Signos Lingüísticos, vol. XVI, no. 31, January-June, 2020, 164-181

Towards construction of the journalistic corpus from Northwestern Mexico (COPENOR)

Manuel Alejandro Sánchez Fernández
ORCID.ORG/0000-0002-5173-0754
El Colegio de México
Centro de Estudios Lingüísticos y Literarios
manuel.sanchez@colmex.mx

Alfonso Medina Urrea
ORCID.ORG/0000-0002-0569-6575
El Colegio de México
Diccionario del Español de México
amedinau@colmex.mx

Abstract: *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 focus on the Mexican Northwest seeks to provide resources for hypothesis testing about this dialect zone. Its digital media composition seeks to support future research on discourse analysis in media studies. 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.*

Keywords: JOURNALISM, PRAGMATICS, ACTIVATION, IDENTIFIABILITY, ACCESSIBILITY

Reception: 9/12/2019 **Acceptance:** 27/05/2020

Signos Lingüísticos, vol. XVI, núm. 31, enero-junio, 2020,164-181

Hacia la construcción del Corpus Periodístico del Noroeste de México (COPENOR)

Manuel Alejandro Sánchez Fernández
ORCID.ORG/0000-0002-5173-0754
El Colegio de México
Centro de Estudios Lingüísticos y Literarios
manuel.sanchez@colmex.mx

Alfonso Medina Urrea
ORCID.ORG/0000-0002-0569-6575
El Colegio de México
Diccionario del Español 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. El enfoque hacia el noroeste del país busca disponer recursos para la comprobación de hipótesis sobre la zona dialectal, y su composición a partir de medios digitales tiene como objetivo apoyar futuras investigaciones sobre el análisis discursivo en ciencias de la comunicación. 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.

Palabras clave: periodismo, pragmática, activación, identificabilidad, accesibilidad

Recepción: 9/12/2019 **Aceptación:** 27/05/2020

Introducción

En este documento se presentan los lineamientos generales para la construcción y el etiquetado del Corpus Periodístico del Noroeste de México (COPENOR), así como una propuesta para el etiquetado de estados informativos (ESIN) —conocidos como identificabilidad y activación—. El corpus fue creado en el marco de una investigación doctoral cuyo propósito es automatizar el proceso de etiquetado de los ESIN. En la primera sección, se habla brevemente de la hipótesis dialectal de la zona noroeste, principal justificación sobre el corte geográfico que abarca el corpus. Le sigue una breve descripción desde las ciencias de la comunicación de cómo se entiende el género periodístico que se trabaja en este corpus. Después se muestran las características técnicas del corpus y las bases que guiaron la captura de las notas periodísticas. En la cuarta 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 ejemplo de la estructura de las notas en el formato XML (Lenguaje de Marcador Extensible) del corpus. Se cierra la nota con una breve conclusión sobre el trabajo por venir y algunos problemas que falta resolver.

1 Un corpus del noroeste

Desde las investigaciones de Pedro Henríquez Ureña (1921) o Juan Miguel Lope Blanch (1970) hasta recientes estudios sociolingüísticos y dialectológicos (Molina Landeros, Crhová y Domínguez Gaona, 2013), se ha demarcado el noroeste de México como zona dialectal del español mexicano. Autores como Dolores Brown (1989), José G. Moreno de Alba (1994), Julio Serrano (2000) y Everardo Mendoza Guerrero (2004, 2006) han documentado diferencias léxicas y fonéticas, como la aspiración de la /x/ y la realización de la /tʃ/ como [ʃ], e, incluso, se ha corroborado la representación subjetiva de esta zona 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) aún falta mucho para la caracterización lingüística completa del "habla del noroeste".

La creación de COPENOR tiene como justificación y punto de partida esta hipótesis de zona dialectal, además de considerar una hipótesis secundaria de trabajo en el problema: no sólo se toma la región del noroeste como Sinaloa, Sonora y la zona sur de la península de Baja California, sino que también se integra lo que se ha denominado como la región "bajacaliforniana septentrional" a partir de las divisiones propuestas por Lope Blanch (1971; 1996) (Martin Butragueño, 2014: 1380). De tal manera, en este corpus se consideran las subzonas noroeste sonorenses, sinaloenses, intermedia y de transición, además de la bajacaliforniana norteña.

De esta manera, COPENOR contiene notas de la zona geográfica del noroeste del país que comprende Baja California, Baja California Sur, Chihuahua, Durango, Sinaloa y Sonora.

2 Un corpus periodístico

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 ESIN. En los estudios 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 ESIN, 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. En las ciencias de la comunicación, la noticia es uno de tantos géneros periodísticos. Éstos se entienden como plantillas lingüísticas que orientan la creación discursiva

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, ya sea por formación o por experiencia profesional, para construir la nota (O'Neill y Harcup, 2009: 161).

encauzada a estructurar la información, la interpretación o la opinión, de forma eficiente (Catenaccio *et al*., 2011). Estos modelos lingüísticos no son accidentales: surgen del oficio moldeado por las demandas de la faena diaria y la interacción con las respuestas de los lectores del medio. El oficio del reportero consiste en dar forma a la información, a partir de convenciones textuales que agilizan la producción de la nota, antes de que su valor noticioso caduque (Salaverría y Cores, 2005). Además, la forma del texto también ayuda al lector a delimitar sus expectativas 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 (*cfr*. Salaverría y Cores, 2005: 150).

3 Características técnicas de COPENOR

3.1 Base de medios digitales

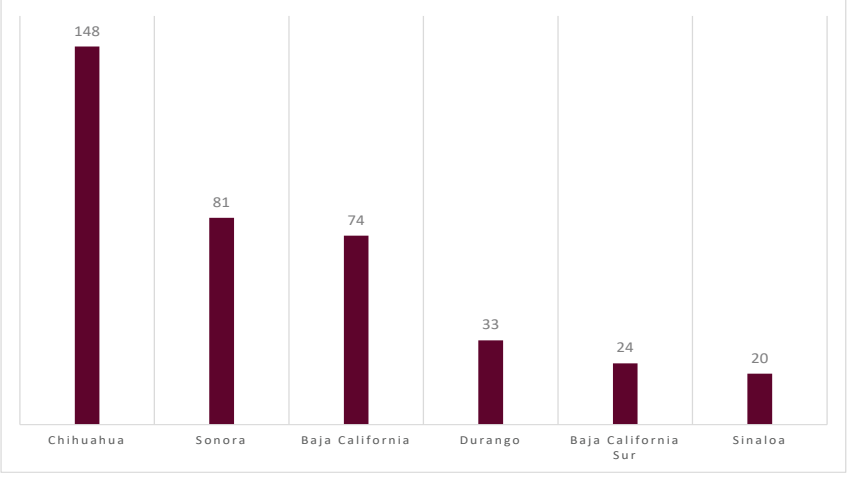
La construcción de la base de medios digitales de COPENOR consistió en tres pasos:

1. **Selección de los medios del noroeste del país**. No existe una sola fuente que posea 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 el director de un medio de comunicación del noroeste), para corroborar los medios y saber si era necesario incluir alguno que no se 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 de la fecha de

la consulta. 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 seguido de un número de tres dígitos entre 001 y 094), el estado y la ciudad de origen del medio, su nombre y la página de internet.

Gráfica 1. Distribución de medios por estado



Fuente: elaboración propia.

2 A lo largo de este trabajo haremos 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 la presente nota.

En la gráfica 1 se muestra la distribución de medios por estado; se puede notar que Chihuahua es el estado con más medios digitales activos (33), seguido de Sonora (21), y Baja California (17).

Debido a que el interés no era cubrir la mayor cantidad de ciudades, para la muestra sólo se toma el nivel estatal como criterio para el conglomerado. En el ambiente profesional del periodismo, se considera que un medio productivo genera aproximadamente diez notas al día. Al tomar como rango de tiempo de captura un periodo de dos meses (del 23 de mayo al 16 de julio del 2019), resultó un universo de 57 000 notas. Para obtener la muestra se utilizó la siguiente fórmula:[3]

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

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.

A la creación de la base de medios digitales le siguieron dos etapas:

1. de captura, ya finalizada, y
2. de análisis y etiquetado manual, la cual está en curso.

La segunda etapa presupone una preparación teórica y técnica que tuvo como resultado la determinación de las etiquetas pertinentes para las propiedades pragmáticas de identificabilidad y activación. 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

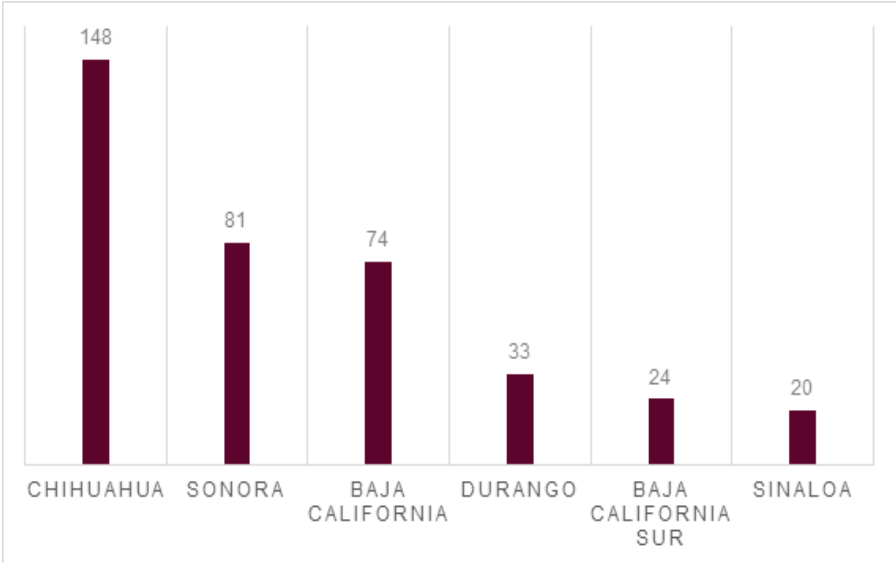
3 N = tamaño de la población; e = margen de error (porcentaje expresado con decimales); z = puntuación z con respecto al nivel de confianza deseado; p = precisión, que en este caso es 0.5 para maximizar el tamaño de la muestra.

y su nivel como subordinada o matriz, (iii) lema de cada ocurrencia y (iv) las propiedades gramaticales de acuerdo con el etiquetador automático Stanza creado por el Stanford NLP Group.

3.2 Captura de las notas

Este periodo 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.

Gráfica 2. Distribución de notas por estado



Fuente: elaboración propia.

Se generó una semilla aleatoria con un código en Python, el cual lleva a cabo un muestreo igualmente al azar, por conglomerado (tomando como conglomerado el Estado y su cantidad de medios activos), lo que también aseguraba que la elección del medio fuera aleatoria.

Para la captura de las notas, de acuerdo con las asignaciones del programa de Python, se consultaba el medio marcado para ese día. Se dio preferencia a aquellas notas del día que fueran firmadas por un periodista. Se consideraron las notas firmadas por la redacción del medio, pero eran descartadas si las firmaba la agencia. Un último criterio fue considerar notas policiacas. De esta manera, si se encontraban dos notas potenciales el mismo día —digamos, con información local y firmadas por la redacción—, aquella que fuera policiaca era la seleccionada. En caso de que ninguno de los criterios anteriores sirviera para determinar la noticia, se seleccionaba otro medio del mismo estado de manera aleatoria.

La estructura de cada nota sigue el etiquetado presentado en la tabla 1, en donde se colocan las consideraciones para cada campo al momento de la captura.

Al momento de la captura se presentaron tres problemas técnicos: (i) cuando se volvió a visitar la página de un medio, ésta se encontraba 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 las 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 ello, 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 como: “ y ” para las comillas que abren y cierran, respectivamente.

Tabla 1. Estructura del xml de la nota en COPENOR[4]

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. Obligatorio.
<titulo>	Título de la nota. Obligatorio.
<subtitulo>	Subtítulo de la nota. En ocasiones, se colocan los llamados *balazos* como subtítulos: entradas consecuentes al título que funcionan como introducciones a la nota, resaltados por un formato distinto al cuerpo de la nota. Opcional.
<medio idm="M001">	Nombre del medio de acuerdo con la base de datos. El identificador del medio 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. Si la nota no tiene fecha, se coloca la fecha de captura. Obligatorio.
<fuente>	Nombre del periodista: sólo primer nombre y apellido. 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. Obligatorio.
<etiquetado>	Contenido textual etiquetado con frase nominal, oración, etiquetado de Stanza, lema y estados informativos. Por el momento opcional.

4 En este contexto, *obligatorio* debe entenderse como que la etiqueta no puede estar vacía, por lo que se espera que un comando de recuperación de texto arroje contenido.

4 Características del etiquetado

En el etiquetado se parte de la noción de *frase nominal* en español como un constituyente sintáctico que tiene un núcleo, el cual establece concordancia de género y número con las unidades léxicas que formen parte de este constituyente. Se sigue, en términos generales, la *Nueva Gramática de la Lengua Española* (Real Academia Española [RAE] / Asociación de Academias de la Lengua Española [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.[5] Los núcleos de una frase nominal pueden ser pronombres, nombres propios, nombres comunes, núcleos elípticos y, en algunos casos, verbos en infinitivo y participio.

En el caso de las *oraciones*, 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.

Para las *propiedades morfosintácticas*, se utilizará el etiquetador automático Stanza (Qi *et al.*, 2020).

4.1 Identificabilidad y estados de activación

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 identificables en la estructura lexicogramatical.

5 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 ADP de Stanza cuando se finalice esa etapa de etiquetado.

Signos Lingüísticos, vol. XVI, núm. 31, enero-junio, 2020, 164-181, ISSN: 1870-4042

De acuerdo con Kibrik (2011), el responsable de dar forma al discurso es el hablante, por lo que sus suposiciones son las que estructuran los dispositivos referenciales. Éstos se dividen en dos grupos: DISPOSITIVOS REFERENCIALES PLENOS, lo cuales tienen nombres propios y comunes como núcleos, y DISPOSITIVOS REFERENCIALES REDUCIDOS, cuyos núcleos son pronombres y marcas cero (Kibrik, 2011: 37). La elección del dispositivo referencial depende de las suposiciones sobre identificabilidad y accesibilidad, pero cuando llega a la activación, el hablante supone que es recuperable del mismo discurso, e, incluso, que está en la atención del hablante.

Tomando como fundamento las investigaciones de los anteriores autores, se sintetizan 10 etiquetas pensadas para el análisis de dispositivos referenciales plenos y su aplicación con algoritmos informáticos. La primera división entre las etiquetas es (1) No-Identificable e Identificable; esta última se divide en recuperable por (2) Memoria a Largo Plazo o (3) por el Registro Discursivo; de identificable también se desprende (4) Accesible por marco y (5) por Origo; y activo. Las etiquetas de activo apelan a variaciones en cuanto a las características sintácticas de la frase nominal analizada, de tal manera, el referente analizado se encuentra activo por ser (6) sujeto; (7) objeto directo/indirecto; u (8) ha sido introducido por una preposición. En el transcurso del análisis se incluyeron dos etiquetas más para considerar ciertos casos especiales que no se expondrán en esta nota, pero que estarán etiquetados en el corpus: (0) No-identificable con baja capacidad de ser recuperado por anáfora e (9) Identificable en situación de aposición.

5 EJEMPLO DE ETIQUETADO EN XML

Este análisis se traduce a un formato que permita su procesamiento informático. Dentro de la etiqueta <etiquetado> del XML en la tabla 1, se agregarían las siguientes etiquetas y atributos:

Tabla 2. Etiquetas y atributos en el xml de copenor

Frase nominal	**Etiqueta**	**Descripción y atributos**
Identificador	<fn … >	
Categoría sintáctica	<ldfn="000">	Número de tres dígitos
	<cs="">	su = sujeto; od = objeto directo; oi = objeto indirecto; at = atributo; pr = preposición; na = no aplica
Estado informativo	<esin="">	Número de 00 a 12
Oración	<or … >	
Identificador	<idor="000">	Número de tres dígitos
Tipo	<tp="">	vm = matriz; vs = subordinada
Stanza	[---]	Código de propiedades gramaticales entre corchetes
Lema	[ ]---	Sin espacios después de la etiqueta de Stanza (forma singular, masculina o infinitiva de la ocurrencia tratada)

Fuente: elaboración propia.

El resultado de la codificación luce como se muestra en la imagen 1.

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. Esto será parte de los recursos que se desarrollarán en un futuro.

Imagen 1. Etiquetado xml en COPENOR

```
<nota idn="253BC">
      <encabezado>
            <titulo>Inició en CICESE verano de la investigación</titulo>
            <subtitulo>Participarán 36 estudiantes de 20 universidades</subtitulo>
            <medio idm="M002">Ensenada.net</medio>
            <url> https://www.ensenada.net/noticias/nota.php?id=57358
            </url>
            <estado>Baja California</estado>
            <ciudad>Ensenada</ciudad>
            <fecha>2019-06-28</fecha>
            <fuente>Elizabeth Vargas</fuente>
      </encabezado>
      <contenido>
      La demanda de estudiantes que eligieron al CICESE ...
      </contenido>
      <etiquetado>
            <ora idora="1" tp="vm">
            <fn cs="su" esin="1_NO_IDENT" idfn="1">
            La[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el demanda[NOUN_Gender=Fem|Number=Sing]demanda
            de[ADP_AdpType=Prep]de
            <fn cs="na" esin="1_NO_IDENT" idfn="2">
            estudiantes[NOUN_Number=Plur]estudiante
            <ora idora="2" tp="vs">
            que[PRON_PronType=Int,Rel]que eligieron[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin]elegir
            a[ADP_AdpType=Prep]a
            <fn cs="od" esin="2_INACTIVO_MLP" idfn="3">
            el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el CICESE[PROPN_]CICESE
            </fn></ora></fn></fn> (...)
            </ora>
      </etiquetado>
</nota>
```

Fuente: Elaboración propia.

6. Conclusiones

Lo presentado hasta aquí es una propuesta de método para capturar y etiquetar notas periodísticas, el cual se desarrolló para investigar el español del noroeste del país desde su producción periodística. El etiquetado está orientado a resolver una necesidad para una investigación particular, pero se considera que COPENOR —cuando esté terminado— será útil para otras investigaciones, ya que la estructura de los datos está pensada para poder integrar otras 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. Esta nota busca ser un pequeño avance —desde la mirada del lingüista— a estos complejos problemas. Aunque el español es una de las principales lenguas en el mundo, se encuentra en desventaja con el avance tecnológico del inglés. Peor aun, los dialectos del español quedan subrepresentados, por lo que es importante hacer investigaciones desde la lingüística computacional y de corpus que tomen en cuenta criterios dialectales, como la delimitación geográfica ofrecida en este corpus. Se espera que este recurso sea útil a la comunidad y que sirva como propuesta para apuntalar otros proyectos basados en teorías lingüísticas para el desarrollo de tecnología.

Bibliografía

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

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

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", *Journal of Pragmatics*, vol. xliii, núm. 7, mayo, pp. 1843–1852.

Chafe, Wallace L. (1994), *Discourse, Consciousness, and Time. The Flow and Displacement of Conscious Experience in Speaking and Writing*, Chicago, University of Chicago Press.

Henríquez Ureña, Pedro (1921), "Observaciones sobre el español en América", *Revista de Filología Española*, vol. viii, pp. 357-390.

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", *Frontiers in Psychology*, vol. VII, pp. 204-224.

Lambrecht, Knud (1994), *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*, Nueva York, Cambridge University Press.

Lope Blanch, Juan Miguel (1996), "México", en Manuel Alvar (ed.), *Manual de dialectología hispánica. El español de América*, Barcelona, Ariel, pp. 81-89.

Lope Blanch, Juan Miguel (1990), *Atlas lingüístico de México*, México, El Colegio de México/Universidad Nacional Autónoma de México/Fondo de Cultura Económica.

Lope Blanch, Juan Miguel (1971), "El léxico de la zona maya en el marco de la dialectología mexicana", *Nueva Revista de Filología Hispánica*, tomo 20, núm. 1, pp. 1-63.

Lope Blanch, Juan Miguel (1970), "Las zonas dialectales de México. Proyecto de delimitación", *Nueva Revista de Filología Hispánica*, tomo 19, núm. 1, pp. 1-11.

Martín Butragueño, Pedro (2014), "La división dialectal del español mexicano", en Rebeca Barriga Villanueva, Pedro Martín Butragueño (eds.), *Historia sociolingüística de México*, vol. 3: *Espacio, contacto y discurso político*, México, El Colegio de México, pp. 1355-1409.

Mendoza Guerrero, Everardo (2004), "Las hablas del noroeste mexicano: una posible zonificación", en Víctor Ml. Sánchez Corrales (ed.), *Memoria del XIII Congreso de la ALFAL*, San José de Costa Rica, Universidad de Costa Rica, pp. 307-314.

Mendoza Guerrero, Everardo (2006), "El español del noroeste mexicano: un acercamiento desde adentro", en Ana María Cestero Mancera, Isabel Molina Martos y Florentino Paredes García (coords.), *Estudios sociolingüísticos del español de España y América*, Madrid, Arco/Libros, pp. 159-167.

Molina Landeros, Rosío, Jitka Crhová y María del Rocío Domínguez Gaona (2013), "El habla de Tijuana: material para el análisis de la variante regional", *Plurilinkgua*, vol. IX, núm. 1, mayo, disponible en [http://idiomas.ens.uabc.mx/plurilinkgua/docs/v9/1/MCD.pdf], consultado: 01 de mayo de 2020.

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

O'Neill, Deirdre y Tony Harcup (2009), "News values and selectivity", en Karin Wahl-Jorgensen y Thomas Hanitzsch (eds.), *The Handbook of Journalism Studies*, Nueva York/Londres, Routledge, pp. 161-174.

Prince, Ellen F. (1981), "Towards taxonomy of given-new information", en Peter Cole (ed.), *Radical Pragmatics*, Nueva York, Academic Press, pp. 223-255.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton y Christopher D. Manning (2020), "Stanza: a python natural language processing toolkit for many human languages", *Association for Computational Linguistics (ACL) System Demonstrations*, disponible en [https://arxiv.org/pdf/2003.07082.pdf], consultado: 01 de mayo de 2020.

Real Academia Española (RAE)/Asociación de Academias de la Lengua Española (ASALE) (2009), *Nueva gramática de la lengua española*, Madrid, Espasa-Calpe.

Rigau, Gemma (1999), "La estructura del sintagma nominal: los modificadores del nombre", en Ignacio Bosque y Violeta Demonte (coords.), *Gramática descriptiva de la lengua española*, vol. I: *Sintaxis básica de las clases de palabras*, Madrid, Espasa-Calpe, pp. 311-362.

Salaverría, Ramón y Rafael Cores (2005), "Géneros periodísticos en los cibermedios hispanos", en Ramón Salaverría y Rafael Cores (eds.), *Cibermedios. El impacto de internet en los medios de comunicación en España*, Sevilla, Comunicación Social Ediciones y Publicaciones, pp. 145-185.

Serrano, Julio (2009), "¿Existe el noroeste mexicano como zona dialectal? Un acercamiento perceptual", en Everardo Mendoza Guerrero, Maritza López Berríos e Ilda Elizabeth Moreno Rojas (eds.), *Lengua, literatura y región*, México, Universidad Autónoma de Sinaloa, pp. 107-130.

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

Manuel Alejandro Sánchez Fernández: Licenciado en Ciencias de la Comunicación y Sociología por la Universidad Autónoma de Baja California, Facultad de Ciencias Administrativas y Sociales; maestro en lingüística por la Universidad de Sonora, Departamento de Letras y Lingüística; doctor en Lingüística por El Colegio de México, Centro de Estudios Lingüísticos y Literarios. Interesado en la descripción de lenguas yumanas, lingüística computacional y de corpus.

Alfonso Medina Urrea: Originario de la Ciudad de México, es Profesor-Investigador de tiempo completo del Diccionario del Español de México (Centro de Estudios Lingüísticos y Literarios, CELL) desde 2012. Obtuvo un Bachelor of Science Degree in Computer Science de la Universidad de Texas en San Antonio (Licenciatura en Matemáticas y Computación, Secretaría de Educación Pública) y un Master of Arts in Latin American Studies de la Universidad de Texas en Austin. En 2003, obtuvo el Doctorado en Lingüística del CELL (promoción 1993-1996). Fue becario sandwich (1998-2000) del Deutscher Akademischer Austauschdienst en la Universidad de Tréveris (Trier), Fach Linguistische Datenverarbeitung. En el Sistema Nacional de Investigadores del CONACYT es Investigador nacional, nivel I, desde 2004.