En Vargas Agudelo, Fabio Alberto, Soto Duran, Darío Enrique y Giraldo Mejía, Juan Camilo, *Investigación e Innovación en Ingeniería de Software - Volumen 4*. Medellín (Colombia): Sello Editorial TdeA.

# Un Método de Etiquetado Semiautomático de Corpus Lingüísticos.

Marín Morales, María Isabel y Molina Mejia, Jorge Mauricio.

Cita:

Marín Morales, María Isabel y Molina Mejia, Jorge Mauricio (2020). *Un Método de Etiquetado Semiautomático de Corpus Lingüísticos. En Vargas Agudelo, Fabio Alberto, Soto Duran, Darío Enrique y Giraldo Mejía, Juan Camilo Investigación e Innovación en Ingeniería de Software - Volumen 4. Medellín (Colombia): Sello Editorial TdeA.*

Dirección estable: https://www.aacademica.org/jorge.mauricio.molina.mejia/58

ARK: https://n2t.net/ark:/13683/pqc6/RpN

# Capítulo 8
# Un Método de Etiquetado Semiautomático de Corpus Lingüísticos

*Maria Isabel Marín Morales - mmarinm2@tdea.edu.co*
*Tecnologico de Antioquia*
*Jorge Molina - Jorge.molina@udea.edu.co*
*Universidad de Antioquia*

## I. INTRODUCCIÓN

La lingüística computacional es un campo interdisciplinar cuyo interés principal es la creación de modelos formales que permitan el procesamiento automático de las lenguas naturales. A partir de estos modelos se procesa el lenguaje natural emulando, desde la inteligencia artificial, la capacidad humana de comprender o reproducir el lenguaje dentro de un entorno comunicativo (producción escrita u oral de datos lingüísticos) [1], [2].

Uno de los mecanismos o enfoques metodológicos de los cuales se sirve la lingüística computacional es la lingüística de corpus. Esta se encarga de estudiar la manera más adecuada de gestionar producciones reales de la lengua, almacenadas en textos u otros formatos (audio, video, imágenes, etc.) con el fin de analizar la propia lengua desde sus distintos niveles: fonético, morfológico, sintáctico, lexical, semántico y pragmático [3]. De ahí que la lingüística de corpus se aplique a lo que tiene que ver con la recolección, procesamiento y análisis de grandes muestras de una determinada lengua [4]. Dichas muestras son definidas como corpus, los cuales son conjuntos de textos digitalizados, reunidos y seleccionados teniendo en cuenta unos criterios lingüísticos, a saber: a) que sean recolectados en entornos naturales; b) que posean rasgos definitorios explícitos; c) con una similar extensión o tamaño; d) que se encuentren disponibles; e) que sean accesibles; f) que sean representativos de una determinada lengua; g) que cuenten con metadatos descriptores, y h) que se encuentren compilados bajo criterios o parámetros de organización [3].

Muchos de los algoritmos para el procesamiento del lenguaje natural no

permiten etiquetar, de manera precisa, fenómenos importantes en el nivel textual como las anáforas, las catáforas o los marcadores lógico-temporales y discursivos, entre muchos otros inherentes al lenguaje escrito u oral [5]. Si bien se han encontrado algunas propuestas, como lo es el caso de "AnCora corpora" [6], o de los modelos de resolución anafórica de Palomar et al. [7] para el español, o de Kundu et al. [8] para el inglés, no ha sido posible, hasta el momento, encontrar un modelo específicamente destinado a los textos para la enseñanza del español como lengua extranjera (ELE), los cuales contienen algunas especificidades que no poseen los textos académicos o literarios. Esto tiene que ver con el carácter referencial de algunos elementos que demandan al lector unos vastos conocimientos del mundo [9]. Por otra parte, es necesario contar con corpus para la enseñanza de la lengua española que se encuentren perfectamente analizados y etiquetados, ya que no es lo ideal enseñar errores a los estudiantes extranjeros de dicha lengua [9].

Las razones antes esbozadas son las que se han tenido en cuenta a lo largo del presente capítulo, en el que se presenta un método que permite etiquetar de manera semiautomática un corpus de documentos textuales dedicados a la enseñanza de ELE. El método cubre las etapas de etiquetado morfosintáctico automáticamente, con la aplicación FreeLing [10], y el etiquetado, de manera manual, de correferencias textuales como las anáforas y catáforas. Las etiquetas finales se expresan sobre un documento XML (eXtensible Markup Language o Lenguaje de Marcado Extensible, en español) extraído de un corpus para la enseñanza de la lengua española y dirigido a un público de estudiantes extranjeros [11].

Si bien el trabajo aquí presentado se encuentra, sobre todo, destinado a la aplicación de la enseñanza del español como lengua extranjera (ELE), como se expone en la discusión, el método puede utilizarse en otros contextos de aplicación como, por ejemplo, el aprendizaje de máquina (Machine Learning).

Este capítulo se distribuye como sigue: en el apartado II se presentan los conceptos y la revisión de literatura con relación al etiquetado de corpus y, en especial, se expone una síntesis de la necesidad de efectuar una aplicación en la enseñanza del español como lengua extranjera; en el apartado III se exhiben los diferentes pasos que constituyen el método propuesto; en el apartado IV se aborda la discusión de la aplicabilidad del método y cómo aquel puede extenderse a otros casos de aplicación y, finalmente, en la sección V se enuncian las conclusiones referentes al presente trabajo.

## II. MARCO TEÓRICO Y REVISIÓN DE LITERATURA

### *A. Sistemas de etiquetado*

Un sistema de etiquetado, o analizador morfosintáctico, es un programa informático que tiene como entrada un texto plano (generalmente en formato *.txt), a partir del cual se crea una versión con etiquetas (*tags*) que dan cuenta de unas características, a la vez, morfológicas y sintácticas que son inherentes al lenguaje natural. En este sentido, el sistema tiene en cuenta la posición de la palabra dentro de la oración al momento de generar las etiquetas que marcan, por una parte, las nociones morfológicas, como el género y el número, o las funciones sintácticas como, por ejemplo, que una palabra actúe como determinante, nombre propio o común, verbo (con sus marcas de modo, tiempo, persona), etc. También son importantes las fases de desambiguado de las unidades etiquetadas (Figura 1), lo que permite que cada elemento etiquetado no se confunda con otro. En este último proceso, la mayor parte de los sistemas se fundamentan en métodos estadísticos, principalmente, en cadenas de Markov que proveen información estadística acerca de la posibilidad de que luego de una determinada unidad gramatical vaya otra íntimamente relacionada con aquella otra; por ejemplo, luego de un determinante es más probable que le siga un sustantivo (DET + NC).

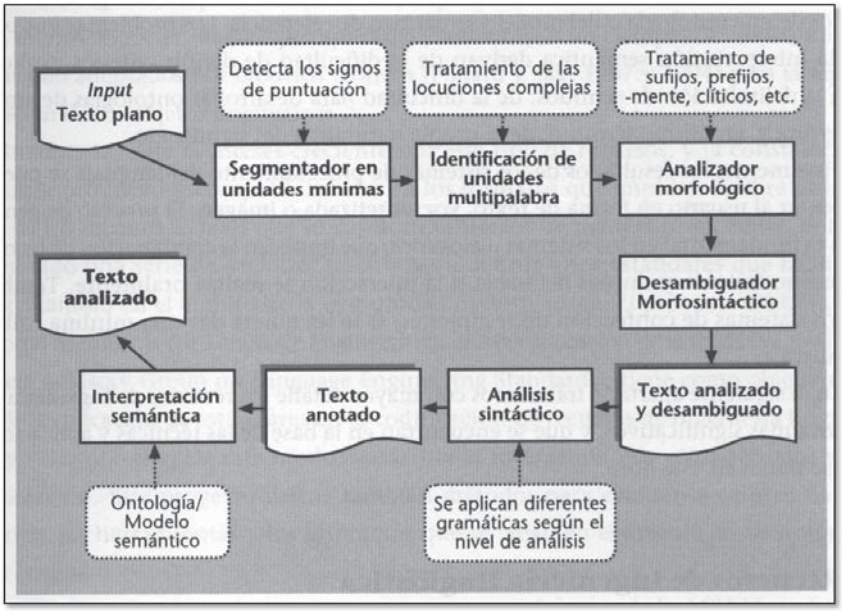


**Fig. 1.** Diferentes etapas en un análisis textual morfosintáctico y semántico [12].

Desde la entrada del texto plano, hasta la extracción del texto completamente analizado, el sistema de etiquetado realiza diversos procedimientos: 1) un tokenizado (segmentación en palabras y signos de puntuación); 2) la identificación de unidades multipalabra o de frasemas (diferentes tipos de locuciones, por ejemplo); 3) el desambiguado de tipo morfosintáctico, etc. Es así como muchos sistemas, en función de su nivel de granularidad y de fineza en el análisis, realizan el etiquetado con un mayor o con un menor grado de efectividad.

En etiquetadores como *FreeLing* [10], las etiquetas contienen rasgos importantes como lo son: si el género de un sustantivo o de un adjetivo es masculino o femenino, o si el número de alguno de estos es singular o plural. Dicha información permite destacar elementos, como es el caso de las entidades nombradas, las cuales permiten, entre otras funciones, el reconocimiento de referentes y correferentes textuales, por ejemplo.

La mayor parte de las herramientas suelen tener un conjunto de etiquetas que proveen la información gramatical más pertinente en lo que tiene que ver con los elementos que constituyen las oraciones. En el caso de *FreeLing* [10], las etiquetas propuestas dan el género y número de elementos o información acerca de los verbos (modo, tiempo, persona), etc.

*B. Corpus en la enseñanza del español como lengua extranjera*

Con respecto al uso de corpus en la Enseñanza del Español como lengua extranjera, la idea, según Boulton y Tyne [13], es que los aprendientes de lenguas adquieran las estructuras lingüísticas de la lengua objeto gracias al empleo de documentos auténticos. Esto ayudaría no solamente a entender mejor la lengua, sino, además, a expresarse en la lengua estudiada de una forma correcta y adecuada, teniendo para ello el modelo lingüístico de primer orden: los textos producidos por hablantes nativos y cuyo objetivo no es la enseñanza sino informar o ser leídos por otros hablantes nativos. La enseñanza a partir de la lingüística computacional y de corpus, y que en lenguas como el inglés y francés tiene un gran desarrollo en países como Estados Unidos, Inglaterra y Francia, ha sido poco desarrollada en el área del español [5], [11].

Por todo ello, se tiene en cuenta que la lingüística computacional proporciona las herramientas que permiten realizar un tratamiento informático de las lenguas naturales de una manera más rápida y con un nivel de precisión altamente aceptable. A partir de dicho tratamiento, se pueden obtener corpus anotados y etiquetados de manera morfológica y sintáctica. En este trabajo inicial se proporciona la base de conocimiento de sistemas

para la enseñanza y aprendizaje de lenguas que funcionan a partir de corpus lingüísticos, como lo señala Antoniadis [14]. Los dispositivos informáticos que resultan de estos procedimientos aplicados sobre corpus se implementan con tres finalidades: ayudar a los docentes en su proceso de planificación de actividades de una forma fácil y automática; hacer que el aprendizaje de la lengua sea más interesante y participativo por parte de los aprendientes, y permitir un aprendizaje a partir de datos auténticos, es decir, de textos que dan cuenta de la lengua real en la que se expresan sus hablantes [11], [14], [15].

*C. Correferencia textual*

La correferencia textual es una categoría problemática en la gramática de las lenguas [16], debido al amplio espectro de mecanismos que intervienen en ella, como la repetición de palabras, la sustitución por pronombres y la elipsis. El manejo correcto de estos mecanismos incluye habilidades de naturaleza tanto gramatical como cognitiva, por lo que se dificulta su sistematización en el procesamiento del lenguaje natural [17]. Existen diversos mecanismos de correferencia, como la repetición de unidades léxicas, su sustitución por medio de elementos fóricos (anáforas y catáforas) y la elipsis.

*D. Revisión de la literatura*

Varios autores que tratan las lenguas naturales, entre otras, el español, han abordado la creación de una metodología específica para el etiquetado morfosintáctico de textos escritos: *TreeTagger* [18]; *Stanford Parser* [19]; *FreeLing* [20]; *TagAnt* [21], entre otras.

En primer lugar, Castellón *et al*. [22] presentan: 1) el etiquetado morfosintáctico automático de un corpus, utilizando para ello la herramienta *FreeLing*; 2) la aplicación de una interfaz de anotación que permite refinar el etiquetado; 3) la realización de pruebas de anotación que permitan definir los criterios de desambiguación, y 4) la extensión de operadores, entre los cuales destacan usos metafóricos, metonimias, multipalabras y partitivos.

Sánchez [23] introduce las siguientes etapas: 1) la compilación semiautomática del corpus con las herramientas *BootCaT* y *WebBootCaT*, y 2) el análisis del corpus con *Sketch Engine*, etapa que se encarga del proceso de etiquetado y que presenta, además, una interfaz de explotación del corpus.

Los autores Martí *et al*. [24] presentan una metodología dividida en cinco etapas: 1) el etiquetado morfosintáctico; 2) el etiquetado sintáctico superficial, es decir,

automático; 3) el etiquetado sintáctico profundo, el cual corresponde a una revisión manual del etiquetado automático; 4) la asignación de unos papeles temáticos de los predicados verbales, y 5) el análisis de sentidos de nombres más frecuentes a partir de *WordNet*.

Anxo, Portela y Xavier [25] especifican el proceso de creación de un corpus paralelo inglés-gallego en las etapas: 1) la adaptación automática al *WordNet 3.0* de las etiquetas semánticas del corpus en inglés; 2) la traducción manual al gallego de los textos; 3) la proyección en los textos en gallego de las etiquetas semánticas del inglés, y 4) la consolidación del corpus paralelo inglés-gallego con el resultado de la anotación semántica de la lengua gallega.

Zafra, Gómez y Navarro [26] presentan la anotación de un corpus en las siguientes etapas: 1) la búsqueda y selección de textos; 2) la anotación manual de metadatos; 3) la anotación automática del idioma con la herramienta *Gate Developer,* y 4) la revisión manual del etiquetado.

Rioja [27] y Vila [28] presentan una metodología que, si bien en algunas etapas coincide con la propuesta que se presentará en la sección siguiente, no ahonda en el detalle referente al proceso de etiquetado. El primero especifica las etapas: 1) la definición de fuentes de información; 2) los criterios de aceptación o rechazo de los documentos; 3) el filtro por tipología textual: novelas o relatos cortos, y 4) el filtro por lengua, únicamente en inglés. El segundo trabajo detalla el carácter de los textos y sus fuentes de consecución, pero deja de lado los elementos puntuales del etiquetado.

## III. RESULTADOS

Con el fin de construir corpus etiquetados que contengan suficiente información sobre las unidades correferenciales (anáforas y catáforas) al interior de un conjunto de textos y que este pueda ser usado como base de conocimiento para sistemas informáticos como los que tienen por objeto la enseñanza del español como lengua extranjera o para el entrenamiento de algoritmos de aprendizaje de máquina, se establece un método que define los pasos y las consideraciones que se deben tener en su ejecución, y además un sistema de etiquetas que reglan la estructura de los archivos trabajados.

Según lo anterior, es menester decidir cuál es la información más relevante que debe acompañar a cada una de las unidades etiquetables y qué factores son los más apropiados para tener en cuenta en el proceso de etiquetado. Es decir, qué datos se deberían priorizar. Al respecto, se piensa en tres factores principales, según la bibliografía antes señalada: los elementos deberían marcarse con

etiquetas que contengan información sobre 1) Estructuras sintácticas. 2) Información léxico-semántica. 3) Correferencias textuales.

Por otro lado, en cuanto a las fases y etapas, en la revisión realizada se encontró que los trabajos a menudo abordaron en las propuestas los aspectos en relación con la obtención de los textos, el etiquetado morfosintáctico y la corrección manual. No obstante, se dejan de lado consideraciones adicionales como la limpieza de los textos una vez obtenidos, y el almacenamiento y el etiquetado de correferencias textuales. Estos últimos se abordan en el método que se presenta en esta sección y que se aplicó en la consolidación de un corpus que sirve como base del conocimiento en una aplicación para la enseñanza del español como lengua extranjera.

El método que se presenta se divide en dos fases principales: la conformación del corpus, y su etiquetado, cada una dividida, a su vez, en tres etapas (ver Figura 2).

*A. Conformación del corpus*

Como acción previa al etiquetado, se hace necesario pasar por una fase de conformación del corpus, la cual consta de las etapas que se presentan a continuación:

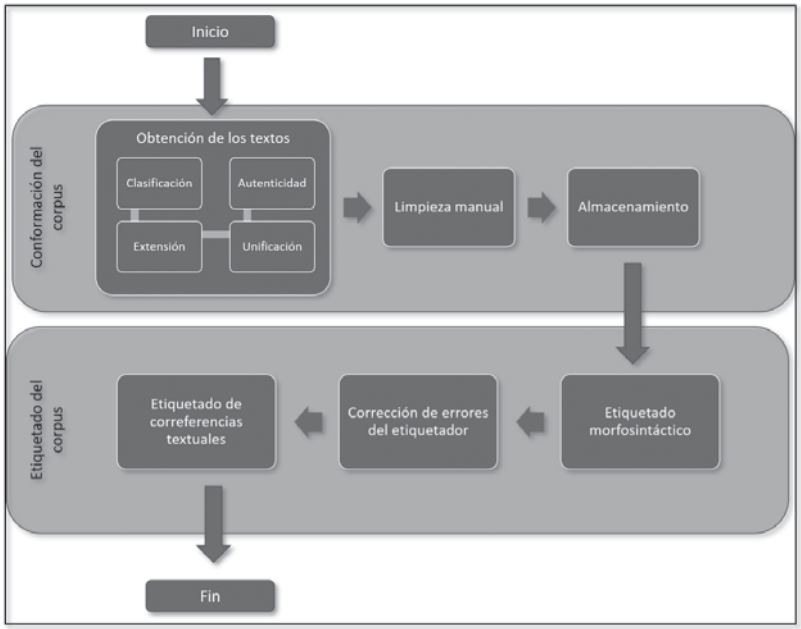


**Fig. 2.** Fases y etapas del método propuesto

*A.1. Obtención de los textos*

La primera etapa corresponde a la definición de las diferentes características que permitirán tener un conjunto de textos homogéneos, a saber:

a) *Clasificación de los textos según el nivel de dificultad y los parámetros planteados por el Marco Común Europeo de Referencia (MCER)*: Conviene, en primer lugar, obtener textos según determinados niveles del MCER (B1, B2 y C1, por ejemplo), puesto que estos serán necesarios para realizar actividades de forma automática en una aplicación basada en corpus. La clasificación de los textos en cada uno de estos niveles se hizo con la asistencia del programa informático presentado por Lamprinos [29], que permite calcular el nivel de dificultad de textos en español dentro del Marco Común Europeo de Referencia para las lenguas (MCER), y se buscó que incorporaran variedad con relación a la tipología: textos argumentativos, explicativos, informativos, etc.; y a las fuentes: periódicos, revistas de divulgación científica, extractos de novelas, ensayos breves, etc.

b) *Definición de la extensión de los textos*: Es importante tener un corpus lo más homogéneo posible en lo que respecta al número de *tokens* (palabras y signos de puntuación). Por ello se estableció un umbral recomendado de entre 500 y 1.000 palabras (alrededor de 800 y 1.200 tokens) para cada uno. Esto se estableció con el fin de que exista un equilibrio de contenido en los tres grupos (B1, B2 y C1), además de que los textos cortos facilitan la ejecución de actividades de enseñanza por la agilidad que permiten en el tiempo de las sesiones de estudio. A partir de la Tabla 1 es posible apreciar los valores aproximados para la cantidad de *types* (entradas únicas de cada palabra en el grupo de textos) y para los *tokens* (cantidad total de palabras de cada subcorpus).

**Tabla 1.** Cantidad de *types* y de *tokens* del corpus DICEELE.

| Nivel MCER | Cantidad *types* | Cantidad *tokens* |
|---|---|---|
| **B1** | 5.844 | 30.326 |
| **B2** | 7.518 | 37.621 |
| **C1** | 7.794 | 43.356 |
| **Total corpus** | 21.156 | 111.303 |

c) *Autenticidad de los textos*: En cuanto a la naturaleza de los textos, es condición que estos sean realizaciones auténticas del español. Se considera importante que los textos no hayan sido previamente didactizados, es decir, construidos artificialmente con

fines de enseñanza u otros fines, ya que esto permite el aprendizaje sobre un entorno más cercano a la realidad. Para ello, es posible hacer un rastreo en Internet y efectuar una selección y extracción de los textos de forma manual.

d) *Unificación del número de textos por nivel a trabajar*: En este caso se hace necesario establecer un número igual por nivel, que, como punto de partida, se recomienda sean 50 documentos por cada uno. No obstante, el corpus debe permitir ser incrementado, de forma igualmente homogénea a futuro, con la consecución de nuevos documentos que sigan los parámetros establecidos en los puntos antes expuestos.

*A.2 Limpieza manual*

Una vez que se tiene todo el corpus recolectado, se hace necesario pasar a la segunda etapa dentro de la primera fase: la limpieza manual. En este sentido, es importante realizar una revisión cruzada que permita validar que los textos cumplen con el grupo de características previamente definidas y que, además, están correctamente categorizados dentro de su tipología. También, que correspondan a realizaciones ejemplares del uso de la lengua en cuanto a los niveles ortográfico, sintáctico, lexical, etc. Es por eso necesario que se defina un número mínimo de errores corregibles, a partir del cual el texto no podría ser usado en el corpus, de tal manera que se conserve la naturaleza del primero. El problema de una intervención demasiado grande de los textos es que estos podrían perder su carácter de auténticos.

*A.3 Almacenamiento*

El almacenamiento es la etapa final de la primera fase, y una condición previa frente al proceso de etiquetado. Los recursos se deben unificar a partir de un formato de texto plano (formato *.txt). Esta unificación se hace necesaria toda vez que en la siguiente etapa se empleará el componente automático del método que consiste en etiquetar morfosintácticamente los recursos, con la herramienta *FreeLing* [10]. Por otra parte, es indispensable que los documentos cumplan con una codificación estandarizada de caracteres; en este caso en particular, se recomienda la codificación UTF-8, la cual permite la impresión de letras del alfabeto español como es el caso de la ñ, de las vocales tildadas y de algunos signos de puntuación.

*B. Etiquetado del corpus*

Una vez que se ha llevado a cabo todo el proceso de recolección de los textos, su limpieza y la codificación, se procede al etiquetado. Para ello, se tienen en cuenta dos niveles de etiquetado: el etiquetado morfosintáctico y el etiquetado de

las correferencias en el nivel de las unidades textuales y discursivas; este último basado en la lingüística textual.

*B.1 Etiquetado morfosintáctico*

Con el fin de realizar el proceso de etiquetado morfosintáctico, se recomienda usar la herramienta *FreeLing*. Esta es una herramienta libre para el procesamiento del lenguaje natural. *FreeLing* puede ser utilizado de dos formas: a) directamente sobre la página web del proyecto, en su versión demo, o b) por consola, es decir, descargando las librerías y ejecutando el etiquetado de manera local. Este analizador morfosintáctico presenta una documentación amplia, con una descripción generosa de cada *token*. El programa ofrece, además, la posibilidad de utilizar el formato XML en el etiquetado, lo que proporciona la estructura principal sobre la cual se puede hacer, *a posteriori*, todo el etiquetado manual textual y discursivo que se realizará en la etapa final.

En el nivel de etiquetado morfosintáctico se obtienen etiquetas como las de nombre, verbo, adjetivo, etc., para las categorías gramaticales, y las de género, número, persona, tiempo, modo y aspecto, etc., que acompañan a las mencionadas categorías gramaticales. Por otra parte, el sistema permite la segmentación en párrafos y en oraciones, anotados en el mencionado formato XML (ver Figura 3).


**Fig. 3.** Salida de texto etiquetado en formato XML utilizando *FreeLing* [10].

*B.2 Corrección de errores del etiquetador*

Por lo general todo sistema para el análisis y etiquetado de las lenguas naturales tiene un margen de error en la asignación de determinadas etiquetas. Esto se debe, fundamentalmente, a la ambigüedad que generan las lenguas naturales. Por esta razón también *FreeLing* incurre en errores de ambigüedad en el etiquetado, aunque con menos frecuencia que otros etiquetadores como *TreeTagger* o *TagAnt*. El nivel de precisión de *FreeLing* en el etiquetado es de 97 a 98 %, gracias a los modelos ocultos de Markov y al *relaxation labelling*, lo que permite la combinación estadística con reglas elaboradas de forma manual [30]. Debido a los márgenes de errores mencionados, se hace necesario que, una vez etiquetados los textos en una estructura XML, se realice una revisión manual de los resultados. Algunos de los errores que este sistema presenta son, *grosso modo*, la confusión entre el uso de la tercera persona y el registro formal de la segunda persona en el singular, el desconocimiento de palabras propias de un dialecto, o la marcación errónea de palabras que comienzan con mayúscula como nombres propios, entre otros.

*B.3 Etiquetado de correferencias textuales*

En virtud de los propósitos de uso del corpus para la generación automática de actividades que permitan la enseñanza del español, se hace necesario etiquetar diversos fenómenos textuales como la correferencia textual (uso de anáforas y catáforas), la progresión temática (tema y rema), los marcadores y conectores lógico-temporales y discursivos, etc. Hasta el día de hoy no existe una aplicación que lo haga de manera automática para el español [11], por lo tanto, esta etapa debe hacerse de forma manual sobre los XML obtenidos en la etapa anterior. Para este fin se hace uso de los atributos ID e IDREF que arroja el etiquetador. Anotar estas nociones permite llevar a cabo todo un trabajo en la lengua española alrededor de las nociones de coherencia y de cohesión textuales.

Finalmente, una vez obtenidos todos los documentos XML de los recursos que conformarán el corpus, se recomienda su validación a partir de la definición del tipo de documento que especifica todos los documentos en su versión XML (DTD: *Document Type Definition*).

*C. Conformación de la DTD*

La DTD es un documento que permite, por una parte, la estructuración, lógica y gramática de todo un grupo de textos o de documentos que previamente han sido etiquetados o anotados bajo el formato XML. Por otra parte, la DTD permite la verificación y la validación de dichos documentos con el fin de

que contengan, efectivamente, todos los elementos que se necesitan para la realización de actividades, o que no contengan, si tal es el caso, elementos innecesarios. Además, una DTD puede mapearse informáticamente en el momento en que se quieren realizar acciones como, en este caso, la generación automática de actividades.

La DTD contendrá, además de los elementos etiquetados de manera morfosintáctica y textual, los metadatos necesarios para la realización de los ejercicios. Esto permite que se puedan buscar los textos por nivel, tipo de texto, autor, variante del español, etc. En la Figura 4 se presenta una parte de la DTD recomendada.[1]

La revisión de literatura permitió conocer otras formas de abordar el etiquetado de un corpus. En el caso del uso para la enseñanza del español como lengua extranjera, surgieron particularidades que demandaron una propuesta, fundamentalmente el dominio característico de los textos elegidos, evidenciado en la segunda fase del etiquetado en donde estos se validan: el almacenamiento y el etiquetado manual de las correferencias textuales.

## IV. DISCUSIÓN

El método presentado facilita la creación de corpus con las condiciones necesarias para ser utilizados en la enseñanza de lenguas. Es conveniente, siempre, que los sistemas para la enseñanza de ELE sean parametrizables, es decir que puedan dinamizarse sus usos con tipos de textos y ejercicios variados y dinámicos.

El método fue aplicado en la construcción del corpus DICEELE: Dispositivo Informático basado en Corpus para la Enseñanza del Español Lengua Extranjera [11]; se trabajaron 150 textos recolectados de internet, originarios de Colombia, España, Venezuela, México, Argentina, Panamá, Honduras, Nicaragua, Cuba, Ecuador, Uruguay y Perú, y han sido organizados de acuerdo con los niveles B1, B2 y C1 del MCER. Las tipologías textuales recuperadas fueron informativa, narrativa, argumentativa y explicativa, y entre sus diferentes microtipologías se destacan: noticias, cartas, relatos cotidianos, columnas de opinión, reseñas, instrucciones y crónicas. Cada texto cuenta con metadatos que dan cuenta de informaciones tales como el título, el autor, el tipo de publicación, la ubicación, el género textual, el país de procedencia y el investigador encargado de su recolección y sistematización. En la Figura 5 se presenta un ejemplo de cómo se verían etiquetados los elementos correferenciales (unidades anafóricas): *veían, la policía, se, el patrullero, esos chapas*, que retoman al referente de tipo GN (grupo nominal) *un carro de la policía*. Adicionalmente,

1 La DTD completa puede ser solicitada a los autores.

se puede observar el marcado morfológico que del referente principal se hace, por ejemplo, la palabra *carro*, se presenta como un *nombre* (sustantivo), *masculino*, *singular*; esta información puede extraerse de la codificación *NCMS0000* realizada por *FreeLing*.

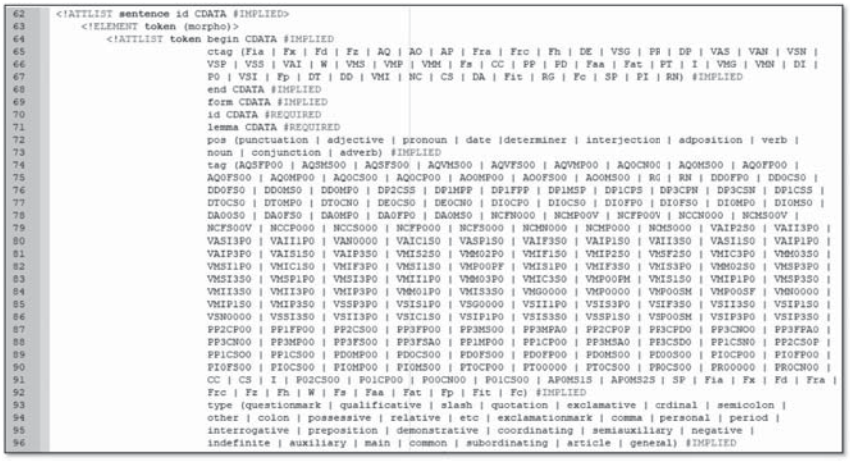


**Fig. 4.** Fragmento de la DTD en la que se aprecian parte de las etiquetas contenidas en los documentos XML

```
<REF ANT="nom" type="SN" id="50" from="268" to="272" chain="8">

    <element morpho="DI0MS0" id="268">un</element>
    <element morpho="NCMS000" id="269">carro</element>
    <element morpho="SP" id="270">de</element>
    <element morpho="DA0FS0" id="271">la</element>
    <element morpho="NCCS000" id="272">policía</element>
</REF>
<chain id="8">
<REF ANT="nom" type="SN" id="50" words="un carro de la policía "/>
<COREF ANA="nom" type="SN" id="53" words="la patrulla policial "/>
<COREF ANA="pron" type="imp" id="54" words="veían "/>
<COREF ANA="nom" type="SN" id="62" words="la policía "/>
<COREF ANA="nom" type="SN" id="85" words="la policía "/>
<COREF ANA="pron" type="PP" id="87" words="se "/>
<COREF ANA="nom" type="SN" id="97" words="esos chapas "/>
<COREF ANA="nom" type="SN" id="98" words="el patrullero "/>
</chain>
```

**Fig. 5.** Etiquetas de correferencia textual: anáforas

Si bien el método se puso en marcha en el contexto mencionado, es decir para un contexto de aplicación de ELE, puede ser extendido a otros contextos como el entrenamiento de algoritmos de aprendizaje de máquina.

En los modelos de aprendizaje de máquina, a menudo implementados a través de redes neuronales, la capacidad del aprendizaje hacia uno profundo (*deep learning*) es posible gracias a la obtención de suficientes datos para

etapas de entrenamiento, validación y verificación. Por eso se hace necesaria la creación de corpus en distintas lenguas, con objetivos de marcaje de la información variados, y que funcionen como bases de conocimiento para la algoritmia tras la inteligencia artificial. El tamaño de las redes neuronales y la cantidad suficiente de datos es, por lo tanto, fundamental para impulsar el avance del aprendizaje profundo y alcanzar, así, un nivel más preciso de la extracción de sentido de las distintas maneras que tienen los humanos de producirlo. En esta línea es esencial avanzar más allá del análisis sintáctico hacia una integración con modelos de conocimiento que permitan establecer una relación lenguaje-realidad que se aproxime más a las características y capacidades humanas, características que se consiguen, en un grado inicial, a partir del marcado de correferencias.

## V. CONCLUSIONES

En este trabajo se presentó un método que conduce a la consolidación de un corpus para su uso en sistemas informáticos que medien la enseñanza del español como lengua extranjera (ELE), que, sin embargo, se puede extender a otros casos de aplicación. Se introduce la herramienta *FreeLing*, que asiste en el etiquetado morfosintáctico en la etapa automatizada. Sobre esta última también se expusieron sus principales limitantes, las cuales se deben tener en consideración siempre que se utilice, y de esta forma se podrán realizar los ajustes, de manera manual, a los que haya lugar.

El método presentado incorpora las fases de consolidación de los textos, así como el proceso de etiquetado como tal, y se sugiere un conjunto de pasos que, llevados a cabo, permitan garantizan la idoneidad del corpus. Se trata de un método semiautomático, puesto que el etiquetado morfosintáctico se efectúa de manera automática, mientras que el etiquetado de correferencias textuales debe realizarse de manera manual o, en su defecto, a partir de algunos procedimientos que permitan agilizar ciertos pasos del proceso de consolidación de las anotaciones, por medio de *scripts* programados, por ejemplo, en *Python* o en *Perl*.

## VI. REFERENCIAS

[1] S. A. Moreno, *Lingüística computacional: Introducción a los modelos simbólicos, estadísticos y biológicos.* Madrid: Editorial Síntesis, 1998.

[2] R. Hausser, *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*. New York: Springer Science & Business Media, 2013.

[3] G. Parodi, *Lingüística de Corpus: de la teoría a la empiria.* Madrid: Iberoamericana Vervuert, 2010.

[4] J. Bernal y D. Hincapié, *Lingüística de corpus*. Colombia: Instituto Caro y Cuervo. 2018.

[5] J. M. Molina, "Lingüística de corpus utilizada en la concepción de plataformas de formación de docentes de lenguas", *CHIMERA: Romance Corpora and Linguistic Studies*, vol. 3, n.° 1, pp. 57-85, 2016.

[6] M. Recasens and M. A. Martí, "AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan", Lang Resources & Evaluation, n.° 44, 2010.

[7] M. Palomar, L. Moreno, J. Peral, R. Muñoz, A. Fernández, P. Martínez-Barco, and M. Saiz-Noeda, "An Algorithm for Anaphora Resolution in Spanish Texts", *Association for Computational Linguistics*, vol. 27, n.°, 4, pp. 545-567, 2001.

[8] G. Kundu, A. Sil, R. Florian, and W. Hamza. "Neural Cross-Lingual Coreference Resolution and its Application to Entity Linking", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Short Papers), Melbourne, Australia, July 15-20, 2018, pp. 395-400.

[9] A. F. Grajales Ramírez y J. M. Molina Mejía, "Problemática actual del procesamiento computacional anafórico: el caso de FreeLing 4.1", *Lenguaje*, vol. 47, n.° 2S, pp. 537-568, 2019. doi: 10.25100/lenguaje.v47i3.8356

[10] L. Padró and E. Stanilovsky, "FreeLing 3.0: Towards Wider Multilinguality", in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, ELRA. Istanbul, Turkey. May 2012.

[11] J. M. Molina, A. F. Grajales, and J. L. Pemberty, "Hacia un dispositivo informático basado en Corpus para la Enseñanza del Español Lengua Extranjera (DICEELE)", *Revista Internacional de Tecnología, Conocimiento y Sociedad*. vol. 7, n.° 1, pp 1-13, 2019. doi: 10.18848/2474-588X/CGP/v07i01/1-13

[12] M. A. Martí, *Introducción Tecnologías del lenguaje*, Barcelona: Editorial UOC, 2003.

[13] A. Boulton et H. Tyne, *Des documents authentiques aux corpus: démarches pour l'apprentissage des langues.* Paris. Didier Coll Langues & didactique, 2014.

[14] G. Antoniadis. "De l'apport pertinent du TAL pour les systèmes d'ALAO. L'exemple

du projet MIRTO", en 2eme Congrès Mondial de Linguistique Française (CMLF-2010). La Nouvelle Orléans, USA, 12-15 juillet 2010.

[15] J. M. Molina, "ELiTe-[FLE]: Un environnement d'ALAO fondé sur la linguistique textuelle, pour la formation linguistique des futurs enseignants de FLE en Colombie », Tesis doctoral, Université Grenoble Alpes, 2015.

[16] R. Mitkov, "Outstanding Issues in Anaphora Resolution", In *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed., Springer, 2001, pp. 110-125.

[17] F. Landragin, "Anaphores et coréférences : analyse assistée par ordinateur", en *Nouvelles perspectives sur l'anaphore: Points de vue linguistique, psycholinguistique et acquisitionnel*, M. Fossard & M.-J. Béguelin, Eds., Peter Lang. 2014, pp. 29-54.

[18] H. Schmid. "Probabilistic Part-of-Speech Tagging Using Decision Trees", in Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994.

[19] D. Klein and C. D. Manning. "Accurate Unlexicalized Parsing", in *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003, pp. 423-430.

[20] X. Carreras, I. Chao, L. Padró and M. Padró. "FreeLing: An Open-Source Suite of Language Analyzers", in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), 2004.

[21] L. Anthony. "TagAnt (Version 1.1.0) [Computer Software]". Tokyo, Japan: Waseda University, 2015. Available from http://www.laurenceanthony.net/software

[22] I. Castellón; S. Climent; M. Coll-Florit; M. Lloberes and G. Rigau. "Constitución de un corpus de semántica verbal del español: Metodología de anotación de núcleos argumentales", *RLA. Revista de lingüística teórica y aplicada*, vol. 50, n.° 1, 2012, pp. 13-38.

[23] M. del M. Sánchez, "Metodología de corpus y formación en la traducción especializada (inglés-español): una propuesta para la mejora de la adquisición de vocabulario especializado", *Revista de Lingüística y Lenguas Aplicadas*, vol. 12, n.° 1, 2017, pp. 137.

[24] M. A. Martí, M. Taulé, L. Màrquez, and M. Bertran. "Anotación semiautomática con papeles temáticos de los corpus CESS-ECE", Procesamiento del Lenguaje Natural, n.° 38, 2017, pp. 67-76.

[25] M. Anxo; S. Portela and G. Xavier. "Diseño y elaboración del corpus SemCor del gallego anotado semánticamente con WordNet 3.0", *Procesamiento del Lenguaje Natural*, N.° 59, 2017, pp. 137-140.

[26] S. Zafra Cremades, J. M. Gómez Soriano, and B. Navarro-Colorado. "Diseño, compilación y anotación de un corpus para la detección de mensajes suicidas en redes sociales", *Procesamiento del Lenguaje Natural*, vol. 59, n.° 0, 2017, pp. 65-72.

[27] M. Rioja Barrocal. "Metodología para la narrativa de traducciones censuradas inglés-español: análisis del Corpus 0 TRACEni", in Estudios Humanísticos. Filología. 317. 10.18002/ehf. v0i29.2821, 2007.

[28] A. R. Vila. "Canon y corpus literario latinoamericano y caribeño. Una metodología de construcción del corpus", in Palabra Clave (La Plata), vol. 7, N.° 2, 2018, pp. 0-12.

[29] D. Lamprinos, "Clasificación automática de textos españoles en los niveles del Marco Común Europeo de Referencia para las Lenguas", Tesis de maestría, Universidad Nacional y Kapodistriaca de Atenas y la Universidad Politécnica Nacional de Atenas, 2014.

[30] L. Padró, "Analizadores Multilingües en FreeLing", in Linguamatica, vol. 3, N.° 2, 2011, pp. 13–20.