Káñina, Rev. Artes y Letras, Univ. Costa Rica XL (Extraordinario): 163-174, 2016 / ISSN:2215-2636

AUTOMATIZACIÓN DEL ANÁLISIS SINTÁCTICO PARA EL ESPAÑOL CON EL FIN DE CREAR UN *TREEBANK* ESTANDARIZADO

AUTOMATION OF SYNTACTIC ANALYSIS FOR SPANISH IN ORDER TO CREATE AN ESTANDARIZED TREEBANK

Minor Sandí Salazar[1]

Gabriela Marín Raventós[2]

Edgar Casasola Murillo[3]

RESUMEN

El crecimiento exponencial en la creación de documentos en la Internet, escritos en español, ofrece variadas oportunidades para el análisis de texto. Debido a su cantidad creciente y a la escasez de herramientas que colaboren en estos procesos, se hace imprescindible la creación de herramientas que los automaticen. Entre estas herramientas los *treebanks* ocupan un papel destacado, puesto que proveen información clave para muchos de los procesos de análisis. Actualmente, existe una tendencia que intenta estandarizar el etiquetado morfológico y sintáctico para crear puntos de contacto entre *treebanks* de distintas investigaciones. A partir de los antecedentes relacionados con el análisis sintáctico de textos, esta investigación propone una metodología para determinar hasta dónde es posible automatizar el proceso de creación de *treebanks*, limitándose a la lengua española.
Palabras clave: análisis sintáctico, *treebanks*, corpus, computación lingüística, estandarización de etiquetados.

ABSTRACT

The exponential growth for creating documents on the Internet using Spanish as their language offers different opportunities for text analytics. But the quantity and the scarcity of tools to support text analysis processes make essential the development of tools for this automation. Among these tools, treebanks occupy a prominent role, as they support many of the key text analysis processes. Currently, there is a tendency to standardize the morphological and syntactic labeling, to create points of contact between treebanks created by different research groups. Based on previous findings, this research proposes a methodology to determine how far it is possible to automate the process of creating treebanks, limiting our results to Spanish.
Keywords: syntactic analysis, treebanks, corpus, Computational Linguistics, tag normalization.

1 Universidad de Costa Rica. Programa de Posgrado en Computación e Informática. Costa Rica. Correo electrónico: minsandi@gmail.com

2 Universidad de Costa Rica. Centro de Investigaciones en Tecnologías de la Información y Comunicación (CITIC). Costa Rica. Correo electrónico: gabrielamarinraventos@gmail.com

3 Universidad de Costa Rica. Escuela de Ciencias de la Computación, Programa de Posgrado en Computación e Informática y Centro de Investigaciones en Tecnologías de la Información y Comunicación (CITIC). Costa Rica. Correo electrónico: edgar.casasola@ucr.ac.cr
Recepción: 15/1/2016. Aceptación: 16/3/2016.

1. Introducción

Uno de los objetivos o tareas más comunes en las Ciencias de la Computación es la automatización de procesos que se realizan, completa o parcialmente, en forma manual y requieren una fuerte inversión de tiempo, debido al nivel de detalle y a la experticia requerida para su correcta ejecución.

Dada la enorme generación de documentos de texto desde el surgimiento de la Internet, se han incrementado la necesidad y las posibilidades de análisis lingüísticos en diferentes lenguajes, incluyendo el español. Según datos actualizados, al menos el ocho por ciento de los documentos presentes en Internet fueron generados utilizando el idioma español (Instituto Cervantes, 2015). Este crecimiento exponencial, que según el reporte del Instituto Cervantes es del 1123% entre los años 2000 y 2013, genera oportunidades para investigar cómo el idioma actualmente adquiere nuevos rasgos o léxico.

El idioma español, a pesar del crecimiento en su utilización en espacios virtuales y haber sido objeto de análisis en diversos campos de la lingüística computacional, mantiene un rezago en cuanto a la abundancia de proyectos de investigación cuando se le compara con otros idiomas, especialmente el inglés. El cuadro que se presenta a continuación, tomado de la investigación de Melero *et al.* (2012), ofrece el grado de avance en las tareas de procesamiento de lenguaje natural de la lengua española a la fecha.

CUADRO 1

Soporte existente a la tecnología linguística para el Español (Melero *et al.* 2012)
Categorización: 1 – Excelente, 2 – Bueno, 3 – Moderado, 4 – Fragmentario, 5 – Escaso

	Cantidad	Disponibilidad	Calidad	Cobertura	Madurez	Sostenibilidad	Adaptabilidad
Tecnologías lingüísticas: herramientas, tecnologías y aplicaciones							
Reconocimiento de voz	2	3	4	2	2	2	4
Síntesis de voz	3	3	4	4	4	3	4
Análisis gramatical	3	3	4	4	4.5	2.5	4.5
Análisis semántico	1.5	2	3	2	2.5	2.5	2.5
Generación de texto	0	0	0	0	0	0	0
Traducción automática	3	2	2	2	4	2	2
Recursos lingüísticos: recursos, datos y bases de conocimiento							
Corpus textuales	3	3	4	4.5	4	4.5	4.5
Corpus de discurso	4	2	4	4	4	3	3
Corpus paralelos	2	4	2	2	2	3	3
Recursos léxicos	3.5	3	4.5	3	4	3	3
Gramáticas	1	4	5	2	2	2	2

Del cuadro anterior, puede concluirse que la disponibilidad actual de recursos léxicos, al igual que la madurez y la sostenibilidad de mecanismos para análisis de textos, ofrece una oportunidad para desarrollar herramientas que permitan el avance de otras investigaciones en el área de la Computación Lingüística.

Según la información anterior, puede inferirse la necesidad de herramientas que permitan acelerar el análisis de textos mediante la automatización de procesos. Para efectos de esta investigación, el interés se ubica en determinar hasta qué grado es posible automatizar en una aplicación el proceso de creación de un *treebank* en español a partir de un corpus anotado para minimizar la intervención de anotadores humanos. Esto con la finalidad de proporcionar un mecanismo que ayude a los lingüistas en su labor de investigación.

En las siguientes secciones, se incluyen definiciones básicas relacionadas con *treebanks*, se presenta el estado del arte y se esboza en forma concisa una propuesta metodológica con la que se pretende alcanzar el objetivo formulado anteriormente.

2. Treebanks

Desde hace algunos años, se ha dado una creciente investigación relacionada con elementos morfológicos y sintácticos presentes en textos escogidos para determinar patrones, facilitar y aumentar la capacidad de procesamiento de los investigadores en el área de la computación lingüística.

Por dicha razón, palabras como «corpus» y «*treebanks*» se han vuelto frecuentes en las investigaciones. En este segmento se dedicará la atención hacia la descripción general acerca de lo que es un *treebank*.

2.1. Concepto

De acuerdo con Hajičová (2010), los *treebanks* son un tipo de corpus anotado que se estructura con la intención de incluir, además del análisis de las palabras de la oración y otros elementos morfológicos, información sobre las relaciones sintácticas entre los componentes de la oración. También es posible que incluyan contenido semántico, si así se requiere. La referencia al concepto de árbol, que el término *treebank* hace, apunta hacia la estructura base que se utiliza para el análisis de la oración. Esta estructura corresponde a la noción de árbol, la cual se encuentra en la teoría formal de grafos que se emplea en las estructuras de datos computacionales.

Las diversas versiones de *treebanks* difieren en cuanto a la codificación que emplean para distinguir las partes de la oración, así como las relaciones sintácticas entre ellas. La tendencia mayoritaria es la de representar la sintaxis contenida en cada oración. La siguiente figura muestra un ejemplo de *treebank*:

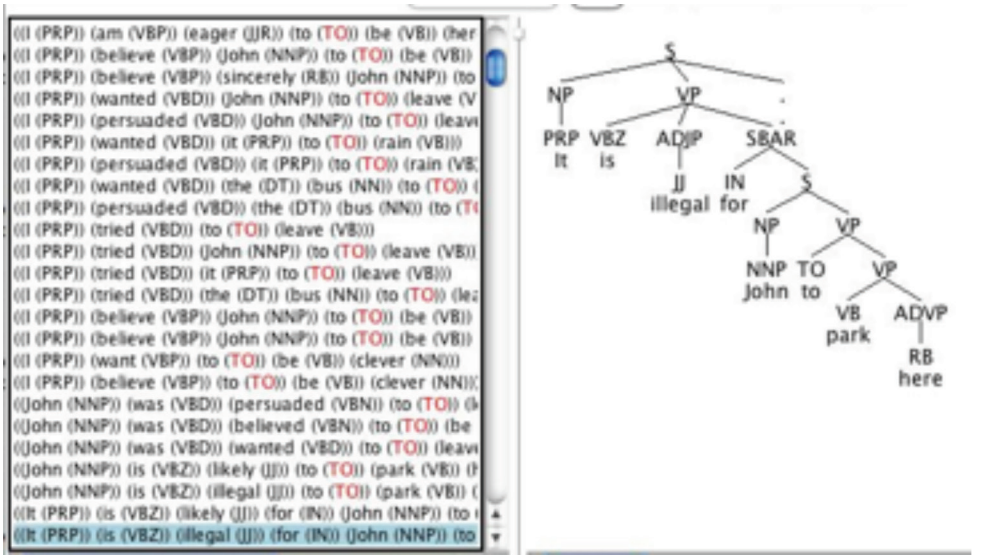

FIGURA 1

Representación de *treebank* (Fong 2015).

2.2. Utilidad de los treebanks

De acuerdo con Jara-Murillo (2013), los *treebanks* pueden ser utilizados para los siguientes fines:

- Estudio de distintos fenómenos propios del idioma de tipo léxico, morfológico y sintáctico.
- Estudio del cambio lingüístico.
- Evaluación de la teoría sintáctica formal.

2.3. Tipos de treebanks

Los *treebanks*, de acuerdo con Jara-Murillo (2013), pueden ser categorizados de acuerdo con su representación:

- Sintagmática: La oración se representa mediante un árbol que contiene los sintagmas.
- Estructura de dependencias: Solamente se representan las palabras incluidas en la oración, no los sintagmas.

2.4. Diferencias entre treebank y corpus

Actualmente, existe alguna confusión entre los conceptos de *treebank* y corpus. Se puede afirmar, partiendo de la definición de *treebank* presentada anteriormente, que aparte del contenido morfológico ya presente en un corpus, se incluya datos sobre la sintaxis de cada oración; lo que evidentemente los diferencia y sugiere que un corpus puede ser un insumo para crear un *treebank*, luego de agregarle información sintáctica de las oraciones que contenga.

Con respecto al corpus, puede decirse que su fin básico es contener información morfológica sobre cada palabra presente en cada oración. El *treebank*, en cambio, como mínimo contiene información morfológica y sintáctica, la cual puede ser utilizado para el estudio de la evolución lingüística de un lenguaje a lo largo del tiempo o incluso para labores educativas.

Estas definiciones sobre *treebanks* y corpus tienen un papel relevante en el desarrollo de las herramientas para análisis de textos. En el estado del arte que se presenta a continuación, se describe el desarrollo histórico de las mismas.

3. Estado del arte

Para presentar el estado del arte se incluye un recorrido histórico que contiene aspectos generales del análisis de textos, específicamente relacionados con textos anotados como corpus y *treebanks*.

Desde los días del teólogo medieval del siglo XIII, Roger Bacon, ha existido interés en la noción de una gramática universal que abarque la mayor cantidad de lenguajes. Nolan y Hirsch (1902: xxv) recogen una frase de Bacon en la que afirmó que "(…) en su sustancia, la gramática es una y la misma en todos los lenguajes, aún si esta accidentalmente varía". Este fervor se ha mantenido constante hasta la época moderna. Lucian Tesnière introdujo la noción de árbol sintáctico de dependencia (Tesnière, 1959) y Noam Chomsky desarrolló otros conceptos teóricos en el área de la lingüística que resultaron claves para el desarrollo de herramientas computacionales que colaboraran en la automatización del análisis de los textos (Lees y Chomsky, 1957). La figura 1 muestra un ejemplo de árbol sintáctico.

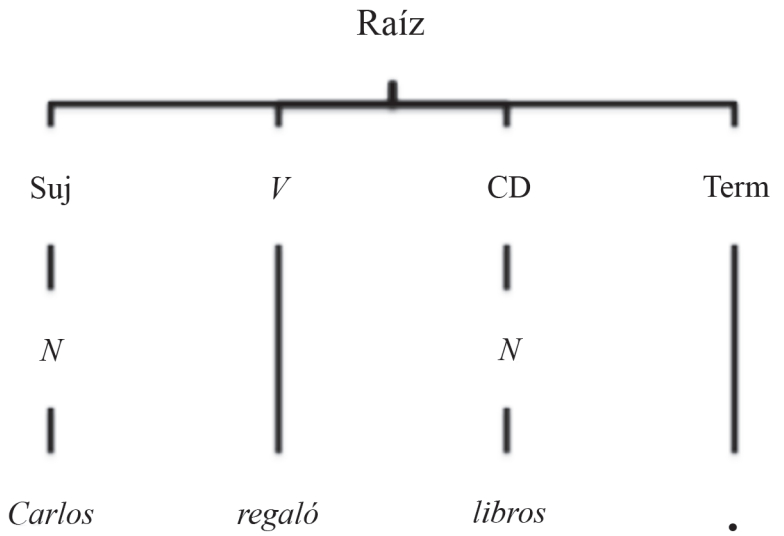

FIGURA 2

Representación de una oración utilizando un árbol sintáctico.

En la época actual, los esfuerzos iniciales por crear textos anotados se remontan a la segunda mitad de la década de los años 60, cuando se completó el primer *treebank*, el cual es conocido como "*Brown Corpus*" y recopiló alrededor de un millón de palabras del inglés de la época (Kučera y Francis, 1967).

Aunque al inicio de los años noventa se crearon varios *treebanks* para la lengua inglesa, el más influyente surgió en 1993. Como resultado de las investigaciones realizadas en la Universidad de Pennsylvania, se desarrolló un *treebank* basado en el inglés americano y, tomando información del *Brown Corpus*, con poco más de 4.5 millones de palabras. Este *treebank* llegó a ser un modelo, debido a que incorporó POS (*part-of-speech*) y más de la mitad de su contenido incluía el esqueleto de una estructura sintáctica, de acuerdo con el reporte de Marcus *et al.* (1993). Es conocido como *Penn*.

Pocos años después, se desarrollaron algunas guías consideradas provisionales para añadir información sintáctica además de las anotaciones morfosintácticas, como por ejemplo, símbolos para indicar las relaciones de dependencia entre palabras. En el mismo documento se señala la

intención de posibilitar que este formato de anotación sintáctica sea multilingüístico. Estas guías son conocidas hoy como EAGLES, y se consideran uno de los primeros pasos para crear etiquetados útiles en diversos idiomas (Leech *et al.,* 1996).

Entre 1990 y 2010 surgieron algunos *treebanks* para español que fueron base para otros que surgieron posteriormente. En 1997 la Universidad PompeuFabra inició el desarrollo del *treebank* que llegó a ser conocido como IULA (Institut Universitari de Lingüística Aplicada, 1998). Montserrat Civit Torruela y Antonin Martí presentaron en el año 2002 el corpus CLiC-TALP, el cual contiene un millón de palabras (Civit Torruela y Martí, 2002). Posteriormente, las autoras tomaron un conjunto de cien mil palabras de CLiC-TALP para crear un *treebank* que es conocido como 3LB (Civit Torruela y Martí, 2004). Los datos generados para 3LB permitieron el desarrollo posterior de otro *treebank* conocido como Áncora, el cual posee la particularidad de orientarse hacia el español y el catalán (Taulé, *et al.* 2008).

En el año 2006, se desarrollaron investigaciones en la Universidad de Stanford que culminaron con la presentación de un *parser* (De Marneffe *et al.,* 2006). Igualmente, Sabine Buchholz y Erwin Marsi (2006) hicieron público un formato para almacenar *treebanks* en trece idiomas, útil para representar la información morfosintáctica, conocido como CoNLL-X.

El *parser,* previamente implementado por de Marneffe *et al.* (2006), es la base para la representación de dependencias basada en árboles sintácticos, la cual toma como fundamento el marco teórico de la gramática léxico funcional propuesta por Joan Bresnan (2001). El trabajo de De Marneffe presenta 48 tipos de relaciones de dependencia que pueden presentarse en una oración. Fue completado en el año 2008. (De Marneffe y Manning, 2008) Ese mismo año, Daniel Zeman se orientó a trabajar en la creación de una herramienta de conversión entre diferentes tipos de etiquetado sintáctico para diversos lenguajes (Zeman, 2008).

En el año 2012, Slav Petrov y su equipo propusieron un etiquetado universal, junto con un conjunto de relaciones entre diversos tipos de etiquetado hacia este formato universal, logrando una anotación capaz de ser empleada en 22 diferentes idiomas y 25 *treebanks*. Para el idioma español, Petrov y su equipo seleccionaron los *treebanks* Áncora y Cast3LB, anteriormente mencionados (Petrov *et al.,* 2012).

En la Universidad de Costa Rica, dada la poca cantidad de *treebanks* que se han desarrollado para el español, se creó el *treebank* IPROCOLDI a partir de documentos cuyo contenido se basaba en discursos presidenciales costarricenses, entre el siglo XVIII y el XX (Jara-Murillo, 2013). El proceso de creación de este treebank en algunas de sus etapas ya contenía elementos automatizados.

El año 2013 fue prolífico para el procesamiento de lenguaje natural, ya que los resultados de varias investigaciones se presentaron:

Ryan McDonald presentó su primera propuesta de un *treebank* universal, cuyo etiquetado puede aplicarse a seis idiomas: alemán, inglés, sueco, español, francés y coreano; facilitando el análisis sintáctico multilingüístico (McDonald *et al*., 2013). Muchos elementos de esta propuesta se basan en los postulados desarrollados por el equipo de De Marneffe.

Tsarfaty (2013) propuso una extensión de las dependencias de Stanford que unifica la anotación de las relaciones sintáctica y morfológica, además de ofrecer dos formas para predecir en forma automática estas anotaciones desde un texto sin procesar.

Cristina Bosco y otros investigadores (Bosco *et al.,* 2013) asociados implementaron la primera versión de un *treebank* para la lengua italiana. Utilizaron las dependencias de Stanford desde un proceso de transformación que tomó como fuentes dos *treebanks* con etiquetados distintos, los cuales requirieron una armonización entre sí, utilizando varios patrones de conversión que se agruparon en dos clases.

Recientemente, De Marneffe presentó una mejora de la representación de dependencias de Stanford con la finalidad de enfatizar en la teoría de la gramática funcional sobre la cual descansa este diseño, así como la habilidad de ser aplicada en diferentes lenguajes (De Marneffe *et al*., 2014).

En el año 2015, Joakim Nivre (2015) presentó la primera versión de las dependencias universales, con el fin de crear guías generales para una anotación gramatical consistente entre diversos lenguajes, así como el desarrollo de un *parser* multilingüístico. Para lograr su objetivo, las dependencias universales tomaron elementos de las dependencias universales de Stanford, extendieron el conjunto de etiquetados definidos por Petrov, adoptaron un subconjunto del inventario definido por Zeman, y crearon una versión revisada del formato CoNLL-X, llamada por este grupo CoNLL-U.

Para concluir, se indica que en el año 2015, Sampo Pyysalo y su equipo asociado (Pyysalo *et al*., 2015) lograron la creación de un *treebank* totalmente compatible con las dependencias universales para el idioma finlandés, a partir de otros *treebanks* creados utilizando otro tipo de etiquetado. Para este fin, se desarrolló un proceso de transformación desde estas fuentes hacia el etiquetado definido por las dependencias universales.

Como se ha presentado, es notable observar el desarrollo de herramientas para el análisis de textos en las que un *treebank* se ha visto involucrado. Sin embargo, muchas de ellas poseen diferencias significativas en cuanto a su etiquetado, lo que dificulta compartir sus resultados o crear otras herramientas que trabajen utilizando resultados de diferentes investigaciones. Igualmente, existe una tendencia hacia la creación de *treebanks* para un idioma o con la capacidad de ser multilingües, además de que utilicen etiquetados estandarizados, cuyas fuentes provengan de diferentes investigaciones.

Otro hecho destacable es que en los diferentes proyectos existen procesos que se realizan manualmente, lo que da oportunidad para la automatización de los mismos. El análisis sintáctico es uno de ellos.

Dado lo anterior, la propuesta que presentamos en este artículo se orienta a responder la siguiente pregunta: «¿Cómo automatizar el proceso de creación de un *treebank* en español reduciendo la intervención de anotadores humanos?»

4. Propuesta

Para lograr nuestro objetivo se ha ideado la siguiente metodología. El objetivo es delinear las actividades necesarias por los investigadores que quieran evaluar la viabilidad de automatizar, en la medida de lo posible, el proceso de creación de *treebanks*.

4.1. Etapa 1: Identificar y caracterizar textos anotados para el español

Como primer paso, se hace necesario seleccionar los elementos u objetos básicos para la implementación de un prototipo para el análisis sintáctico automatizado:

i. Identificar mediante búsqueda literaria al menos cinco corpus utilizados para el idioma español.
ii. Determinar cuáles características de estos son relevantes para la investigación. Dado que caracterizar es el acto por el que se determinan los atributos o cualidades que distinguen a un objeto de los demás, se entiende que habrá una selección de las características que atañen a este proceso de investigación.
iii. Crear un cuadro comparativo con las características de los corpus encontrados.
iv. Seleccionar el corpus que más se ajuste a los propósitos de la investigación.

4.2. Etapa 2: Establecer la equivalencia entre anotaciones

Para cumplir este objetivo, se requiere llevar a cabo los siguientes pasos, tomando como base el corpus seleccionado previamente:

i. Hacer una lista de las categorías gramaticales consideradas en la anotación morfológica seleccionada por el corpus.
ii. Recopilar los etiquetados morfológicos empleados para los accidentes gramaticales.
iii. A partir de la información anterior, establecer equivalencias con el modelo estándar de representación de dependencias elegido.

iv. Preparar un cuadro que represente adecuadamente las equivalencias existentes entre las anotaciones morfológicas del corpus de entrada con aquellas que se utilizarán como parte del *treebank* generado.

4.3. Etapa 3: Proponer un modelo de creación automática de un treebank

Para alcanzar este objetivo se hace necesario el desarrollo de una aplicación que agrupe en un solo proceso la invocación de estos módulos específicos, como mínimo, para asegurar la obtención de resultados aceptables.

i. En primera instancia, se implementa y valida un módulo de extracción capaz de reconocer en forma individual cada oración que pertenezca al corpus seleccionado para el análisis sintáctico. Este, requiere haber sido anotado morfológicamente.
ii. Luego, la oración es procesada empleando un analizador sintáctico (existen algunos «*open source*») que reconozca el formato de anotación morfológica utilizado por el corpus seleccionado para la investigación.
iii. Como tercera etapa, se implementa un módulo de transformación del etiquetado morfosintáctico que emplee las equivalencias definidas previamente entre el etiquetado del corpus y el etiquetado utilizado por el formato estándar para la representación de dependencias.
iv. Como última etapa, se desarrolla un modelo de almacenamiento en el que cada oración procesada en los módulos anteriores se incluya en un documento de texto. Dicho módulo requiere reconocer el formato de almacenamiento CoNLL-U y el formato UTF-8. Cada oración es tratada de forma individual y almacenada de manera conjunta con las otras que pertenecen al *treebank*, creando un documento de texto que puede ser procesado por otras herramientas que reconozcan este formato.

La siguiente figura ilustra el proceso por el que se propone la automatización de la creación de un *treebank*.

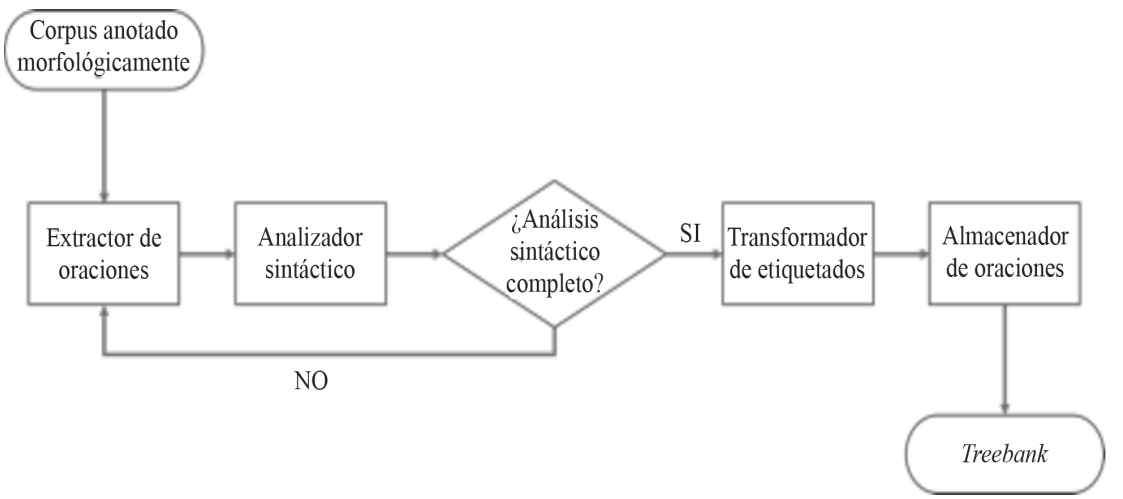

FIGURA 3

Proceso sugerido para analizar sintácticamente un corpus anotado morfológicamente.

4.4. Etapa 4: Evaluar en qué proporción puede ser automatizado el proceso de creación de un treebank

Si bien es cierto que el objetivo principal de esta investigación es sugerir un modelo para la automatización del análisis de textos anotados morfológicamente para obtener un *treebank*, se hace necesario validar que esta automatización ofrezca los resultados adecuados. Para lograr este cometido se propone lo siguiente:

1. Seleccionar un *treebank* al que llamaremos «*treebank* dorado», que utilice el etiquetado seleccionado como representación estándar de dependencias y cuyo idioma sea el español. Su fin es validar la eficiencia del modelo.
ii. Hacer una copia del «*treebank* dorado» para eliminar las anotaciones sintácticas sin afectar las morfológicas, esto con la finalidad de generar un corpus con el mismo conjunto de oraciones.
iii. Utilizar el corpus creado en el paso anterior para generar un *treebank* que contenga el etiquetado seleccionado como representación estándar de dependencias.
iv. Evaluar el último *treebank* generado de acuerdo con las anotaciones sintácticas presentes en el «*treebank* dorado», tomando en cuenta las variables que se explican posteriormente.
v. Redactar las conclusiones obtenidas luego que el proceso de evaluación haya sido completado.

4.4.1. Métricas para evaluación de treebanks

Para validar la eficiencia del modelo propuesto se utilizarán las métricas que se presentan a continuación. Además, se aclara que para todas ellas los signos de puntuación no son considerados. Estas métricas que fueron tomadas de la investigación desarrollada por Buchholz y Marsi (2006), se emplearán exclusivamente para la etapa cuatro del diseño metodológico:

i. *Labeled attachement score* (LAS): Esta variable mide el porcentaje de palabras en *treebank* generado a las que se asignó correctamente su etiquetado sintáctico, así como la palabra con la que se relacionan gramaticalmente, con respecto al «*treebank* dorado». Se calcula utilizando esta fórmula:

$$LAS = \frac{número\ de\ palabras\ con\ tag\ y\ relación\ correctamente\ asignada}{cantidad\ total\ de\ palabras\ en\ el\ treebank}$$

ii. *Labeled attachement score* 2 (LAS2): Esta variable evalúa el porcentaje de palabras en el *treebank* generado que fueron anotadas sintácticamente, de forma correcta, al ser comparadas con las anotaciones sintácticas del «*treebank* dorado». Puede ser representada con esta fórmula:

$$LAS2 = \frac{número\ de\ palabras\ con\ tag\ correctamente\ asignado}{cantidad\ total\ de\ palabras\ en\ el\ treebank}$$

iii. *Unlabeled attachement score* (UAS): El propósito de esta variable es indicar el porcentaje de palabras en el *treebank* generado, a las cuales se asignó correctamente su relación con la palabra indicada, de acuerdo a la comparación con el «*treebank* dorado»:

$$UAS = \frac{número\ de\ palabras\ con\ relación\ correctamente\ asignada}{cantidad\ total\ de\ palabras\ en\ el\ treebank}$$

5. Conclusiones

Luego de haber expuesto algunos conceptos básicos sobre *treebanks* y corpus, un estado del arte y una propuesta para nuestra investigación, se ha llegado a las siguientes conclusiones:

En primer lugar, los *treebanks,* corpus *y parsers* son herramientas con muchas utilidades para el análisis de textos y validación de herramientas en la Lingüística Computacional, así como para reconocer patrones en la evolución de un lenguaje. Pero, en su proceso de creación y recopilación carecen de herramientas computacionales que ayuden al lingüista en su labor de anotación sintáctica. Esto deriva en la inversión de tiempo que podría ser reducido si se desarrolla un mecanismo que automatice en lo posible el anotado sintáctico.

Otro hecho que se concluye es la existencia de distintos sistemas de etiquetado morfológico que representan las categorías y accidentes gramaticales presentes en un idioma determinado. La situación mencionada, aplica también para el etiquetado sintáctico. Esta diversidad de codificaciones causa que *treebanks* creados por diferentes investigaciones para un mismo idioma enfrenten dificultades a la hora de crear equivalencias entre los etiquetados que utilizan.

Paralelamente, en cuanto al almacenamiento, en los *treebanks* se han implementado diversas formas de estructuración, lo cual genera una necesidad de crear mecanismos de interpretación entre diversos sistemas.

Todas las situaciones presentadas con anterioridad se dan en la lengua inglesa, esto debido a que es la más analizada actualmente. Sin embargo, dado que muchas de estas herramientas y teorías han sido utilizadas para analizar el español, los patrones de diversidad y divergencia se mantienen, con el agravante de que el español no posee un nivel tan avanzado de análisis como el inglés.

Igualmente se pudo observar, a partir del estado del arte, que desde años recientes una de las tendencias actuales es la creación de mecanismos que automaticen y estandaricen el procesamiento de los textos. En primera instancia, mecanismos creados para cada idioma, con el objetivo de posibilitar el análisis multilingüe como una segunda etapa. En todos ellos los *treebanks* juegan un papel notable, por lo que el aporte de esta investigación sería importante, puesto que proporciona una herramienta que colabora con estas aplicaciones.

Finalmente, dada la diversidad de etiquetados y estructuras de almacenamiento para corpus y *treebanks*, este entorno ofrece oportunidades para investigar cómo el proceso de análisis sintáctico de un texto en español puede ser automatizado. Lo anterior, no sustituye la labor del lingüista, sino que complementa su trabajo, dada la enorme cantidad de textos que se generan en el presente.

Al emplear los elementos necesarios para que el conocimiento pueda ser transmitido más fácil y claramente, se utilizarán los etiquetados morfológicos y sintácticos de las dependencias universales, los cuales fueron desarrollados para dicho fin.

6. Bibliografía

Bosco, Cristina *et al*. 2013. "*Converting Italian treebanks: Towards an Italian Stanford dependency treebank*". En: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*: 61-69.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Wiley-Blackwell.

Buchholz, Sabine y Erwin Marsi. 2006. "CoNLL-X shared task on multilingual dependency parsing". En: *Proceedings of the Tenth Conference on Computational Natural Language Learning*: 149-64.

Civit Torruela, Montserrat y Antonin Martí. 2002. "Design principles for a Spanish treebank". En: *Proceedings of TLT.*

Civit Torruela, Montserrat y Antonin Martí. 2004. "Building Cast3LB: a Spanish treebank". En: *Research on Language and Computation* II (4): 549–574.

De Marneffe, Marie-Catherine *et al*. 2006. "Generating typed dependency parses from phrase structure parses". En: *Proceedings of Language Resources and Evaluation* VI: 449-454.

De Marneffe, Marie-Catherine y Christopher Manning. 2008. "The Stanford typed dependencies representation". En: *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*: 1-8.

De Marneffe, Marie-Catherine *et al*. 2014. "Universal Stanford Dependencies: A cross-linguistic typology". En: *Proceedings of Language Resources and Evaluation*: 4585-4592.

Fong, Sandiway. 2015. "TreeBank Search". Recuperado de http://dingo.sbs.arizona.edu/~Sandiway/treebanksearch/index.html [Consulta 25 enero. 2016].

Hajičová, Eva *et al*. 2010. "Treebank Annotation". En: Nitin Indurkhya and Fred J. Damerau (eds.): 167-188.

Indurkhya, Nitin y Fred J. Damerau. 2010. *Handbook of natural language processing*. 2. Boca Ratón, FL: CRC Press.

Institut Universitari de Lingüística Aplicada. 1998. "El Corpus de L'IULA: Etiquetaris". Recuperado de http://www.iula.upf.edu/repositori/98inf018.pdf. [Consulta 22 de julio 2015].

Instituto Cervantes. 2015. "El Español: Una lengua muy viva". Informe 2015. Recuperado de http://elnuevosol.net/wp-content/uploads/2016/05/espanol_lengua-viva_20151.pdf. [Consulta 12 diciembre. 2015].

Jara-Murillo, Carla. 2013. "El treebank del español IPROCOLDI: componente anotado del corpus CODIMEP-CR". En: *Revista de Filología y Lingüística de la Universidad de Costa Rica* XXXIX (2): 143-171.

Kučera, Henry y Nelson F. 1967. *Computational analysis of present-day American English*. Providence, United States: Brown University Press.

Leech, G. *et al*. 1996. "Guidelines for the standardization of syntactic annotation of corpora". En: *EAGLES Document EAG-TCWG-SASG/1.8*.

Lees, Robert y Noam Chomsky. 1957. "Syntactic Structures". En: *Language* XXXIII (3): 375-408.

Marcus, Mitchell *et al*. 1993. "Building a large annotated corpus of English: The Penn Treebank". En: *Computational linguistics* XIX (2): 313-330.

McDonald, Ryan et al. 2013. "Universal Dependency Annotation for Multilingual Parsing". *Association for Computational Linguistics* (2): 92-97.

Megyesi, Beáta. 2015. Nordic Conference of Computational Linguistics NODALIDA 2015. Suecia: Linköping University Electronic Press.

Melero, Maite. *et al*. 2012. *The Spanish language in the digital age*. Berlín: Springer.

Nivre, Joakim. 2015. "Towards a Universal Grammar for Natural Language Processing". In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 3-16). Springer International Publishing.

Nolan, Edmond y Samuel Abraham Hirsch. 1902. *The Greek Grammar of Roger Bacon and a Fragment of his Hebrew Grammar*. Cambridge: Cambridge University Press.

Petrov, Slav *et al*. 2012. "A universal part-of-speech tagset". En LREC.

Pyysalo, Sampo *et al*. 2015. "Universal Dependencies for Finnish". En: Megyesi, Beáta: 163.

Taulé, Mariona *et al*. 2008. "AnCora: Multilevel Annotated Corpora for Catalan and Spanish". En: *LREC*.

Tesnière, Lucien. 1959. *Élements de syntaxe estructurale*. Paris: C. Klincksieck.

Tsarfaty, Reut. 2013. "A Unified Morpho-Syntactic Scheme of Stanford Dependencies". En: *Association for Computational Linguistics* (2): 578-584.

Zeman, Daniel. 2008. "Reusable Tagset Conversion Using Tagset Drivers". En: *LREC 2008: 28-30.*