada oración permitió generar las métricas generales (LAS, LAS2 y UAS) para el *treebank* y alimentar una base de datos de la cual se extrajeron otros análisis para determinar la eficacia del proceso de análisis sintáctico, clasificando los datos por categoría gramatical y función sintáctica, con el fin de generar conclusiones.

Dichos análisis permitieron distinguir la eficacia del proceso propuesto, ya que los resultados indican, en primer lugar, porcentajes de acierto superiores al 75% para las métricas generales del *treebank*. Igualmente, el incluir nuevas reglas gramaticales muestra tendencia al mejoramiento de las métricas en forma sostenida y constante.

Los análisis realizados permiten distinguir un buen reconocimiento en oraciones activas y pasivas, ya que sus funciones se identifican adecuadamente Además, la eficacia para detectar las raíces de las oraciones sobrepasó el 90%. Ello demostró que la estrategia utilizada, --- el uso de la voz verbal para separar el análisis en casos para verbos activos y pasivos---, es una solución viable para este tipo de fenómenos lingüísticos. Además, la consideración de casos específicos para las estructuras pasivas colaboró en la delimitación de las reglas.

En resumen, dadas las evidencias y aportes señalados, podemos afirmar que es posible crear un analizador sintáctico automatizado basado en reglas para la creación de un *treebank* en español reduciendo la intervención de anotadores humanos.

VI. TRABAJO FUTURO

Como trabajo futuro, existe la opción de combinar análisis de reglas gramaticales y análisis estadístico, para mejorar el rendimiento de acuerdo con las métricas propuestas.

Las reglas gramaticales pueden ser mejoradas, para que las métricas incrementen en su porcentaje de acierto. Esto implica una futura revisión de las reglas creadas en esta investigación, aplicándolas para otros textos.

RECONOCIMIENTOS

El autor del presente artículo agradece al Comité de Tesis involucrado durante el proceso de preparación y defensa de esta propuesta. Su labor insistente en la rigurosidad fue de mucha ayuda.

REFERENCIAS

[1] E. Nolan, S. Hirsch (eds), "The Greek Grammar of Roger Bacon and a Fragment of his Hebrew Grammar". Cambridge University Press, 1902

[2] A. Aho, M. Lam, R. Sethi y J. Ullman, "Compilers: Principles, Techniques, and Tools". Pearson Educación, 2006.

[3] M. Marcus, M. Marcinkiewicz y B. Santorini, "Building a large annotated corpus of English: The Penn Treebank". Computational Linguistics, 19 (2), 313-330, 1993

[4] M. Taulé, M. Martí y M. Recasens, ¨AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Language Resources and Evaluation¨. In Language Resources and Evaluation, 2008.

[5] M. Sandí, "Automatización del análisis sintáctico para el español americano con el fin de crear un *treebank* estandarizado" Tesis de Maestría, Universidad de Costa Rica, 2017.

[6] R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev y J. Lee, "Universal Dependency Annotation for Multilingual Parsing". Association for Computational Linguistics (2) (pp. 92-97).

[7] N. Chomsky, "Aspects of the Theory of Syntax". MIT Press, 1965.

[8] L. Tesnière, "Élements de syntaxe structurale". Paris, 1959.

[9] S. Buchholz y E. Marsi, "CoNLL-X shared task on multilingual dependency parsing". In Proceedings of the Tenth Conference on Computational Natural Language Learning (pp. 149-164). Association for Computational Linguistics, 2006.