

UN ENFOQUE BASADO EN CORPUS PARA DESCUBRIR RELACIONES SEMÁNTICAS ENTRE ENTIDADES NOMBRADAS

A corpus-based approach to discovering semantic relationships between named entities

Lic. Reynier Ávila Peña *¹, <https://orcid.org/0000-0002-1523-8404>

Dr. C. Celia María Pérez Marqués ², <https://orcid.org/0000-0002-0243-8159>

¹ Empresa de desarrollo de aplicaciones, tecnología y sistemas (Datys), Cuba

² Universidad de Oriente, Cuba

*Autor para correspondencia. email avilascorpio@gmail.com

Para citar este artículo: Ávila Peña, R. y Pérez Marqués, C. M. (2023). Un enfoque basado en corpus para descubrir relaciones semánticas entre entidades nombradas. *Maestro y Sociedad*, (Número Especial), 23-31. <https://maestrosociedad.uo.edu.cu>

RESUMEN

Introducción: El objetivo de este estudio es analizar un texto noticioso relacionado con la identidad cultural, que forma parte de un corpus lingüístico etiquetado, con el fin de anotar las relaciones sintácticas y semánticas entre las entidades nombradas en el texto. Materiales y métodos: Se presenta una clasificación de las relaciones semánticas establecidas entre las entidades nombradas y cómo funcionan en un formato XML etiquetado, utilizando el etiquetado gramatical y el análisis sintáctico. Se etiquetaron 20 entidades nombradas, 13 relaciones gramaticales y 36 relaciones semánticas. Resultados: La propuesta presentada en este artículo resulta ser útil para desarrollar y evaluar nuevos sistemas abiertos de extracción de información en español. Discusión: Los corpus lingüísticos, la lingüística de corpus y la lingüística computacional son herramientas valiosas en el proceso de aprendizaje automático de las computadoras para entender el lenguaje natural. El análisis de relaciones semánticas y sintácticas entre entidades nombradas en un texto noticioso es crucial para la extracción de información relevante y la identificación de patrones lingüísticos. Conclusiones: Este estudio destaca la relevancia de los corpus lingüísticos etiquetados y la lingüística de corpus en el análisis del lenguaje natural y en el desarrollo de sistemas de procesamiento de lenguaje natural capaces de comprender y analizar el lenguaje humano en diferentes contextos. La importancia de este trabajo radica en la necesidad de desarrollar sistemas de procesamiento de lenguaje natural que permitan a las computadoras comprender y analizar el lenguaje humano en diferentes contextos.

Palabras clave: lingüística de corpus, entidades nombradas, corpus lingüísticos, lingüística computacional.

ABSTRACT

Introduction: The objective of this study is to analyze a news text related to cultural identity, which is part of a labeled linguistic corpus, in order to annotate the syntactic and semantic relationships between the named entities in the text. Materials and methods: A classification of the semantic relationships established between the named entities and how they function in a labeled XML format is presented, using grammatical tagging and syntactic analysis. 20 named entities, 13 grammatical relationships, and 36 semantic relationships were tagged. Results: The proposal presented in this article proves to be useful for developing and evaluating new open information extraction systems in Spanish. Discussion: Linguistic corpora, corpus linguistics, and computational linguistics are valuable tools in the process of machine learning for natural language understanding. The analysis of syntactic and semantic relationships between named entities in a news text is crucial for relevant information extraction and linguistic pattern identification. Conclusion: This study highlights the relevance of labeled linguistic corpora and corpus linguistics in the analysis of natural language and in the development of natural language processing systems that are capable of understanding and analyzing human language in different contexts. The importance of this work lies in the need to develop natural language processing systems that enable computers to understand and analyze human language in different contexts.

Keywords: corpus linguistics, discourse analysis, named entities, linguistic corpus, computational linguistics.

INTRODUCCIÓN

La Lingüística de Corpus (LC) es un enfoque metodológico que se basa en datos obtenidos a partir de corpus lingüísticos, y no una disciplina lingüística en sí misma. Este enfoque no solo es útil para describir lenguas o variedades de lenguas que no conocemos, sino también para aquellas que ya poseen una larga tradición descriptiva como el inglés o el español. La lingüística de corpus ha aportado una importante cantidad de datos que antes eran impensables, lo que ha modificado la percepción sobre dichas lenguas de las que aparentemente se tenía todo el conocimiento posible. Esto ha llevado a muchos a ver una especie de revolución en la lingüística computacional. Se enfoca en la construcción de modelos de lenguaje "entendibles" para las computadoras, es decir, en la realización de aplicaciones informáticas que imiten la capacidad humana de hablar y entender. Esto implica transformar el conocimiento sobre los fenómenos lingüísticos que ya se conocían y aplicarlos a la creación de sistemas computacionales capaces de procesar y entender el lenguaje natural. En resumen, la lingüística de corpus ha permitido obtener una gran cantidad de datos lingüísticos que antes no estaban disponibles, lo que ha llevado a una revolución en la lingüística computacional y a la creación de sistemas informáticos que imitan la capacidad humana de hablar y entender el lenguaje natural.

La Lingüística de Corpus ha generado una serie de métodos de investigación que buscan conectar los datos y la teoría. Wallis y Nelson (2001) introdujeron la perspectiva de las tres A (3A perspective): anotación, abstracción y análisis. La anotación consiste en la aplicación de un esquema para textos, que puede incluir etiquetado estructural, etiquetado gramatical y análisis sintáctico. La abstracción implica la traducción o mapeo de términos a un conjunto de datos, lo que incluye la búsqueda lingüística dirigida y puede admitir la regla de aprendizaje para los analizadores. El análisis implica sondear, manipular y generar el conjunto de datos, todo ello de manera estadística. Esto podría incluir evaluaciones estadísticas, optimización de bases de reglas o métodos de descubrimiento de conocimiento. En resumen, la perspectiva de las tres A proporciona un marco para la realización de investigaciones de lingüística de corpus. La anotación, abstracción y análisis se refieren a las etapas principales del proceso de investigación y son fundamentales para la extracción de información y la generación de conocimiento a partir de los datos obtenidos a partir de corpus lingüísticos.

En la Lingüística de Corpus, el etiquetado de palabras se refiere a la asignación de categorías sintácticas a cada palabra de un texto o corpus, lo que se conoce como Part Of Speech Tagging (POS Tagging) en inglés. Este proceso requiere de un conjunto predefinido de etiquetas (tagset) y un algoritmo de asignación de etiquetas.

Un corpus es una recopilación de muestras reales de una lengua, que pueden ser novelas, obras de teatro, guiones de cine, noticias de prensa, ensayos, transcripciones de programas de radio o televisión, conversaciones o discursos. Otras definiciones de corpus son: "una colección de texto lingüístico de ocurrencia natural seleccionada para caracterizar un estado o variedad de una lengua" (Sinclair, 1991), "una colección de textos reunidos según criterios precisos, eventualmente estructurados y enriquecidos con información adicional, en vista de una explotación teórica o práctica" (Mercado, 2008), y "una recopilación de textos seleccionados según criterios lingüísticos, codificados de modo estándar y homogéneo, con la finalidad de poder ser tratados mediante procesos informáticos y destinados a reflejar el comportamiento de una o más lenguas" (Torruela & Llisterra, 1999).

El corpus puede aparecer en línea o en formato electrónico debido a su gran tamaño. Las muestras se seleccionan a partir de criterios objetivos que se establecen previamente y que buscan ofrecer una representación de algún aspecto de la lengua. La representatividad es la piedra angular de la Lingüística de Corpus, ya que de ella depende que se puedan extraer conclusiones fiables a partir de los datos estadísticos (Cruz, 2012, p. 36).

Una Entidad Nombrada (NE) es una frase que identifica un elemento de un conjunto de otros elementos que tienen atributos similares. En términos generales, puede referirse a cualquier cosa con un nombre propio, como una persona, un lugar o una organización. El término también se extiende para incluir fechas, tiempos y otros tipos de expresiones temporales, así como expresiones numéricas como precios (Jurafsky y Martin, 2017).

El objetivo de este trabajo es anotar las relaciones sintácticas y semánticas entre entidades nombradas en un texto noticioso. Este tema es importante porque estas anotaciones son útiles para evaluar y desarrollar nuevos sistemas abiertos de extracción de información en español.

Para llevar a cabo este trabajo, se siguió un procedimiento que constó de varias etapas. En primer lugar, se realizó una revisión ortográfica del texto. A continuación, se segmentó el texto y se etiquetó gramaticalmente

cada una de las palabras. Además, se detectaron y resolvieron las correferencias para identificar las entidades nombradas y asignarles una etiqueta correspondiente.

Posteriormente, se procedió a analizar los patrones de relaciones entre las diferentes entidades nombradas. De esta manera, se pudieron identificar las diversas asociaciones entre las entidades de un conjunto y las entidades de otro conjunto.

MATERIALES Y MÉTODOS

Etiquetado de un texto noticioso del ámbito cultural

Las principales Etiquetas usadas para clasificar las entidades nombradas son:

PERSON(PER)

ORGANIZATION(ORG)

LOCATION(LOC)

EVENT(EVN)

MATTER(MAT)

DOCUMENT(DOC)

MISCELANEOUS(MIS)

QUANTITY(QNT)

PERCENTAGE(PRC)

MONETARY_QTY(MNQ)

DATE(DAT)

TIME(TIM)

PERSON_GROUP(G_PER)

ORGANIZATION_GROUP(G_ORG)

LOCATION_GROUP(G_LOC)

EVENT_GROUP(G_EVN)

DOCUMENT_GROUP(G_DOC)

MISCELANEOUS_GROUP(G_MIS)

USER_TWITTER(U_TWT)

TAG_TWITTER(T_TWT)

EMAIL(EMAIL)

URL(URL)

PRODUCT(PRO)

PRODUCT_GROUP(G_PRO)

FACILITY(FAC)

FACILITY_GROUP(G_FAC)

La clasificación de las entidades en un tipo particular depende de la naturaleza del elemento que se está identificando. Por ejemplo, existen tipos de entidades definidos por las competencias de evaluación, tales como persona, lugar, organización, evento, miscelánea, fecha, cantidad monetaria, entre otros.

En los nombres de entidades detectados puede haber relaciones o conexiones predefinidas o no según Culotta *et al.* (2006). La tarea de Extracción de Relaciones (ER, por sus siglas en inglés) ha sido reconocida como un problema importante y difícil para los investigadores de las ramas de la lingüística, filosofía y ciencias de la computación. Esta tarea tiene como objetivo encontrar y clasificar relaciones semánticas entre

las entidades nombradas del texto. A menudo, estas relaciones son binarias, como "cónyuge-de", "hijo-de", "empleo", "parte-de", "membresía" y relaciones geoespaciales (según Jurafsky y Martin, 2017). La Extracción de Relaciones se puede aplicar en tareas como Búsqueda de Respuestas, Bioinformática, Construcción de Ontologías y otras áreas.

El ejemplo a analizar es el siguiente:

Cuba celebra 80 años del debut de la legendaria bailarina Alicia Alonso

El Ballet Nacional de Cuba celebra este jueves el 80 aniversario del debut artístico de Alicia Alonso, una de las mejores bailarinas de todos los tiempos, fundadora y directora de esta compañía desde 1948, informó la prensa local. Con una gala nocturna en el Gran Teatro de La Habana, meca del ballet en la isla, "la compañía recordará aquella primera vez en que Alicia, el 29 de diciembre de 1931, salió a escena en una función de la escuela de ballet de la Sociedad Pro Arte Musical", señaló el diario oficial Granma. "Esta noche baila Alicia", pero "la gran bailarina y coreógrafa", que cumplió hace ocho días 91 años, no "lo hará físicamente", sino "con el alma y el corazón", añadió el periódico. Según Granma, el "programa conmemorativo" incluye la vuelta a escena de varias "obras coreografiadas" por Alonso, entre ellas selecciones de "La bella durmiente del bosque" y de la "Flauta mágica", así como "Preciosa y el aire". Alonso, quien ostenta el rango de "prima ballerina assoluta", el más alto al que puede aspirar un artista de la danza, está casi ciega y tiene problemas para caminar, pero dirige activamente a su compañía y la acompaña a cada una de sus presentaciones internacionales. Embajadora cultural de la revolución cubana y muy respetada por su talento y entrega al arte, Alonso estudio ballet en La Habana, pero luego fue a Estados Unidos, donde comenzó su carrera con el New York City Ballet. Se convirtió en estrella mundial, como figura del American Ballet Theatre (ABT). En 2010, la leyenda cubana de la danza recibió una serie de tributos de prestigiosas compañías del mundo por su contribución al ballet, entre ellas del Teatro Bolshoi de Moscú y el ABT .

El texto trata sobre la celebración del 80 aniversario del debut artístico de la Prima Ballerina Alicia Alonso. Con su estilo, la bailarina marcó al ballet cubano de tal manera que ha influido sobremanera en los cinco continentes, creando una originalidad en cada una de sus presentaciones. Se destaca la disciplina y el rigor que ha llevado a la escuela cubana de Ballet. Aunque ya no puede bailar físicamente, su legado perdurará eternamente.

En este caso, la noticia es un tipo de texto que puede ser escrito, auditivo o audiovisual, que consiste en narrar de manera precisa algún hecho novedoso con interés público, como lo es el debut de Alicia Alonso. Dentro de los géneros discursivos, se encuentran el científico, publicitario, epistolar, judicial y periodístico, siendo este último el abordado en este trabajo.

En el análisis se puede observar que la noticia solo se limita a los hechos y no emite una opinión excesiva ni toma una posición en particular. El lenguaje utilizado es divulgativo y no especializado, lo que permite que cualquier persona pueda acceder a la información de manera clara y coherente. Además, la noticia es directa y no abusa de recursos lingüísticos como la metáfora o la jerga popular, ni utiliza adjetivos relacionados con juicios de valor o morales. Esto se hace para evitar que el lector interprete la información de manera diferente a como se pretendía. Además, no se utilizan exclamaciones.

Las frases y oraciones son, en su mayoría, concisas y breves, utilizando una construcción sintáctica simple: sujeto + verbo + complementos. Se utiliza la voz activa en lugar de la pasiva, frases afirmativas en lugar de negativas, y se evitan subordinaciones e incisos. La temática de la noticia es cultural. Consta de ocho oraciones gramaticales y 20 entidades nombradas que son:

Ballet Nacional de Cuba

jueves

80 aniversario del debut artístico de Alicia Alonso

1948

diario oficial *Granma*

Gran Teatro de La Habana

Alicia Alonso

29 de diciembre de 1931

escuela de ballet de la Sociedad Pro Arte Musical

ocho días

91 años

La bella durmiente del bosque

Flauta mágica

Preciosa y el aire

La Habana

Estados Unidos

New York City Ballet

American Ballet Theatre

2010

Teatro Bolshoi de Moscú.

En un texto, las palabras no se presentan de manera aislada, sino que están interrelacionadas para transmitir un mensaje, ya sea de forma explícita o implícita. Lo mismo sucede con las entidades nombradas presentes en una oración o texto, las cuales son frases que pueden incluir cualquier tipo de palabra, pero principalmente están representadas por sustantivos.

Para comprender toda la información que se brinda acerca de estas entidades, es necesario realizar una correcta interpretación semántica. Sin embargo, detectar las relaciones semánticas entre ellas puede ser complejo debido a la diversidad de significados de las palabras o frases y a las distintas formas de expresar una misma idea, especialmente por la ambigüedad semántica del lenguaje.

Por esta razón, se utiliza una clasificación de relaciones semánticas basada en el análisis sintáctico-gramatical de las entidades nombradas.

Correferencia: Relaciones donde las entidades significan o representan el mismo concepto. Este se presenta fundamentalmente a partir de sustantivos en aposición o expansión de siglas. Los pronombres personales yo, tú, usted, él, ella, sus variantes pronominales me, te, la, le, lo y las estructuras a sí mismo, a mí mismo con sus variantes en género y número, los pronombres posesivos su, sus, siempre que estén separados del verbo (el pronombre se no se incluye).

Ejemplo: <relation_expresion SOURCE="Américan Ballet Theatre" TARGET="ABT" TYPE="SUST"REAL_REL="("SEMANTIC_TYPE="COREF"></relation_expresion><phrase explicit="FALSE" coreferente="Américan Ballet Theatre">ABT</phrase>).

Físico_Ubicación: Describe una localización física que se establece entre una persona o evento, en un lugar.

Ejemplo: <relation_expresion SOURCE="Alonso" TARGET="La Habana" TYPE="VERB" REAL_REL="estudiar_en"SEMANTIC_TYPE="PHIS_Loc">estudió </relation_expresion>ballet en<phrase explicit="TRUE">La Habana</phrase>

Temporal: Describe la relación que se establece entre entidades de tipo: persona, evento, organización y lugar con la entidad de tipo fecha. Así como la relación entre las entidades de tipo persona, evento y lugar con la entidad de tipo hora. En esta relación el Argumento 2 siempre será una entidad tipo fecha y hora.

Ejemplo: <relation_expresion SOURCE="Ballet Nacional de Cuba" TARGET="jueves" TYPE="VERB"REAL_REL="celebrar"SEMANTIC_TYPE="TEMP">celebra </relation_expresion>este <phrase explicit="TRUE">jueves </phrase>

Parte-Todo_Geográfico: representa la ubicación de una instalación o un lugar como parte de otro centro o lugar. Describen relaciones de entidades que se pueden encontrar en un mapa o plano. Estas son permanentes, aunque puede haber excepciones. En la relación, las dos entidades aparecerán en sintagmas nominales diferentes.

Ejemplo: <relation_expresion SOURCE="Gran Teatro" TARGET="La Habana" TYPE="PREP-pertenencia"REAL_REL="de" SEMANTIC_TYPE="PART_WHL_Geo">de</relation_expresion><phrase explicit="TRUE" NESTED="TRUE">La Habana</phrase>

Parte-Todo_Subsidiario: Representa la relación entre entidades de tipo Organización y Evento con las

entidades de tipo Organización y Lugar. Esta incluye la relación entre una empresa y su empresa matriz, así como entre un departamento de una organización y esa organización. También incluye la relación entre las organizaciones y el gobierno de una localidad.

Ejemplo: <relation_expresion SOURCE="Ballet Nacional" TARGET="Cuba" TYPE="PREP-pertenencia"REAL_REL="de" SEMANTIC_TYPE="PART_WHL_Sub">de</relation_expresion>
<phrase explicit="TRUE" TYPE="LOC" NESTED="TRUE">Cuba</phrase>

Organización-Afiliación_Empleo: Se establece entre personas y sus empleadores. Es solo etiquetada cuando puede ser razonablemente asumido que una persona es pagada por una organización o un lugar.

Ejemplo: <relation_expresion SOURCE="Alonso" TARGET="New York City Ballet" TYPE="VERB"REAL_REL="comenzar_con" SEMANTIC_TYPE="ORG_AFF_Emp"/><phrase explicit="TRUE">New York City Ballet</phrase>

Otra-Relación: Describe las relaciones entre entidades que no se encuentran en las clasificaciones anteriores. El texto está etiquetado con el formato xml ¹.

Ejemplo: <relation_expresion SOURCE="Ballet Nacional de Cuba" TARGET="80 aniversario del debut artístico de Alicia Alonso" TYPE="VERB" REAL_REL="celebrar" SEMANTIC_TYPE="OTHER_RELATION">el </relation_expresion>

<phrase explicit="TRUE">80 aniversario del debut artístico de<phrase explicit="TRUE" NESTED="TRUE">Alicia Alonso</phrase>

Las primeras etiquetas del esquema son title, topic y date. En el caso de la primera, se introduce en ella el título de la noticia; la segunda, se utilizaría, en el futuro, para realizar estadísticas sobre qué tipo de relaciones entre entidades nombradas tiene una mayor o menor ocurrencia en determinadas temáticas; y la tercera se muestra la fecha. Ejemplo:

<title>Cuba celebra 80 años del debut de la legendaria bailarina Alicia Alonso</title>
<topic>Cultura</topic>
<date>29/12/2011</date>

Bajo la etiqueta entities se enumeran las representaciones léxicas más extensas de cada una de las entidades que aparecen en el texto. La etiqueta entity tiene los atributos: type para el tipo de la entidad (persona, organización, lugar, evento, documento, materia, miscelánea, fecha, hora, por ciento, cantidad...); globalPolarity describe la polaridad de la entidad y admite los valores positivo, negativo o neutro; y formalEntity que expresa el nombre real de la entidad.

En caso de que un atributo no tenga valor se suprime de la etiqueta. Ejemplo:

<entity type="PER" globalPolarity="POS">Alicia Alonso</entity>
<entity type="ORG" globalPolarity="NONE" formalEntity="Periódico Granma">diario oficial Granma</entity>

En las oraciones (sentence) se especifican los atributos: id que es el número de la oración; related para expresar si en la oración aparece alguna entidad nombrada o no; mediante los valores TRUE/FALSE y subjectivity que indican el valor de objetivo o subjetivo de la frase de acuerdo con lo que expresa.

La etiqueta original_text contiene la oración como se presenta en la noticia y la etiqueta text tiene la oración etiquetada por los especialistas; phrase marca las entidades nombradas y contiene los siguientes atributos: explicit (si está explícitamente la entidad en la oración), coreferente para señalar que la frase es una correferencia, TYPE describe el tipo de la entidad y NESTED para especificar si la entidad está anidada dentro de otra frase. Si las entidades se etiquetan como anidada, significa que una frase puede contener otras frases. Por ejemplo, en la entidad de Organización: Ballet Nacional de Cuba, la entidad de Lugar: Cuba está anidada:

<phrase explicit="TRUE"><phrase explicit="TRUE" TYPE="ORG"NESTED="TRUE">Ballet Nacional </phrase><relation_expresion SOURCE="Ballet Nacional" TARGET="Cuba" TYPE="PREP-pertenencia"REAL_REL="de"SEMANTIC_TYPE="PART_WHL_Sub">de</relation_expresion><phrase explicit="TRUE" TYPE="LOC" NESTED="TRUE">Cuba</phrase>

1 Es un lenguaje de marcado que define un conjunto de reglas para la codificación de documentos. El lenguaje de marcado es un conjunto de códigos que se pueden aplicar en el análisis de datos o la lectura de textos creados por computadoras o personas.

Se decidieron anotar relaciones de tipo sintáctico, es decir, que partieran de elementos compositivos de la oración como son preposiciones, verbos y sustantivos, que relacionan una entidad con otra a nivel de sintaxis.

Para marcar las relaciones se usa la etiqueta `relation_expresion` con los atributos: `SOURCE` y `TARGET` que muestran, a través de los representantes léxicos más extensos definidos en la etiqueta `entities`, las entidades relacionadas; `TYPE` para describir el tipo de relación que se determina, en dependencia de la categoría gramatical de las palabras que se encuentran entre los argumentos, `REAL_REL` para especificar la o las palabras que vinculan el par de entidades en análisis, `SEMANTIC_TYPE` para clasificar la relación de manera semántica y `DIRECTION` para indicar que el orden en que aparecen las entidades está invertido.

Ejemplo: Embajadora cultural de la revolución cubana y muy respetada por su talento y entrega al arte, Alonso estudió ballet en La Habana, pero luego fue a Estados Unidos, donde comenzó su carrera con el New York City Ballet...

Texto etiquetado

```
<phrase explicit="FALSE" coreferente="Alicia Alonso">Embajadora cultural de la revolución cubana </phrase>y muy respetada por <phrase explicit="FALSE" coreferente="Alicia Alonso">su </phrase>talento y entrega al arte, <phrase explicit="FALSE" coreferente="Alicia Alonso">Alonso </phrase><relation_expresion SOURCE="Alonso" TARGET="La Habana" TYPE="VERB" REAL_REL="estudiar_en" SEMANTIC_TYPE="PHIS_Loc">estudió </relation_expresion>ballet en<phrase explicit="TRUE">La Habana</phrase>, pero luego <relation_expresion SOURCE="Alonso" TARGET="Estados Unidos" TYPE="VERB" REAL_REL="ir_a" SEMANTIC_TYPE="PHIS_Loc">fue a </relation_expresion><phrase explicit="TRUE">Estados Unidos</phrase>, donde comenzó <phrase explicit="FALSE" coreferente="Alicia Alonso">su</phrase>carrera con el <relation_expresion SOURCE="Alonso" TARGET="New York City Ballet" TYPE="VERB" REAL_REL="comenzar_con" SEMANTIC_TYPE="ORG_AFF_Emp"/><phrase explicit="TRUE">New York City Ballet</phrase>.
```

Después de analizar el texto que conforma la oración, se determina la polaridad de las entidades dentro de la etiqueta `sentiment`. Cada entrada (entry) tiene los atributos: `source`, para indicar la persona que expresa un criterio en torno a la entidad marcada, en el cual se incluye al escritor del artículo; `entity` que señala la entidad analizada; `relativePolarity` que expresa la polaridad (POS, NEG) en caso de poseerla o NONE si hay falta de polaridad; y `degree` para mostrar la intensidad de polaridad expresada, mediante los valores: MIDDLE, STRONG y WEAK.

Ejemplo:

```
<entry source="WRITER" entity="Alicia Alonso" relativePolarity="POS" degree="STRONG"/>
```

```
<entry source="WRITER" entity="Estados Unidos" relativePolarity="NONE" degree="NONE"/>
```

Después de haber hecho el análisis del texto con la propuesta presentada, se obtuvieron los siguientes resultados, mostrados a continuación en las siguientes tablas que presentan la frecuencia con que aparecen los tipos de relaciones gramaticales y clasificación semántica.

RESULTADOS

Descripción de los resultados

En la tabla 1 y 2 se muestran las cantidades de relaciones teniendo en cuenta la clasificación gramatical y semántica.

Tabla 1. Cantidad de relaciones teniendo en cuenta la clasificación gramatical

Tipo de relación gramatical	Cantidad
Verbal	8
Sustantiva	2
Preposicional	3
Total de relaciones gramaticales	13

Tabla 2. Cantidad de relaciones teniendo en cuenta la clasificación semántica

Clasificación semántica	Cantidad
Parte_Todo_Subsidiario	1
Temporal	3

Otras_relaciones	2
Parte_Todo_Geográfico	1
Correferencia	25
Físico_Ubicación	3
Organización-Afiliación_Empleo	1
Total de relaciones semánticas	36

DISCUSIÓN

Con el trabajo se logró identificar las relaciones sintáctico-gramaticales y semánticas entre las entidades nombradas presentes en el texto analizado, utilizando una nueva propuesta que emplea etiquetas predefinidas. En total, se lograron etiquetar 20 entidades nombradas. Durante el análisis, se observó que hubo un predominio de las relaciones verbales, mientras que las relaciones sustantivas fueron las menos afectadas. En términos de relaciones semánticas, la correferencia fue la más predominante, mientras que las relaciones Parte-Todo-S subsidiario, Parte-Todo-Geográfico y Organización-Afiliación-Empleo tuvieron menor frecuencia.

La ventaja de esta propuesta es que una vez que los equipos de cómputo sean entrenados con estas etiquetas, serán capaces de detectar las relaciones de manera automática, lo que será muy útil para desarrollar y evaluar nuevos sistemas abiertos de extracción de información en español. Sin embargo, la desventaja es que se requiere que los textos sean etiquetados manualmente por especialistas para evitar la diversidad de significados que pueden tener las palabras o frases, así como la variedad de formas de expresar una idea similar.

CONCLUSIONES

El enfoque basado en corpus utilizado en este estudio ha demostrado ser una herramienta práctica para anotar y analizar relaciones sintácticas y semánticas entre entidades nombradas en un corpus lingüístico etiquetado. Los resultados sugieren que este enfoque puede ser beneficioso para la identificación de patrones lingüísticos y coocurrencias entre los términos utilizados en el corpus, lo que ha permitido generar una red de relaciones sintácticas y semánticas entre las entidades nombradas. Si bien es necesario tener en cuenta algunas limitaciones, esta metodología puede ser aplicada en diferentes áreas de investigación para mejorar nuestra comprensión de las relaciones sintácticas y semánticas entre entidades nombradas, lo que puede tener importantes aplicaciones prácticas en el desarrollo de tecnologías lingüísticas.

REFERENCIAS BIBLIOGRÁFICAS

1. Alonso, L. (1998). El análisis sociológico de los discursos: una aproximación desde los usos concretos. Ed. Fundamentos.
2. Análisis del Discurso. (2015). <https://metodosdeinvestigaciondcdgunefa.wordpress.com/2015/07/04/analisis-del-discurso/>
3. Arredondo Toledo, L. M. (2018). Extracción de relaciones entre las entidades nombradas en el idioma español [Tesis de Maestría].
4. Bernal Chávez, J. A. y Hincapié Moreno, D. A. (2018). Lingüística de corpus. <http://bibliotecadigital.caroycuervo.gov.co/1703/1/Linguistica-de-corpus-2018.pdf>
5. Boillos Pereira, M. M. (2018). La elaboración de un corpus del profesorado de español (copele): ¿utopía o realidad? Disponible en: https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-48832018000200153
6. Cruz Piñol, M. (2017). Lingüística de corpus y enseñanza del español como 2/L. Arco/Libros. https://www.arcomuralla.com/detalle_libro.php?id=872
7. Culotta, A., & Sorensen, J. (2004). Dependency tree kernels for relation extraction. In Proceedings of the 42nd annual meeting on association for computational linguistics (p. 423). Association for Computational Linguistics.
8. Culotta, A., McCallum, A. & Betz, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (pp. 296-303). Association for Computational Linguistics.
9. Filología e informática. (1999): nuevas tecnologías en los estudios filológicos (pp. 45-77). Milenio.
10. Jurafsky, D., & Martin, J. H. (2017). Vector Semantics. Speech and Language Processing: An Introduction to

11. Lyons, John. (1997). *Semántica lingüística*. Paidós.
12. Martín Peris, Ernesto. (coord.) (2008). *Diccionario de términos clave de ELE*. SGEL.
13. Mercado, H. (2008). *Fundamentos de la lingüística de corpus*. (s.e.).
14. Pardo Abril, N. G. (2002). El contexto y el discurso público. <https://revistas.udistrital.edu.co/index.php/enunc/article/view/2465/3432>.
15. Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
16. Torruela, J. & Llisterri, J. (1999). Diseño de corpus textuales y orales. En *Filología e informática: nuevas tecnologías en los estudios filológicos* (pp. 45-77). Milenio.
17. Wallis, S. and Nelson G. (s.f.). Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery*, 5: 307–340.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.