# Hacía el etiquetado de estados informativos en el Corpus periodístico del noroeste de México (COPENOR)

Mtro. Manuel Alejandro Sánchez Fernández[0000-0002-5173-0754]
Centro de Estudios Lingüísticos y Literarios, El Colegio de México.
manuel.sanchez@colmex.mx

Dr. Alfonso Medina Urrea[0000-0002-0569-6575]
Diccionario del Español de México, El Colegio de México.
amedinau@colmex.mx

**Resumen:**
El presente artículo presenta la creación de COPENOR y su etiquetado lingüístico para analizar estados informativos, lo que podrá contribuir al desarrollo de diversos tipos de tecnologías del lenguaje. La principal aportación radica en la exposición del método de captura de notas periodísticas y la propuesta de taxonomía para el etiquetado de las propiedades pragmáticas de identificabilidad y activación. El enfoque al noroeste del país busca disponer recursos para la comprobación de hipótesis sobre la zona dialectal, y su composición a partir de medios digitales busca apoyar futuras investigaciones sobre el análisis discursivo en ciencias de la comunicación.
**Palabras clave:** periodismo, pragmática, activación, identificabilidad, accesibilidad

**Summary:**
This paper presents the steps for the creation of COPENOR and its linguistic tagset for the analysis of information states, which may contribute to the development of various types of language technologies. The main contribution lies in the presentation of a method for capturing notes and a taxonomy proposal for the labeling of the pragmatic properties of identifiability and activation. The approach to the Mexican Northwest seeks to provide resources for hypothesis testing about this dialect zone, and its composition from digital media seeks to support future research on discoursive analysis in media studies.
**Keywords:** journalism, pragmatics, activation, identifiability, accessibility

## 0. Introducción

En este documento se presentan las características generales del Corpus Periodístico del Noroeste de México (COPENOR), así como una propuesta para el etiquetado de estados informativos –también conocidos como identificabilidad y activación. El corpus fue creado en el marco de una investigación doctoral que tiene como propósito automatizar el proceso de etiquetado de los estados informativos. En una primera sección, se habla brevemente de la hipótesis dialectal de la zona noroeste, principal justificación sobre el corte geográfico que abarca el corpus. Después se muestra las características técnicas del corpus y las bases que guiaron la captura de las notas periodísticas, mientras que en la tercera parte se exponen las bases teóricas que guían la taxonomía de etiquetas, y el diagrama para el etiquetado de las propiedades pragmáticas de identificabilidad y activación. Se concluye esta sección con un análisis de un fragmento de una de las notas del corpus y un ejemplo de cómo está estructurada la nota en XML. El último apartado es una breve conclusión sobre el trabajo por venir y algunos problemas que falta por resolver.

## 1. Un corpus del noroeste

Desde los trabajos de Henríquez Ureña (1921) o Lope Blanch (1970) hasta recientes estudios sociolingüísticos y dialectológicos (Molina, Crhová, y Domínguez 2013), se ha demarcado el noroeste de México como potencial zona dialectal del español mexicano. Autores como Brown (1989), Mendoza Guerrero (2004, 2006), Moreno de Alba (1994) y Serrano (2000) han documentado la existencia de no sólo diferencias léxicas en esta zona –como distintas palabras tomadas del yaqui, mayo u ópata– sino también diferencias fonéticas como la

aspiración de la /x/ y la realización de la /tʃ/ como [ʃ]; e incluso se ha corroborado la existencia de esta zona como representación subjetiva por los mismos hablantes (Serrano 2009). Aunque estas investigaciones han apoyado los hallazgos plasmados en el *Atlas lingüístico de México* (ALM) (Lope Blanch, 1990-2000) –gran empresa que tuvo como objetivo capturar la variación del español mexicano– aún falta mucho para la caracterización lingüística completa del "habla del noroeste".

En este trabajo, tomo como justificación y punto de partida esta hipótesis de zona dialectal, además de considerar una hipótesis secundaria de trabajo en el problema: no sólo tomo la región del noroeste como Sinaloa, Sonora y la zona sur de la península de Baja California, sino que integro lo que Lope Blanch llamaba "bajacaliforniana septentrional" (BSP) y que Everardo Mendoza Guerrero (2006) después delimitara como un área que abarca los municipios de Ensenada, Tijuana, Tecate y Mexicali de Baja California. Siguiendo el trabajo de zonificación de Mendoza Guerrero, considero dentro de mi investigación las sub-zonas noroestes *sonorenses*, *sinaloenses*, *intermedia* y *de transición*; además de la zona BSP. Este trazo coincide con otra propuesta de Lope Blanc (1996) redefiniendo la Zona Noroeste. Por lo anterior, la muestra se toma de la zona geográfica del noroeste del país que comprende Baja California, Baja California Sur, Chihuahua, Durango, Sinaloa y Sonora. Realizar este corte geográfico permite que COPENOR pueda aportar a futuras investigaciones con respecto a la siguiente hipótesis: primero, que la producción mediática –en este caso, notas periodísticas– se ve afectada por la variación dialectal. De manera consecuente, la hipótesis nula es que no existe un efecto de la variación dialectal en la producción mediática. Sin embargo, también se tiene que considerar que, si no

se encuentra variación, esto sólo nos puede decir que en este tipo de discurso no aparece, pero lo contrario otorgaría evidencia a favor de la zona dialectal en general. En todo caso, este trabajo busca ofrecer un corpus que pueda apoyar en la investigación sobre la zona noroeste del país, en particular, sobre su producción mediática.

## 2. Características de COPENOR

Debido a que la creación de COPENOR está en función de una investigación doctoral, una de las principales razones que motiva la compilación de este corpus es evaluar las posibilidades y limitaciones de un etiquetador automático de estados informativos. En los trabajos del área computacional, las primeras investigaciones usualmente se basan en este tipo de corpus debido a la facilidad de su acceso para pruebas y experimentos –como el trabajo de Kibrik *et al*. (2016). Además, las nociones de información nueva y dada, básicas para los estados informativos desarrollados más delante, tienen una relación natural con la *noticiosidad (newsworthiness*[1]*)*: se espera que cada nota nos presente la estructuración lingüística de información nueva relacionada con información dada. La noticia es uno de tantos formatos en la comunicación periodística, que se clasifican en géneros periodísticos. Estos se entienden como plantillas lingüísticas que orientan la creación textual (o en otros formatos) encauzada a estructurar la información, la interpretación o la opinión, de forma eficiente (Catenaccio *et al*., 2011). Estos modelos lingüísticos no son accidentales; no surgen de la repetición de una acción aislada, sino del oficio moldeado por las

[1] En los estudios sobre periodismo, la *noticiosidad* se ha entendido como una serie de valores intrínsecos a ciertos hechos que los destacan de la cotidianidad y los vuelven dignos de ser *noticia*. Estos valores son detectados y aprovechados por los periodistas para construir la nota, ya sea por formación o por experiencia profesional (O'Neill y Harcup, 2009: 161).

demandas de la faena diaria y la interacción con las respuestas de los lectores del medio. El oficio del reportero consiste en formular una pieza de información –un texto, en este caso– antes de que su valor noticioso caduque a partir de convenciones textuales que agilizan la producción de la nota (Salavarría y Cores 2005). Estos modelos lingüísticos no sólo son estructuras aprendidas por los periodistas sino también por sus lectores. La forma del texto ayuda al lector a delimitar las expectativas que tenga sobre la información que se le presenta.

Para COPENOR se consideró sólo el género informativo, en particular la noticia, por lo que los géneros periodísticos interpretativos, dialógicos y argumentativos se encuentran descartados (cf. Salavarría y Cores 2005: 150).

### 2.1 Base de medios de comunicación digitales

El preámbulo de la construcción de COPENOR consistió en tres pasos:

1) **Selección de los medios del noroeste del país**. Cabe mencionar que no existe una sola fuente que tenga una lista exhaustiva de los medios vigentes. Se recurrió a dos páginas de internet que tienen listados de medios de comunicación en el mundo para construir esta base: http://www.prensaescrita.com/ y http://www.abyznewslinks.com/. Esta primera base fue conformada por 125 medios.

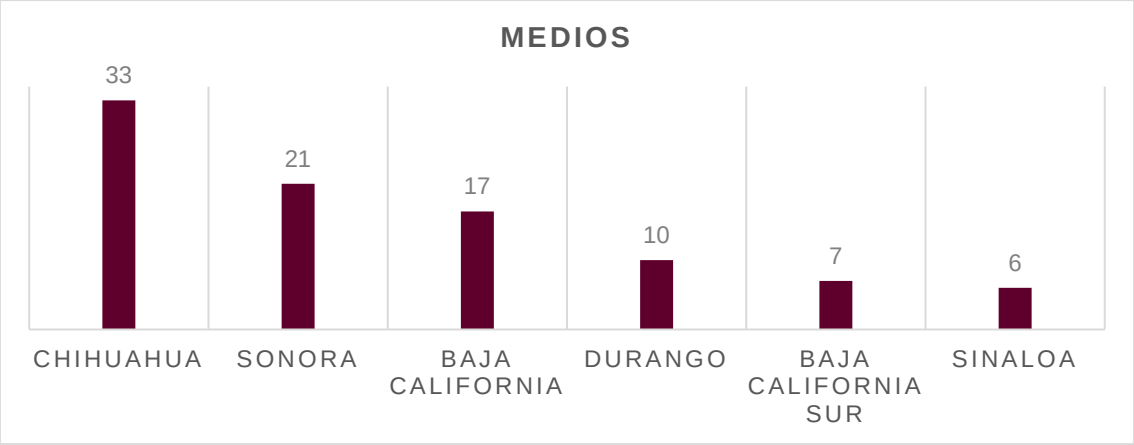
2) **Verificación con expertos.** En un segundo momento, se recurrió al conocimiento de dos expertos (una periodista y un director de un medio de comunicación del noroeste) para corroborar los medios y si era necesario incluir alguno que no estuviera considerando.

3) **Verificación del sitio en línea**. Finalmente, se revisó cada medio para verificar que tuvieran página de internet vigente y que produjeran por lo menos

una nota en la última semana. Esta base final se encuentra en un repositorio creado en Gitlab[2].

Al final se obtuvo una base de 94 medios los cuales se etiquetaron con un ID único (empezando con la letra M y un número de tres dígitos entre 001-094), el estado y ciudad de origen del medio, su nombre y la página de internet.

A continuación, la gráfica (1) muestra la distribución de medios por estado, en donde se puede notar que Chihuahua es el estado con más medios activos (33), seguido de Sonora (21), y en tercer lugar Baja California (17).




Gráfica 1. Distribución de medios por estado.

En cuanto a las ciudades con más medios activos, se muestran los primeros seis lugares en la tabla (1):

| Ciudad | Medios | | Ciudad | Medios |
|---|---|---|---|---|
| Chihuahua (CH) | 15 | | Durango (DU) | 10 |
| Ciudad Juárez (CH) | 14 | | Hermosillo (SO) | 8 |
| Tijuana (BC) | 11 | | Culiacán (SN) | 5 |

Tabla 1. Ciudades con más medios representadas en COPENOR[3].

[2] A lo largo de este trabajo haré referencia "al repositorio" de COPENOR cuya página de internet puede encontrarse en https://opencor.gitlab.io/es/lista-de-corpus/; una ventaja de utilizar este tipo de repositorio es que quedan registrados los cambios a los archivos. Si se desea utilizar algún documento del repositorio, citar como se sugiere en el archivo "Readme.md" presente en la misma página.

[3] Las abreviaturas usadas, tanto en este texto como en COPENOR fueron: BC = Baja California; BS = Baja California Sur; Chihuahua = CH; Sinaloa = SN; Sonora = SO; Durango = DU.

Por mencionar algunos ejemplos de las ciudades con sólo un medio en la muestra, se encuentra Nuevo Casas Grandes (CH); Puerto Peñasco (SO) y Mazatlán (SN). Todos los medios consultados del estado de Durango pertenecen a su capital del mismo nombre.

Debido a que el interés no era cubrir la mayor cantidad de ciudades, para la muestra se toma como criterio de conglomerado el nivel estatal. No obstante, con lo anterior se hace notar el hecho de que no todas las ciudades de los estados están representadas en COPENOR.

En promedio, un medio productivo genera diez notas al día. Al tomar como rango de tiempo de captura un periodo aproximado de dos meses (del 23 de mayo al 16 de julio del 2019), resultó un universo de 57,000 notas. Para obtener la muestra tome la siguiente fórmula[4]:

$$n = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + (\frac{z^2 \times p(1-p)}{e^2 N})}$$

De esta manera, una muestra representativa del noroeste del país, con 95% de confianza y 5% de margen de error, implica la captura de 380 notas periodísticas. Para mantener la proporción de medios por estado, se realizó un muestreo aleatorio por conglomerado que explicaré más adelante.

Lo anterior conformó la base para la construcción del corpus a la cual le siguieron dos etapas: 1) de CAPTURA, la cual está finalizada; y 2) de ANÁLISIS Y ETIQUETADO MANUAL, la cual empezará en enero del 2020. La segunda etapa presupone una preparación teórica y técnica que resultó en la determinación de las etiquetas pertinentes para las propiedades pragmáticas de identificabilidad y activación.

[4] N = tamaño de la población; e = margen de error (porcentaje expresado con decimales); z = puntuación z con respecto al nivel de confianza deseado; p = precisión, que en este caso es 0.5 para maximizar el tamaño de la muestra.

De la misma manera, se necesitó determinar qué otras propiedades era necesario etiquetar manualmente que pudieran servir como predictores de estas propiedades pragmáticas. Con la intención de intervenir el corpus lo menos posible, se decidió etiquetar sólo cuatro factores más: (i) frase nominal y su categoría sintáctica, (ii) oración y su nivel como subordinada o matriz, (iii) lema de cada ocurrencia y (iv) las propiedades gramaticales de acuerdo con el conjunto de etiquetas propuesta por EAGLES[5]. El propósito de la segunda mitad de este artículo es presentar los criterios de etiquetado de estas propiedades, lo cual se desarrollará a partir de la §3.

### 2.2 Captura de las notas

Para generar una semilla aleatoria de la muestra se elaboró un código en Python que lleva a cabo un muestreo aleatorio por conglomerado para tomar también de manera aleatoria uno de los medios. Para ejemplificar el procedimiento, siguiendo la distribución de la tabla 2, supongamos que el número aleatorio de inicio es 25 (entre 1 y 94). El estado en el rango es Chihuahua. Posteriormente, se toma otro número aleatorio entre 1 y 33 –rango que corresponde de la cantidad de medios de ese estado. Supongamos que el resultado de esta nueva tirada es 2.

| Estado | Rango de medios |
|---|---|
| Baja California | 1-17 |
| Baja California Sur | 18-24 |
| Chihuahua | **25-57** |
| Durango | 58-67 |
| Sinaloa | 68-73 |

[5] Para más información sobre este recurso visitar: http://blade10.cs.upc.edu/freeling-old/doc/tagsets/tagset-es.html

| | |
|---|---|
| Sonora | 74-94 |

Tabla 2. Suma acumulativa de los medios por estado.

Esto significa que la primera nota sería el segundo medio en la lista de medios de Chihuahua, que en este caso resulta ser Acento Noticias.

Para la captura de las notas, se dio preferencia a aquellas del día que fueran firmadas por un periodista. Aunque sí se consideraron los casos firmados por la redacción del medio, aquellos firmados por agencias fueron descartados. Un último criterio fue considerar notas policiacas. De esta manera, si se encontraban dos notas potenciales ese día, las dos con información local y firmadas por la redacción, aquella que fuera policiaca era la seleccionada. En el caso en donde ninguno de los criterios anteriores sirviera para determinar la noticia, se seleccionaba la nota de manera aleatoria, aunque estos casos fueron los menos. COPENOR está codificado con el Lenguaje de Marcador Extensible (XML). En este lenguaje es necesario declarar un formato de nombres únicos. En este caso, se realizó a partir de la base proporcionada por el Diccionario del Español de México (DEM).

La estructura de cada nota sigue el etiquetado presentado en la tabla (3) en donde se coloca las consideraciones para cada campo al momento de la captura.

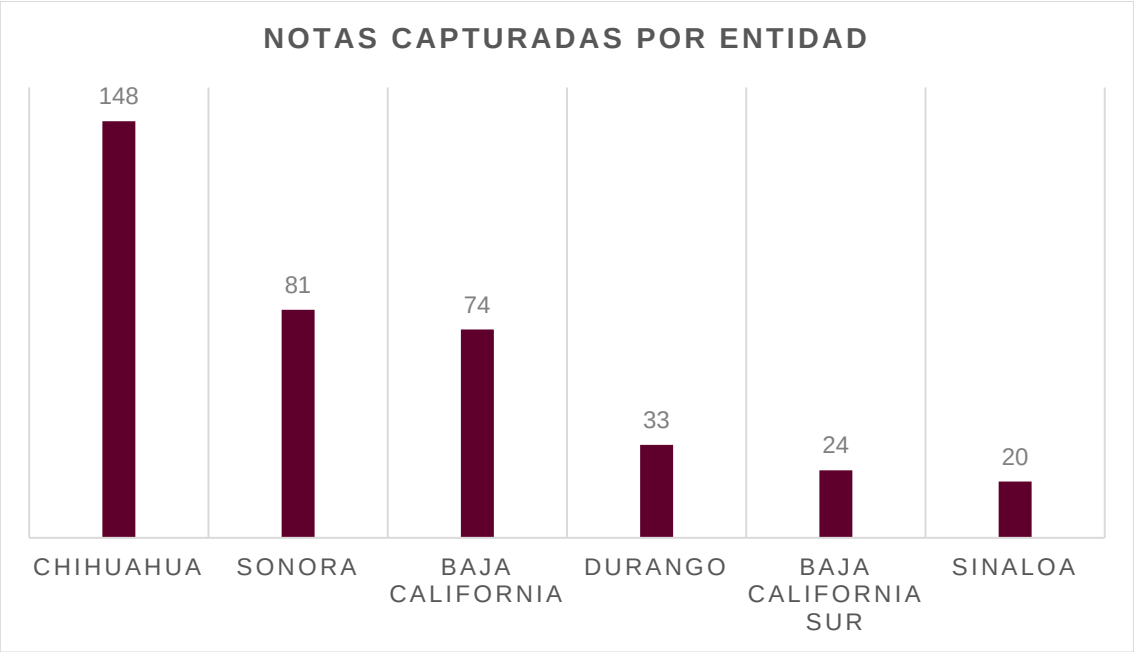
| **XML** | **Descripción** |
|---|---|
| <nota idn=”001CH”> | Cabeza de la estructura de datos de toda la nota capturada en COPENOR. Existen 380 instancias de nota. El Identificador de cada nota (IDN) corresponde a los primeros tres dígitos, seguido de la abreviatura del estado. |
| <titulo> | Título de la nota, obligatorio. |
| <subtitulo> | Subtítulo de la nota, opcional. Hay veces que se colocan los llamados “balazos” como subtítulos: entradas consecuentes |

| | |
|---|---|
| | al título que funcionan como introducciones a la nota, pero resaltadas por un formato distinto al cuerpo de la nota. |
| <medio idm="M001"> | Nombre del medio de acuerdo con la base de datos. El identificador del medio, como se mencionó, inicia con la letra M seguida de tres dígitos. Obligatorio. |
| <URL> | Página de internet de la nota. Obligatorio |
| <estado> | Uno de los seis estados considerados. Obligatorio. |
| <ciudad> | Ciudad de la nota. Si la nota no tiene la ciudad expresada de manera explícita, se coloca la ciudad del medio. Obligatorio. |
| <fecha> | Fecha de la nota. Obligatorio. Si la nota no tiene fecha, se coloca la fecha de captura sólo si se verifica que el medio produjo otra nota ese mismo día. |
| <fuente> | Nombre del periodista: sólo primer nombre y primer apellido[6]. Si no existe este dato, o si son siglas del nombre, se deja en blanco, asumiendo que la redacción firma. Opcional. |
| <contenido> | Contenido textual de la nota en crudo en formato de codificación iso-8859-1. |
| <etiquetado> | Contenido textual etiquetado con frase nominal, oración, EAGLES, lema y estados informativos. |

Tabla 3. Estructura del XML de la nota en COPENOR.

Después del periodo de captura que abarcó del 23 de mayo al 16 de julio del 2019, se capturaron las 380 notas que muestran una distribución por estado como aparece en la gráfica 2.

[6] Para mantener la confidencialidad de las personas mencionadas dentro de cada nota, se cambian los nombres y se coloca sólo la primera letra de su apellido y un punto.




Gráfica 2. Distribución de notas por estado.

Al momento de la captura se presentaron tres problemas técnicos: (i) cuando se volvió a visitar la página de un medio, la página estaba temporalmente dada de baja; para solucionar esto, se escogió otro medio de manera aleatoria y se marcó ese medio en la base de datos, pero no se eliminó de otras posibles tiradas en que volviera a aparecer; (ii) algunos medios tenían una producción muy baja de contenido; en esos casos, como la condición fue sólo que publicaran una nota a la semana, si resultaba que aparecían en el itinerario dos veces en una semana particular, se buscaba una nota más allá del rango de tiempo de captura, por lo que en COPENOR, algunas notas tienen fechas de creación fuera del tiempo de captura; (iii) se debe tener en cuenta que no fue criterio la extensión de la nota ni tampoco la inclusión o no de citas directas, por lo que estas propiedades deberán ser ponderadas al momento del análisis y cálculo estadístico; en todo

caso, las citas directas son colocadas con código HTML: “ y ” para las comillas que abren y cierran respectivamente[7].

## 3. Metodología del etiquetado de estados informativos

Antes de proceder a explicar las propiedades lingüísticas, eje de este corpus, es necesario mencionar algunos detalles de las otras etiquetas también utilizadas. Típicamente, en trabajos de lingüística computacional (p. e. Hempelmann *et al*. 2005; McCarthy *et al*. 2012), la noción de frase nominal se toma como dada y resuelta, lo que dificulta la reproducibilidad de los experimentos desarrollados. En esta investigación entenderé frase nominal en español como un constituyente sintáctico que tiene un núcleo que establece concordancia de género y número con las unidades léxicas que formen parte de este constituyente; se sigue en términos generales, las propuestas de la *Nueva Gramática de la Lengua Española* (Real Academia Española y Asociación de Academias de la Lengua Española (RAE y ASALE) 2009) y la *Gramática Descriptiva de la Lengua Española* (Brucart 1999; Rigau 1999). Por lo anterior, se considera que dentro de la frase nominal se podrán encontrar determinantes que encabezan las FN pero no preposiciones[8]. Los núcleos de una frase nominal pueden ser pronombres, nombres propios, nombres comunes, núcleos elípticos y, en algunos casos, verbos en infinitivo y participio[9]. Para ejemplificar esto, las frases de (1a-c) a continuación –tomadas de Recio Diego (2015:49), excepto el ejemplo

[7] Cuando se finalice el etiquetado de las propiedades gramaticales, se podrán extraer estas secciones mediante el atributo F que contiene las etiquetas EAGLES.

[8] Aunque las frases preposicionales no se encuentran etiquetadas en el corpus, se puede acceder a la extracción de esta estructura a partir de buscar el atributo SP de EAGLES cuando se finalice esa etapa de etiquetado.

[9] En COPENOR, aquellos casos en donde un verbo sea núcleo de FN se etiquetan como verbos pero con la letra N al final, por ejemplo, *el*[DA0MS0] *ser*[VSN0000**N**] *alto*[AQ0CS0].

en (1g)– se encontrarán etiquetadas como FN, pero no los constituyentes que aparecen en negritas de (1d-g):

(1). a. Quiero **pan.**
b. Tienen **un precioso gato persa de dos años.**
c. Dame **la que está en el armario grande.**
d. Decías **que no vendrían tan pronto.**
e. Preguntó **si lo conocíamos.**
f. Confesó **con quién había cometido el crimen.**
g. **Que tu hayas venido** provocó **que se pusieran de acuerdo.**

En el caso de la oración, se toma una visión amplia en la que se consideran como “unidades mínimas de predicación, es decir, segmentos que ponen en relación un sujeto con un predicado” (RAE y ASALE 2009: 1.13a). Se identifican oraciones MATRICES que pueden llegar a ser complejas y oraciones SUBORDINADAS en las que se agrupan las sustantivas, adverbiales y adjetivales. Mientras que existen muchos proyectos de etiquetado de frases nominales, oraciones y propiedades gramaticales, no hay corpus etiquetados con las propiedades pragmáticas de identificabilidad, accesibilidad y activación, que son el núcleo de este artículo y que se desarrollan a continuación.

### 3.1 Antecedentes de las nociones de información nueva y vieja

En lingüística, dos áreas han utilizado de manera general las nociones de información nueva y dada. Por un lado, la semántica, a partir de las teorías de la definitud, establece que esta propiedad se estructura en las lenguas como una forma que instruye al oyente a identificar del referente, ya sea porque existe sólo una entidad a la cual la descripción le es pertinente o porque el referente le es familiar a los interlocutores (Aguilar-Guevara, Pozas Loyo, y Vázquez-Rojas

Maldonado, 2019). Por otro lado, desde la pragmática, y en especial, desde la Estructura de la Información, se parte de dos duplas: la identificabilidad y activación; y el tópico y foco. De acuerdo con la propuesta teórica de Lambrecht (1994), el tópico y el foco operan a partir de las presuposiciones y aserciones pragmáticas. Las primeras, conforman el cuerpo de conocimiento que se presupone en la enunciación de una oración, mientras que la aserción es la relación establecida de manera efectiva en una oración determinada.

En el presente ejercicio de etiquetado no se tiene como objetivo partir de las aserciones y presuposiciones pragmáticas, las cuales dan lugar al análisis de tópico y foco (Lambrecht, 1994), sino sólo considerar la identificabilidad y la activación. Estas propiedades operan tanto en proposiciones como en referentes, entendidos éstos como representaciones mentales de entidades, eventos, estados –por lo que la discusión sobre la existencia real de un referente es irrelevante en este nivel; mientras que las proposiciones se entienden como relaciones estructuradas de manera preponderante por predicaciones verbales –en el caso del español.

De acuerdo con Prince (1981), Chafe (1994), Lambrecht (1994) y Kibrik (2011), la identificabilidad se refiere a la suposición del hablante de que determinado referente puede ser recuperado por el oyente. Durante el intercambio comunicativo, pueden intervenir muchas suposiciones sobre el conocimiento que tiene el oyente, pero las que interesan en la Estructura de la Información son aquellas que pueden ser identificadas en la estructura lexicogramatical. Esto no resulta una tarea fácil ya que existen varios factores que pueden ayudarnos a determinar que un referente se supone identificable en un discurso. La propuesta de este etiquetado y su motivo, apoyado por Kibrik (2011), descansa en que no

existe una forma particular que nos señale esta suposición, sino que entra en juego el DISCURSO en donde opera la frase nominal que estructura la referencia, entendiendo discurso como lo dicho y estructurado a partir de la concatenación de oraciones, la unidad máxima de análisis lingüístico.

Si el referente es identificable, puede deberse a que es ACCESIBLE o tiene algún grado de ACTIVACIÓN. La propuesta de Prince (1981) y Lambrecht (1994) señalan distintas maneras en que el referente puede ser accesible o semi-activo. Kibrik (2011), por su cuenta, ahonda en las maneras en que un referente puede estar activo. En el etiquetado propuesto, parto de los conceptos psicológicos expuestos por Kibrik (2011) para entender no sólo la noción de activación sino también integrar la noción de accesibilidad y distinguir algunas diferencias que parecen problemáticas con la propuesta de Lambrecht (1994).

Para Kibrik (2011), la activación es un ejercicio mental que realiza tanto el hablante como el oyente. Sin embargo, es el hablante el responsable de dar forma al discurso, por lo que son sus suposiciones las que estructuran los dispositivos referenciales. Éstos se dividen en dos grupos (Kibrik, 2011: 37): DISPOSITIVOS REFERENCIALES PLENOS, cuyos núcleos son nombres propios y comunes[10], y DISPOSITIVOS REFERENCIALES REDUCIDOS, cuyos núcleos son pronombres y marcas cero.

La elección referencial que tiene el hablante se circunscribe a las posibilidades lexicogramaticales que le otorga su lengua: desde marcas cero, pronombres ligados, pronombres libres hasta frases nominales con variados modificadores.

[10] Como ya se vio, las frases nominales en español pueden tener otros tipos de núcleos. Se extenderá la visión de Kibrik (2011) de los dispositivos referenciales plenos con las desarrolladas en este trabajo.

En este artículo, al igual que Kibrik (2011: 32), suponemos que la referencia está concentrada de manera preponderante en frases nominales con referentes específicos definidos. Puede existir una variedad de funciones referenciales, entre las cuales se encuentran:

a) Definidas, como *yo, el libro, este libro, mi libro, el libro que me diste, Yusnai*; indefinidas, con ejemplos como *alguien, un extranjero, unas personas*;
b) Existenciales, *Pásame* ***cualquier cubeta***;
c) Universales, ***Todos los niños*** *adoran el helado;*
d) Atributivas, ***El gato más bonito*** *ganará el concurso*;
e) Genéricas, ***El perro*** *es el compañero de la humanidad*;
f) Predicativas, Mi mejor amiga es **violinista**; y
g) Autónimas, *Mi sobrino fue nombrado* ***Matías***.

Hay que notar que la especificidad, de acuerdo con Kibrik (2011), es interpretada como un continuo que abarca todos los usos referenciales mencionados. Con respecto a esto, se entenderá ESPECIFICIDAD como "la intención del hablante de comunicar y hacer manifiesto que pretende referirse a una entidad determinada" (Leonetti, 1999: 858) sin importar si él mismo conoce a la entidad referida. Es precisamente la identificabilidad, en los términos de la Estructura de la Información, que se refiere a si existe la suposición de compartir la especificidad con el oyente. Se reconoce que este concepto es complejo y abarca distintas tradiciones. Sin embargo, debe tenerse en cuenta que el ejercicio que aquí se presenta busca acotarlo para su operacionalización en tecnologías del lenguaje.

Primero que nada, siguiendo de cerca lo expuesto por Kibrik (2011), las menciones de los referentes son formalizaciones de las suposiciones que el hablante tiene sobre el estado de activación que tiene el oyente. Esto, a su vez, supone identificabilidad. Por ejemplo, considérese la siguiente nota inventada:

(2) a. **Un hombre** fue arrestado por **policías ministeriales** en Ensenada.

b. Fue puesto en libertad a las pocas horas.

La frase nominal *un hombre* supone no-identificabilidad, pero su mención basta para activar al referente. Esto también ocurre con la frase nominal escueta, sin determinante, *policías ministeriales* que tampoco es identificable; para este caso, se puede establecer que el hablante no considera relevante que el oyente identifique cuáles policías fueron los que arrestaron, aunque se supone que fueron unos policías en particular. En el caso de la frase *un hombre* muestra que, aunque se trate de un dispositivo referencial pleno que supone la no-identificabilidad del referente, se activa, lo que, de acuerdo con Kibrik (2011), se observa al tener un dispositivo referencial reducido en la segunda oración –el sujeto tácito estructurado en la forma del verbo *fue,* tercera persona singular.

Apelando a bases psicológicas y neurológicas, Kibrik (2011) propone que la activación se da en la MEMORIA DE TRABAJO (MT), un módulo estudiado en las áreas cognitivas como interfaz entre distintas capacidades: atención, percepción y recuperación de memoria a largo plazo. Lo que interesa en Estructura de la Información es la manera en que estas nociones psicológicas pueden ayudarnos a explicar el uso de determinados dispositivos referenciales y también a explicar la elección del hablante por uno de ellos en el discurso.

Dada una oración, en la memoria de trabajo podemos encontrar a los referentes que las frases nominales mencionan. Tomando la primera oración del ejemplo anterior (2), en la memoria de trabajo se encuentran *un hombre* y *policías ministeriales*. En este punto, es necesario notar que la mención de un referente a partir de un dispositivo referencial deja ver la suposición del estado informativo, es decir, si es identificable, accesible o se encuentra activo en la MT.

Si bien, la mención en una oración anterior establece la activación de un referente en la MT, esta activación decae hasta que el referente queda INACTIVO en un discurso dado. Para ello, se emplea el concepto de Registro Discursivo: todo referente mencionado al dejar de estar activo, pasa a este registro desde donde puede ser recuperado. El decaimiento de la activación es gradual y será posible probar esto con COPENOR cuando esté completado a partir de proponer valores a los dispositivos referenciales utilizados a lo largo de un texto para el mismo referente.

Estos conceptos sobre identificabilidad, inactividad y activación están expresados en las distintas taxonomías presentadas por Prince (1981: 237), Lambrecht (1994: 109) y por las explicaciones presentadas por Kibrik (2011). El objetivo es tomar estas nociones y construir un método para etiquetar estas propiedades en COPENOR. Para lograrlo, fue necesario aclarar ciertas diferencias notadas en la revisión bibliográfica sobre lo que se entiende por accesibilidad.

En la propuesta que se presenta, ACCESIBILIDAD se entenderá como la suposición del hablante de que un referente es identificable por el oyente a partir de un razonamiento e inferencia, ya sea por anáfora asociativa, por coherencia discursiva o apelar al Origo – el yo/aquí/ahora (Bühler, 2011). De acuerdo con Noordman y Vonk (2015), tales procesos son complejos. En ellos intervienen no

sólo la memoria a largo plazo, la memoria de trabajo, los referentes activos y los Marcos de los referentes –entendido Marco a partir de Fillmore (1982)–, sino también la coherencia discursiva razonada en un momento dado. Esto complejiza el análisis: si el propósito es tener un corpus desde donde se pueda obtener información probabilística para etiquetar de manera automática los estados informativos, las propiedades que intervienen para identificar la accesibilidad harían la tarea demasiado grande. Tal como mencionan Ziai, De Kuthy y Meurers (2016: 209), son precisamente estos casos los más complejos para las tecnologías del lenguaje. No obstante, nuestra intención es etiquetarlos para aislarlos. En otro momento, se podrían agregar otras etiquetas u otros métodos para mejorar la identificación automática de estos estados.

### 3.2 Propuesta de etiquetado

Con lo anterior, inspirado en los trabajos de Prince (1981), Lambrecht (1994) y Kibrik (2011), se propone la siguiente TAXONOMÍA DE ETIQUETADO para COPENOR sobre estados informativos expuesta en la figura (1).

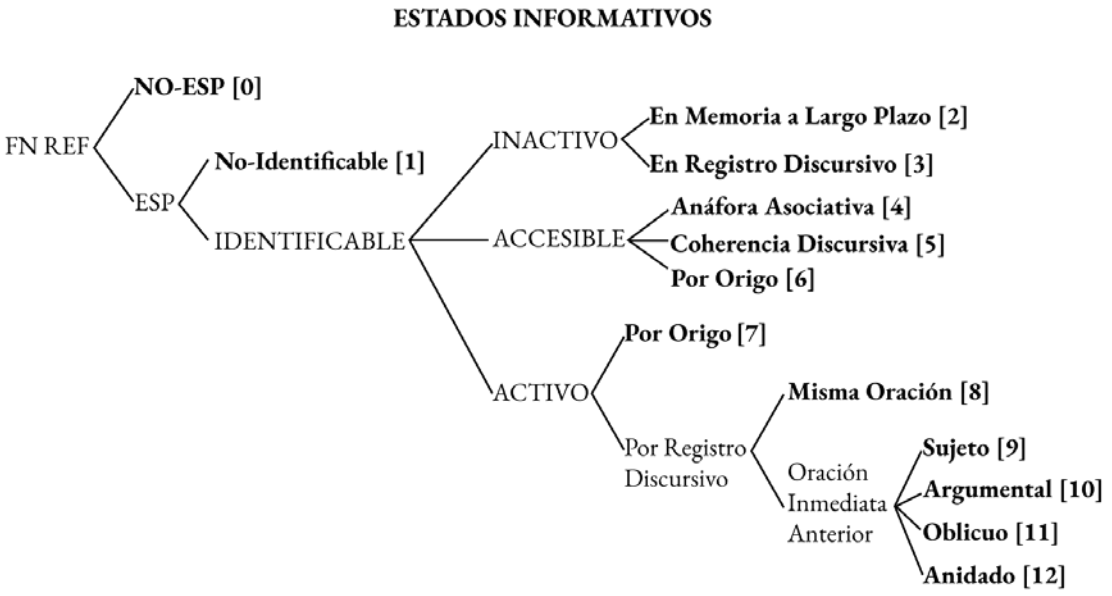


Figura 1. Propuesta de Estados Informativos etiquetados en COPENOR.

El primer paso en el análisis de estados informativos es el determinar si la frase nominal en cuestión es específica. Como se señaló líneas arriba, se acude a una noción casi intuitiva de lo que es especificidad, la cual sólo se puede capturar al comprender el contexto de enunciación. Si bien la estructura de la frase nominal puede dar pistas –se esperarían, por ejemplo, frases encabezadas con determinantes definidos–, cada frase deberá ser evaluada por separado. Aquellas frases que no se consideren específicas no implican inespecificidad total; sólo serán aisladas bajo la etiqueta **[0]** que se nombra **NO-ESPECÍFICO**.

Para determinar si el referente no es identificable, primero se evalúan algunas condiciones ¿La FN hace referencia a un referente que ya fue mencionado antes en el discurso, es decir, se supone acceso a la memoria del **Registro Discursivo?** Si el referente no se encuentra en el Registro Discursivo, significa que en ese texto particular no se ha mencionado, por lo que se desprenden tres opciones: la suposición sobre el acceso a Memoria a largo plazo, a la situación inmediata de enunciación o a algún tipo de inferencia.

En el caso de que la suposición radique en el **acceso a Memoria a largo plazo**, se coloca el número **[2]** y se toma como INACTIVO. En esta categoría, se consideran casos de nombres propios y acrónimos como primera mención sin modificadores.

En el caso de identificabilidad por **inferencia**, se distinguen tres tipos: aquella que supone identificación por relación conceptual, que se nombra Accesibilidad por Anáfora Asociativa, la cual sólo se etiquetará en caso de poder establecer el Marco desde un referente activo. A este estado se le asigna el **[4]** y se referirá a él como ACCESIBLE AA. El otro tipo es la Accesibilidad por Origo, que se refiere al

uso de demostrativos y adverbios que suponen el razonamiento del aquí/ahora del oyente. A este segundo tipo se le asigna el **[6]** y se considera como ACCESIBLE OR. Lo importante de ambos casos es que el referente de la FN no está activo en la Memoria de Trabajo. Es primera mención, pero guarda un tipo de relación con otro referente y se sostiene en la búsqueda de coherencia en el discurso –tema, que como se mencionó, no se tratará en esta investigación, pero que se identifica para aislarse. Es por esto que, si se supone que una frase nominal puede ser accesible por la coherencia en el discurso dado (por ejemplo, que de un aparente *ex nihilo* surja un referente del que puede razonarse una relación con el tema global de la nota periodística), entonces esos casos serán etiquetados como **[5]** y se nombrarán como ACCESIBLE CD. Cabe recalcar que etiquetar estas frases servirá más para aislarlas en el análisis estadístico proyectado que para poder predecirlas, por lo menos, en este punto de la investigación.

La siguiente opción por evaluar es: si se trata de un pronombre de primera o segunda persona, o se usan demostrativos plenos –aquellos que usan el gesto para indicar al referente–, es decir, si los referentes se encuentran activos debido a la participación en el acto de habla o de manera perceptual. Para estos casos se usa el nombre ACTIVO OR. En las notas periodísticas, se busca distinguir si la primera persona corresponde con el autor del texto y si la segunda con interlocutor, debido a que hay casos en donde se codifica el pronombre de primera persona, pero no se hace referencia al escritor. Se esperan pocos o nulos casos de demostrativos plenos. A este conjunto se le asigna el número **[7]**.

Si ninguna de las opciones anteriores se cumple, es decir, si el referente no está mencionado en el discurso anterior, si no puede establecerse la suposición de conocimiento previo, sino es un pronombre de primera o segunda persona, los

cuales refieran a los interlocutores, o si no se puede determinar que interviene el Marco de un referente anterior, el tema global o las coordenadas espaciotemporales, entonces, en este análisis, el referente **no es identificable** y se le asigna el **[1].** Para Lambrecht (1994) y Prince (1981), los referentes no-identificables se dividen entre los no-anclados y los anclados. Estos últimos implican modificadores. En esta propuesta, ambos casos se reúnen en esta etiqueta.

Lo anterior abarca estados de identificabilidad, pero sin aparición en el Registro Discursivo. Para los casos en donde sí exista mención en el texto, la suposición está en que el oyente es capaz de recuperar tal referente por la atención colocada en el presente discurso. Se dividen en seis posibilidades de mención, inspiradas en los factores señalados por Kibrik (2011):

- El referente fue mencionado en una oración más allá de la inmediata. Esto corresponde al estado INACTIVO RD, marcado con **[3]**. Se esperan frases nominales plenas o nombres propios.
- El referente fue mencionado en la misma oración. Para este caso, no es relevante la categoría sintáctica, por lo que es suficiente condición que aparezca la mención en cualquier punto de la oración partiendo del verbo matriz. Este caso se etiqueta como **[8]** y se nombra ACTIVO MO (misma oración).
- El referente fue mencionado en la oración inmediata anterior (OAN). Existe la posibilidad de que haya sido mencionado en distintas partes, una de ellas corresponde al sujeto del verbo matriz de esa oración. Si cumple esta condición, se nombra ACTIVO S, marcado con **[9]**. El antecedente no necesita

ser primera mención en el discurso. La forma del dispositivo referencial se espera también como pronombre o sujeto tácito.

- Mención en la oración inmediata anterior como objeto directo o indirecto del verbo matriz. En este caso se nombra ACTIVO O marcado con **[10]**. Se esperan pronombres clíticos o dispositivos introducidos por la preposición *a*.
- Mención en la oración inmediata como oblicuo del verbo matriz, llamado ACTIVO B y marcado con **[11]**. Se esperan frases nominales plenas, nombres propios, y en menor medida, pronombres o demostrativos.
- Mención como argumento u oblicuo de una oración anidada de la oración inmediata anterior. Se nombra ACTIVO N y se etiqueta como **[12]**. Se esperan frases nominales plenas y nombres propios. En especial, para este estado no se esperan dispositivos referenciales reducidos.

Debe recalcarse que lo anterior se considera una síntesis orientada a ser una guía para el análisis computacional. No se agotan las posibilidades ni se profundiza en los detalles con respecto a conflictos teóricos. Para ilustrar el procedimiento, se presenta la siguiente y última sección de esta primera parte del trabajo como ejemplo de análisis de estados informativos en COPENOR.

### 3.3 Ejemplo de etiquetado

El primer paso para el análisis es dividir el texto en Unidades Elementales del Discurso (EUD) o en oraciones. En este caso, todas las notas se dividen en oraciones matrices y sus correspondientes oraciones subordinadas. Luego de esto, se segmentan las frases nominales, tanto las argumentales como las anidadas. Para la ejemplificación se toman las primeras dos oraciones de la nota COPENOR-253BC. Partamos de la primera oración:

(OR1) 1. [La demanda de [estudiantes que [eligieron]$^{VS}$ a [el CICESE]$^{2}$

2. a través de [los programas de [Verano de la Investigación

3. Científica]$^{1}$]$^{1}$]$^{1}$]$^{1}$ **[aumentó]$^{VM}$** casi a [el doble]$^{0}$ en relación a

4. [el año pasado]$^{6}$, a [el [ser]$^{VS}$ [36 estudiantes de

5. [licenciatura]$^{0}$]$^{8}$ [los seleccionados que [cuentan]$^{VS}$ con [el

6. apoyo de [la Academia Mexicana de Ciencias]$^{1}$ [(AMC)]$^{8}$ y

7. [el Programa Delfín]$^{1}$]$^{0}$]$^{8}$]$^{0}$.

La primera estrategia que puede notarse en el análisis es que se rompen las contracciones del tipo *al* y *del,* como se observa en las L(íneas) 3 y 4. Ahora, partiendo del verbo matriz *aumentó* de la L3, podemos distinguir que la frase nominal argumental inicia en *la demanda* y termina en *científica*. Notamos que es una FN específica, en tanto no se puede sustituir el determinante *la* con *cualquier.* Sin embargo, no hay Registro Discursivo en este punto, por lo que tendremos que evaluar si se apela a la Memoria a Largo Plazo. Debido a que no es un nombre propio ni un acrónimo sin modificador, queda descartado; tampoco notamos que sea necesario recurrir a las coordenadas espaciotemporales del momento de la enunciación, no es núcleo un pronombre de primera o segunda persona, ni tampoco encontramos un demostrativo pleno, o incluso determinante; además, ya que es la primera frase mencionada en la nota, queda descartado el que se vincule con una anáfora asociativa o por coherencia discursiva al no haber discurso en este punto. Por lo anterior, se analiza como **[1]** no-identificable y se coloca este número como superíndice al final de los corchetes que la abarcan. Examinemos ahora las frases nominales anidadas. La siguiente frase inicia con un nominal escueto *estudiantes* y una relativa que termina en *científica* en la L3. El mismo examen de la frase anterior lo llevamos

a cabo y, de hecho, llegamos a la misma conclusión. Lo interesante es que un nominal escueto se nos presenta específico. Esto se considera así porque es posible colocar un definido, *los estudiantes*, pero no un indefinido (por ejemplo, *cualquier número de estudiantes*), sin cambiar el sentido de la oración, y porque el verbo en la relativa, *eligieron*, está en indicativo, lo cual es otra pista para la especificidad. La siguiente frase nominal es *el CICESE* en L1, la cual es un acrónimo como primera mención. Por la decisión metodológica que se ha tomado anteriormente, esta frase nominal encabezada con artículo definido se etiqueta como **[2]**, se supone identificable, recuperable de MLP. La siguiente frase está encabezada por *los programas* y termina en *científica*. El mismo examen para la primera frase nominal es aplicable, sin embargo, ahora podemos examinar el RD: ¿esta frase codifica al referente de *el CICESE*? La respuesta es negativa, por lo que, en efecto, esta otra frase nominal estructura un referente nuevo en el discurso, así que se etiqueta como **[1]**, no identificable. La siguiente frase es un nombre propio *Verano de la Investigación Científica*, está extendido, en el mismo sentido que, por ejemplo, *Juan* no está extendido, pero sí *Juan Pérez de la Oz*. Esta pista es la que orienta a descartar la suposición de recuperación por MLP. Debido a que ninguna de las otras opciones para referentes no antes mencionados es analizable, se etiqueta como **[1]**.

Continuado con el análisis, la siguiente frase nominal que se nos presenta es *el doble* que no es específica. En este caso está funcionando como modificador del verbo, por lo que se etiqueta como **[0].** Luego tenemos la frase *el año pasado*, su modificador nos indica que es necesario acudir al ahora para determinar el referente de esta frase nominal (el año pasado con respecto a este año, 2019). Este referente no fue mencionado antes, por lo que la evaluación pertinente

parece ser la Accesibilidad por Origo, etiquetado con **[6]**. Aunque *el doble* no establece un referente, al ser una mención, posibilita su aparición en el RD. Por lo que, *el doble* y *el año pasado* se integran al RD. Finalmente, se nos presenta la nominalización del verbo *ser* introducido por la preposición *a* lo que nos apunta a la nominalización de una oración subordinada causal, parafraseable como *la demanda aumentó **porque** 36 estudiantes fueron seleccionados.* El verbo *ser* es el núcleo de esta frase nominal que mantiene su estructura argumental en el hecho de que tenemos las dos frases nominales de la cópula ecuativa: *36 estudiantes…* y *los seleccionados…* Las frases nominales con núcleo verbal en infinitivo por lo general se analizan como no específicas, a menos que se tengan pistas de otra función, por lo que en este caso se coloca un **[0]**. La FN *36 estudiantes de licenciatura* se analiza como referente mencionado en la misma oración con el número **[8]**, debido a que seguimos dentro de la oración matriz. Su referente es el mismo que la mención *estudiantes que eligieron…* La siguiente frase es *licenciatura* que en este caso es un atributo por lo que se marca como **[0].** La frase *los seleccionados que cuentan…* tiene la última mención de su referente en la misma oración, por lo que se etiqueta como **[8].** Dentro de esta última frase, encontramos *el apoyo de…* que es una frase no específica, por lo que aparece con **[0]**. Al interior de esta frase, encontramos tres frases coordinadas: la *Academia Mexicana de Ciencias,* que es etiquetada como NO IDENTIFICABLE **[1]** debido a que es la primera mención de un nombre propio extendido, y luego tenemos su acrónimo en una aposición *(AMC)* que se etiqueta como ACTIVO MO **[8]** debido a que el referente de esta mención se encuentra en la misma oración; inmediatamente antes. La última frase de esta oración es el

*Programa Delfín*, la cual, al ser también un nombre propio extendido, se etiqueta como **[1]**. En este punto, el Registro Discursivo luce de la siguiente manera:

1. La demanda de estudiantes que eligieron a el CICESE a través de los programas de Verano de la Investigación Científica
2. estudiantes que eligieron… ; 36 estudiantes…; los seleccionados que cuentan…
3. el CICESE
4. los programas de Verano de la Investigación Científica
5. Verano de la Investigación Científica
6. el doble
7. el año pasado
8. el ser 36 estudiantes de licenciatura los seleccionados que…
9. el apoyo de…
10. Academia Mexicana de Ciencias; (AMC)
11. el Programa Delfín

Hay que notar que todos ellos, de acuerdo con lo planteado en el modelo, se encuentran activos en la MT en este punto de discurso. La forma de estos dispositivos referenciales plenos –en donde el acrónimo *(AMC)* sería el único caso de uno reducido, sin considerar a los que no se etiquetan como FN, como los sujetos tácitos– parte de las suposiciones que tiene el hablante sobre el estado informativo que tiene el oyente de cada uno de ellos.

De esta manera, pasamos a la siguiente oración mostrada en OR2 a continuación.

(OR2) 1. [Los estudiantes]$^{11}$ **[provienen]**$^{VM}$ de [20 [universidades]$^{0}$ y
2. [tecnológicos]$^{0}$ de [México]$^{2}$]$^{1}$, entre [ellos]$^{11}$ [la UNAM]$^{2}$,

3. [la Universidad de [Guanajuato]$^{2}$]$^{1}$, [ITESO]$^{2}$ y
4. [las universidades autónomas de [Nayarit]$^{2}$, [Sinaloa]$^{2}$
5. y [Baja California Sur]$^{2}$]$^{1}$, por [mencionar]$^{VS}$ [algunas]$^{12}$.

Para este caso, no se mostrará el análisis individual que se siguió. Sólo se resaltará que *los estudiantes* son sujeto de esta oración en donde el verbo matriz es *provienen*. Se observa que las menciones de nuevos referentes desplazan a los referentes activos en la MT de la oración anterior. Parece ser que *los estudiantes*, a pesar de estar activo, con tres menciones distribuidas en la oración inmediata anterior, no se presenta con un dispositivo referencial reducido sino con uno pleno sin modificador. En este caso se analiza como **[12]** que significa que aparece anidado en la oración inmediata anterior, esto abarca sus tres menciones, pero si, por ejemplo, una de ellas estructurara al sujeto, se daría preferencia a ese caso para el etiquetado. No obstante, no es el caso, y es interesante notar el detalle en el patrón: a ausencia de sujeto en la mención anterior, el dispositivo se presenta pleno.

### 3.4 Ejemplo de etiquetado en XML

Para completar la exposición anterior, se ilustra la primera oración etiquetada tal y como aparece en el código XML, dentro de la etiqueta <etiquetado> explicado en los metadatos, considerando las siguientes etiquetas y atributos:

| | **Etiqueta** | **Descripción y atributos** |
|---|---|---|
| **Frase nominal** | <fn … > | Definido como se señaló en §3. |
| Identificador | <Idfn=”000”> | Número de tres dígitos. |
| Categoría sintáctica | <cs=””> | su = sujeto; od = objeto directo; oi = objeto indirecto; ob = oblicuo; |

| | | |
|---|---|---|
| | | na = no aplica. |
| Estado informativo | <esin=””> | Número de 01 a 12, de acuerdo con lo señalado en §3.2. |
| **Oración** | <or … > | Definido como se señaló en §3. |
| Identificador | <idor=”000”> | Número de tres dígitos. |
| Tipo | <tp=””> | vm = matriz; vs = subordinada. |
| EAGLES | [---] | Código de propiedades gramaticales entre corchetes. |
| Lema | []--- | Colocado, sin espacios, inmediatamente después de la etiqueta de EAGLES (forma singular, masculina o infinitiva de la ocurrencia tratada). |

Tabla 4. Etiquetas y atributos en el XML de COPENOR.

El resultado de la codificación luce de la siguiente manera:

```
<or orid="001" tp="vm">
    <fn fnid="001" cs="su" esin="01">la[DA0FS0]el demanda[NCFP000]demanda de[SPS00]de
        <fn fnid="002" cs="na" esin="01">estudiantes[NCMP000]estudiante
            <or orid="002" tp="vs">que[PR0CN000]que eligieron[VMIS3P0]elegir a[SPS00]a
                <fn fnid="003" cs="od" esin="02">el[DA0MS0]el cicese[NP000OC]</fn> a[SPS00-]a través[SPS00-] de[SPS00+]de
                <fn fnid="004" cs="ob" esin="01">los[DA0MP0]el programas[NCMP000]programa de[SPS00]de
                    <fn fnid="005" cs="na" esin="01">verano[NCMS000-]verano de[SPS00-]de la[DA0FS0-]el investigación[NCFS000-
                    ]investigación científica[NCFS000-]científico [NP000V0+]</fn>
                </fn>
            </or>
            aumentó[VMIS3S0]aumentar casi[RG]casi a[SPS00]a
            <fn fnid="006" cs="na" esin="00">el[DA0MS0]el doble[RGN]</fn> en[SPS00-]en relación[NCFS000-]relación a[SPS00+]a
            <fn fnid="007" cs="ob" esin="06">el[DA0MS0]el año[NCMS000] pasado[AQ0MSP]pasado</fn>,[Fc], a[SPS00]a
            <fn fnid="008" cs="ob" esin="00">el[DA0MS0]el ser[VSN0000N]ser
                <fn fnid="009" cs="na" esin="08">36[Z]36 estudiantes[NCMP000]estudiante de[SPS00]de
                    <fn fnid="010" cs="na" esin="00">licenciatura[NCFS000]licenciatura</fn>
                </fn>
                <fn fnid="011" cs="na" esin="08">los[DA0MP0]el seleccionados[AQ0MPPN]seleccionado
                    <or orid="003" tp="vs">que[PR0CN000]que cuentan[VMIP3P0]contar con[SP00]con
                        <fn fnid="012" cs="ob" esin="00"> el[DA0MS0]el apoyo[NCMS000]apoyo de[SP00]de
                            <fn fnid="013" cs="na" esin="01">la[DA0FS0]el academia[NCFS000]academia mexicana[AQ0FS0]mexicano
                            de[SPS00]de ciencias[NCFP000]ciencia [NP000O0+]</fn>
                            <fn fnid="014" cs="na" esin="08">([Fpa]( amc[NP000OC]amc )[Fpt])</fn> y[CC]y
                            <fn fnid="015" cs="na" esin="01">el[DA0MS0-]el programa[NCMS000-]programa delfín[NCMS000-]delfín
                            [NP000O0+]</fn>
                        </fn>
                    </or>
                </fn>
            </fn>
</or>
```

Ilustración 1. Etiquetado XML en COPENOR.

Como se puede observar, en este nivel del etiquetado resulta necesaria una interfaz que permita, por un lado, volver más legible el texto original, y por otro,

la búsqueda de información particular –para, por ejemplo, obtener todas las frases nominales sujeto. Esto será parte de los recursos que se desarrollarán en un futuro y que se colocarán en el repositorio.

## 4. Conclusiones

Lo presentado hasta aquí es una propuesta de método para capturar y etiquetar notas periodísticas, que se desarrolló para investigar el español del noroeste del país desde su periodismo. El etiquetado está orientado a resolver una necesidad para una investigación particular, pero consideramos que COPENOR, cuando esté terminado, será útil para otras investigaciones, porque podrán etiquetarse otros niveles, ya que la estructura de los datos está pensada para poder integrar otros etiquetados y propiedades.

La creación de este tipo de corpus aporta en la construcción de puentes entre áreas con objetivos comunes. En especial, el trabajo en tecnologías del lenguaje se beneficia al tener corpus de variados orígenes creados con la supervisión de lingüistas. A su vez, la aplicación de diversas tecnologías puede servir para notar patrones lingüísticos complejos, que a veces resultan invisibles al ojo del analista.

Sumado a esto, con una visión más aplicada, la explosión en producción mediática ha vuelto imperante apoyar las investigaciones sobre, por ejemplo, resumidores y extractores de información. Este trabajo busca ser un pequeño avance, desde la mirada del lingüista, a estos complejos problemas. A pesar de que el español es una de las principales lenguas en el mundo, se encuentra en desventaja con el avance tecnológico del inglés. Peor aún, los dialectos del español quedan subrepresentados, por lo que es importante hacer investigaciones que tomen en cuenta criterios dialectales, como la delimitación

geográfica ofrecida en este corpus. Esperamos que este recurso sea de utilidad a la comunidad y que sirva como propuesta para apuntalar otros proyectos basados en teorías lingüísticas para el desarrollo de tecnología.

## 5. Bibliografía


Aguilar-Guevara, Ana, Julia Pozas Loyo, y Violeta Vázquez-Rojas Maldonado. (2019), "Definiteness across languages: An overview" en Ana Aguilar-Guevara, Julia Pozas Loyo y Violeta Vázquez-Rojas (eds.) *Definiteness across languages*, Berlin, Language Science Press, pp. iii–x.

Brown, Dolores, (1989), "El habla juvenil de Sonora, México: la fonética de 32 jóvenes", en *Nueva Revista de Filología Hispánica,* 37,1, pp. 43–82.

Brucart, José M. A. (1999), "La estructura del sintagma nominal: las oraciones del relativo", en Ignacio Bosque y Violeta Demonte (eds.), *Gramática descriptiva de la lengua española*. *Vol. 1. Sintaxis básica de las clases de palabras,* Madrid, Espasa Calpe, pp. 395–522.

Bühler, Karl (2011), *Theory of language the representational function of language*, Amsterdam, John Benjamins Publishing.

Catenaccio, Paola, Colleen Cotter, Mark De Smedt, Giuliana Garzone, Geert Jacobs, Felicitas Macgilchrist, Lutgard Lams, Daniel Perrin, John E. Richardson, Tom Van Hout, y Ellen Van Praet (2011), "Towards a linguistics of news production", en *Journal of Pragmatics,* 43, 7, pp. 1843–52.

Chafe, Wallace L. (1994), *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*, Chicago, University of Chicago Press.

Fillmore, Charles J. (1982), "Frame semantics", en Linguistic Society of Korea (eds.), *Linguistics in the morning calm*, Seoul, Hanshin Pub. Co., pp. 111–38.

Hempelmann, Christian F., David F. Dufty, Philip M. McCarthy, Arthur C. Graesser, Zhiqiang Cai, y Danielle S. McNamara (2005), “Using LSA to Automatically Identify Givenness and Newness of Noun Phrases in Written Discourse”, en *Proceedings of the Annual Meeting of the Cognitive Science Society*, 27, pp. 941–46.

Henríquez Ureña, Pedro (1921), “Observaciones sobre el español en América”, en *Revista de Filología Española,* 8, pp. 357–90.

Kibrik, Andrej A. (2011), *Reference in Discourse,* Oxford, Oxford University Press.

Kibrik, Andrej A., Mariya V. Khudyakova, Grigory B. Dobrov, Anastasia Linnik, y Dmitrij A. Zalmanov (2016), “Referential Choice: Predictability and Its Limits”, en *Frontiers in Psychology,* 7, pp. 204–24.

Lambrecht, Knud (1994), *Information structure and sentence form: topic, focus, and the mental representations of discourse referents*, New York, Cambridge University Press.

Leonetti, Manuel (1999), “El artículo”, en Ignacio Bosque y Violeta Demonte (eds.), *Gramática descriptiva de la lengua española. Vol. 1. Sintaxis básica de las clases de palabras*, Madrid, Espasa Calpe, pp. 787–890.

Lope Blanch, Juan M. (1970), “Las zonas dialectales de México. Proyecto de delimitación”, en *Nueva Revista de Filología Hispánica*, 19, 1, pp. 1–11.

Lope Blanch, Juan M. (ed.) (1990-2000), *Atlas lingüístico de México*, Ciudad de México, El Colegio de México-Universidad Nacional Autónoma de México-Fondo de Cultura Económica.

Lope Blanch, Juan M. (1996), “México”, en Manuel Alvar (ed.) *Manual de dialectología hispánica. El español de América*, Barcelona, Ariel, pp. 81–89.

McCarthy, Philip M., David F. Dufty, Christian F. Hempelmann, Zhiqiang Cai, Danielle S. McNamara, y Arthur C. Graesser (2012), “Newness and Givenness of Information”, en M. McCarthy y C. Boonthum-Denecke (eds.)

*Applied Natural Language Processing*, Pennsylvania, IGI Global, pp. 457–78.

Mendoza Guerrero, Everardo (2004), “Las hablas del noroeste mexicano: una posible zonificación”, en *Memoria del XIII Congreso de la ALFAL*, Universidad de Costa Rica.

Mendoza Guerrero, Everardo (2006), “El español del noroeste mexicano: un acercamiento desde adentro.” en A. M. Cestero Mancera, I. Molina Martos, y F. Paredes García (eds.), *Estudios sociolingüísticos del español de España y América*, Madrid, Arco Libros, pp. 159–67.

Molina, Rosío Landeros, Jitka Crhová, y María del Rocío Gaona Domínguez (2013), “El habla de Tijuana: material para el análisis de la variante regional”, en *Plurilinkgua,* 9, 1.

Moreno de Alba, José G. (1994), *La pronunciación del español en México*, Ciudad de México, El Colegio de México.

Noordman, Leo G. M. y Wietske Vonk (2015), “Inferences in Discourse, Psychology of”, en James D. Wright (ed.) *International Encyclopedia of the Social & Behavioral Sciences 2ed*, Elsevier, pp. 37–44.

O’Neill, D. y T. Harcup. (2009) “News values and selectivity”, en K. Wahl-Jorgensen y T. Hanitzsch (eds), *The handbook of journalism studies*, New York: Routledge, pp. 161–74.

Prince, Ellen F. (1981), “Towards taxonomy of Given-New Information”, en P. Cole (ed.), *Radical Pragmatics*, New York, Academic Press, pp. 223–55.

Real Academia Española y Asociación de Academias de la Lengua Española (2009), *Nueva gramática de la lengua española*, Madrid, Espasa.

Recio Diego, Álvaro (2015), *La estructura argumental del sintagma nominal en español*, tesis de doctorado en lengua española, Salamanca, Facultad de Filología-Universidad de Salamanca.

Rigau, Gemma (1999, “La estructura del sintagma nominal: los modificadores del nombre”, en Ignacio Bosque y Violeta Demonte (eds.), *Gramática*

*descriptiva de la lengua española. Vol. 1. Sintaxis básica de las clases de palabras*, Madrid, Espasa Calpe, pp. 311–62.

Salavarría, Ramón y Rafael Cores (2005), “Géneros periodísticos en los cibermedios hispanos”, en R. Salaverría y R. Cores *Cibermedios, El impacto de internet en los medios de comunicación en España*, Sevilla, Comunicación Social Ediciones y Publicaciones, pp. 145–85.

Serrano, Julio César (2000), “Contacto dialectal (¿y cambio lingüístico?) en español: el caso de la /tʃ/ sonorense”, en Pedro Martin Butragueño (ed.), *Estructuras en contexto. Estudios de variación lingüística*, Ciudad de México, El Colegio de México, pp. 45–59.

Serrano, Julio César. (2009), “¿Existe el noroeste mexicano como zona dialectal? Un acercamiento perceptual”, en Everardo Mendoza Guerrero, M. López Berríos, y I. E. Moreno Rojas, *Lengua, literatura y región*, México, Universidad Autónoma de Sinaloa, pp. 107–30.

Ziai, Ramon, Kordula De Kuthy, y Detmar Meurers (2016), “Approximating Givenness in Content Assessment through Distributional Semantics”, en *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, Berlin, Germany, pp. 209–18.