Anales de Lingüística. Segunda época
Nro. 11 | julio - diciembre 2023 | Mendoza, Argentina
ISSN 0325-3597 (impreso) - ISSN 2684-0669 (digital)
http://revistas.uncu.edu.ar/ojs/index.php/analeslinguistica
Recibido: 08.03.2023 | Aceptado: 06.11.2023 | pp. 95-134

Relaciones entre entidades nombradas en textos noticiosos

Relations between Entities Named in News Texts

Reynier Ávila Peña
Datys - Cuba
https://orcid.org/0000-0002-1523-8404
ravila81@nauta.com

Celia María Pérez Marqués
Universidad de Oriente - Cuba
https://orcid.org/0000-0002-0243-8159
celiapm@uo.edu.cu

Yaney Bourzac Álvarez
Datys - Cuba
https://orcid.org/0000-0001-6924-2815
yaney8305@gmail.com

Daymara López Cordero
Datys - Cuba
https://orcid.org/0000-0001-8051-9926
daylopezcordero@gmail.com

Resumen

El artículo se enfoca en el análisis de textos de noticias dentro de un corpus etiquetado, abordando temas diversos. Se presenta una clasificación de las relaciones semánticas entre las entidades nombradas en un texto etiquetado en formato XML, junto con una descripción de las etiquetas utilizadas. El objetivo principal es desarrollar una solución para extraer relaciones entre entidades nombradas en noticias en español, basándose en el etiquetado gramatical, la detección de entidades y la resolución de correferencias. El método empleado consiste en el etiquetado gramatical y el análisis sintáctico, y abarca tanto entidades nombradas como relaciones gramaticales y semánticas. Esta propuesta puede considerarse útil para el desarrollo y evaluación de nuevos sistemas de

extracción de información en español, concluyendo que los corpus, la lingüística de corpus y la lingüística computacional son herramientas valiosas en el proceso de enseñanza automática a las computadoras para comprender el lenguaje natural.

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

Abstract

The article focuses on the analysis of news texts within a tagged corpus, addressing various topics. It presents a classification of semantic relationships between named entities in the tagged text in XML format, along with a description of the used tags. The main objective is to develop a solution for extracting relationships between named entities in Spanish news, based on grammatical tagging, entity detection, and coreference resolution. The method employed consisted of grammatical tagging and syntactic analysis, covering both named entities and the grammatical and semantic relationships. This proposal is considered useful for the development and evaluation of new information extraction systems in Spanish, concluding that corpora, corpus linguistics, and computational linguistics are valuable tools in the process of teaching computers automatic understanding of natural language.

Keywords: corpus linguistics, named entities, language corpora, computational linguistics.

Introducción

La Lingüística de Corpus (LC) se define como “una rama de la lingüística que basa sus investigaciones en datos obtenidos a partir de corpus” (Martín Peris et al., 2008). Ciertamente, no se refiere a una disciplina lingüística en sentido estricto, sino de un enfoque metodológico que puede ser adoptado desde diversas disciplinas. No solo es útil para describir lenguas o variedades de lenguas que no conocemos, sino también para la descripción de aquellas que ya poseen una larga tradición descriptiva como el inglés o el español; por tanto, lo que la lingüística de corpus ha aportado es que, frente a los métodos tradicionales, se dispone de una importante cantidad de datos que resultaba impensable décadas atrás. Esta situación ha

modificado la percepción sobre dichas lenguas de las que aparentemente se tenía todo el conocimiento posible, lo que conlleva que pueda notarse una especie de revolución en la lingüística computacional. Esta última se ocupa de la construcción de modelos de lenguaje «entendibles» para las computadoras, es decir, la realización de formalismos del funcionamiento del lenguaje para aplicaciones informáticas que imiten la capacidad humana de hablar y entender, que transforman el conocimiento sobre los fenómenos que ya se conocían.

Problema

El objetivo de este trabajo es desarrollar una solución para extraer relaciones entre entidades nombradas en noticias en español. Para ello se recurrirá al etiquetado gramatical, la detección de estas entidades y la resolución de correferencias. El tema en cuestión resulta relevante puesto que estas anotaciones son útiles para evaluar y desarrollar nuevos sistemas abiertos de extracción de información en español.

Antecedentes y Marco teórico

La LC ha generado una serie de métodos de investigación, a fin de trazar un camino de datos a la teoría. Wallis y Nelson (2001) introdujeron por primera vez lo que ellos llamaron la perspectiva de las tres A (3A perspective): anotación, abstracción y análisis. La anotación consiste en la aplicación de un esquema para textos. Las anotaciones pueden incluir etiquetas estructurales, etiquetado gramatical, análisis sintáctico. La abstracción consiste en la traducción (mapeo) de términos en el esquema de un conjunto de datos; incluye la búsqueda lingüística dirigida y puede admitir, por ejemplo, la regla de aprendizaje para los analizadores. El análisis consiste en sondear, manipular y generar el conjunto de datos, todo ello de manera estadística. Este puede incluir evaluaciones estadísticas, optimización de bases de reglas o métodos de descubrimiento de conocimiento.

En la LC se llama 'etiquetado de palabras' a la asignación de categorías sintácticas a cada palabra de un texto o corpus, denominado en inglés *Part Of Speech Tagging (POS Tagging)*. Requiere de un conjunto de etiquetas pre-definidas (*tagset*) y un algoritmo de asignación de etiquetas.

Un corpus consiste en una recopilación de muestras reales de una lengua, novelas, obras de teatro, guiones de cine, noticias de prensa, ensayos, transcripciones de programas de radio o televisión, conversaciones o incluso discursos. Otras definiciones son: "(...) corpus es una colección de texto lingüístico de ocurrencia natural seleccionada para caracterizar un estado o variedad de una lengua" (Sinclair, 1991). También es definido como: "Colección de textos, reunidos según unos criterios precisos, eventualmente estructurados y enriquecidos con información adicional, en vista de una explotación teórica o práctica" (Mercado, 2008, p. 7). Otra definición es: "(...) 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 & Llisterri, 1999a, p. 7). 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. De esta manera, la representatividad se convierte en "la piedra angular de la Lingüística de Corpus, pues de ello depende que se puedan extraer conclusiones fiables a partir de los datos estadísticos" (Cruz Piñol, 2012: 36).

La Entidad Nombrada (NE, por las siglas en inglés de *Named Entity*) es la frase que identifica un elemento de un conjunto de otros elementos que tienen atributos similares. Es, en términos generales, cualquier cosa a la que se pueda referir con un nombre propio, como una persona, un lugar, una organización. El término se extiende comúnmente para incluir cosas que no son entidades *per se*, así tal que se pueden incluir fechas, tiempos y otros tipos de expresiones temporales, e incluso expresiones numéricas como precios (Jurafsky y Martin, 2017).

El procedimiento empleado para realizar este trabajo fue el siguiente: primero se realizó una revisión ortográfica; luego, a partir de la segmentación del texto, el etiquetado gramatical, la detección y la resolución de correferencias se identificaron las entidades nombradas y se le asignó una etiqueta a cada una de ellas. Luego, se detectaron los patrones de relaciones entre las entidades nombradas. Este trabajo busca aplicar estas técnicas en el contexto de textos noticiosos en español. Si bien existen enfoques similares en otros idiomas y dominios, la adaptación y el ajuste de estos métodos al español y a las características propias de los textos noticiosos representan una contribución de un valor considerable. De esta manera se representarían las diferentes asociaciones entre las entidades de un conjunto, con entidades de otro conjunto.

Etiquetado de textos noticiosos

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 identifica. Existen tipos de entidades definidos por las competencias de evaluación[1], por ejemplo, persona, lugar, organización, evento, miscelánea, fecha, cantidad monetaria.

Entre los nombres de entidades detectados pueden existir relaciones o conexiones que aparezcan predefinidas o no (Culotta et al., 2006). La tarea de Extracción de Relaciones (ER, *Relation Extraction*, por las siglas en inglés) ha sido reconocida como un problema importante y difícil para los investigadores de las ramas de la lingüística, filosofía y ciencia de la computación; esta encuentra y clasifica relaciones semánticas entre las entidades nombradas del texto. A menudo son relaciones binarias como cónyuge-de, hijo-de, empleo, parte-de, membresía y relaciones geoespaciales (Jurafsky y Martin, 2017). Se puede aplicar en tareas como Búsquedas de respuestas, Bioinformática, Construcción de ontologías y otras.

La investigación actual se origina de la tesis de maestría de Liané Mercedes Arredondo Toledo en el 2018. Esta tesis, titulada: *Extracción de relaciones entre las entidades nombradas en el idioma español*, presentó una solución para identificar relaciones entre entidades nombradas a nivel de oración

[1] Esta definición de tipos de entidades se realiza con la guía de anotaciones de las entidades en el idioma inglés creado en la competencia ACE del año 2008. Disponible en: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

utilizando el enfoque de Extracción Abierta de Información[2]. La propuesta se basó en el análisis de etiquetado gramatical y patrones para identificar distintos tipos de relaciones gramaticales en español, así como en la resolución de correferencias[3] para establecer vínculos entre entidades. Sin embargo, se encontró una limitación en el corpus utilizado, lo cual dificultó el entrenamiento de los equipos computacionales.

Teniendo en cuenta esta base, el presente estudio se apoyó en el enfoque anteriormente mencionado. No obstante, se construyó un nuevo corpus que incorpora métodos adicionales de resolución de correferencias para detectar relaciones entre entidades nombradas, incluyendo pronombres y, además, se llevó a cabo la tarea de etiquetar manualmente un mayor número de documentos con el fin de mejorar la clasificación de las relaciones.

De esta manera, se ampliaron los resultados con respecto a la investigación anterior. Las muestras que se analizarán pertenecen a dicho corpus compuesto por 90 noticias en español, extraídas de los sitios digitales cubanos Granma, Cubadebate, Juventud Rebelde, Trabajadores y el mexicano elEconomista.es, en el periodo comprendido entre 2007 y 2018. Estas noticias, abarcan una amplia variedad de temas, tales como política, cultura, deporte, economía, medio ambiente, salud, sociedad y religión, entre otros. Aunque este artículo forma parte de una investigación mayor, se ha seleccionado una muestra más acotada como ejemplo, debido a limitaciones de espacio.

[2] La extracción abierta de información (Open Information Extraction en inglés) es la tarea de extraer afirmaciones del texto, sin especificación previa de la relación o dominio y sin vocabulario pre-especificado o un corpus de entrenamiento etiquetado manualmente. Zhila, Alisa y. Gelbukh, Alexander (19 de agosto de 2013): Conferencia La extracción abierta de información para el español. VI Coloquio de Lingüística Computacional. Facultad de Filosofía y Letras. Universidad Nacional Autónoma de México.

[3] Relaciones donde las entidades significan o representan el mismo concepto.

Ejemplo 1:

> 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[4].

El texto trata de la celebración del 80 aniversario del debut artístico de la Prima Ballerina, quien con su estilo marcó al ballet cubano a tal punto que ha influido sobremanera en los cinco continentes, marcando una originalidad en cada una de sus presentaciones. La distinguen la disciplina y el rigor que ha llevado a la escuela cubana de Ballet, y aunque ya no pueda bailar físicamente, *Giselle* seguirá danzando eternamente.

[4] Ecodiario.eleconomista.es, 2011. Disponible en: https://www.elcomercio.com/tendencias/cultura/cuba-celebra-80-anos-del.html

La noticia es un tipo de texto, tanto escrito como auditivo o audiovisual, que consiste en una narración precisa de algún hecho novedoso, con interés público, en este caso el debut de Alicia Alonso. En cuanto a los géneros discursivos están el científico, publicitario, epistolar, judiciales y el periodístico del que se ocupa este trabajo.

En el análisis se puede constatar que se remite solo a los hechos, no se opina en demasía, ni se toma ninguna posición. La noticia está escrita en un lenguaje divulgativo, no especializado, de tal forma que permite a cualquier persona acceder a la información. Es coherente, directa, no abusa de recursos lingüísticos tales como la metáfora o la jerga popular, ni adjetivos relacionados a juicios de valoración o morales, pues puede ocurrir que el lector interprete la información explicada de una manera diferente a cómo ha querido el emisor. Tampoco hay uso de exclamaciones.

Las frases en su mayoría son concisas y breves, al igual que las oraciones. Así, de esta manera, la construcción sintáctica será la más simple: sujeto acompañado del verbo y de complementos. Se usa la voz activa en lugar de la pasiva, las frases afirmativas en sustitución de las negativas, las que no se acompañan de subordinaciones ni de incisos. De acuerdo con su temática, 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

Las palabras en un texto no se presentan de forma aislada, sino que existe una estrecha relación entre sus significados, para así transmitir un mensaje de manera explícita o implícita. Lo mismo ocurre con las entidades nombradas presentes en la oración o texto, pues estas son frases que incluyen cualquier tipo de palabra, aunque principalmente están representadas por sustantivos, y para poder conocer toda la información que se brinda de ellas es necesario hacer una correcta interpretación semántica, pero detectar estas relaciones semánticas es complejo debido a la diversidad de significados que tienen las palabras o las frases y a la variedad de formas de expresar una misma idea, sobre todo por la ambigüedad semántica del lenguaje. Por tanto, se utiliza una clasificación de relaciones semánticas[5] a partir del 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).

[5]Esta definición de tipos de relaciones entre nombres de entidades se realiza con la guía de anotaciones de las relaciones en el idioma inglés creado en la competencia ACE del año 2008. Disponible en https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-relations-guidelines-v6.2.pdf

Ejemplo: <relation_expresion SOURCE="**American Ballet Theatre**" TARGET="**ABT**" TYPE="SUST"REAL_REL="("**SEMANTIC_TYPE**="**COREF**">(</relation_expresion><phrase explicit="FALSE" coreferente="**American 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 Balle**t</phrase>

Otra-Relación: Describe las relaciones entre entidades que no se encuentran en las clasificaciones anteriores. El texto está etiquetado con el formato xml[6].

> **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 *tittle*, *topic* y *date*. En el caso de la primera, se introduce en ella el título de la noticia*;* la segunda, se utilizaría,

[6]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.

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, porcentaje, 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 (*sentences*) 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>

Se decidió 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, resumidos a continuación en las siguientes tablas que presentan la frecuencia con que aparecen los tipos de relaciones gramaticales y clasificación semántica.

Datos Estadísticos

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

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

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

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

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

Ejemplo 2:

Alejo Carpentier, eterno genio literario

Amigos cercanos de Alejo Carpentier lo recuerdan en cansado pero vehemente paso por las calles de París aquel abril de 1980, como despidiéndose de toda la luz del mundo en la tierra de ancestros suyos. Por aquellos viejos muros de la capital francesa debió de estar la morada de tantos y de tantos fantasmas que cobraron vida en sus novelas, y que se hicieron definitivamente de vida en el oficio de narrar de aquel genio. Es posible que, en aquella postrera vuelta por plazas parisienses, Alejo sellara un pacto de inmortalidad con esa energía de Revolución, que llegaría con sueños y excesos a lo real maravilloso del Caribe. Carpentier se disponía a remontar la muerte sin otras armas que no fueran la inspiración y los recuerdos. El presentimiento suyo era demasiado cierto como para no tenerlo en cuenta: así, como realización puntual de una corazonada, falleció el gran escritor, periodista y musicólogo cubano el 24

> de abril de 1980, hace hoy exactamente 33 años. Alejo Carpentier devino profeta en su tierra y fuera de ella. Se verifica una buena forma de inmortalidad. Vive en esa dimensión una circunstancia que él mismo deparó a sus personajes: el fin es una fuga y no la muerte. En Viaje a la semilla, Capellanías pasa en contra de las leyes del tiempo físico para escapar de las tinieblas. En El reino de este mundo, Mackandal queda en la conciencia colectiva de sus hermanos como el héroe prófugo que nunca, jamás podrá morir. En El siglo de las luces, desolada queda la casa de Sofía y de Esteban tras la carga de Madrid contra los mamelucos de Napoleón el 2 de mayo de 1808.En la empresa musicológica de Carpentier, también encontraríamos sorpresas, sobre todo para salvar en juventud plena a sus amigos Amadeo Roldán y Alejandro García Caturla. Toda la obra de Alejo Carpentier, incluido su paso enigmático por las calles de la capital francesa, fue una pretensión de remontar la muerte[7].

El texto es un tributo a Alejo Carpentier, reconocido escritor, periodista y musicólogo cubano, quien falleció en 1980. Describe cómo amigos cercanos de Carpentier lo recuerdan durante su última visita a las calles de París, un lugar cargado de fantasmas que cobran vida en sus novelas. Se sugiere que Carpentier selló un pacto de inmortalidad con la energía revolucionaria, que llevó a cabo sueños y excesos en lo real maravilloso del Caribe. Destaca la capacidad de fuga y escape de los personajes de Carpentier, quienes logran liberarse de las sombras. También menciona su trabajo como musicólogo y su esfuerzo por rescatar la obra de sus amigos Amadeo Roldán y Alejandro García Caturla. El texto es emotivo y poético, utilizando lenguaje evocador para describir la vida y obra de Carpentier, resaltando su importancia y legado en la literatura latinoamericana. La estructura del texto es sólida y utiliza un lenguaje claro y conciso para transmitir la información. De acuerdo con su temática es cultural. Consta de trece oraciones gramaticales y 20 entidades nombradas que son*:*

Alejo Carpentier
París
abril de 1980
Revolución

[7] Disponible en: https://www.radiocamoa.icrt.cu/alejo-carpentier-eterno-genio-literario/

Caribe
24 de abril de 1980
Hoy
33 años
Viaje a la semilla
Capellanías
El reino de este mundo
Mackandal
El siglo de las luces
Sofía
Esteban
carga de Madrid
mamelucos de Napoleón
2 de mayo de 1808
Amadeo Roldán
Alejandro García Caturla

El procedimiento que se sigue en el análisis es el mismo utilizado con el primer ejemplo, de tal forma que solo se presentarán los fragmentos concretos sin la explicación anterior.

Relaciones semánticas en el texto:

> Temporal: <relation_expresion SOURCE="**gran escritor, periodista y musicólogo cubano**" TARGET="**24 de abril de 1980**" TYPE="VERB" REAL_REL="fallecer" **SEMANTIC_TYPE="TEMP**"/><phrase explicit="TRUE">24 de abril de 1980</phrase>
>
> Otra-Relación: <relation_expresion SOURCE="**Mackandal**" TARGET="**El reino de este mundo**" TYPE="VERB" REAL_REL="quedar" SEMANTIC_TYPE="**OTHER_RELATION**"/><phrase explicit="TRUE">Mackandal </phrase>queda en la conciencia colectiva de <phrase explicit="FALSE" coreferente="Mackandal">sus</phrase> hermanos como el héroe prófugo que nunca, jamás podrá morir. </text>

En relación con el etiquetado del título, tópico y fecha queda de esta manera.

Ejemplo: <**title**>Alejo Carpentier, eterno genio literario</**title**>

<topic>Cultura**</topic>**

<date>24/04/2013**</date>**

En el caso de la Polaridad global y la Entidad formal quedan etiquetadas así.

Ejemplo: <entity type="PER" **globalPolarity**="POS">Alejo Carpentier</entity>

<entity type="PER" globalPolarity="NEG" **formalEntity**="Don Marcial, Marqués de Capellanías">Capellanías</entity>

Las entidades anidadas y el tipo de entidad se etiquetaron de la siguiente forma

Ejemplo: <text>En <phrase explicit="**TRUE**">El siglo de las luces</phrase>, desolada queda la casa de <phrase explicit="**TRUE**">Sofía </phrase>y de<phrase explicit="TRUE">Esteban </phrase>tras la <phrase explicit="TRUE">carga de <phrase explicit="TRUE" **TYPE**="LOC" **NESTED**="TRUE">Madrid</phrase> </phrase>

Las anotaciones de tipo sintáctico quedan de la siguiente forma.

Ejemplo: El presentimiento suyo era demasiado cierto como para no tenerlo en cuenta: así, como realización puntual de una corazonada, falleció el gran escritor, periodista y musicólogo cubano el 24 de abril de 1980, hace hoy exactamente 33 años.

Texto etiquetado:

<text>El presentimiento <phrase explicit="FALSE" coreferente="Alejo Carpentier">suyo </phrase>era demasiado cierto como para no tenerlo en cuenta: así, como realización puntual de una corazonada, falleció el <phrase explicit="FALSE" coreferente="Alejo Carpentier">gran escritor, periodista y musicólogo cubano </phrase>el <relation_expresion **SOURCE**="gran escritor, periodista y musicólogo cubano " **TARGET**="24 de abril de 1980" TYPE="**VERB**" REAL_REL="fallecer"

SEMANTIC_TYPE="TEMP"/><phrase explicit="TRUE">24 de abril de 1980</phrase>, hace <phrase explicit="TRUE">hoy </phrase>exactamente <phrase explicit="TRUE">33 años</phrase>. </text>

La polaridad de las entidades queda de este modo.

Ejemplo: <entry source="WRITER" entity="El reino de este mundo" relativePolarity="NONE" degree="NONE"/>

<entry source="WRITER" entity="Mackandal" relativePolarity="POS" degree="MIDDLE"/>

Datos Estadísticos

En las tablas 3 y 4 se muestran las cantidades de relaciones teniendo en cuenta la clasificación gramatical y semántica.

Tipo de relación gramatical	Cantidad
Verbal	3
Sustantiva	0
Preposicional	1
Total de relaciones gramaticales	4

Tabla 3. Cantidad de relaciones teniendo en cuanta la clasificación gramática

Clasificación semántica	Cantidad
Otras_relaciones	3
Temporal	1
Total de relaciones semánticas	4

Tabla 4. Cantidad de relaciones teniendo en cuenta la clasificación semántica

Ejemplo 3:

> El Caribe en García Márquez
>
> El intenso erotismo del Caribe, su olor a salitre y a monte, su gente marcada por el sol y las desinhibiciones, han sido parte de la materia prima de uno de los autores latinoamericanos más leído: Gabriel García Márquez. Revelar algunas de las trazas de esa presencia caribeña en la obra del Gabo es la intención de Casa de la Américas, que entre el 14 y el 15 de noviembre dedicará al premio Nobel de Literatura una jornada de homenajes, donde no faltarán las exposiciones, talleres y proyecciones de audiovisuales. Según revelaciones de Gustavo Adolfo Bell, embajador de Colombia en La Habana, el programa incluirá una exposición con ilustraciones basadas en obras del novelista y otra con fragmentos de sus novelas en las que se verifica la presencia del Caribe en el "universo" literario del autor de El amor en los tiempos del cólera. Para la fecha, la Casa se llenará de las mariposas amarillas que revelaban la presencia de Mauricio Babilonia, personaje de Cien años de soledad, texto que ha devenido ícono del boom de la literatura latinoamericana del siglo XX y que fuera escrita por el Gabo hace casi medio siglo. También se espera la presencia de Conrado Zuluaga, profesor y ensayista colombiano de reconocido prestigio por sus estudios acerca de la vida y la obra de García Márquez. Los que prefieren los audiovisuales podrán disfrutar en esta ocasión de "todos" los audiovisuales que se le han dedicado al autor, tal como aseguró el diplomático colombiano a la prensa. Gabriel García Márquez es un amigo de Cuba y de la Casa de la Américas, institución que cada dos años organiza uno de los premios literarios más prestigiosos de la región, en el que el prestigioso escritor y novelista ha participado como jurado. Famosa es también su relación con Fidel Castro, de quien se considera "amigo personal[8]".

El texto describe una jornada de homenaje a Gabriel García Márquez en la Casa de las Américas en La Habana, Cuba. La intención es revelar algunas de las trazas de la presencia caribeña en la obra del autor, que es conocido por su intensa descripción del erotismo y la cultura del Caribe. La jornada incluirá exposiciones, talleres y proyecciones de audiovisuales, así como la

[8] Disponible en: http://www.trabajadores.cu/temas/gabriel-garcia-marquez/

presencia de Conrado Zuluaga, un ensayista colombiano reconocido por sus estudios sobre la vida y obra de García Márquez. El texto destaca la relación de amistad entre García Márquez y Cuba, así como su participación como jurado en uno de los premios literarios más prestigiosos de la región, organizado por la Casa de las Américas. El texto es informativo y destaca la importancia de la obra de García Márquez en la literatura latinoamericana, está bien estructurado y utiliza un lenguaje claro y conciso para transmitir la información, su temática es cultural. Consta de ocho oraciones psicológicas y 11 entidades nombradas que son*:*

Caribe
Gabriel García Márquez
Casa de las Américas
entre el 14 y el 15 de noviembre
Gustavo Adolfo Bell
Mauricio Babilonia
siglo XX
Conrado Zuluaga
Cuba
dos años
Fidel Castro

Se expondrán a continuación las relaciones que aparecen en el texto etiquetado:

Correferencia

Ejemplo: <relation_expresion SOURCE="**uno de los autores latinoamericanos más leído**" TARGET="**Gabriel García Márquez**" TYPE="SUST" REAL_REL=":" **SEMANTIC_TYPE="COREF**">:</relation_expresion> <phrase explicit="TRUE">Gabriel García Márquez</phrase>.

Otra Relación

Ejemplo: <relation_expresion SOURCE="**Casa de las Américas**" TARGET="**premio Nobel de Literatura**" TYPE="VERB"

REAL_REL="dedicar_a" **SEMANTIC_TYPE="OTHER_RELATION**">dedicará al </relation_expresion><phrase explicit="FALSE" coreferente="Gabriel García Márquez">premio Nobel de Literatura </phrase>una jornada de homenajes, donde no faltarán las exposiciones, talleres y proyecciones de audiovisuales.

Organización-Afiliación_Empleo

Ejemplo**:* <relation_expresion SOURCE="Gustavo Adolfo Bell***" TARGET="***La Habana***" TYPE="PREP-pertenencia" REAL_REL="en" ***SEMANTIC_TYPE="ORG_AFF_Emp***">en </relation_expresion><phrase explicit="TRUE" TYPE="LOC" NESTED="TRUE">La Habana</phrase>

El etiquetado del título, tópico y fecha quedan de este modo:

Ejemplo: **<title>**El Caribe en García Márquez**</title>**

<topic>Cultura**</topic>**

<date>30/10/2013**</date>**

En el caso de la Polaridad global y la Entidad formal quedan etiquetadas así:

Ejemplo: <entity type="PER" **globalPolarity**="POS">Gabriel García Márquez</entity>

<entity type="LOC" globalPolarity="NONE" **formalEntity**="República de Cuba">Cuba</entity>

Se etiquetaron de la siguiente manera las entidades anidadas y el tipo de entidad.

Ejemplo: <text>Para la fecha, la <phrase explicit="FALSE" coreferente="Casa de las Américas">Casa </phrase>se llenará de las mariposas amarillas que revelaban la presencia de <phrase explicit="TRUE">Mauricio Babilonia</phrase><relation_expresion SOURCE="Mauricio Babilonia" TARGET="personaje de Cien años de soledad" TYPE="SUST" REAL_REL="," SEMANTIC_TYPE="COREF">,</relation_expresion><phrase explicit="FALSE" coreferente="Mauricio Babilonia">personaje de

<phrase explicit="TRUE" **TYPE**="DOC" **NESTED**="TRUE">Cien años de soledad</phrase></phrase>, texto que ha devenido ícono del boom de la literatura latinoamericana del <phrase explicit="TRUE">siglo XX </phrase>y que fuera escrita por el <phrase explicit="FALSE" coreferente="Gabriel García Márquez">Gabo </phrase>hace casi medio siglo.</text>

De esta manera, se presentan las anotaciones de tipo sintáctico.

Ejemplo:

También se espera la presencia de Conrado Zuluaga, profesor y ensayista colombiano de reconocido prestigio por sus estudios acerca de la vida y la obra de García Márquez.

Texto etiquetado:

<text>También se espera la presencia de <phrase explicit="TRUE">Conrado Zuluaga</phrase> <relation_expresion **SOURCE**="Conrado Zuluaga" **TARGET=**"profesor y ensayista colombiano" TYPE="SUST" REAL_REL="","**SEMANTIC_TYPE="**COREF">,</relation_expresion><phrase explicit="FALSE" coreferente="Conrado Zuluaga">profesor y ensayista colombiano </phrase>de reconocido prestigio por <phrase explicit="FALSE" coreferente="Conrado Zuluaga">sus </phrase>estudios acerca de la vida y la obra de <phrase explicit="FALSE" coreferente="Gabriel García Márquez">García Márquez</phrase>. </text>

La polaridad de las entidades se determina de esta manera:

Ejemplo: <entry source="WRITER" entity="Gabriel García Márquez" relativePolarity="POS" degree="WEAK"/>

<entry source="WRITER" entity="Cuba" relativePolarity="NONE" degree="NONE"/>

Datos Estadísticos

En la tabla 5 y 6 se muestran las cantidades de relaciones teniendo en cuenta la clasificación gramatical y semántica.

Tipo de relación gramatical	Cantidad
Verbal	3
Sustantiva	5
Preposicional	1
Total de relaciones gramaticales	9

Tabla 4. Cantidad de relaciones teniendo en cuenta la clasificación gramatical

Clasificación semántica	Cantidad
Correferencia	5
Otras relaciones	3
Organización-Afiliación_Empleo	1
Total de relaciones semánticas	9

Tabla 5. Cantidad de relaciones teniendo en cuenta la clasificación semántica

Ejemplo 4:

Francia aprueba proyecto de ley contra el odio en Internet

La Asamblea Nacional de Francia aprobó hoy en primera lectura un proyecto de ley destinado a luchar contra el discurso de odio en Internet, iniciativa que genera cuestionamientos en algunos sectores de la sociedad. En la cámara baja, el proyecto presentado por la diputada Laetitia Avia, del partido gobernante La República en Marcha (LREM), contó con 434 votos a favor, 33 en contra y 69 abstenciones. El texto analizado en el Palacio de Borbón irá al Senado, donde el oficialismo y sus aliados no tienen el control como en la Asamblea, sin embargo, debe pasar sin muchas dificultades, sobre todo por el apoyo al mismo de la fuerza dominante en la cámara alta, Los Republicanos (LR), aunque estos son opositores al presidente Emmanuel Macron. De aprobarse el proyecto, las redes sociales, como Facebook, Twitter y YouTube, y los motores de búsqueda en Internet, Google, por ejemplo, estarían obligados a eliminar dentro de las 24 horas mensajes reportados por su contenido nocivo. Según el proyecto de ley adoptado en primera lectura, si las plataformas no eliminan estos contenidos, se expondrían a fuertes multas, que pudieran superar el millón de euros. Por su actualidad e intención, el tema no parecería generador de mucha polémica, sin

> embargo, la plataforma política de izquierda Francia Insumisa advierte desde la Asamblea sobre su potencial amenaza para la libertad de expresión y el riesgo de confiar en operadores privados la responsabilidad de actuar. Otros sectores de la sociedad llaman la atención acerca del peligro de censura de una iniciativa que en la práctica funcionaría como una especie de 'botón' visible en Internet, que los usuarios accionarían en caso de considerar un mensaje portador de odio, aunque correspondería en última instancia a la justicia definir si lo es o no[9].

El texto presenta información precisa sobre la aprobación en primera lectura de un proyecto de ley en la Asamblea Nacional de Francia para combatir el discurso de odio en Internet. Proporciona detalles sobre los resultados de la votación en la cámara baja y menciona la posible facilidad con la que el proyecto puede pasar en el Senado debido al apoyo de la fuerza política dominante en esa cámara. Se plantea que, de aprobarse el proyecto, las redes sociales y los motores de búsqueda estarían obligados a eliminar mensajes reportados por contenido nocivo en un plazo de 24 horas. También se menciona la posibilidad de fuertes multas en caso de no cumplir con esta obligación. Se muestra la preocupación expresada por la plataforma política de izquierda Francia Insumisa sobre los posibles riesgos para la libertad de expresión al confiar en operadores privados la responsabilidad de actuar. Además, se resalta el temor de otros sectores de la sociedad sobre el peligro de censura y la posibilidad de que la iniciativa se convierta en una forma de "botón" en Internet para denunciar mensajes de odio, dejando en última instancia la definición de tal contenido a la justicia. En general, el análisis del texto es objetivo y proporciona una visión equilibrada de los diferentes puntos de vista y preocupaciones relacionadas con el proyecto de ley. De acuerdo con su temática es social. Consta de siete oraciones gramaticales y 17 entidades nombradas que son*:*

Asamblea Nacional de Francia
hoy
Internet

[9] Disponible en: http://www.cubadebate.cu/noticias/2019/07/10/francia-aprueba-proyecto-de-ley-contra-el-odio-en-internet/#.XSi4cXbBDcc

Laetitia Avia
partido gobernante La República en Marcha
434 votos
Palacio de Borbón
Senado
Los Republicanos
presidente Emmanuel Macron
Facebook
Twitter
YouTube
Google
24 horas
millón de euros
plataforma política de izquierda Francia Insumisa

Relaciones semánticas existentes en el texto:

Parte_todo_subsidiario

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

Temporal

Ejemplo: <relation_expresion SOURCE="**Asamblea Nacional de Francia**" TARGET="**hoy**" TYPE="VERB" REAL_REL="aprobar"**SEMANTIC_TYPE="TEMP**">aprobó</relation_expresion><phrase explicit="TRUE">hoy</phrase>

Se explicará este tipo de relación al no estar incluida en los primeros ejemplos

Organización-Afiliación_Política

Representa la relación entre una persona y la organización política a la cual pertenece. Esta es una relación permanente por lo general.

> **Ejemplo**: <relation_expresion SOURCE="**Laetitia Avia**" TARGET="**partido gobernante La República en Marcha**" TYPE="PREP-pertenencia" REAL_REL="de" **SEMANTIC_TYPE="ORG_AFF_Pol**">del</relation_expresion><phrase explicit="TRUE">partido gobernante La República en Marcha</phrase>.

Correferente

> **Ejemplo**: <relation_expresion SOURCE="**partido gobernante La República en Marcha**" TARGET="**LREM**" TYPE="SUST" REAL_REL="(" **SEMANTIC_TYPE="COREF**">(</relation_expresion><phrase explicit="FALSE" coreferente="partido gobernante La República en Marcha">LREM</phrase>, contó con <phrase explicit="TRUE">434 votos </phrase>a favor, 33 en contra y<phrase explicit="TRUE">69 abstenciones</phrase>.

Esta relación también será explicada a continuación:

Personal-Social_Oposición

Representa la conexión entre dos entidades que están en posiciones contrarias.

> **Ejemplo**: <relation_expresion SOURCE="**Los Republicanos**" TARGET="**presidente Emmanuel Macron**" TYPE="VERB" REAL_REL="ser_opositor" **SEMANTIC_TYPE="PER-SOC_Opp">**son opositores</relation_expresion>al<phrase explicit="TRUE">presidente Emmanuel Macron</phrase>

De esta manera queda el etiquetado del título, tópico y fecha.

> **Ejemplo**: **<title>**Francia aprueba proyecto de ley contra el odio en Internet**</title>**
>
> <**topic**>Ciencia y Tecnología**</topic**>
>
> **<date>**10/07/2019</**date>**

La etiqueta de la Polaridad global y la Entidad formal queda de esta manera. (En esta noticia no se etiquetó ninguna Entidad formal)

> **Ejemplo**: <entity type="LOC" **globalPolarity**="NONE">Palacio de Borbón</entity>

Las anotaciones de carácter sintáctico tienen el siguiente formato

Ejemplo: La Asamblea Nacional de Francia aprobó hoy en primera lectura un proyecto de ley destinado a luchar contra el discurso de odio en Internet, iniciativa que genera cuestionamientos en algunos sectores de la sociedad.

Texto etiquetado:

> <text>La <phrase explicit="TRUE"><phrase explicit="TRUE" TYPE="ORG" NESTED="TRUE">Asamblea Nacional </phrase><relation_expresion **SOURCE=**"Asamblea Nacional " **TARGET="**Francia" TYPE="PREP-pertenencia" REAL_REL="de" **SEMANTIC_TYPE="**PART_WHL_Sub">de</relation_expresion><phrase explicit="TRUE" TYPE="LOC" NESTED="TRUE">Francia</phrase></phrase><relation_expresion **SOURCE="**Asamblea Nacional de Francia" **TARGET=**"hoy" TYPE="VERB" REAL_REL="aprobar" **SEMANTIC_TYPE**="TEMP">aprobó </relation_expresion><phrase explicit="TRUE">hoy</phrase>en primera lectura un proyecto de ley destinado a luchar contra el discurso de odio en<phrase explicit="TRUE">Internet</phrase>, iniciativa que genera cuestionamientos en algunos sectores de la sociedad.</text>

Así queda establecida la polaridad de las entidades.

> **Ejemplo**: <entry source="WRITER" entity="Asamblea Nacional de Francia" relativePolarity="NONE" degree="NONE"/>
>
> <entry source="WRITER" entity="hoy" relativePolarity="NONE" degree="NONE"/>

Datos Estadísticos

En la tabla 7 y 8 se muestran las cantidades de relaciones teniendo en cuenta la clasificación gramatical y semántica.

Tipo de relación gramatical	Cantidad
Verbal	3
Sustantiva	2
Preposicional	2
Total de relaciones gramaticales	7

Tabla 6. Cantidad de relaciones teniendo en cuenta la clasificación gramatical

Clasificación semántica	Cantidad
Parte_Todo_Subsidiario	1
Temporal	1
Organización-Afiliación_Política	2
Correferencia	2
Personal-Social_ Oposición	1
Total de relaciones semánticas	7

Tabla 7. Cantidad de relaciones teniendo en cuenta la clasificación semántica

Ejemplo 5:

> El único hábitat de la vaquita marina fue declarado por la Unesco Patrimonio Mundial en Peligro
>
> El único hábitat de la vaquita marina, en México, fue declarado este miércoles como Patrimonio en Peligro por el Comité del Patrimonio Mundial de la Organización de las Naciones Unidas para la Educación, Ciencia y Cultura (Unesco, por sus siglas en inglés). Dentro del marco de la celebración de la reunión del Comité del Patrimonio Mundial, las Islas y Áreas protegidas del Golfo de California, en México, fueron ingresadas en la Lista del Patrimonio Mundial en Peligro. El portal oficial de la Unesco señala que esta decisión se debe al peligro de extinción de la vaquita marina, de la cual, escasamente quedan 10 especímenes de casi 300 que

> habían sido censados en 2005. La disminución de la vaquita marina se debe a la pesca ilegal con redes de enmalle, lo que ha arrasado con la especie en peligro de extinción. Por su parte, el Comité instó y alentó a México al fortalecimiento del monitoreo de la zona para evitar que se sigan llevando a cabo este tipo de actividades ilegales que están acabando con parte del valor universal excepcional del sitio, de acuerdo con lo publicado por el portal web. Adicionalmente, el Comité retiró de la Lista del Patrimonio Mundial en Peligro al Lugar de Nacimiento de Jesús en Belén, en Palestina, y a las Oficinas salitreras de Humberstone y Santa Laura en Chile, esto se debe a las restauraciones realizadas en el lugar y el mantenimiento proporcionado, en el caso de Palestina. En el caso de Chile, se debe al esfuerzo de conservación realizado por las autoridades correspondientes para garantizar la conservación del lugar. La celebración número 43 de la reunión del Comité del Patrimonio Mundial se está llevando a cabo en la ciudad de Baku, en Azerbaiyán, inició el pasado 30 de junio y tiene pautado culminar el 10 de julio[10].

El texto informa sobre la declaración del hábitat de la vaquita marina en México como Patrimonio en Peligro por parte del Comité del Patrimonio Mundial de la Unesco. Se destaca la importancia de la vaquita marina, una especie en peligro de extinción, y cómo la pesca ilegal con redes de enmalle ha disminuido su población. El Comité instó a México a fortalecer el monitoreo de la zona para evitar actividades ilegales que están acabando con parte del valor universal excepcional del sitio. Además, se menciona la retirada de la Lista del Patrimonio Mundial en Peligro de otros lugares debido a las restauraciones y conservación realizada. El texto resalta la importancia de la conservación de la biodiversidad y el patrimonio cultural, y destaca el papel de la Unesco en la protección del patrimonio mundial. El lenguaje utilizado es claro y conciso, transmitiendo la información de manera efectiva y destacando la importancia de la preservación de la vaquita marina y su hábitat. También se enfatiza la importancia de la cooperación internacional en la protección del patrimonio mundial. De

[10] Disponible en: https://www.cubadebate.cu/noticias/2019/07/03/el-unico-habitat-de-la-vaquita-marina-fue-decalarado-patrimonio-en-peligro-unesco/

acuerdo con su temática es medioambiental. Consta de ocho oraciones gramaticales y 20 entidades nombradas que son*:*

México
miércoles
Patrimonio en Peligro
Comité del Patrimonio Mundial
Organización de las Naciones Unidas para la Educación, Ciencia y Cultura
reunión del Comité del Patrimonio Mundial
Islas y Áreas protegidas del Golfo de California
Lista del Patrimonio Mundial en Peligro
10 especímenes
en 2005
Lugar de Nacimiento de Jesús en Belén
Palestina
Oficinas salitreras de Humberstone
Santa Laura
República de Chile
celebración número 43
ciudad de Baku
Azerbaiyán
pasado 30 de junio
10 de julio

Las relaciones semánticas que aparecen en el texto son:

Parte-Todo_Subsidiario

Ejemplo: <relation_expresion SOURCE="**Comité del Patrimonio Mundial**" TARGET="**Organización de las Naciones Unidas para la Educación, Ciencia Cultura**" TYPE="PREP-pertenencia" REAL_REL="de" **SEMANTIC_TYPE="PART_WHL_Sub**">de</relation_expresion>la<phrase explicit="TRUE">Organización de las Naciones Unidas para la Educación, Ciencia y Cultura </phrase>

Correferencia

Ejemplo: <relation_expresion SOURCE="**Organización de las Naciones Unidas para la Educación, Ciencia y Cultura**" TARGET="**Unesco**" TYPE="SUST" REAL_REL="(" **SEMANTIC_TYPE="COREF**">(</relation_expresion><phrase explicit="FALSE" coreferente="**Organización de las Naciones Unidas para la Educación, Ciencia y Cultura** ">**Unesco**</phrase>

Parte-Todo_Geográfico

Ejemplo: <relation_expresion SOURCE="**Islas y Áreas protegidas**" TARGET="**Golfo de California**" TYPE="PREP-pertenencia" REAL_REL="de" **SEMANTIC_TYPE="PART_WHL_Geo**">del </relation_expresion><phrase explicit="TRUE" TYPE="LOC" NESTED="TRUE">**Golfo** </phrase>

Otra-Relación

Ejemplo: <relation_expresion SOURCE="**Islas y Áreas protegidas del Golfo de California**" TARGET="**Lista del Patrimonio Mundial en Peligro**" TYPE="VERB" REAL_REL="ser_ingresar" **SEMANTIC_TYPE="OTHER_RELATION**">fueron ingresadas </relation_expresion>en la<phrase explicit="TRUE">Lista del Patrimonio Mundial en Peligro</phrase>

Físico_Ubicación

Ejemplo: <relation_expresion SOURCE="**celebración número 43**" TARGET="**ciudad de Baku**" TYPE="VERB" REAL_REL="llevar_a_cabo_en" **SEMANTIC_TYPE="PHIS_Loc**">llevando a cabo en </relation_expresion>la<phrase explicit="TRUE">ciudad de <phrase explicit="TRUE" TYPE="LOC" NESTED="TRUE">Baku</phrase>

El título, tópico y fecha quedan anotados de la siguiente manera

Ejemplo: **<title>**El único hábitat de la vaquita marina fue declarado por la Unesco Patrimonio Mundial en Peligro**</title>**

<topic>Medio Ambiente**</topic>**

<date>03/07/2019**</date>**

La Polaridad global y la Entidad formal son etiquetadas de la siguiente forma.

> **Ejemplo**: <entity type="LOC" globalPolarity="NONE">ciudad de Baku</entity>
>
> <entity type="LOC" globalPolarity="NONE" formalEntity="República de Azerbaiyán">Azerbaiyán</entity>

Se realizó el etiquetado de las entidades anidadas y el tipo de entidad de la siguiente manera.

> **Ejemplo**: <text>Dentro del marco de la celebración de la <phrase explicit="TRUE">reunión del <phrase explicit="TRUE" **NESTED**="TRUE">Comité del Patrimonio Mundial</phrase></phrase>, las<phrase explicit="TRUE"><phrase explicit="TRUE" **TYPE**="LOC" **NESTED**="TRUE">Islas y Áreas protegidas</phrase><relation_expresion SOURCE="Islas y Áreas protegidas" TARGET="Golfo de California" TYPE="PREP-pertenencia" REAL_REL="de" SEMANTIC_TYPE="PART_WHL_Geo">del </relation_expresion><phrase explicit="TRUE" **TYPE**="LOC" NESTED="TRUE">Golfo </phrase><relation_expresion SOURCE="Golfo " TARGET="California" TYPE="PREP-pertenencia" REAL_REL="de" SEMANTIC_TYPE="PART_WHL_Geo">de </relation_expresion><phrase explicit="TRUE" **TYPE**="LOC" **NESTED**="TRUE">California</phrase></phrase>

Las anotaciones de tipo sintáctico se presentan de la forma siguiente.

> **Ejemplo:**
>
> Adicionalmente, el Comité retiró de la Lista del Patrimonio Mundial en Peligro al Lugar de Nacimiento de Jesús en Belén, en Palestina, y a las Oficinas salitreras de Humberstone y Santa Laura en Chile, esto se debe a las restauraciones realizadas en el lugar y el mantenimiento proporcionado, en el caso de Palestina.

Texto etiquetado:

<text>Adicionalmente, el <phrase explicit="FALSE" coreferente="Comité del Patrimonio Mundial ">Comité</phrase><relation_expresion **SOURCE**="Comité" **TARGET**="Lista del Patrimonio Mundial en Peligro" TYPE="**VERB**" REAL_REL="retirar" **SEMANTIC_TYPE**="OTHER_RELATION">retiró</relation_expresion>de la <phrase explicit="TRUE">Lista del Patrimonio Mundial en Peligro</phrase>al<phrase explicit="TRUE">Lugar de Nacimiento de Jesús</phrase> <relation_expresion SOURCE="Lugar de Nacimiento de Jesús" TARGET="Belén" TYPE="PREP-ubicación_espacial" REAL_REL="en" **SEMANTIC_TYPE**="PART_WHL_Geo">en</relation_expresion><phrase explicit="TRUE">Belén</phrase>,<relation_expresion **SOURCE**="Belén" **TARGET**="Palestina" TYPE="**PREP-ubicación_espacial**" REAL_REL="en" **SEMANTIC_TYPE**="PART_WHL_Geo">en</relation_expresion><phrase explicit="TRUE">Palestina</phrase>, y a las<phrase explicit="TRUE"> <phrase explicit="TRUE" TYPE="LOC" NESTED="TRUE">Oficinas salitreras </phrase><relation_expresion **SOURCE**="Oficinas salitreras" **TARGET**="Humberstone" TYPE="**PREP-posterioridad**" REAL_REL="de" **SEMANTIC_TYPE**="PART_WHL_Geo">de</relation_expresion><phrase explicit="TRUE" TYPE="LOC" NESTED="TRUE">Humberstone</phrase></phrase>y<phrase explicit="TRUE">Santa Laura</phrase> <relation_expresion **SOURCE**="Santa Laura" **TARGET**="Chile" TYPE="**PREP-ubicación_espacial**"REAL_REL="en"**SEMANTIC_TYPE**="PART_WHL_Geo" >en</relation_expresion><phrase explicit="TRUE">Chile</phrase>, esto se debe a las restauraciones realizadas en el lugar y el mantenimiento proporcionado, en el caso de <phrase explicit="TRUE">Palestina</phrase>. </text>

De este modo se presenta la polaridad de las entidades.

Ejemplo: <entry source="WRITER" entity="celebración número 43" relativePolarity="NONE" degree="NONE"/>

<entry source="WRITER" entity="reunión del Comité del Patrimonio Mundial" relativePolarity="NONE" degree="NONE"/>

Datos Estadísticos

En la tabla 9 y 10 se muestran las cantidades de relaciones teniendo en cuenta la clasificación gramatical y semántica.

Tipo de relación gramatical	Cantidad
Verbal	5
Sustantiva	1
Preposicional	10
Total de relaciones gramaticales	16

Tabla 9. Cantidad de relaciones teniendo en cuenta la clasificación gramatical

Clasificación semántica	Cantidad
Parte-Todo_Subsidiario	1
Correferencia	1
Parte-Todo_Geográfico	8
Otra-Relación	4
Temporal	2
Total de relaciones semánticas	16

Tabla 10. Cantidad de relaciones teniendo en cuenta la clasificación semántica

Datos Estadísticos generales

En las tablas que siguen a continuación (11 y 12) se muestra el total de clasificaciones gramaticales y relaciones semánticas que aparecen en los ejemplos analizados

Tipo de relación gramatical	Cantidad
Verbal	22
Sustantiva	10
Preposicional	17
Total de relaciones gramaticales	49

Tabla 11. Total de relaciones teniendo en cuenta la clasificación gramatical

Clasificación semántica	Cantidad
Parte_Todo_Subsidiario	2
Temporal	7
Otras_relaciones	5
Parte_Todo_Geográfico	5
Correferencia	33
Físico_Ubicación	3
Organización-Afiliación_Empleo	2
Organización-Afiliación_Política	2
Personal-Social_ Oposición	1
Total de relaciones semánticas	72

Tabla 12. Cantidad de relaciones teniendo en cuenta la clasificación semántica

Discusión

Los datos presentados en las tablas 11 y 12 permiten analizar y discutir las relaciones gramaticales y semánticas identificadas.

En cuanto a las relaciones gramaticales, se observa que existen un total de 49 relaciones en el texto analizado. Se observa que la relación verbal es la más frecuente, con un total de 22 casos. Esto sugiere que los textos analizados se centran en la descripción de acciones y procesos. Por otro lado, se identificaron 17 relaciones preposicionales, lo que indica que se establecen conexiones entre diferentes elementos a través de preposiciones. Al considerar la clasificación semántica de las relaciones, se observa que hay un total de 72 relaciones diferentes. La relación de correferencia es la más predominante, con un total de 33 casos. Esto implica que los textos analizados hacen referencia a un mismo elemento en varias instancias, lo que contribuye a la coherencia del contenido. Debe destacarse que las relaciones temporales también son relevantes, con un total de 7 casos identificados. Esto sugiere que se establezca una secuencia o narrativa coherente en su desarrollo. Además, se identificaron 3

relaciones de físico-ubicación, que indican la posición o ubicación física de los elementos mencionados. Estas relaciones pueden ayudar a visualizar y comprender la disposición espacial de la información presentada. Por otro lado, las relaciones de parte-todo, tanto en su forma subsidiaria como geográfica, indican una conexión entre un elemento y una parte o elemento constituyente. Esto es relevante para comprender la estructura y organización de los análisis abordados en el artículo. Finalmente, las dos relaciones de organización-afiliación-empleo identificadas señalan la conexión entre un elemento y una entidad organizativa o de empleo. Esta relación puede ser importante para comprender las conexiones y roles de los diferentes actores mencionados. El análisis de los datos estadísticos revela una variedad y complejidad de relaciones gramaticales y semánticas en el artículo científico. Estas relaciones son fundamentales para la coherencia y comprensión del contenido, y nos permiten obtener una visión más completa de la estructura y organización de los textos.

Conclusiones

En este trabajo, se utilizó el enfoque basado en corpus como una herramienta práctica para anotar y analizar las relaciones sintácticas y semánticas entre entidades nombradas en un corpus lingüístico etiquetado. Los resultados obtenidos demostraron que este enfoque puede ser beneficioso para identificar patrones lingüísticos y coocurrencias entre los términos utilizados en el corpus. Gracias a esta metodología, se pudo generar una red de relaciones sintácticas y semánticas entre las entidades nombradas. Sin embargo, es importante tener en cuenta algunas limitaciones que pueden afectar los resultados. Estas limitaciones pueden incluir la variabilidad en la calidad y la cantidad de datos disponibles en el corpus, así como posibles sesgos lingüísticos en la selección de términos y entidades. A pesar de estas limitaciones, el enfoque basado en corpus puede ser aplicado en diversas áreas de investigación para mejorar nuestra comprensión de las relaciones sintácticas y semánticas entre las entidades nombradas. Esto puede tener importantes aplicaciones prácticas en el desarrollo de tecnologías lingüísticas, como la mejora de herramientas de

traducción automática, reconocimiento de voz y procesamiento del lenguaje natural.

Referencias Bibliográficas

Alonso, L. (1998). El análisis sociológico de los discursos: una aproximación desde los usos concretos. Ed. Fundamentos.

Arredondo Toledo, L. M. (2018) Extracción de relaciones entre las entidades nombradas en el idioma español (Tesis presentada en opción al Título Académico de Máster en Ciencia de la Computación).

carmitada77, (4 julio, 2015) Análisis del Discurso. *Métodos de investigación* URL: https://metodosdeinvestigaciondcgunefa.wordpress.com/ 2015/07/04/analisis-del-discurso/

Bernal Chávez, Julio Alexánder y Diana Alejandra Hincapié Moreno (2018): Lingüística de corpus. URL: http://bibliotecadigital.caroycuervo.gov.co/1703/1/Linguistica-de-corpus-2018.pdf

Boillos Pereira, Mari Mar. (2018) La elaboración de un corpus del profesorado de español (copele): ¿utopía o realidad? URL: https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-48832018000200153

Culotta, A., & Sorensen, J. (2004, July). Dependency tree kernels for relation extraction. In Proceedings of the 42nd annual meeting on association for computational linguistics (p. 423). Association for Computational Linguistics.

Culotta, A., McCallum, A., & Betz, J. (2006, June). 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.

Filología e informática (1999): nuevas tecnologías en los estudios filológicos (pp. 45-77). Milenio.

Jurafsky, D., & Martin, J. H. (2017) Vector Semantics. Speech and Language Processing: An Introduction to Natural Language Processing. Computational Linguistics, and Speech Recognition (3rd ed draft chapter 15-16).

Lyons, John. (1997): Semántica lingüística. Paidós.

Mar Cruz Piñol. Lingüística de corpus y enseñanza del español como 2/L. Madrid: Arco/Libros, 2017. pp 189. URL: https://www.arcomuralla.com/detalle_libro.php?id=872

Martín Peris, Ernesto. (coord.) (2008). Diccionario de términos clave de ELE. SGEL.

Mercado, H. (2008). Fundamentos de la lingüística de corpus.

Pardo Abril, Neyla Graciela (2002). El contexto y el discurso público. URL: https://revistas.udistrital.edu.co/index.php/enunc/article/view/2465/3432

Sinclair, J.M. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Torruela, J. & Llisterri, J. (1999a). Diseño de corpus textuales y orales. En Filología e informática: nuevas tecnologías en los estudios filológicos (pp. 45-77). Milenio

Wallis, S. and Nelson G. 'Knowledge discovery in grammatically analysed corpora'. *Data Mining and Knowledge Discovery*, 5: 307–340.

Nota biográfica

Reynier Ávila Peña es Licenciado en Letras. Desempeña su labor en la empresa Desarrollo de Aplicaciones, Tecnología y Sistemas (Datys) (Cuba). Se dedica a la construcción de corpus para entrenamiento y evaluación de las herramientas de Procesamiento de Lenguaje Natural. Trabaja en la construcción de ontologías, en la definición de Actos de diálogos en proyectos de asistentes virtuales, en la identificación de Entidades de todo tipo y trabajos de edición, transcripción, corrección ortográfica tanto de documentos, como de textos en la web.

Celia María Pérez Marqués es Doctora en Ciencias Filológicas y Profesora Titular en la Universidad de Oriente, Cuba. Se ha especializado en estudios léxico-estadísticos a partir de corpus textuales. Actualmente coordina la construcción de un corpus de conversaciones de Santiago de Cuba siguiendo la metodología del corpus AMERESCO, creado en la Universidad de Valencia, España.

Yaney Bourzac Álvarez es egresada de un Máster en Ciencias de la Educación. Ha realizado diversos cursos de posgrado: Tutoría y Oponencia; Máster en Ciencias de la Educación; Quién, Qué, Cuándo, Dónde: Etiquetado automático de roles semánticos en el procesamiento del Lenguaje; Freeling 4.0 al descubierto: Uso avanzado de la librería, entre otros. Es especialista en servicios, procesamiento y análisis de la información, en la empresa Datys (Cuba). Cuenta con experiencia en la construcción de corpus lingüísticos para el entrenamiento y la evaluación de las herramientas de Procesamiento de Lenguaje Natural, en la construcción de Ontologías, definición de Actos de diálogos en proyectos de Asistentes virtuales y en la detección y clasificación de entidades nombradas en idioma español, así como sus relaciones semánticas. Además, ha llevado a cabo trabajos de edición, transcripción y corrección ortográfica de textos y audios digitales.

Daymara López Cordero posee experiencia en construcción de Corpus para el entrenamiento y la evaluación de las herramientas de Procesamiento de Lenguaje Natural, y en la identificación de Entidades de todo tipo. Asimismo, ha realizado trabajos de edición, transcripción y corrección ortográfica, tanto de documentos como de textos en la web. Cuenta con experiencia como Tester y Analista de datos.