EL COLEGIO DE MÉXICO

CENTRO DE ESTUDIOS LINGÜÍSTICOS Y LITERARIOS

Hacia el etiquetado de Estados Informativos: experimentación en notas periodísticas del español del noroeste de México

TESIS

que para optar por el grado de

DOCTOR EN LINGÜÍSTICA

presenta

MANUEL ALEJANDRO SÁNCHEZ FERNÁNDEZ

Asesor: Dr. Alfonso Medina Urrea

Ciudad de México Septiembre, 2021

RESUMEN

Esta tesis tiene como objetivo general identificar la propiedad pragmática del Estado Informativo en frases nominales (FFNN) por medio de herramientas de representación vectorial para su etiquetado semisupervisado. Los Estados Informativos son una propuesta que subsume dos categorías desarrolladas en distintas teorías lingüísticas conocidas como *identificabilidad cognitiva* y *activación*. Dichas categorías ayudan a responder, desde una perspectiva lingüística, a la interrogante inicial de esta investigación: ¿Cómo se expresa la información nueva/dada en un texto y en la gramática? El objetivo anterior está fundamentado en buscar un método para la identificación automática de estas propiedades.

Las hipótesis que funcionan como pivotes versan alrededor de evaluar que el Estado Informativo corresponde a una medida que es el resultado de aplicar una función al vector de una FN. Tal vector se obtiene a partir del contexto de la FN, su estructura morfosintáctica, sus elementos léxicos y las acepciones asociadas a estos elementos provenientes del *Diccionario del español de México* (DEM).

Se utilizaron dos estrategias para obtener los vectores de las FFNN y sus medidas: el Análisis de Semántica Latente (LSA por sus siglas en inglés) y una variación llamada SPAN, las cuales demostraron buenos resultados para FFNN en inglés, pero carecían de ensayos en español.

La fase de experimentación se dividió en 3 momentos. El primero, la creación del *Corpus periodístico del noroeste de México* (COPENOR); el segundo, el análisis de este corpus y el tercero, el desarrollo de una propuesta sobre la forma de incluir las acepciones del DEM en el proceso de identificación automática. El corpus final de esta investigación contiene 2 388 FFNN. Tanto la extracción de las FFNN de un corpus inicial de 380 notas, como el etiquetado de sus Estados Informativos se realizó de forma manual.

Durante la fase de experimentación se buscó comparar cuatro contextos: primero, los resultados de los trabajos previos con los logrados en esta tesis; segundo, la medida obtenida por medio de LSA contra la obtenida por SPAN; tercero, el usar las palabras interiores de la FN frente a las palabras que se encuentran en su exterior (para la creación del vector de la FN); y cuarto, la presencia y la ausencia de las acepciones en el interior de la FN.

Las pruebas estadísticas empleadas para evaluar a las medidas como predictores de los Estados Informativos fueron: i) una exploración descriptiva de los datos, ii) correlación *r* de Pearson entre las medidas y las distintas agrupaciones de Estados Informativos, iii) un análisis de varianza y iv) un clasificador de regresión múltiple utilizando distintos conjuntos de predictores. También se probó la normalidad de las distribuciones y se realizó otro análisis de varianza con las pruebas no paramétricas Kruskal-Wallis y Conover-Iman. Finalmente, con el propósito de comparar los resultados, se empleó una estrategia preliminar de clasificación con bosques aleatorios.

Los resultados de esta tesis muestran que las medidas de LSA y SPAN de las FFNN son buenas predictoras para identificar las propiedades que asocian al referente como no mencionado en el discurso o texto. La estrategia de LSA es la que mostró mejor desempeño. Con el clasificador de bosques aleatorios se alcanzó un 81 % de Exactitud para los experimentos que dividen los Estados Informativos en dos grupos: las etiquetas Fuera de texto [0] y Activo por texto [1], en donde la etiqueta Fuera de texto [0] logró una Métrica F_β de 0.87. Cabe mencionar que la regresión múltiple también registra los mismos porcentajes de Exactitud. No obstante, en otras agrupaciones de Estados Informativos, ni la regresión ni los bosques logran identificar los estados activos y accesibles.

Por otro lado, se observó que usando sólo las medidas —lo que es relevante para procesos automáticos— se puede alcanzar una Exactitud que ronda el 80 %. En cuanto a las bolsas de palabras, utilizar las unidades léxicas de la frase mostró mejores resultados que el usar las unidades léxicas que rodean a la frase. La bolsa interior que integró las acepciones del DEM no mejoró la identificación, aunque tampoco la entorpeció. En algunos casos se observó una ligera mejora, lo que podría sugerir resonancia entre usar las bolsas con y sin acepciones. Finalmente, este método muestra que con un corpus de FFNN relativamente pequeño se pueden obtener resultados relevantes para trabajos posteriores en el área.

ABSTRACT

The general aim of this dissertation is to identify the pragmatic property of the Informative States in noun phrases (NPs) by means of vector representation tools for their semi-supervised labeling. Informative States are a proposal that subsumes two categories developed in different linguistic theories known as cognitive identifiability and activation. These categories help to answer, from a linguistic perspective, the initial question of this research: How is new/given information expressed in a text and in grammar? The above objective is guided on the search for a method for automatic identification of these properties. The hypotheses that work as pivots for this dissertation revolve around evaluating that the Informative State corresponds to a measure that results from applying a function to the vector of a NP. This vector is obtained from the context of the NP, its morphosyntactic structure, its lexical elements and the meanings associated with these elements taken from the Diccionario del español de México (DEM).

Two strategies were used to obtain the NPs vectors and their measures: Latent Semantic Analysis (LSA) and a variation called SPAN. Both showed good results for NPs in English but lacked tests in Spanish.

The experimentation phase was divided into three stages: 1) the creation of the Corpus Periodístico del Noroeste de México (COPENOR); 2) the analysis of this corpus; 3) the development of a proposal on how to include the meanings of the DEM in the automatic identification process. The final corpus of this research contains 2,388 NPs. Both the extraction of NPs from an initial corpus of 380 notes and the labeling of their Informative States were carried out manually.

During the experimentation phase, four contexts were compared: 1) the results of previous works with those achieved in this dissertation; 2) the measure obtained by means of LSA versus that obtained by SPAN; 3) the use of the words inside the NP versus the words found outside it (at the creation of the NP vector); and 4), the presence and absence of the DEM meanings inside the NP.

The statistical work used to evaluate the measures as predictors of the Informative States were: i) a descriptive exploration of the data, ii) Pearson's r correlation between the measures and the different groupings of informative states, iii) an analysis of variance and iv) a multiple

regression classifier using different sets of predictors. The distribution's normality was also tested, and another analysis of variance was performed using the Kruskal-Wallis and Conover-Iman nonparametric tests. Finally, for the purpose of comparing the results, a preliminary classification strategy with random forests was employed.

The results of this dissertation show that LSA and SPAN measures are good predictors for identifying labels that associate the referent as not mentioned in the discourse or text. The LSA method is the one that showed the best performance. With the random forest classifier, 81% Accuracy was achieved for the experiments dividing the Informative States into two groups: The Out-of-text [0] and Active-by-text [1] labels, where the Out-of-text [0] label achieved an Fβ Metric of 0.87. It is worth mentioning that the multiple regression also records the same Accuracy percentages. However, in other Informative State groupings, neither the regression nor the forests manage to identify the active and accessible states.

On the other hand, it was observed that by using only these measures -which is relevant for automatic processes- an Accuracy around 80% can be achieved. As for the bags-of-words, using the lexical units of the sentence showed better results than using the lexical units surrounding the sentence. The inner bag that integrated the meanings of the DEM did not improve identification, although it did not hinder it either. In some cases, a slight improvement was observed, which could suggest resonance between using the bags with and without meanings. Finally, this method shows that with a relatively small corpus of NPs it is possible to obtain relevant results for further work in the area.

COMISIÓN LECTORA

Dr. Sergio Bogard Sierra

Centro de Estudios Lingüísticos y Literarios
El Colegio de México

Dr. Iván Vladimir Meza Ruiz

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México

Dr. Juan-Manuel Torres-Moreno

Laboratoire Informatique d'Avignon
Université d'Avignon

Esta investigación fue posible gracias al apoyo del Consejo Nacional de Ciencia y Tecnología (CONACYT)

AGRADECIMIENTOS

A mi madre y mi hermana.

Siempre les estaré agradecido. A ustedes les dedico este sueño materializado.

A mi padre y mi familia.

A Roberto Salazar y Elizabeth Vargas, por su acompañamiento y sensibilidad en todo el proceso, porque fueron guías y amigos que estuvieron ahí desde el inicio y hasta el último brindis, escuchando mis miedos y mis sueños.

A la Dra. Pamela Munro, por apoyarme desde aquella tarde en Los Ángeles, hasta la fecha, y motivarme a seguir preparándome. De la misma manera, a la Dra. Karen Dakin, que me motivó a sumergirme en el mundo de la lingüística.

A la Dra. Ana Lidia Munguía, porque en medio de mis dudas, nunca dejó de creer en mí, y en cada paso que me comprometía tomar, ella me apoyaba desde su trinchera.

A la Dra. Nina Martínez, que junto con las mujeres investigadores que he enunciado, son de esas personas que me apoyaron incluso con mi irreverencia. Que veían en mí lo que podía ser. Que sin pedir nada a cambio y sin conocerme, me dieron su mano para subir un peldaño más.

Me es difícil encontrar palabras que capturen lo profundamente agradecido que estoy con mis compañeros del doctorado Irasema Cruz, Carmen Fajardo, Francisco Chincoya, Juan Ubiarco y Verónica Luna. Gracias por todas las veces que nos reunimos, las pláticas en el camión de regreso a nuestras casas; las veces que nos quedábamos en los salones resolviendo ejercicios, discutiendo, apoyándonos. Las veces en la biblioteca, por lo menos saludándonos y compartiendo algunas palabras para sacar una sonrisa. Hermanes de comedor, resistencia y vigilia. Me siento profundamente afortunado, ya que, sin ustedes, no sé en qué circunstancias hubiera terminado esta etapa de mi vida.

A todos los profesores del doctorado, cada uno aportó en mí perspectivas que desconocía, temas con los que se siguió avivando el interés académico por el área. En particular, le agradezco el tiempo que me dedicaron Niktelol Palacios y Erick Franco. Desde el primer día

me trataron como un igual y me animaron a no despreciar mi curiosidad y mis ganas de crear puentes.

A Violeta Vázquez-Rojas. Sin tu paciencia y sabiduría difícilmente hubiera podido continuar en este camino.

A mi Mentor, el doctor Alfonso Medina. Me has enseñado mucho, y en cada oportunidad que tuve, te lo dije de frente: me siento afortunado de haber encontrado a alguien como tú en este camino. Tú y Violeta me recobraron la esperanza por una academia que no pierde el suelo. Sólo diré que, de todo lo que me enseñaste, me ayudaste a no olvidar que es importante respirar.

A mis lectores, al Dr. Iván Meza por sus asesorías, su apertura a trabajar este tema y a recibirme en el IIMAS. Al Dr. Bogard por saber canalizar mis arranques de efusividad, por ayudarme a recuperar el norte. Su guía y claridad condujeron mi formación en todo momento. Al Dr. Juan Manuel, porque incluso en estas circunstancias de pandemia, siguió apoyando el proyecto y no olvidaré aquellas palabras del seminario en las que dijo con todas sus letras "estamos formando un investigador".

A los matemáticos del CIMAT, y a los economistas del Colegio que me dieron espacio, sus consejos y paciencia para que discutiera y comprendiera ideas.

A los miembros del Diccionario del español de México, que también me brindaron otro espacio de discusión y diálogo, así como ánimos para que no desistiera.

A todas las secretarias del CELL. Todas ellas se empaparon, muchas de las veces de manera accidental, de las sesiones que tenía con Alfonso sobre la tesis, y recuerdo que llegamos a platicar sobre este momento. Gracias Jair y Tania por hacer esos momentos en la dirección tan agradables.

A otros lingüistas fuera del Colegio en los que encontré otros interlocutores. Nunca estuve solo. En especial a Sebas y Francisco.

Esteban, carajo, qué hubiera hecho sin ti, hermano. Jorge Andrés, tampoco te escapas.

A todos, ya buscaremos un espacio para brindar por estos años de trabajo culminado. Pero, como diría una venerable sabia de Hermosillo: esto apenas inicia.

Lista de figura

Lista de tablas

Contenido

Introducción

0.1 Fundamento en una Lingüística Computacional

En nuestra época cada vez es más sencillo el acceso a grandes bases de datos, lo que requiere de conocimientos técnicos especializados. En este contexto, la lingüística computacional parece presentarse como sólo un conjunto de herramientas a disposición de una lingüística de corpus. Esta visión minimiza sus posibles funciones y aportaciones e incluso su historia. En el marco de desarrollar esta tesis doctoral en lingüística he tenido la oportunidad de consultar a investigadores de distintas áreas. Estas conversaciones han nutrido mi búsqueda por delimitar el campo desde donde realizo mi investigación y el área que concuerda mejor con mis intereses personales y profesionales. Antes de entrar de lleno a la exposición del presente trabajo de investigación desarrollo a continuación una serie de reflexiones que se desprenden de estas conversaciones, las cuales enmarcan mi trabajo.

Parto de la posibilidad de ver a la LINGÜÍSTICA COMPUTACIONAL desde tres vértices:

La Ingeniería Lingüística: un área de la ingeniería que trata con procesamiento del lenguaje natural (PLN) cuyo objetivo principal es desarrollar tecnologías para encontrar aplicaciones concretas y realizar la investigación pertinente. Su origen lo podemos rastrear a las primeras traducciones automáticas y, de acuerdo con Kay (2003), correspondería a la evolución de una de las motivaciones de la lingüística computacional, la cual…

> … came from the desire to create a technology, based on sound scientific principles, to support a large and expanding list of **practical requirements** for translation, information extraction, summarization, grammar checking, and the like. In none of these enterprises is success achievable by linguistic methods alone (p. xviii, negritas propias).

La lingüística de máquina: área que en este momento es ciencia ficción, pero de la cual podríamos no estar tan lejos. Desde la Ingeniería Lingüística se han desarrollado modelos que generan lenguaje, tal como GPT-2 de OpenAI[1], Watson de IBM[2] o Google Duplex[3]. Debido a que el propósito de estos recursos es desarrollar una inteligencia artificial (IA) que solucione problemas específicos de lenguaje, no se considera prioritario develar la estructura de reglas lingüísticas. La máquina aprende, de manera no supervisada, los patrones adecuados a partir de grandes cantidades de datos. Si bien es debatible el que la máquina realmente *aprenda algo*, lo cierto es que los modelos actuales generan lenguaje de tal manera que *parece* que la máquina *habla*; modelos que parecen superar el *Test de Turing* (Turing 1950) o los planteamientos de John Searle y su Caja China (Searle 1980).

Dado este escenario, propondría que el siguiente paso para los lingüistas fuese rastrear las estructuras que captura una inteligencia de este tipo, haciendo "preguntas" de manera similar a las que se realizan en la elicitación de lenguas humanas. Un ejercicio de este tipo implicaría cuestionar las suposiciones antropocéntricas de la relación entre lengua y pensamiento. La antigua pregunta de Turing sobre si las máquinas pueden pensar será un cuestionamiento constante en los años venideros dado el auge de las tecnologías del lenguaje, en donde, me parece, el lingüista, deberá replantearse la posibilidad de que este nuevo "interlocutor" *usa* una lengua. Esto lo presento como un pequeño vistazo desde la ciencia ficción hacia un

[1] Radford et. al (2018) https://openai.com/blog/better-language-models/; para 2021 existe GPT-3, un sistema que supera en creces a este primer modelo en apenas tres años.
[2] https://www.ibm.com/mx-es/watson
[3] https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

posible rol del método descriptivo del lingüista, parecido al que tenía la psicóloga de robots Susan Calvin en los relatos de Isaac Asimov.

La Lingüística Computacional: cuyo pivote, planteo, es el axioma de Método Computacional.

> **Axioma de Método Computacional:** un método para refutar una hipótesis en lingüística consiste en probar su capacidad para implementarse en algoritmos informáticos. Para evaluar esta capacidad, la teoría debe brindar los pasos suficientes para identificar los elementos que la conforman; tales pasos se convertirían en reglas, probabilísticas o discretas, susceptibles a ser programadas. Entre más elementos no estén delimitados de manera explícita por la teoría más problemas existirán en su implementación computacional, llegando al extremo en el que la teoría sería inoperable con alguna tecnología del lenguaje[4].

Por ejemplo, la teoría establece la existencia de unidades llamadas “frases nominales”. Para su identificación existe un cuerpo de supuestos sobre su funcionamiento. Una forma de probar la cientificidad de la teoría sería a partir de evaluar que este cuerpo es suficiente para delimitar a la entidad en estudio; esto podría lograrse al medir la capacidad de implementación en un algoritmo informático y, además, que tal algoritmo llegue a la misma identificación que un lingüista humano. Tal como lo manejan autores como Kay (2003) o Hausser (2014), la motivación teórica de una área así provenía de la creciente percepción de

[4] No es mi intención adjudicarme la autoría de un axioma de este tipo, el cual ha sido mencionado de diversas maneras en textos del área computacional. No obstante, sí es mi responsabilidad volver explícita la motivación que guía mi trabajo y método. Considero necesario que de una vez por todas se mencione este supuesto para que esté a merced de la crítica de lingüistas que no participan en el área computacional.

que se podían alcanzar avances en las propuestas teóricas al evaluar la capacidad de estos sistemas a ser modelados de manera electrónica: “requiring that a formal system be implementable helped to ensure its internal consistency and revealed its formal complexity properties” (Kay 2003, xvii).

Evidentemente este axioma metodológico representa una versión fuerte de la base con la que se trabaja. En realidad, por ejemplo, ni entre humanos existe un acuerdo al 100% al momento de realizar análisis, usando misma teoría y mismos pasos para la identificación de entidades lingüísticas. Además, la capacidad de una teoría a transformarse en un algoritmo informático de identificación no es una condición suficiente para la cientificidad de esa teoría. En todo caso, sostengo, que este axioma de método es un marco de trabajo válido para realizar investigación científica, pero, obviamente, no el único.

Lo anterior va de la mano con que, en nuestra época, la ciencia ya no puede ignorar las ventajas de los avances en ciencias computacionales para realizar investigación. Estos recursos nos permiten encontrar patrones estructurales que de otra manera serían invisibles. Esta aproximación está emparentada, por ejemplo, con la informática médica y biológica, las cuales buscan utilizar inteligencias artificiales capaces de proponer hipótesis novedosas a través de grandes volúmenes de información (desde proteínas, ADN, placas de rayos-x, patrones de frecuencia acústica, hasta artículos de investigación en medicina que necesitan una rápida síntesis).

Una versión ligera del axioma consistiría en sólo buscar identificar entidades lingüísticas que la teoría predice por el método computacional que fuese. Mi trabajo se circunscribe a esta versión ligera del axioma metodológico. Dado este preámbulo, continuo con el planteamiento de la presente investigación.

0.2 Planteamiento

El documento que a continuación se desarrolla es una investigación enmarcada en lingüística computacional. Tiene como objetivo preliminar utilizar, en un algoritmo informático, los conceptos de lo nuevo/dado desarrollados en lingüística en frases nominales referenciales con el fin de identificarlos y evaluar las posibilidades de etiquetarlos de manera automática. La forma del discurso y determinadas unidades léxicas, tanto de contenido funcional como conceptual, nos hablan de relaciones con conocimiento anterior. Para cualquier hablante enfrentado a un texto o discurso, distinguir estas nociones de lo nuevo/dado no resulta difícil. De hecho, es parte de la competencia en su lengua poder manejar con soltura tanto la codificación como decodificación de las relaciones entre la información que se le presenta con información anterior, tanto de manera explícita en el texto como supuesta. En esta investigación, el primer problema que me interesa afrontar es la aparente paradoja en la que, para un hablante, esto resulta intuitivo y asequible, pero en la teoría lingüística es problemático delimitar estos conceptos. Mucho más complejo resulta tratar de utilizarlos con ánimos de crear un identificador automático.

Investigaciones como las de Prince (1981; 1992), Brown y Yule (1983), Chafe (1994), Leonetti (1999; 1990), Lambrecht (1994), Kibrik (2011; 2016), o Abbott (2010) han tratado de hacerle frente a la dicotomía nuevo/dado como nociones lingüísticas. Sumado a estos esfuerzos, trabajos como los de la escuela de Praga (Hajičová, Sgall, y Skoumalová 1993; Sgall, Hajičová, y Panevová 1986; Hajičová, Sgall, y Skoumalová 1995) han tratado de establecer una teoría lingüística propia que ayude a predecir qué piezas dentro de un texto

proveen información nueva/dada, además de buscar de manera activa que su teoría pudiera implementarse en un algoritmo informático.

Para Lambrecht (1994), lo nuevo/dado se estudia a través de las suposiciones de conocimiento compartido que se manifiesta en los estados de activación e identificabilidad y es relevante para proseguir el análisis de tópico y foco en una oración. Pero, como señala Prince (1981), lo nuevo y lo dado, en lingüística, no son categorías resueltas, es decir, con una descripción que prevea todas sus formas y sus contextos de aparición. Esta autora señala que se necesita una taxonomía clara, con fronteras bien delimitadas y un conjunto de pruebas estructurales que ayuden a identificar tales propiedades, aunque lamentablemente ella no dispone esas pruebas.

Lo anterior muestra que, al preguntarle a distintas propuestas teóricas cuál es la mejor manera de obtener la información nueva/dada de un texto, la respuesta no es unificada. Propongo que en este problema se utilicen los conceptos de estados de activación e identificabilidad en lugar de "lo nuevo/dado". En mi trabajo, estos conceptos los subsumo en un conjunto de 10 categorías diferenciadas a partir de la revisión teórica que he realizado, las cuales llamo ESTADOS INFORMATIVOS (EI). De esta manera, sostengo que, para identificar la novedad en un texto, se deben observar sus DISPOSITIVOS REFERENCIALES, específicamente sus frases nominales (cf. Capítulo 1). De ellas se analizan su estructura morfosintáctica y el contexto de aparición, lo que nos permite determinar su dependencia a conocimiento compartido.

La complejidad del fenómeno advierte algunos problemas, entre otras cosas porque el Estado Informativo es una propiedad a merced de distintas variables. Es decir, no es un fenómeno que podamos establecer unívocamente: por ejemplo, la forma del artículo definido en español no está vinculada con sólo un Estado Informativo, ya sea la identificabilidad, la activación o

la accesibilidad del referente. Tal y como menciona Kibrik (2011), la relación de un dispositivo referencial con aquellos mencionados anteriormente en el discurso se resuelve de manera multifactorial. Este autor menciona algunos de estos factores, como la distancia retórica del antecedente, su protagonismo, la sujetividad, el rol sintáctico y semántico, entre otras propiedades lingüísticas.

Me parece que la presente investigación resulta una aportación justo en este aspecto. El problema teórico y computacional sobre las relaciones entre dispositivos referenciales se ha concentrado en sus versiones reducidas, específicamente en los pronombres como en los trabajos de Kibrik (2011) y Givon (1983a), así como Kehler (2018) y Morales Carrasco (2004); pero no en dispositivos referenciales plenos o frases nominales.

Esta investigación aporta a la discusión teórica sobre la novedad y lo dado en dispositivos referenciales plenos, y dado el axioma del Método Computacional, vuelve evidente que los criterios para determinar las propiedades pragmáticas no son consistentes, pero podrían serlo desde una óptica multifactorial y probabilística más que discreta y reducida a la mera presencia de una u otra forma. Además, es precisamente esta clase de fenómenos complejos los que, me parece, deberían atenderse asistidos por herramientas computacionales: de la misma manera que es imposible analizar una proteína de manera manual, así también hay fenómenos lingüísticos que necesitan apoyo informático para su exploración. La asistencia por computadora puede automatizar el análisis, asistir en la identificación en grandes corpus e incluso señalar nuevas hipótesis.

Por lo que, al cuestionamiento inicial, la teoría nos proporciona una respuesta distinta: si no es posible implementar un algoritmo para “lo nuevo/dado”, tal vez sí un algoritmo para identificar los Estados Informativos, siempre y cuando se articule con ello una teoría sobre

el significado que pueda ser susceptible a la formalización en computadora; en este caso, una teoría de la representación vectorial del significado.

Para crear el algoritmo informático de identificación y etiquetado automático de Estados Informativos me apoyaré en la propuesta de Análisis de Semántica Latente (Landauer et al. 2007). Esta teoría del significado sostiene que la representación vectorial de un pasaje/palabra es reflejo de un sistema de relaciones que retratan diferencias pertinentes en un texto o corpus.

Esta implementación implicaría tener un **corpus** etiquetado con frases nominales y oraciones, así como las propiedades morfosintácticas de todos los elementos del texto en cuestión. Este algoritmo buscaría tanto dentro de la frase nominal como en su alrededor y realizaría las comparaciones adecuadas con otras frases nominales para determinar el Estado Informativo.

Por lo que, el planteamiento anterior se justifica desde tres áreas:

- En lo práctico, servirá para el análisis lingüístico de propiedades, automatizar este proceso permitiría abonar en estudios sobre la correferencia y en agilizar análisis de corte estadístico, en lingüística de corpus o en sociolingüística.
- En lo teórico, debido a que permitirá ver en dónde se carece de una definición lo suficientemente explícita para delimitar conceptos, y refinar estas definiciones en aras de que hasta una computadora pueda seguir las instrucciones, tanto si la identificación se logra por métodos probabilísticos o discretos. Además, la mayoría de los trabajos se concentran en dispositivos referenciales reducidos, por lo que orientar la investigación a dispositivos referenciales plenos aportaría al estudio del fenómeno de manera general.

- En lo metodológico, abonará en la investigación y al propósito de la lingüística computacional. Además, la implementación de Conjuntos de Sentidos (que explicaré más adelante), en particular, el inventario que ofrece el Diccionario del español de México, resulta novedoso para la investigación en lexicografía computacional.

El tipo de corpus que planeo construir, además, ayuda a futuras investigaciones en lingüística de corpus y sociolingüística: un corpus etiquetado con frase nominal, oración, categorías sintácticas de los argumentos y propiedades morfosintácticas de las palabras de notas periodísticas del noroeste mexicano.

0.3 Objetivo e hipótesis

Entrando en detalle sobre la implementación propuesta, y siguiendo el planteamiento anterior, el objetivo de mi investigación es:

(O_1) Identificar la propiedad pragmática del Estado Informativo en frases nominales a partir de herramientas de representación vectorial, para su etiquetado de manera semisupervisada.

Para exponer las hipótesis que le sigue a este objetivo, supongamos que tenemos un discurso (un texto o un diálogo) que contiene una serie de dispositivos referenciales. Cada uno de estos dispositivos hacen referencia a una entidad; varios pueden hacer referencia a una misma entidad y ninguna entidad carece de un dispositivo referencial en el discurso. De acuerdo con lo expuesto por Kibrik (2011), podemos encontrar dos tipos de dispositivos referenciales: plenos y reducidos. Los mejores ejemplos de los primeros son frases nominales con extensas

descripciones (1a), mientras que los reducidos pueden ser tan cortos como un pronombre o las marcas de persona en el verbo (en el caso del español) (1b)[5]:

(1) a. [**El perro que me regaló mi tía en el año en que se dio aquella enfermedad extraña**] aún vive conmigo.

b. ([**Aquel**]) aún vive conmigo.

La estructura morfosintáctica de los dispositivos referenciales[6], de acuerdo con las predicciones de Prince (1981), Lambrecht (1994) y Kibrik (2011), depende de las suposiciones del hablante sobre la **identificabilidad** y el **estado de activación** del referente. Los dispositivos referenciales reducidos implican que el hablante supone que el referente del dispositivo es **identificable** y está **activo** en un momento dado de la conversación/texto, lo cual no supone gran problema en un etiquetador automático, aunque merece su propia comprobación empírica. No obstante, los dispositivos referenciales plenos traen consigo otras dinámicas un poco más complejas.

El primer supuesto que articula la hipótesis de este trabajo es que los Estados Informativos son parte de las propiedades pragmáticas de un dispositivo referencial. Para identificar esta propiedad, se debe observar la forma del dispositivo referencial pleno y su contexto de aparición. En esta investigación sólo me concentro en analizar los Estados Informativos de dispositivos referenciales plenos, y en particular, sólo las frases nominales que no contengan pronombres en sus núcleos.

[5] Uso los corchetes [] para marcar los constituyentes que me interesa ilustrar en los ejemplos. En los casos de citas directas, los llego a utilizar para insertar comentarios propios o cortar fragmentos.
[6] A la estructura morfosintáctica la llamaré como la *forma* del dispositivo referencial pleno.

El segundo supuesto del que parto se fundamenta en que para esta investigación me es necesaria una representación que sea eficiente en un algoritmo informático pero que además parta de una teoría del significado. Supongo, entonces, basado en los trabajos de Landauer (1997; 2007), que las palabras en un texto tienen una representación vectorial, y que esta misma idea puede ser extendida a constituyentes de órdenes superiores. El significado de un constituyente está dado por su contexto y su forma, lo que se verá reflejado en su representación vectorial.

Por lo que la lógica de este razonamiento deductivo versa de la siguiente manera:

1. El Estado Informativo (EI) es la suposición que tiene el hablante sobre el estado de un referente en la mente de su interlocutor codificada en la forma del dispositivo referencial usado en un momento dado del texto/diálogo (T).
2. El EI y el contexto son propiedades que influyen en la forma de la estructura morfosintáctica de una Frase Nominal (FN). Analizando esta estructura morfosintáctica podemos determinar *parcialmente* el EI. Es necesario incluir el análisis del contexto, que influye tanto en el EI como en la estructura morfosintáctica.
3. La representación vectorial de una FN (representado con una flecha sobre la variable, en este caso $\overrightarrow{FN}$) es producto de su estructura morfosintáctica que a su vez se ve influenciada por su contexto.

Se sigue que la forma de un vector $\overrightarrow{FN}$ también es influenciada por el EI de la FN en T.

En cuanto a las propiedades morfosintácticas de las frases nominales analizadas, me limitaré a las que se pueden identificar a través del *Universal Tag Set* de Stanza (cf. Capítulo 2). A su

vez, me limitaré a entender discurso o texto (T) como todas las FFNN anteriores a la FN analizada y una cantidad de palabras alrededor de cada una[7].

Algo importante a señalar es que el proceso por el cual se obtiene un vector de FN no es el mismo proceso para descomponer ese vector y obtener una medida que represente un EI, es decir, la función no es biyectiva. De hecho, parto del supuesto de que no es posible tal función.

Por lo anterior, mi hipótesis general consiste en la siguiente afirmación:

(H_1) El Estado Informativo corresponde a una medida que es resultado de aplicar una función a $\overrightarrow{FN}$, vector que se obtiene de una FN en T, su estructura morfosintáctica y un contexto.

Si $\overrightarrow{FN}$ es el resultado de un contexto y una estructura morfosintáctica en T, entonces, al aplicar la función adecuada debería ser posible extraer una medida que se asocie con un Estado Informativo, lo que conduzca a su identificación y etiquetado automático.

Para esta investigación propongo dos funciones a evaluar: la primera es el cálculo de coseno con respecto las FFNN anteriores; la segunda es el cálculo de la proyección de $\overrightarrow{FN}$ al hiperplano generado por los vectores de las FFNN anteriores (llamado SPAN en Hu et al. (2003) y Hempelmann et al. (2005)). Además, se tendrá que evaluar la posibilidad de no poder descartar la hipótesis nula, entendida como que estas dos funciones no nos proporcionen una medida que se pueda asociar al EI de la FN evaluada.

[7] Esta noción de discurso es sólo para el algoritmo informático. Para la discusión de este trabajo utilizo la noción de discurso que desarrollo en la §1.2.2.

Una hipótesis que se desprende de las dos anteriores explora el uso de las acepciones del Diccionario del español de México en la función. Un método para poder obtener el vector es recurrir a un inventario de sentidos que nos permita determinar el valor de una secuencia de palabras. La hipótesis que incluye estas acepciones sostendría que:

(H_2) El Estado Informativo corresponde a una medida que es resultado de aplicar una función a $\overrightarrow{FN}$, vector que se obtiene de una FN en T, su estructura morfosintáctica, un contexto y **acepciones asociadas a las palabras contenidas en la FN.**

La variación de esta última hipótesis respecto a la que presento en (H_2) es la manera en que se obtiene el vector de donde se determina el EI[8].

0.4 Estructura de la investigación

El siguiente trabajo está dividido en tres grandes capítulos. En el **Capítulo 1 Teoría lingüística sobre lo nuevo/dado** busco responder las siguientes interrogantes: ¿qué significa nuevo o dado para los lingüistas? la estructura lingüística de un texto ¿refleja su novedad? ¿se puede crear un índice de novedad de un texto? ¿es la novedad una propiedad lingüística? ¿alguna teoría lingüística proporciona los pasos para la identificación de esta propiedad?

Estas primeras preguntas me llevaron a consultar diversos textos (Du Bois 2003; Halliday 1967; Chafe 1976; 1994; Lambrecht 1994; Kibrik 2011; Abbott 2010; Prince 1981; García

[8] Los pasos precisos para la integración de las acepciones del DEM pueden consultarse a detalle en §2.4.3.

Fajardo 2016; Aguilar-Guevara, Pozas Loyo, y Vázquez-Rojas Maldonado 2019; Leonetti 1999; 1990; Ariel 1990; Brown y Yule 1983; Alcina Caudet 1999) que dan evidencia que en lingüística las nociones de lo dado/lo nuevo no están resueltas, pero al mismo tiempo, no se duda de su utilidad para describir fenómenos que afectan la estructura de la lengua. En particular, el tópico/foco, la referencia, la especificidad y la definitud son fenómenos lingüísticos que necesitan de estas nociones para delimitar sus alcances descriptivos.

La conclusión a la que llego de esta revisión (como se verá en el Capítulo 1) es que, en las lenguas, en el nivel pragmático-léxico, la **estructura de la información** es en donde se observan las relaciones estructurales en un texto que dan forma a los **dispositivos referenciales** a partir de su dependencia con distintas fuentes de información (Lambrecht 1994). La forma del dispositivo es consecuencia de la elección que realiza el hablante dado un conjunto de opciones a las que tiene acceso gracias a su lengua y que son pertinentes en un momento particular del discurso (Chafe 1994). Estas opciones deben su forma a las suposiciones que el hablante hace constantemente sobre si el referente forma parte del **conocimiento** del oyente y si tal referente se encuentra en su **atención**. Esta revisión teórica me permite suponer que no existe una regla asociada a una forma (ya sea el artículo definido o el modo indicativo, por ejemplo) para determinar que el referente es información que el oyente tiene y que además atiende, es decir, que coloca su atención en un momento del discurso determinado[9]. Si queremos recurrir a alguna teoría lingüística para determinar un tipo de "índice de novedad" de un texto, mi propuesta es observar los dispositivos referenciales del texto y analizar su forma y su relación con otros dispositivos referenciales

[9] Aunque si existen reportes de una tendencia de la sujetividad (*subjecthood*) y la definitud que no ha sido medida con precisión (Prince 1992; Givón 1983b; 2001).

del mismo texto. Cabe señalar que esta clase de fenómenos necesitan no sólo de alguna teoría lingüística para ser tratados: en ellos se entrelazan fenómenos cognitivos que dejan ver lo complejo de una investigación lingüística. Al igual que Wallace I. Chafe (1994) pienso que *no se puede entender el lenguaje humano sin entender la mente humana*.

Para ello, basado en la visión multifactorial y cognitiva de Kibrik (2011), e inspirado en trabajos anteriores sobre el tema (Prince 1981; Chafe 1994; Lambrecht 1994) propongo **un conjunto de etiquetas** para realizar el análisis de Estados Informativos, entendidos como un conjunto que cubre identificabilidad y estados de activación. Estas propiedades pragmáticas me parecen que son lo más cercano a una noción intuitiva de lo nuevo/lo dado desde la teoría lingüística. En el ejercicio de síntesis para crear este conjunto, busqué mantener un *compromiso cognitivo*[10] además de apegarme al axioma de Método Computacional. Esto con el fin de lograr utilizar las etiquetas en un algoritmo informático.

Me queda señalar que, aunque hay ciertas categorías lingüísticas que parece que obvio en el trabajo, especialmente la de especificidad y definitud, no es mi intención problematizarlas y desarrollarlas por completo. Tomo una perspectiva pragmática sobre lo que se entiende por especificidad y la separo del concepto de identificabilidad. Sólo retomo la definitud para

[10] Establecido por Lakoff (1990) y retomado por Kibrik (2011, 16–17) como "… the commitment to coordinate linguistic research with what is known about the mind and brain from the neighbouring sciences also exploring cognition, in particular psychology and neuroscience" siguiendo con sus exposición, coincido plenamente en que el fenómeno que trato en esta tesis, al tratarse de lengua en discurso, la única manera de entender adecuadamente los procesos discursivos es a través de entender los elementos subyacentes como son la memoria, la atención, la consciencia, la representación del conocimiento o la categorización. Aunque esto podría parecer como "accesorio" a un trabajo de naturaleza computacional, no lo es, como lo ha demostrado el mismo Kibrik al buscar implementar sus hallazgos a algoritmos informáticos (Kibrik 2011, cap. 14; Kibrik et al. 2016).

señalar que no es garantía de un Estado Informativo particular, con lo que me alejo de las posturas lógico-filosóficas que han tratado el tema. Espero esto quede claro en el Capítulo 1.

En el **Capítulo 2 Metodología para el etiquetado automático de Estados Informativos** inicio con una revisión general sobre los métodos empleados para la identificación de lo nuevo/dado y repaso, de manera breve, el trabajo de la escuela de Praga. Después, desarrollo con detalle los pasos para el cálculo de LSA y SPAN que seguí en la investigación, los cuales describo para que puedan ser susceptibles a reproducibilidad y revisión. Le sigue la presentación del tratamiento del Diccionario del español de México como Conjunto de Sentidos y las adecuaciones que le realicé para implementarlo en el análisis. Continúo con una descripción sobre la manera en que se construyó el Corpus Periodístico del Noroeste de México (COPENOR), su captura, estructuración en XML, etiquetado de frases nominales y oraciones (con lineamientos estipulados en el Anexo A), y el algoritmo manual seguido para el etiquetado de Estados Informativos (EI); después de este proceso le sigue un etiquetado automático de propiedades morfosintácticas y tipos de palabra (*Part-Of-Speech tag*) a partir de Stanza. En este punto, describo con mayor detalle el conjunto de experimentos a realizar y los análisis estadísticos que me permiten establecer relaciones entre las medidas y los Estados Informativos. Me concentro en utilizar las pruebas estadísticas utilizadas en trabajos anteriores, además de incluir pruebas estadísticas no paramétricas dado el tipo de datos que presento.

En el **Capítulo 3 Resultados de LSA y SPAN como predictores** muestro los resultados de cada una de las pruebas, dejando ver qué modelo predice mejor la presencia del EI de las FFNNs. Para ello, primero expongo las matrices de contingencia que muestran las correlaciones entre las distintas medidas y los Estados Informativos, luego realizo una

regresión múltiple, primero con todas las medidas y luego con aquellas que mostraron la mejor capacidad predictora. Lo anterior me permite comparar mis resultados con los de las últimas investigaciones que siguen un método similar. Por la misma razón muestro los resultados de un Análisis de Varianza de Una sola vía (ANOVA). A lo anterior le agrego una serie de pruebas estadísticas para medir la normalidad de los datos ya que, aunque las anteriores pruebas me permiten comparar resultados en el área, y, aunque el análisis de ANOVA es lo suficientemente robusto para otra clase de distribuciones, me parece adecuado ajustar las pruebas para la clase de datos particular que tengo. Por ello, realizo un análisis de varianza con el método Kruskal-Wallis a través de la prueba de Conovar-Iman. Sumado a esto, utilizo un algoritmo para realizar bosques aleatorios y determinar cuáles medidas funcionan mejor como clasificadores de las variables, lo que sería la antesala a la construcción de un etiquetador automático.

Para cerrar la investigación, en el **Capítulo 4 Conclusiones** muestro las limitaciones y posibilidades de un etiquetador de este tipo, una evaluación de en qué medida alcancé el objetivo y cómo se logró corroborar la hipótesis general. Cierro con una vuelta a la reflexión inicial sobre la visión de una lingüística computacional y las investigaciones que quedan pendientes.

Capítulo 1. Teoría lingüística sobre lo nuevo/dado

1.1 Introducción

En este capítulo desarrollo el marco teórico que guía el análisis de mi corpus, así como los conceptos de identificabilidad, accesibilidad y activación que pretendo identificar de manera automática en frases nominales. En la §1.2 abordo el problema de las nociones de información nueva/dada tanto en lingüística como en tecnologías del lenguaje. Señalo que, aunque ha sido deseable identificar patrones lexicogramaticales que puedan orientarnos hacia estas propiedades, no ha habido éxito de establecer claramente a qué se refieren, a pesar de recurrir a ellas para definir funciones en la lengua. Las áreas en lingüística desde donde se han abordado estos conceptos han sido la semántica, con el concepto de definitud, y la pragmática, con la estructura de la información. Si bien, los trabajos en lingüística computacional han buscado partir de los términos lingüísticos de tópico/foco para aproximarse a la información dada/nueva, no han sido fructíferos, entre otras razones, porque en lingüística no es posible relacionar una forma con la función de *lo dado*, y porque en los trabajos de lingüística computacional que he explorado se nota poco compromiso por afianzar nociones psicológicas relacionadas con el lenguaje.

En la §1.2.2 señalo que, desde el trabajo de Lambrecht (1994), las bases de la estructura de la información son la identificabilidad y la activación, antes que el tópico y el foco, y que se necesita el reconocimiento de sus bases psicológicas para ser definidas con mayor cuidado. Hipotetizo, al igual que Lambrecht (1994), que existen patrones lexicogramaticales para estos estados mentales en los que se encuentran los referentes discursivos y muestro que desde la estructura de la información existen por lo menos cuatro maneras distintas de entender *lo dado*, por lo que acoto mi estudio a los estados de identificabilidad, accesibilidad

y activación. La frase nominal es el punto de partida para el análisis de la referencia y de los estados mentales en los que se encuentran los referentes, entendidos como representaciones mentales, por lo que en la §1.3 defino la manera en que entiendo frase nominal en español, y en la §1.4 adelanto la exposición sobre su función referencial entendida desde una perspectiva pragmática (Alcina Caudet 1999).

El modelo que sigo para analizar la activación de los referentes discursivos proviene de Kibrik (2011) y lo desarrollo en la §1.5, en donde también expongo las bases psicológicas a las que apela este autor para explicar conceptos como grados de activación, atención, memoria de trabajo y memoria a largo plazo. Me detengo en §1.6 a revisar con un poco más de detalle la noción de accesibilidad, y en esta sección propongo una definición que resulte operativa para una investigación en lingüística computacional como la que presento. En la §1.7 continuo la exposición del modelo de Kibrik (2011), con énfasis en los conflictos de activación y ayudas referenciales. En la §1.8 busco mostrar cómo este autor analiza los dispositivos referenciales que codifican a los referentes discursivos y sus estados de activación a partir de observarlos como productos de un fenómeno multifactorial y señalo cuáles factores podré incorporar en mi análisis.

En este punto de mi trabajo llego a proponer un conjunto de categorías que llamo *Estados Informativos* a partir de los trabajos de Prince (1981) y Lambrecht (1994) e inspirados por el fundamento psicológico en lingüística dispuesto por Chafe (1994) y Kibrik (2011), síntesis que desarrollo en la §1.9. Le sigue en §1.10 una descripción más fina de los alcances para la caracterización de la referencialidad en frases nominales y las etiquetas que propongo para el análisis de Estados Informativos en dispositivos referenciales plenos. En §1.11 expongo un ejemplo de análisis de una nota de mi corpus. A manera de vínculo entre este capítulo y

el capítulo metodológico, en §1.12 hablo sobre las bases de la teoría de semántica latente que fundamentan el trabajo computacional con el mismo nombre y cierro esta sección con las conclusiones del capítulo.

1.2 Estados mentales de los referentes

1.2.1 Preámbulo: información nueva/dada, definitud y referencia

Considérese el siguiente fragmento de noticia:

(2) La demanda de estudiantes que eligieron al CICESE a través de los programas de Verano de la Investigación Científica aumentó casi al doble en relación al año pasado, al ser 36 estudiantes de licenciatura los seleccionados que cuentan con el apoyo de la Academia Mexicana de Ciencias (AMC) y el Programa Delfín (…).

COPENOR-253BC[11]

No necesito ser lingüista para advertir que se me informa algo como lector. Con sólo ser hablante de español me es suficiente. Existe información nueva que puedo destacar e información que el escritor omite pero que supongo como compartida, como la ausencia de descripción en algunas partes. De tal manera, aunque podría desconocer por completo qué es *el CICESE* o l*os programas de Verano de Investigación Científica*, y el fragmento no me hace el favor de describirme qué son con mayor precisión, puedo saber que *la demanda aumentó casi el doble con respecto al año pasado*, y que el número de estudiantes de

[11] Las características del corpus que construí para esta investigación, COPENOR por Corpus Periodístico del Noroeste de México pueden revisarse en §2.6.

licenciatura es de *36*. No es de sorprender que el área de procesamiento de lenguaje natural y recuperación de información parta de estas mismas intuiciones para proceder en sus investigaciones, con lo que asumen que existe una correlación entre la información nueva o dada y la estructura gramatical.

En diversos trabajos sobre lingüística computacional se ha hablado acerca de los términos "información nueva" e "información vieja" (Hempelmann et al. 2005; McCarthy et al. 2012; Ziai, De Kuthy, y Meurers 2016). Se asume que tales ideas, tan intuitivas como se presentan a los legos, deben tener un correlato en la estructura de la lengua y el discurso. De estos trabajos, podemos desprender dos incógnitas que han guiado su agenda de investigación: en un texto ¿qué piezas lingüísticas son información nueva/dada? y ¿cuáles son las características gramaticales que permiten determinar tal situación informativa? Aunque estos planteamientos son genuinos y podrían tener un alcance importante en las tecnologías del lenguaje, no han germinado tan rápido como otros temas de investigación en el área computacional (Jones 1994; Nadkarni, Ohno-Machado, y Chapman 2011). En lo que procede de este preámbulo mi objetivo es mostrar la complejidad de los términos *información dada/nueva* desde la lingüística y cómo es que tampoco están del todo definidos. Se han adecuado a las explicaciones de fenómenos lingüísticos obviando su base psicológica. Como es de esperar, no busco generalizar: precisamente, me apoyo en Chafe (1994) y Kibrik (2011) para sostener que la explicación de algunos fenómenos gramaticales se encuentra en los avances de la psicología y las neurociencias, una explicación plausible, mucho más robusta que sólo apelar a las intuiciones de lingüistas que recurren a términos cognitivos como "memoria", "atención" o "conocimiento compartido", que, de acuerdo con Kibrik (2011, 53), no son utilizados de manera consistente y precisa.

En algunas teorías lingüísticas, las nociones de información dada y nueva se han integrado en sus aparatos explicativos con algunas diferencias. Uno de los trabajos que ilustra esta complejidad es el de Ellen F. Prince (1981)[12], quien propone que el concepto de información dada (*givenness*) en lingüística se replantee como FAMILIARIDAD ASUMIDA[13] (*asummed familiarity*), entendido a partir de tres tendencias en la teoría lingüística de su momento que funcionan como explicaciones a diversos mecanismos gramaticales. Estas tendencias son: (i) PREDICTIBILIDAD (*predictability*), que parte de que el hablante tiene hipótesis sobre las creencias del oyente y elabora estrategias comunicativas con respecto a ellas (Halliday 1967); (ii) PROMINENCIA (*saliency*), que se refiere a la asunción que tiene el hablante sobre que cierta entidad se encuentre en la consciencia del oyente al momento de la enunciación (Chafe 1994); y (iii) CONOCIMIENTO COMPARTIDO (*shared knowledge*) en el que el hablante asume que el oyente conoce una entidad particular o la puede inferir (Krifka 2008; Givón 1983a).

Al final, para el concepto de Familiaridad Asumida, la autora descarta el acercamiento desde el Conocimiento Compartido, entre otras razones debido a que no se puede hablar directamente de algo *compartido*. En el mejor de los casos, la estructura lingüística sólo nos deja ver las suposiciones del hablante. No podemos decir mucho sobre lo que el oyente tenga en la mente hasta que éste produzca un enunciado, momento en el cual será el nuevo hablante en el intercambio comunicativo. Este apego a la estructura lingüística es crucial: en la evaluación de las intenciones de nuestro interlocutor podrían funcionar otro tipo de operaciones cognitivas pero nuestro quehacer como lingüistas se circunscribe a lo que quede

[12] La autora, de manera chusca pero acertada menciona que el *workshop* sobre información nueva/dada llevado a cabo en Urbana, Estados Unidos en 1978, fue renombrado como el "*Mushy Information Workshop*".
[13] Utilizaré versalitas dentro del texto para indicar conceptos nuevos y luego haré referencia a estos conceptos utilizando iniciales en mayúsculas.

atestiguado en la lengua. Por lo anterior, la Familiaridad Asumida parte de que sólo estamos interesados en lo que un individuo pueda hipotetizar acerca de las creencias o conocimiento de otro individuo cuando esas hipótesis se manifiesten en la producción de una estructura lingüística (Prince 1981, 233).

Aunque el trabajo de Prince (1981) da cuenta de lo escurridizo de los conceptos lingüísticos acerca de lo dado (*givenness*), la misma autora no pudo escapar del todo a su propio señalamiento. Su propuesta carece de una orientación explícita sobre cómo analizar las nociones de lo dado y lo nuevo en el lenguaje. Si bien, las construcciones dislocadas, mencionadas por la autora, son de las evidencias lingüísticas más estudiadas al respecto (p. e. Prince (1978) y Lambrecht (1986)), en su propio análisis no parece ser el único mecanismo que ayuda a determinar o asociar esa Familiaridad Asumida, y aunque propone su *taxonomía* (término usado por la propia autora), no así un inventario lingüístico o de factores gramaticales que intervengan en la identificación de sus propias categorías. No obstante, el señalamiento de Prince (1981) deja ver lo extendido del concepto y cómo ha ayudado a describir el comportamiento de algunas formas en determinados contextos. Esto último lo podemos encontrar en la descripción del artículo definido e indefinido en español. Por ejemplo, Andrés Bello (2002, sec. 267), en su gramática de 1848, menciona que al encabezar un sustantivo con un artículo definido, “damos a entender que el objeto es determinado, esto es, consabido de la persona a quien hablamos, la cual, por consiguiente, oyendo el artículo, mira, por decirlo así, *en su mente al objeto* que se le señala” (énfasis propio).

Amado Alonso (1957, 151–52) con respecto a *el/un*, menciona que la diferencia entre estas dos formas radica en la determinación, “que separa a un individuo de entre sus congéneres;

las gramáticas suelen acomodar esta idea (y en muy legítima dirección, como luego veremos) a la de que el objeto nombrado sea o no *consabido del hablante y del oyente*" (énfasis propio).

En la *Nueva Gramática de la Lengua Española* (Real Academia Española y Asociación de Academias de la Lengua Española (RAE y ASALE) 2009, sec. 14.1a) se menciona que el artículo *un* es indeterminado y funciona como primera mención para "presentar *entidades nuevas* en el discurso", mientras que el artículo determinado o definido *el* supone que la entidad es identificable por el oyente; esto se logra o porque el hablante "la haya escrito, porque se haya hablado de ella previamente, porque forme parte del *conocimiento compartido* por ambos interlocutores o por cualquier otro motivo que permita activar su presencia en la *mente del receptor*" (énfasis propio).

Es decir, con Prince (1981) no existía una formalización directa de la Familiaridad Asumida, entendida como la síntesis lingüística de "la información dada/nueva". Al mismo tiempo, la tradición hispánica —por mencionar un ejemplo pertinente para este trabajo— describe el artículo definido y el indefinido como manifestación en la lengua de estas nociones.

En los estudios sobre **definitud** se ha discutido ampliamente sobre los alcances de describir esta propiedad como "conocimiento compartido". Se entiende que una DESCRIPCIÓN DEFINIDA es una frase nominal encabezada por un artículo definido, y en algunos casos, sólo singular (Rivero 1975). Se analiza la definitud de estas descripciones desde dos perspectivas: o asumiendo que la descripción hace referencia a una entidad única, desarrollada desde la perspectiva de UNICIDAD; o a cierta idea de FAMILIARIDAD, que se asemeja mucho a lo discutido por Prince (1981) —y desarrollado por Heim (2008) como teoría de la definitud— en cuanto a que esta propiedad depende del bagaje común (*common ground*) entre los interlocutores (Aguilar-Guevara, Pozas Loyo, y Vázquez-Rojas Maldonado 2019; Pozas

Loyo 2016). Esta discusión en la teoría sobre la definitud sigue en pie, y no es mi objetivo desarrollarla en estas páginas, pero sí el notar que persisten aquellas tendencias que ya señalaba Ellen F. Prince en los años ochenta del siglo pasado. Como corolario, Givón (2001, 437) señala que no es sorprendente observar por qué no se puede dar cuenta coherente de la definitud sin trascender los confines de la proposición aislada de contexto.

Sin embargo, son comprensibles las parcelas en la investigación lingüística para atender la explicación de este fenómeno. La división por niveles de la lengua es una operación analítica que nos permite acercarnos a variados fenómenos lingüísticos y analizarlos en detalle. Hay casos en donde es difícil e inadecuado la separación, pero en otros casos, como el que se presenta, resulta pertinente y necesaria. Las nociones de información nueva/dada han sido utilizadas desde dos niveles de la lengua: por un lado, desde el nivel pragmático, lo que ha permitido explicar ciertos mecanismos y decisiones que estructura el hablante en el intercambio comunicativo, es decir, en el **uso** de la lengua; y, por otro lado, desde el nivel semántico, que está reflejado en el estudio de formas gramaticales particulares y la manera en que están asociadas a ciertas funciones, lo que permite aislarlas y con ello "prescindir" del contexto amplio de enunciación para su descripción. En esta última perspectiva destacan las aproximaciones desde la lógica, con el objetivo de dilucidar errores en inferencias, sin centrarse en la descripción de las posibilidades comunicativas de la lengua (Ariel 1988). El nivel desde donde me interesa observar este fenómeno de información nueva/dada es el pragmático, dejando para otras exploraciones el semántico. Al respecto, un trabajo computacional parecido al que presento, pero desde el nivel semántico y sobre la definitud, es el realizado por Bhatia, Lin, et al. (2014), doy un breve comentario de él en el Capítulo 2.

Sin lugar a duda, otro concepto que se cruza con lo tratado hasta este momento es el de la REFERENCIA. En el nivel semántico, la capacidad de hacer referencia a través de la lengua depende del significado instruccional dado por una forma, por ejemplo, del artículo definido, el cual se describe como la identificación de una entidad a la que se le pueda aplicar una descripción (García Fajardo 2016)[14]. Esto se refiere al concepto de Unicidad de la teoría sobre la definitud, en donde se privilegia la idea de que la definitud apela a la existencia de una sola entidad. Por otro lado, en el nivel pragmático, la referencia implica las suposiciones que el hablante realiza sobre la capacidad del oyente para *identificar* un referente y para establecer la atención sobre él (Abbott 2010; Kibrik 2011; Hopper y Thompson 1984). Aunque esto último podría asociarse a las nociones de Familiaridad descritas desde las teorías de la definitud, creo más pertinente su relación con lo que abarca la Familiaridad Asumida dada por Prince (1981). En mi trabajo sólo me acercaré al nivel pragmático. Trataré con más detalle esta perspectiva de referencia en la §1.4 y §1.10, y cómo impacta mi análisis. Adelanto que entiendo a la referencia como un ejercicio amplio ejecutado a través de la mera presencia de una frase nominal, independientemente de sus características morfosintácticas. En estos términos, *identificar* algo no se refiere a que la entidad exista, que sea específica, sino que, en términos comunicativos, se busca destacar o, como menciono más adelante, a *figurar*, algo para hablar de ello. Este "fragmento conceptual" presentado y construido en la conversación o en el discurso, no necesita ni de la presuposición de existencia ni de que se trate de una entidad; tampoco se deja de lado que las frases nominales que apelan a clases o propiedades

[14] Un aspecto constante en mi exposición es que trato de no usar, en la medida de lo posible, el término de Universo del Discurso. Me parece que este concepto ha sido definido de distintas maneras abarcando fenómenos de variada naturaleza (Lambrecht 1994, 36; García Fajardo 1994). En todo caso, parto de la línea psicológica planteada por Kibrik (2011), debido a que me parece suficiente para el fenómeno que trato además que este autor muestra una manera de modelarlo en computadora.

no refieran: introducen referentes discursivos de los cuales se puede hablar. De esta manera, la referencia la entiendo como un fenómeno discursivo.

Si bien, existen estructuras lingüísticas y factores gramaticales que favorecen ciertas lecturas o nos ayudan a determinar las suposiciones del hablante, no es posible determinar una estructura o forma lingüística **única** para tratar la información nueva/dada. A continuación, retomo de Kibrik (2011) su tratamiento sobre la referencia y la activación desde una perspectiva comprometida con los avances en psicología y neurología. Al mismo tiempo parto de la propuesta de Lambrecht (1994) sobre la estructura de la información desde donde me parece que se explican mejor los conceptos de información nueva/dada.

1.2.2. Estructura de la información

El problema conceptual que se nos presenta es triple; la referencialidad, la definitud y las nociones acerca de lo dado y lo nuevo han dado lugar a una plétora de estudios. Cada campo ha delimitado líneas y programas de investigación independientes. Es por ello por lo que resulta un poco sorprendente la ligereza con la cual se tratan estos temas desde el área computacional. No obstante, cuando se busca sobre información nueva/dada en procesamiento del lenguaje natural (PLN), usualmente los primeros estudios están orientados a la identificación de lo que se conoce como TÓPICO y FOCO. Mostraré, a continuación, que partir en PLN de esta dupla como pivote no es del todo afortunado.

En este trabajo, y con el objetivo de simplificar lo más que lo permita el fenómeno estudiado, establezco como primer eje teórico el trabajo de Knud Lambrecht (1994). Retomo de este

autor la parcela desde donde se estudian estas dos nociones de información nueva e información vieja, nombrada ESTRUCTURA DE LA INFORMACIÓN la cual se entiende como

> El componente de la gramática de la oración en el que las proposiciones como representaciones conceptuales del estado de las cosas se alinean con estructuras lexicogramaticales de acuerdo con los **estados mentales de los interlocutores** quienes usan e interpretan estas estructuras como unidades de información en determinados contextos discursivos [énfasis propio][15].

Esta área se circunscribe a lo que el autor llama PRAGMÁTICA LÉXICO-DISCURSIVA, la cual está enfocada en la relación entre la forma y el significado de determinadas estructuras en el discurso y su convencionalización gramatical. Se me podría señalar que he dicho antes que mi interés no es una forma particular como el artículo definido o la descripción definida, y en efecto, no es mi interés ni la oración ni la frase nominal ni una partícula en específico, sino su interacción en el DISCURSO. En este trabajo utilizo los términos *discurso* y *texto* de manera intercambiable. En ambos casos, hago referencia a la unidad máxima de análisis lingüístico, una secuencia de oraciones en la que existen mecanismos lexicogramaticales que se observan sólo en este nivel. El hablante organiza los componentes de un texto, así como sus formas, a partir de ciertos principios, lo que permite distinguir un texto de un grupo de oraciones aleatorias (Kibrik et al. 2016; Kruijff-Korbayová y Steedman 2003; Van Dijk 1980; Brown y Yule 1983). Aunque la determinación y análisis de los principios de la textualidad es amplia, se suele demarcar a la luz de por lo menos siete criterios: cohesión,

[15] Todas las citas directas que son traducciones mías terminan con una nota a pie de página, en el cual coloco la versión original para su examinación si fuese necesario. “That component of sentence grammar in which propositions as conceptual representations of states of affairs are paired with lexicogrammatical structures in accordance with the mental states of interlocutors who use and interpret these structures as units of information in given discourse contexts” (Lambrecht 1994, 5).

coherencia, informatividad, intencionalidad, aceptabilidad, situacionalidad e intertextualidad (de Beaugrande y Ulrich 1997).

> Whereas a language is the virtual system of available choices that can be made but which have not yet been selected, **the text is a specific organization that has already been realized: an actualized relationship between elements in which certain possible selections have been made and implemented** (Giuffrè 2017, 54 énfasis propio).

Dado este marco conceptual, entiendo que la pragmática presentada por Lambrecht puede corresponder con este sentido de discurso, en el que partir de funciones comunicativas es imperante. De esta manera, la descripción de los estados mentales del referente, a decir, la identificabilidad y la activación, es parte del análisis del texto. Aunque la estructura de la información se concibe como parte de la gramática de la oración, esto no entra en conflicto en entender que una parte de su análisis descansa en analizar el texto en donde aparece la oración en cuestión.

Es usual que en los estudios sobre estructura de la información se tomen como primitivos los conceptos de TÓPICO y FOCO. En términos generales, el Tópico es de lo que se habla en una oración, mientras que el Foco ha tenido una amplia variedad de definiciones en la literatura: foco contrastivo, psicológico, semántico e identificacional, por mencionar algunos[16]. Todas ellas coinciden en que en cierta sección de la oración se trata información relativamente *nueva* (Gutiérrez Bravo 2008; Krifka 2008; Awaihara 2018). Por lo mismo, en distintos trabajos de lingüística computacional que han recurrido a alguna teoría lingüística para desarrollar identificadores de información nueva/dada, el primer intento se concentra en este

[16] Más adelante trataré con especial atención el foco psicológico ya que considero que no forma parte de esta dupla sino de la que conforman la accesibilidad/activación.

par de primitivos discursivos, sin embargo, no es difícil encontrar, o variaciones que se adecuen a las intenciones originales y traten de manera laxa la complejidad de esta dupla, o en un escenario extremo, abandonar la intención de etiquetar tales primitivos y optar por posturas que no parten de propuestas lingüísticas (Endriss y Klabunde 2000; Hajičová, Sgall, y Skoumalová 1993; Mírovský et al. 2013; Ziai y Meurers 2018; Kruijff-Korbayová y Kruijff 2004)[17]. El Tópico y el Foco en general se retoman como formas discursivas que tienen un correlato en la estructura oracional. No es parte de mi objetivo ahondar en estos conceptos o ejemplificarlos, aunque en otro apartado hablaré un poco sobre los identificadores automáticos construidos a partir de ellos en los trabajos de Praga (Hajičová, Sgall, y Skoumalová 1995). No obstante, estos primitivos los retomo debido a que desde la propuesta de Lambrecht (1994), el Tópico y el Foco presuponen la existencia de otra dicotomía más básica: la identificabilidad y la activación de referentes. Estos conceptos apelan a que los interlocutores, al hacer referencia ya sea una entidad o una proposición, *tienen* la representación mental de tal referente[18]. Tales entidades, como referentes, encuentran su formalización en la lengua como categorías argumentales o adjuntas, ya sea en frases nominales, pronombres, y varios tipos de oraciones subordinadas finitas y no-finitas, así como frases adverbiales.

[17] Una de las razones por las que no se han implementado rigurosamente algoritmos que permitan determinar factores gramaticales para la detección de la información nueva/dada se debe a las agendas de investigación en computación: se busca el desarrollo de tecnología, pero poco se atiende a la tecnología *que sirva para* la investigación en lingüística (aunque, véase Baumann et al.(2004) y Kibrik (2011, cap. 14)).

[18] En este sentido, coincido con García Fajardo (2016, 227), quien menciona que "los referentes son las representaciones que construimos de las cosas y de sus estados, no las cosas externas ni un valor de verdad". En la misma línea, Kibrik (2011, 31) menciona que es más razonable entender la referencia como una relación de palabras con cosas en la mente, en vez de con entidades en el mundo; así también encontramos esta misma idea en Givón (2001, 437–39).

La IDENTIFICABILIDAD se entiende como el conocimiento que el hablante tiene de un referente, y que supone que comparte con el oyente. La importante diferencia con respecto a otras definiciones de este tipo de "conocimiento compartido" es que Lambrecht (1994) señala de manera explícita que se trata de la capacidad cognitiva de recuperación de conocimiento en memoria.

El hecho que un hablante enuncie una frase *suponiendo* identificabilidad, significa que apela a la memoria del oyente y a su capacidad de recuperar esa información. Como bien señala el autor, esta noción se relaciona con la definitud (Lambrecht 1994, 79–87) pero no directamente: no existe una correlación uno a uno entre las categorías de definitud y la de identificabilidad en la gramática. Las lenguas pueden carecer de una forma que marque definitud, pero se asume que la capacidad mental de identificar un referente —acotarlo, recuperarlo de memoria y establecer su atención o especificar un referente nuevo— es igual para cualquier hablante de cualquier lengua.

El concepto de identificabilidad que utilizaré en este trabajo está más acotado que el presentado por Lambrecht (1994) y lo separo del de especificidad. Para este autor, la no-identificabilidad de un referente no implica inespecificidad, pero la inespecificidad sí implica no-identificabilidad:

> [o]ne way of describing the specific/non-specific distinction in pragmatic terms is to say that a "specific indefinite NP" is one whose referent is *identifiable* to the speaker but not the addressee, while a "non-specific indefinite NP" is one whose referent neither the speaker nor the addressee can *identify* at the time of utterance (Lambrecht 1994, 80–81; énfasis propio).

Más adelante, Lambrecht mencionará que este concepto de identificabilidad se refiere a una capacidad cognitiva y opera en grados. Demostrará que la identificabilidad no tiene un correlato con la definitud ni con lo específico. Sostengo que, el concepto identificabilidad

cognitiva debe entenderse en este contexto como la suposición del hablante de la capacidad de su interlocutor de encontrar una representación mental de un referente, y no la identificabilidad en términos de especificidad. Más adelante ahondaré en la propuesta de Kibrik (2011), pero por lo pronto es relevante mencionar que su concepto de identificabilidad coincide con el de Lambrecht en la medida en que ambos retoman lo descrito por Chafe (1994, cap. 8). De tal manera, es pertinente repasar lo que supone la identificabilidad para este autor. Para Chafe (1994, 94), la identificabilidad se analiza a partir de tres componentes. Un referente identificable es aquel que (1) se asume estar compartido directa o indirectamente por el oyente; (2) verbalizado en una forma adecuada para su identificación; y (3) contextualmente relevante. Se puede observar que estos componentes se relacionan con lo que Prince (1981) ya había enunciado sobre lo nuevo/dado y las tendencias lingüísticas sobre *shared knowledge* y *saliency,* pero ninguno de estos componentes habla sobre la especificidad o necesidad de existencia de alguna entidad nombrada: seguimos en el terrenos sobre las suposiciones que tiene el hablante sobre el estado mental de los referentes. Ahora bien, el análisis de la identificabilidad de Chafe (1994) procede de manera en que un referente es **no identificable** porque el hablante asume que el oyente no puede ubicar el referente por su memoria, por inferencia, o no se encuentra en su atención o en el momento de la enunciación.

Para evitar el uso ambiguo entre identificable y específico, los utilizaré por separado y entenderé al primero como al concepto identificabilidad cognitiva, en la cual lo relevante es la suposición del hablante sobre la capacidad del oyente de recuperar un referente; de esta manera, el término no identificable apela sólo a aquellos casos en donde no aplica la descripción de la identificabilidad cognitiva. No obstante, hay casos en los que queda

constancia lingüística de que el hablante supone que el oyente no puede recuperar un referente, con lo que se busca introducir uno nuevo referente en el discurso. Por ejemplo, algunas frases nominales encabezadas con indefinido o aquellas con variados modificadores y relativas, las cuales son, además, primeras menciones de un referente en el discurso. Las formas lingüísticas que analizo como no identificables las expongo con más detalle en §1.10. En el caso de identificable, implica entonces una escala más compleja que expongo en las siguientes secciones.

La ACTIVACIÓN, siguiendo lo expuesto por Lambrecht (1994), se asocia con una escala de atención a los referentes. Si la identificabilidad tiene que ver con las suposiciones sobre el acceso a memoria, la activación se refiere a las suposiciones sobre la atención que se tenga de un referente ya identificado. Ilustraré lo anterior con el siguiente fragmento de una nota inventada:

(3) a. **Un hombre** fue arrestado por **policías ministeriales** en **Ensenada**.
b. **Fue** puesto en libertad a **las pocas horas**.

En (3) el referente de *un hombre*, frase nominal encabezada por un artículo indefinido, formaliza la suposición del escritor de la no-identificabilidad de tal entidad por el lector: se asume, por lo tanto, que el referente es nuevo. En su caso, *policías ministeriales* es introducida como una frase nominal escueta, aunque se infiere que fueron unos policías en particular, el periodista no asume ni considera pertinente que el lector identifique a estos referentes. Esto no significa, claro, que no se construya tal referente, pero no se apela a la memoria del lector para identificar a esa entidad en ese momento del discurso. Finalmente, *Ensenada*, como nombre propio, se presenta sin modificador, casi de la misma manera que se presentara un nombre como *Alfonso*, y con ello, evidenciara que supongo que el oyente

sabe a qué *Alfonso* hago referencia. Si se trata de una primera mención, la frase nominal *Ensenada*, en este contexto, evidencia que este referente es identificable. Por lo pronto, con (3a) se observa la aplicación de la identificabilidad, mientras que en (3b) tenemos un ejemplo de activación. El verbo *fue* tiene la forma para tercera persona singular. La ausencia de una frase nominal que encabece esta oración nos indica que el sujeto debe ser una entidad activa, es decir, el hablante supone que el oyente tiene en su atención o foco psicológico a la entidad de quien se habla. En el caso de la frase *las pocas horas*, si bien la presencia del artículo definido podría indicarnos que el hablante asume que el referente es identificable, no es el caso. El referente sólo está siendo *figurado* –se establece como entidad referencial, pero no como suposición de identificabilidad. En este contexto, el artículo definido sólo legitima la frase nominal que forma parte de una frase adverbial. Más adelante mostraré con mayor detalle el análisis que seguí. Por el momento, solo es importante notar que no es necesario recurrir al concepto de definitud para explicar la manera en que se establece la referencia: la identificabilidad y la activación pueden ser suficientes, apelando, claro está, a la evidencia en la estructura gramatical[19].

Como mencioné antes, el hablante puede formalizar a los referentes a partir de distintos recursos de la gramática. En mi trabajo sólo me circunscribiré a analizar los referentes de las frases nominales. Sigo las mismas ideas de Prince (1981), Abbott (2010) y Kibrik (2011) quienes sostienen que las frases nominales son las principales expresiones lingüísticas que

[19] Un análisis de la variedad de usos que puede tener una descripción definida que "escapa" a las perspectivas de la teoría sobre la definitud se puede encontrar en García Fajardo (1994).

se involucran pragmáticamente en la referencia, y, por lo tanto, a través de ellas es posible analizar la identificabilidad y la activación de los referentes.

Se debe notar que he usado el concepto de referente de manera un tanto laxa. Sigo la misma idea de Kibrik (2011, cap. 1) al mencionar que lo que se puede construir como referente es de diversa variedad: entidades, lugares, tiempos, estados y eventos. En el discurso, puede ser un individuo —sin requisito de existencia en el mundo real— un conjunto de individuos, el mejor ejemplar del conjunto, una substancia sin límites distinguibles, un concepto sin correlato preciso en el mundo perceptual; o como entidades "gancho" a las cuales colgar propiedades (Prince 1981, 235)[20]. Aunque, el uso de *las pocas horas* de (3b) es interesante, no es mi objetivo detenerme mucho en estos casos. Los resuelvo como la individuación del tiempo; en este contexto, es irrelevante su identificación como un referente en memoria. Los etiquetaré en el corpus para tomarlos en cuenta en el tratamiento estadístico y computacional sólo con el fin de contrastar y diferenciar contextos de uso.

Existe otra relación planteada en lo dado/nuevo que es necesario mencionar, pero en la que no ahondaré en este trabajo. La estructura de la información en Lambrecht (1994) plantea otra dupla para hablar sobre Tópico y Foco: las presuposiciones y aserciones pragmáticas. En esta área cobra otro sentido el concepto de Información. Para este autor, una PRESUPOSICIÓN PRAGMÁTICA se entiende como el conjunto de proposiciones **evocadas** lexicogramaticalmente las cuales el hablante asume que el oyente ya sabe o está listo para tomar por sentadas en el momento de la enunciación. Coincide de manera parcial con lo que

[20] En la sección 1.4 hablaré con más detalle sobre la referencia y el hecho de que discursivamente cualquier frase nominal es referencial, en 1.10 se encontrará el análisis.

he trazado como información dada. A su vez, la ASERCIÓN PRAGMÁTICA se refiere a la proposición expresada en una oración la cual se espera que el oyente dé por dada como resultado de la enunciación, integrando con ello información nueva. Lambrecht pretende delimitar con estos conceptos aquellos que hacen referencia al "conocimiento" y "representación del mundo" dentro de la terminología lingüística. La Aserción, en estos términos, se refiere a la relación entre conocimiento dado o presupuesto y la proposición que se establece como nueva. Es a esta relación a la que él llama INFORMACIÓN —estructurada en la oración— en donde se agrega o modifica conocimiento a las representaciones mentales existentes que tenga el oyente. En esta misma línea, Lambrecht señala que se ha comparado en la bibliografía de manera desafortunada que "información nueva" equivale a "nuevo constituyente" o "nueva/primera mención", pero esto no es posible sostenerlo desde la aserción y presuposición pragmática: la información, entendida como relaciones entre proposiciones, no se puede determinar en un solo constituyente de la oración.

Lambrecht establece una diferencia entre los estados pragmáticos (el estado de los referentes) y relaciones pragmáticas. Las nociones de información nueva/dada que él plantea las acota a las relaciones pragmáticas entre proposiciones, entre *cosas* y el *estado de las cosas* (*state of affairs*) (Lambrecht 1994, 50). Por otro lado, es en las suposiciones de los estados mentales de los referentes en donde sí se puede apuntar a constituyentes de las oraciones. Para este autor, esto no es nombrado "información nueva/dada" sino que sólo es una diferencia entre los estados asumidos de "la designata de varios constituyentes de la oración en la mente del oyente al momento de la enunciación" (Lambrecht 1994, 49). El análisis de las relaciones entre proposiciones que menciona Lambrecht llegará a su estudio de Tópico y Foco. En mi trabajo, como ya mencioné, no partiré de estos dos primitivos. Mi intención es concentrarme

en los constituyentes que son expresiones de referentes, en donde las nociones de información nueva/dada tienen otro matiz, el de los estados mentales de los referentes.

Resumo lo expuesto de la siguiente manera. Las nociones de información nueva/dada cubren de manera extensa distintos trabajos, tanto lingüísticos como de procesamiento de lenguaje natural. El lugar común para atender estas nociones se concentra en la estructura de la información. Sin embargo, desde esta área de la pragmática, dejando de lado las implicaciones de Tópico y Foco, existen por lo menos cuatro tipos de información nueva y vieja —que resumo en la Tabla 1.

Tabla 1. Variaciones con respecto a la información nueva/dada

Concepto	**Descripción**	**Capacidades cognitivas**[21]	**Lo nuevo**	**Lo dado**
Identificabilidad de referente	Suposición del hablante sobre el **conocimiento** de un referente o marco conceptual por el oyente.	Memoria 1	No identificable	Identificable (en memoria a largo plazo).
Accesibilidad de referente		Memoria 2	Referente establecido por inferencia o perceptible.	Conocimiento necesario para la inferencia o por percepción.
Activación de referente	Suposición del hablante sobre la **atención** de un referente por el oyente.	Atención	Referente **no** se encuentra activo (en el foco psicológico).	Referente se encuentra activo (en el foco psicológico).
Aserción y presuposición pragmática	Es necesaria la estructuración gramatical por parte del hablante.	(Las capacidades anteriores)	Información resultado de la aserción pragmática.	Presuposición pragmática

[21] La *memoria 1* hace referencia a lo que en psicología se llama Memoria a largo plazo (Ariel 1988). Para objeto del resumen que presento en la tabla, con *memoria 2* me refiero a tanto a la memoria a largo plazo como a la capacidad de obtener la información del contexto inmediato o de inferencias. Más adelante detallaré la naturaleza de estos conceptos.

La identificabilidad, accesibilidad y activación apelan sólo a los referentes discursivos. Esto se debe a las asunciones del marco teórico que desarrollo en el presente trabajo sobre la preponderancia de las entidades por encima de las proposiciones. La primera de estas nociones se refiere a la diferencia entre utilizar una frase nominal que instruya sobre el supuesto de no-identificación del referente, contrapuesta con frases nominales que suponen conocimiento del oyente. Si el referente es identificable, podrá ser debido al acceso a memoria, a la relación conceptual o inferencia, a la percepción del oyente, o como se muestra en el tercer caso de la Tabla 1, se refiere a los referentes activos o en el foco psicológico del oyente, los cuales suponen "lo dado", aunque en este último caso, ya no como información accesible por memoria sino por el momento del discurso. Finalmente, el cuarto grupo es el inicio de la instrumentalización lingüística de estas nociones psicológicas para el manejo complejo de la oración, que incluye no sólo frases nominales y referentes, sino al juego entre proposiciones y al concepto de información de Lambrecht (1994, 43), que de acuerdo con él, implica un análisis fino de las diversas relaciones entre tópico y foco[22].

En este trabajo trataré de utilizar en lo mínimo el concepto de *información nueva/dada* así como el de Conocimiento Compartido o Familiaridad Asumida; sólo utilizaré identificabilidad, accesibilidad y activación, nombrados por Lambrecht (1994) como "estados mentales de los referentes". Como la intención de este trabajo ha sido sintetizar los conceptos de información nueva/dada en lingüística, he optado por nombrar el grupo como

[22] El primer boceto de esta investigación partía de buscar la manera de etiquetar tópico y foco. Sin embargo, esto resultó demasiado ambicioso para el estado del arte sobre PLN en español. No contamos con los recursos tecnológicos para desarrollar un etiquetador automático de estas propiedades, por lo menos entendidas desde Lambrecht (1994). Lo que noté es que antes tenía que construirse un predictor de identificabilidad/activación, partiendo de que estas nociones son básicas.

Estados Informativos. Se entenderá por lo tanto que un referente es nuevo cuando la estructura gramatical permite determinar que el hablante supone que el oyente no puede identificar —recuperar de memoria— a tal referente en un discurso dado; mientras que lo contrario aplica para los referentes dados o conocidos, es decir, que el hablante supone que el oyente sí lo puede identificar; nótese que utilizo el concepto de identificabilidad cognitiva, en donde es irrelevante el carácter específico del referente. El conocimiento necesario para la identificación las supone el hablante de cuatro áreas principalmente: conocimiento general, inferido, inmediato y textual. En términos de Lambrecht (1994, 109), estos conocimientos se refieren al tipo de accesibilidad, llamados respectivamente como: (i) referentes inactivos (pero recuperables de memoria), (ii) accesibles por inferencia, (iii) por contexto inmediato y (iv) por el texto[23]. Finalmente, un quinto subtipo de suposición de identificabilidad no apela a memoria o inferencia sino a la atención que el oyente tenga de un referente: (v) el referente se encuentra activo. Estas cinco categorías forman parte de la gradación de la activación en términos de Lambrecht (1994) y son parte del núcleo de mi análisis. No obstante, y a manera de preámbulo de la exposición que sigue, en mi trabajo distinguiré las categorías II y III, como accesibles; la I y la IV como inactivos, y sólo para la quinta utilizaré el término activo.

Antes de proseguir, es necesario realizar un par de precisiones. La primera, tiene que ver con la perspectiva desde donde se realiza el análisis. Como ya he mencionado, parto de la posición del hablante para analizar las suposiciones que se estructuran en la lengua. Pero esta posición no es la única en este tipo de estudios. Se puede partir desde la posición del oyente,

[23] Nótese que para Hawkins (1978, 107) estas maneras de acceder a los referentes se clasifican en los usos: anafórico e inmediato (activación alta); situaciones amplias y anáforas asociativas (accesible por situación e inferencia). Aunque es evidente el paralelismo, seguiré utilizando los conceptos de Lambrecht (1994) que después apuntalo con lo trabajado por Kibrik (2011).

en donde se asume que es él quien resuelve la referencia y calcula la correferencia. Esta perspectiva es la que se toma en general en los análisis en PLN. El objetivo en esos estudios es resolver la referencia de manera automática (Recasens 2010; Sukthanker et al. 2018; Kehler y Rohde 2018). Al igual que Kibrik (2011, 48), tomo la perspectiva orientada al hablante/escritor —términos que utilizo de manera intercambiable a lo largo de mi trabajo, al igual que oyente/lector— ya que la manera más eficiente de explicar la estructuración del discurso es a partir de reconstruir y modelar procesos cognitivos del hablante que dan lugar a fenómenos discursivos observables; los cuales son susceptibles de ser analizados por un lingüista.

La segunda precisión tiene que ver con estos fenómenos que observaremos y analizaremos. La propuesta de Kibrik (2011) descansa en la idea de ELECCIÓN REFERENCIAL (*referential choice*), la cual coloca la atención en el hablante y su responsabilidad de dar forma al discurso. Para ello, el hablante tiene a su disposición un inventario de recursos referenciales de dónde elegir, recursos regidos por su lengua, pero orientados por las suposiciones de identificabilidad y activación de los referentes. La evidencia tipológica, de acuerdo con lo mostrado por Kibrik (2011, caps. 3–9), sugiere que estas posibilidades referenciales colapsan en dos grupos: dispositivos referenciales plenos y reducidos. Los primeros se estructuran en frases nominales de distintas dimensiones y características morfosintácticas, y las segundas tienden a ser pronombres, marcas en el verbo y ceros morfológicos. Es debido a esto, y antes de continuar con la exposición de los Estados Informativos, que es importante establecer lo que entenderé por frase nominal en mi análisis, a lo cual le dedicaré la siguiente sección.

1.3 Frase nominal en español

Para proceder en el análisis que dé cuenta de cómo el hablante formaliza la identificabilidad y la activación de los referentes en su discurso, delimitaré lo que entenderé por frase nominal en español. Me apoyaré en dos gramáticas para este fin: la *Gramática Descriptiva de la Lengua Española* en donde Rigau (1999), Picallo (1999) y Brucart (1999) desarrollan descripciones del sintagma nominal, y la gramática de la RAE y ASALE (2009). De igual manera, me guio en la exposición y descripción de la frase nominal que realiza Recio Diego (2015). A parte de que la noción de frase nominal es importante para el modelo de Kibrik (2011) y Lambrecht (1994), es importante recalcar que busco, ante todo, una noción manejable a nivel de corpus y procesable con tecnologías del lenguaje.

Una FRASE NOMINAL es un constituyente que tiene como núcleo un nominal, nombre propio o pronombre. Este núcleo establece la concordancia de género y número, la cual demanda sobre otras unidades, como son los modificadores y determinantes. El nominal también llega a influir en la selección de la forma de las palabras que introducen oraciones subordinadas de relativo como los pronombres *que, quien* o *cual,* adverbios como *donde, como, cuando,* y adjetivos como *cuyo*. La frase nominal es la que formaliza roles sintácticos y semánticos en la oración, y por extensión, es la manera en que los nominales se manifiestan en el discurso; no es pues el nominal el que adquiere un rol gramatical sino la frase nominal. “However, within the noun phrase, a noun is typically the syntactic and semantic head, defining the type of entity involved. All other elements in the noun phrase are modifiers of that head noun” (Givón 2001, 59).

Aunque distintas teorías analizan la preposición y el determinante como núcleos de frase, en mi trabajo no las abordaré. Esto debido a que mi objetivo se centra en los constituyentes que funcionan para establecer un referente discursivo (Rijkhoff 2004, 19).

Para ejemplificar lo anterior, obsérvese el análisis sugerido para los casos de (4):

(4) a. **El** cuerp**o** calcinad**o** de un hombre fue encontrado al interior de un vehículo (…).
COPENOR-147SO[24]

b. Familiares y amigos en redes sociales piden la ayuda de la ciudadanía para localizar a Jorge. **Él** vestía una camisa verde, con pantalón beige y zapatos de color negro (…).
COPENOR-178CH

c. El acto es a las 13:30 horas, en la Ciudad Deportiva, en la cancha de atletismo, **la que** está enfrente del estadio 'Emilio Ibarra'; (…).
COPENOR-108SN

En (4a) tenemos un nominal *cuerpo* que es núcleo de la frase *el cuerpo calcinado de un hombre*. Podemos distinguir la concordancia que ejerce sobre el determinante *el* debido a sus propiedades de número y género: masculino singular. Esto mismo se observa en el modificador *calcinado* que acompaña al nominal; además de la frase preposicional *de un hombre* que modifica el núcleo y sus dependientes. Toda esta FN formaliza el argumento que asume el rol de sujeto sintáctico en el predicado de *fue encontrado*.

Así mismo, en (4b), tenemos el pronombre de tercera persona singular *él* que resuelve su referencia por la primera mención *Jorge* que aparece en la primera oración. Este pronombre

[24] Las notas periodísticas capturadas en COPENOR cubren distintos eventos trágicos. El manejo de los fragmentos de estos textos no busca faltar al respeto de las víctimas retratadas: en todo momento debe interpretarse mi respeto hacia los casos reales que se traten y se pide prudencia y sensibilidad al lector.

es núcleo de una frase nominal, pero en este caso no se encuentra ningún modificador. En algunos casos tenemos frases nominales como en (4c) *la que está enfrente del estadio 'Emilio Ibarra'* en donde no tenemos un núcleo explícito, sin embargo, las analizo como frase nominal por el hecho de tener un modificador y determinante que es coherente con la flexión de un sustantivo; es decir, se espera que una frase nominal rija los modificadores de su núcleo sólo en género y número; si buscamos alguna palabra para "llenar" esa posición vacía, podemos notar que en efecto, sólo cambian esas propiedades; si colocamos algún otro tipo de palabra, la oración estaría mal formada como se observa en el siguiente contraste:

(5) La (cancha/*canchas/*campo) que está enfrente del estadio.

Con lo anterior, entenderé frase nominal como un constituyente que establece dependencias internas de concordancia de género y número con un núcleo que puede ser un nombre común, un nombre propio, un pronombre o incluso un cero (núcleo elíptico, Rigau (1999)) en algunos casos. Además de esto, su núcleo puede ser modificado a la izquierda y a la derecha. El grupo izquierdo está conformado por lo general, por los determinantes, mientras que el derecho por los modificadores que pueden abarcar adjetivos, participios y frases preposicionales con distintas funciones.

Uno de los problemas en la identificación automática de la frase nominal es determinar algún límite en su extensión. Teóricamente pueden ser modificadas por *n* número de unidades que a su vez pueden contener elementos subordinados. Dejando de lado esta extensión imaginaria, analizaré los límites del constituyente cuando éste pueda ser sustituido completamente por pronombre. Por ejemplo:

(6) a. **[El Agente del Ministerio Público de la Unidad de Investigación de Saucillo adscrita a la Fiscalía de Distrito Zona Centro]^{i}**, obtuvo **[una sentencia condenatoria]^{j}** en el Juicio Oral número 15/2018 (…)

COPENOR-001CH

b. **[Él]^{i}** obtuvo **[eso]^{j}** en el Juicio Oral Número 15/2018.

En este caso, lo que en (6a) se encuentra dentro de los corchetes del superíndice *i* puede ser sustituido por el pronombre *él* como aparece en (6b); de igual manera, lo que en (6a) aparece dentro de los corchetes del superíndice *j* podría ser sustituido por el pronombre demostrativo *eso*.

En la tradición hispánica existen amplios tratados sobre la descripción de la frase nominal en español, por ejemplo, Company Company (2006a; 2006b). No es mi propósito repasar estas características ya que mi trabajo no se centra en la descripción de la FN en español sino en su capacidad de establecer referencia a entidades discursivas. Aunque existen tendencias observables por los lingüistas sobre cuáles estructuras y distribuciones dentro de la frase nominal favorecen la capacidad referencial, ninguna de ellas parece ser exclusiva de esta función. Prácticamente cualquier combinación de unidades léxicas y frases dentro de una frase nominal puede tener esa función.

Para dejar clara la diferencia entre lo que entenderé en este trabajo como FN y lo que en la teoría funcional abarca la función de FN, obsérvense algunos ejemplos que tomo de Recio Diego (2015, 49) en (7) a continuación, además de uno extra que he agregado en (7f) a manera de contraste. En estos casos, entenderé que los fragmentos en negritas de (7a-b) son frases nominales por tener un núcleo nominal: *pan* y *gato*; el caso de (7d) representa a las frases nominales con núcleo cero pero que obedecen reglas de concordancia que ya mencioné antes.

(7). a. Quiero **pan.**

b. Tienen **un precioso gato persa de dos años.**

c. Decías **que no vendrían tan pronto.**

d. Dame **la que está en el armario grande.**

e. Preguntó **si lo conocíamos.**

f. **Que tu hayas venido** provocó **que se pusieran de acuerdo.**

Por otro lado, aunque cumple función argumental, no analizaré como FN los casos de (7c), debido a que no es una frase nominal sino una oración subordinada sustantiva de complemento directo. De la misma manera, tampoco etiquetaré los constituyentes que funcionan como argumentos en (7e) y (7f) debido a que no tenemos FN sino oraciones completivas introducidas con "que" y "si"[25].

1.4 Dispositivos referenciales

Una vez teniendo claro que por dispositivo referencial pleno me referiré a la frase nominal en español, y que son estos dispositivos el centro de mi investigación, procederé a detallar el marco que guiará mi análisis de los Estados Informativos. En este apartado sigo de cerca lo expuesto en *Reference in Discourse* por Andrej A. Kibrik (2011). Este autor toma como parámetro para el análisis de dispositivos referenciales la clasificación de la referencia de Elena V. Paducheva (1985 *apud* Kibrik 2011, 32). La reproduzco a continuación con ejemplos y traducción propia:

1. Específico

    a. Definido: *yo, el libro, este libro, mi libro, el libro que me diste, Yusnai*

[25] El único caso excepcional que tomaré en mi corpus es el análisis de verbos en infinitivo como núcleo de frase nominal.

b. Indefinido: *alguien, un extranjero, unas personas*

2. Inespecífico

    a. Existencial: *Pásame <u>cualquier cubeta</u>.*

    b. Universal: *<u>Todos los niños</u> adoran el helado.*

    c. Atributivo: *<u>El gato más bonito</u> ganará el concurso.*

    d. Genérico: *<u>El perro</u> es el compañero de la humanidad.*

3. Predicativo: *Mi mejor amiga es <u>violinista</u>.*

4. Autónimo: *Mi sobrino fue nombrado <u>Matías</u>.*

En el caso de las frases definidas, también se incluyen pronombres personales, frases nominales con determinantes posesivos y nombres propios, por lo que esta es la mayor divergencia en cuanto a las DESCRIPCIONES DEFINIDAS.

Un primer problema para realizar análisis desde esta clasificación sobre la referencia lo señala Kibrik (2011): las frases nominales inespecíficas, predicativas y autónimas no carecen de referente, sólo no es el mismo que en los casos específicos.

> … en algunos sistemas terminológicos (incluyendo la terminología original en ruso de Paducheva), el uso específico de una frase nominal es llamado "referencial" o "referenciador", provocando con ello que todos los otros estados de la clasificación anterior sean no referenciales. Esto probablemente sugiere que tales frases nominales no refieren o no tienen referentes. No obstante, sugiero que ellas sí refieren y sí tienen referentes, pero tales referentes no son específicos[26].

[26] "… in some terminological systems (including Paducheva's original Russian-language terminology) the specific use of an NP is called 'referential' o 'referring', thus rendering all other statuses in (2.2) no referential. This probably suggests that the corresponding NPs do not refer, or do not have referents. I rather suggest that they do refer and do have referents, but those referents are not specific" (Kibrik 2011, 33).

La anterior clasificación supone una descripción y análisis de la definitud y la especificidad que ha tenido distintas soluciones en la investigación lingüística (Ariel 1988; Alcina Caudet 1999). Ya señalé en §1.2.2 que Lambrecht (1994) y Chafe (1994) coinciden en que la identificabilidad cognitiva no tiene por qué coincidir con una marca gramatical de definitud, y no tiene una relación uno a uno con la especificidad. Me parece que es un problema conceptual y de tradiciones en el área —por ejemplo, el equiparar identificabilidad cognitiva con específico, o la necesidad de definir conceptos semánticos con conceptos lógico-filosóficos, o la incapacidad de asegurar que *the murderer of Smith* es referencial, cuando lo es (Kibrik 2011, 33). No busco resolver estas aparentes asimetrías en mi trabajo, por lo que no partiré de esta clasificación de la referencia.

En el contexto de mi análisis, **cualquier frase nominal puede establecer un referente discursivo**. De tal manera, el estatus referencial que podría diferenciarse dada una clasificación como la anterior no tiene fundamento discursivo. Uno puede introducir un referente discursivo como *cualquier cubeta*, *la persona que gane la carrera* o *color rojo* y seguir predicando sobre ella en las siguientes oraciones independientemente de su existencia real o si tal concepto es una entidad.

> No importa que *los unicornios*, *los fantasmas* o *el primer bebé que nazca en el año 2100* sean o no entidades que existan en el mundo real. Lo que importa, desde el punto de vista comunicativo, es el hecho de que los hablantes pueden referirse a ellos y que su interlocutor comprenda esta referencia. De este modo, los conceptos lógicos de verdad y existencia carecen de interés (Alcina Caudet 1999, 72).

Como se ha visto hasta este momento, he utilizado los términos *referente* y *entidad* de manera casi intercambiable. Por *referente* aludo a *referente discursivo* y por *entidad* aludo, no a la cosa que pueda existir en la realidad, sino a su representación mental. Sin embargo, para propósitos del análisis que propongo, los dos conceptos se refieren a *algo* de lo cual se *habla*,

por lo que suspendo la diferencia "referencial" que pudiera haber entre la propiedad que implica la frase *color rojo* en *la madera de color rojo*, el atributivo como *la persona que gane la carrera*, o el genérico como *los caballos* en *los caballos tienen cuatro patas*. De todos estos **referentes discursivos** se puede seguir predicando en oraciones posteriores. En §1.10 ahondo más sobre este análisis.

Por lo pronto, acotando estos términos y siguiendo la exposición del marco teórico de Kibrik (2011), se puede observar que en el discurso es normal que un mismo **referente** tenga distintas **menciones**. Cada vez que un referente es introducido al discurso a partir de una primera mención, este se presenta en el REGISTRO DISCURSIVO de ese diálogo o texto particular (Lambrecht 1994). En algunos casos, tal referente no vuelve a ser mencionado, en otros, se vuelve constante su presencia. Considérese el siguiente fragmento:

(8) El Banco Inmobiliario de México (BIM) abrió este 23 de mayo de 2019 un centro financiero en Tijuana, Baja California. Con ello, la institución financiera tendrá presencia física en esta ciudad fronteriza como parte de una estrategia de consolidación y crecimiento. Leonardo Arana de la Garza, director general de BIM, comentó a ZETA que actualmente la cartera del banco en Tijuana asciende a alrededor de 150 millones de pesos (…).

COPENOR-002BC

En (8) tenemos varios referentes mencionados. En particular, aquellas frases subrayadas son menciones de un mismo referente. En este sentido, se dirá que, por ejemplo, la segunda mención de *BIM* —la que actúa como complemento de la frase preposicional que modifica *director general*— tiene una relación **anafórica** con un referente ya establecido en el discurso, enunciado en la primera frase nominal del fragmento, en donde aparece *BIM* como aposición de *el Banco Inmobiliario de México*. Con el análisis anterior también busco dejar

claro que no entenderé anáfora sólo como un fenómeno sintáctico, sino desde una visión amplia de seguimiento de referencia a partir de distintos dispositivos referenciales, entre los cuales están frases nominales plenas.

Para Kibrik (2011), el hablante tiene la posibilidad de escoger de un repertorio de DISPOSITIVOS REFERENCIALES para introducir y mencionar entidades en el discurso. Las expresiones referenciales que se formalizan en frases nominales se agrupan en dos conjuntos de acuerdo con el tipo de núcleo:

1. Dispositivos referenciales plenos (DRP)
    a. Nombres propios.
    b. Nombres comunes.
2. Dispositivos referenciales reducidos (DRR)
    a. Pronombres.
    b. Marca cero.

La decisión de un hablante de optar por un DRR depende principalmente del estado de activación que tenga tal referente en el discurso. Por lo general, la marca cero corresponderá a un referente que se encuentra en el grado más alto de activación[27]. Aunque de manera usual la posición preferente para la frase nominal con uso referencial es la argumental, me apoyaré en el análisis de otros tres tipos de dispositivos referenciales: el determinante posesivo, los

[27] Con respecto a esta terminología, marca cero podría sonar contradictorio pero sigo la misma idea que Kibrik (2011) al mencionar que es una decisión práctica: al "marcar" el cero señalamos la importancia de esa ausencia para construir significado —y en este caso, establecer referencia. En este trabajo, y si llega a ser necesario, sólo marcaré el cero morfológico en español para la tercera persona singular. Los sujetos tácitos los entenderé como marcados en el verbo por una forma pronominal que se ha afijado a él —a menos que sean verbos en infinitivo. Tal visión es coherente con la perspectiva tipológica desarrollada por Kibrik (2011).

complementos de frases prepositiva, frases nominales apositivas y pronombres demostrativos. Los ejemplifico a continuación.

(9) a. **Mi** perro juega con la pelota.
b. La crema de **cacahuate** es nutritiva.
c. El Banco Inmobiliario de México **(BIM)** abrió este 23 de mayo de 2019 un centro financiero en Tijuana (…)
d. El **mío** juega con la caja.

En (9a) se analiza que el determinante posesivo de primera persona singular determina la referencia de *perro* y su relación se esperaría "accesible", en este caso, con una entidad presente en el acto comunicativo: la primera persona. En (9b) la frase prepositiva *de cacahuate* restringe la referencia de *la crema*. Esto, dependiendo del contexto, podría ayudar a relacionar ese referente con algún otro referente mencionado antes en el discurso. Si ese fuese el caso, el referente estaría formalizando la suposición del hablante de establecer esta relación. En (9c) tenemos una secuencia de dos FN, en donde *BIM* hace referencia a *El Banco Inmobiliario de México*. Debido a que *BIM* guarda una relación con un referente en el Registro Discursivo, se supone identificable, pero, además, se supone que es posible inferir la relación de estos dos nombres. Finalmente, en (9d), el pronombre posesivo *mío* podría funcionar como respuesta y seguimiento a (9a) en donde sustituye a *perro*. De esta manera, la FN *el mío* guarda una relación con el referente *el perro* pero no por ser un mismo referente sino por apelar a la capacidad de inferir la relación entre las frases. En este último caso, de nuevo, se trataría de un tipo de accesibilidad.

Otra característica de mi trabajo que se desprende de analizar texto escrito es que no tomo en consideración la prominencia prosódica para el análisis. Esto da lugar a una de las variaciones más importantes con respecto a los trabajos sobre estructura de la información, como los de

Lambrecht (1994) y Schwarzschild (1999), aunque, de acuerdo con lo establecido por Kibrik (2011), la prominencia afecta la estructura de la información a nivel de tópico y foco, pero no la referencia discursiva.

Como ya he señalado, escoger uno u otro dispositivo depende de dos restricciones de fundamento cognitivo: las suposiciones de identificabilidad y de activación del referente. En el caso de los participantes del acto de habla, la primera y segunda persona se encuentran permanentemente activos en un intercambio comunicativo por lo que los pronombres personales tienen resuelta su referencia por defecto.

De las opciones de dispositivos referenciales que un hablante puede escoger, se supondría que los DRR son los más oscuros debido a la variabilidad de la posible referencia, mientras que las frases nominales léxicamente plenas nos muestran información que es más transparente para resolver la referencia, pero esto no es así. Los DRR resultan ser mucho más eficientes al apelar, precisamente, a la capacidad cognitiva de activar representaciones mentales. En una oración como *Un perro lo mordió* la frase *un perro* parecería ser mucho más clara que el pronombre *lo* con respecto a establecer una referencia, pero esta visión es una ilusión provocada por atomizar la construcción: en el contexto es mucho más claro para los interlocutores establecer a qué se refiere el pronombre *lo*. La propuesta de Kibrik (2011) busca predecir que las frases nominales léxicamente plenas aparecen en casos en donde el referente no se haya mencionado antes, y las frases nominales reducidas o DRR, aparecen cuando exista un referente activo. Aunque esto, como se verá, no es categórico: los DRP muestran un comportamiento complejo que no se reduce a la falta de activación de un referente.

Las opciones pertinentes de dispositivos referenciales en un momento dado del discurso para determinado referente dependen de su grado de activación en la MEMORIA DE TRABAJO (MT) (*working memory*). Este concepto no se trata de una metáfora, sostiene Kibrik (2011), sino de una capacidad estudiada por psicólogos y psicolingüistas a partir de un módulo presente en el funcionamiento del cerebro. Esta capacidad parece tener un correlato y función en las maneras en que se presentan distintos dispositivos referenciales en el discurso. Como menciona Chafe (1994, 72), "es en última instancia imposible entender la distinción entre información nueva y dada sin tomar en consideración la consciencia"[28] por lo que, siguiendo el señalamiento de Kibrik (2011), es necesario poner atención al uso de conceptos psicológicos para tratar estos temas en lingüística; atención que pongo en la siguiente sección.

1.5 Base cognitiva de los estados mentales

Conforme a lo que he planteado en las secciones anteriores podemos sostener que la estructura gramatical codifica las suposiciones del hablante sobre conocimiento acumulado en la MEMORIA A LARGO PLAZO (MLP) del oyente (Chafe 1994); a esta primera suposición corresponde la (no)identificabilidad del referente, entendiendo identificabilidad como la suposición de que el oyente puede recuperar cierta representación mental de algún referente. Si resulta estar identificado, esto podría deberse a diversas razones, pero en todas se supone cierta información anterior. Ahora, un referente puede no estar identificado, pero su mera

[28] "It is ultimately impossible to understand the distinction between given and new information without taking consciousness into account" (Chafe 1994, 72)

mención en el discurso lo activa, es decir, lo plantea tanto en la MT como en el Registro Discursivo. Esta diferencia es importante en lo que respecta a mi trabajo y que no está presente en Kibrik (2011). Para fines cognitivos, el Registro Discursivo y la MLP son lo mismo: la base cognitiva del registro se encuentra en la MLP. Sin embargo, pienso importante, para los fines del análisis computacional que realizaré, el que se establezca la diferencia entre Registro Discursivo, MT y MLP. El primero lo entenderé, de manera heurística, como la memoria establecida en el discurso corriente, una lista de conjuntos de FFNN que hacen referencia al mismo referente[29]; mientras que con la MT me referiré al espacio psicológico en donde se coloca al referente activado por la mención, y desde donde se supone tal activación para decidir si una consecuente mención debe ser formalizada como un DRP o un DRR. De esta manera, sólo me referiré a la MLP para la recuperación de información supuesta por parte del hablante en la memoria del oyente.

Veamos con más detalle la cuestión de las menciones en el discurso. Guiándome en la Figura 1, la mención de una entidad en un momento t_2 coloca la atención sobre tal referente y al traerse a la MT es más probable que se mencione en lo que siga del discurso, además de que ciertos DRR son probables. Como ya lo señalé, esto parte de las suposiciones del hablante de "traer a la mente" del oyente a un referente o apelar a su memoria.

[29] Es importante señalar que la MLP no es un terreno resuelto en las áreas cognitivas. De acuerdo con Nyberg (1996) se pueden clasificar hasta 23 tipos de funciones, apelando a subsistemas de MLP. Reconozco la simplificación realizada aquí, que no es detallada en Kibrik (2011), la cual resulta operativa, ya no a nivel computacional, sino para efectos del trabajo en lingüística.

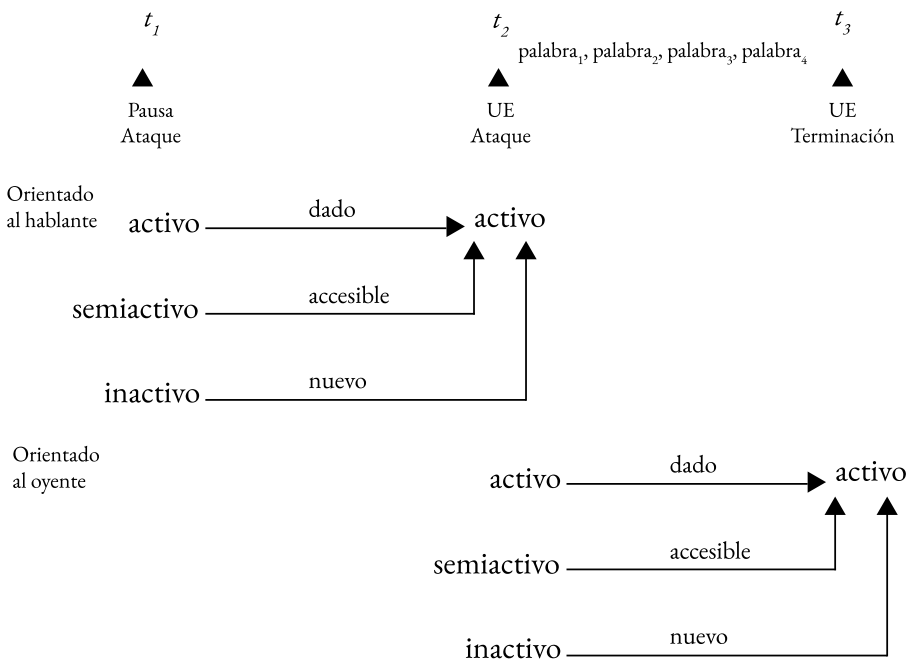


Figura 1. Sincronización de activación de costos en relación con el hablante y el oyente[30]

En el fondo, estas ideas tienen un correlato muy explorado en psicología. Primero que nada, el hecho de que en un intercambio comunicativo exista la disposición del hablante de construir un modelo sobre lo que sabe o está en la atención del oyente se ha estudiado como Teoría de la Mente. Esta propuesta parte de la empatía que tiene un hablante al elaborar sus construcciones con su interlocutor (Tomasello 2008). Tal como lo plasma Kibrik (2011, 57), el hablante es consciente de que, en algún momento dado, su oyente no piensa en lo mismo que él. Lo que Chafe (1994) nos muestra en la Figura 1 es esta relación de coordinación para lograr la comunicación, llamada *sincronización de la activación*. Lo que señala el esquema es que, en la construcción del enunciado, representado en el momento t_1, el hablante puede

[30] Tomado de Chafe (1994, 74); traducción propia.

partir de distintas suposiciones de activación[31]; en todo caso, en t_2, al enunciar el referente, este es activado y el oyente ejecuta su propia resolución, presentada en t_3, a partir de sus suposiciones sobre tal mención y el porqué es presentada de una determinada manera; esta resolución trae a la MT al referente pertinente: lo activa.

Si bien, esta coordinación sería ideal en un intercambio comunicativo, no siempre es así. Podemos clasificar el resultado plasmado en el intercambio de tres maneras distintas partiendo del hablante: ejecución *egocéntrica*, *óptima* y *sobreproductiva*. En el caso de la egocéntrica, el hablante falla en la suposición y el referente no puede ser recuperado, lo que desencadena, o en una mala elección del referente por parte del oyente, o sencillamente en la pregunta ¿quién/qué? El caso óptimo sería el efecto reflejo del proceso por parte de los interlocutores (el caso de la Figura 1) en donde la suposición de identificabilidad y activación del referente fue exitosa y se codifica en un dispositivo referencial que guía al oyente a una representación mental similar a la del hablante o en todo caso, adecuada. Finalmente, la sobreproductiva es aquella en donde el dispositivo referencial usado es, por lo general, un DR PLENO en un caso en donde el oyente esperaría un DR REDUCIDO. Aunque el análisis de mi corpus podría ayudar a la identificación de estos perfiles a partir de una cuantificación de las FFNN plenas y reducidas, partiré del supuesto de que el escritor de las notas periodísticas busca una ejecución óptima.

[31] Lo interesante de este modelo es que estas suposiciones son tanto del oyente como de la propia capacidad del hablante: se supone que el referente también se activa en la MT del hablante independientemente de las suposiciones de identificabilidad y activación en el oyente. Esto supone otro planteamiento del problema que no sigo en este trabajo.

La base cognitiva de Kibrik (2011) descansa, como se ha podido observar, en tres conceptos de fundamento psicológico y neurocientífico: MLP, MT y Atención. Resumo su perspectiva en la Tabla 2. En el caso de la ATENCIÓN, Kibrik (2011), a partir de varios autores en psicología, la considera como una propiedad emergente del sistema cognitivo que discrimina la información que es percibida. De entre los distintos subtipos de atención que se han trabajado en psicología, es la atención *selectiva*, *voluntaria* y *orientada a objetos* la pertinente para explicar fenómenos de referencia (Levelt 1989).

Tabla 2. Síntesis de las capacidades cognitivas inmiscuidas en la referencia

Término usado:	**Capacidad cognitiva inmiscuida:**
Identificabilidad/accesibilidad	Memoria de largo plazo
Activación	Memoria de trabajo
Mención	Atención

La Memoria a Largo Plazo es requerida para la identificabilidad o en algunos tipos de accesibilidad; en cuanto a la MEMORIA DE TRABAJO, Kibrik (2011, cap. 10) ofrece una amplia base de estudios psicológicos, entre los que destacaré la mención de Woodward et al. (2006, 317), quien a su vez sintetiza lo planteado por el psicólogo británico Alan Baddeley (1992):

> La memoria de trabajo permite a los individuos mantener y manipular una limitada cantidad de información en un estado activo por un pequeño periodo de tiempo. Puede operar en una variedad de representaciones cognitivas, tales como gustos, sonidos, imágenes, fonemas, conceptos, locaciones, patrones y colores[32].

Algo que me parece importante de esta mención es que la MT implica mucho más que lenguaje en su operación. Lo que Kibrik (2011) plantea, y en lo que también coincido, es una

[32] "Working memory allows individuals to maintain and manipulate a limited amount of information in an active state for a brief period of time. It can operate on a variety of cognitive representations, such as tastes, sounds, images, phonemes, concepts, locations, patterns, and colors" (Woodward et al. 2006, 317).

base cognitiva para el acto referencial pero no exclusiva a ella ni al lenguaje. Esta MT pertenece a una dicotomía que también se ha nombrado "memoria de corto/largo plazo". Jensen (2006), apoyado en evidencia neurológica, menciona que la MT funciona a suerte de interfaz múltiple en donde operan distintas capacidades cognitivas, tales como la MLP, la percepción y la atención, por lo que es deseable dejar de lado un modelo binario; en la misma línea Baddeley (2007, 7) señala que, en sus investigaciones, el término de *memoria a corto plazo* lo ha reservado para tareas de recuperación de información en pequeñas cantidades, mientras que la MT implica otras operaciones cognitivas. Para él, se trata de un modelo multicomponencial en donde la atención y la percepción entran en juego.

La Atención y la Memoria de Trabajo se han manejado como similares, pero son distintas y operan en conjunto. Kibrik (2011) menciona que, debido al manejo descuidado de los conceptos psicológicos de atención, consciencia y MT, en donde prominencia (*saliency*) y predictibilidad (*predictibility*) se mezclan, no queda claro cuáles, en qué medida y cómo podrían implementarse estas nociones en las explicaciones lingüísticas sobre la referencia. En el tema que compete este trabajo, un referente que se encuentre activo (en la MT) no garantiza su mención (Atención) en las siguientes construcciones lingüísticas, pero sí supone el que haya sido mencionado en algún momento anterior en el discurso o supone su disponibilidad a la percepción en el momento de la enunciación. Además, se podría predecir el tipo de dispositivo referencial si es que se vuelve a mencionar.

Para ejemplificar lo anterior, obsérvense las oraciones en (10) a continuación. La mención de *el perro* en la oración inicial activa este referente, además que formaliza el rol gramatical de sujeto de la oración, lo que lo coloca en el FOCO PSICOLÓGICO (concepto que trataré más adelante). Sin embargo, esto no garantiza su mención en la siguiente oración, pero debido a

su alto grado de activación se puede predecir con **cierta probabilidad** que el dispositivo referencial será reducido.

(10) El perro come un pedazo de pollo.
   a. Qued**ó** satisfecho.
   b. **Él** quedó satisfecho.
   c. **Un perro** quedó satisfecho.
   d. **El canino** quedó satisfecho.

Nótese que los ejemplos de (10a-b) nos parecen (como hablantes de español) más probables como secuencia a la oración inicial: el sujeto tácito y el pronombre; en español, de hecho, esperaríamos sólo (10a) reservando para contextos contrastivos (10b). Por otro lado, un DRP como *un perro* en (10c) es poco afortunado si quisiese usarse para referirse al mismo referente *el perro* de la oración inicial. Ahora, el caso de (10d) es interesante, el uso de *el canino* es afortunado debido a una relación de accesibilidad, pero es un DRP, lo que va en contra de la predicción. En este caso, primero, la hipótesis de trabajo está en que este tipo de DRP es poco probable, y segundo, que esto supone una ejecución sobreproductiva por parte del hablante.

Un planteamiento interesante de Kibrik (2011) corresponde a la cantidad de entidades en el MT en un momento dado. Sugiere, tanto por los estudios en psicología como en sus propias investigaciones, que se trata de una capacidad que se limita a entre cuatro y siete referentes. De estas entidades, una se encuentra en el FOCO PSICOLÓGICO. Este foco, como podrá notarse, no tiene que ver con lo establecido como Foco en la relación binaria con el Tópico, sino como un subtipo de MT en donde se encuentra el grado más alto de activación. De tal manera, en el centro de la MT se encontraría este referente de activación alta, mientras que aquellos que tienen activación cero están fuera de la MT por completo. Conforme transcurre el tiempo y

el discurso, aquel referente en el centro de la MT —*foco psicológico* de ahora en adelante— se iría desactivando. Este proceso es gradual y no debe suponerse un desplazamiento inmediato de los referentes.

El correlato lingüístico básico de la relación fuera/dentro de la MT lo encontramos entre los DRP y DRR; los primeros presentarían información con activación cero mientras que los segundos, referentes activos. Kibrik (2011) destaca que el referente en el foco psicológico será preferentemente aquel que es formalizado como el sujeto de una oración (Tomlin 1995). A su vez, la capacidad de la MT coincide con el hecho de que en una oración encontremos por lo general no más de tres o cuatro argumentos, estando otros elementos adjuntos en una situación desfavorable de activación.

1.6 Accesibilidad como estado mental

En sus conclusiones, Kibrik (2011) admite que en su trabajo dedica poca atención a lo que se conoce como ACCESIBILIDAD, por lo que me es importante ahondar más en ella y afinar su base cognitiva. En lingüística el concepto de accesibilidad se ha utilizado para abarcar lo que en mi trabajo nombro Estados Informativos (Ariel 1988; Keenan y Comrie 1977; Ariel 1990). En este caso, los restringiré a un conjunto de situaciones en los que el hablante supone capacidad para relacionar o inferir temas. Lambrecht (1994), al mencionar la accesibilidad, coloca en un mismo grupo la accesibilidad por situación, texto e inferencia. Además, establece que son variaciones del estado de activación. Sin embargo, difiero en este punto, en particular, para los referentes que son accesibles gracias al conjunto de expectativas

asociadas con un marco conceptual establecido. Son los casos de las anáforas asociativas o por evento que ejemplifico a continuación:

(11) a. Leí **una columna** ayer. **El escritor** es brillante.
b. **Llegar** a Cuernavaca es fácil. **El viaje** es rápido.

En (11a) el artículo definido es afortunado debido a la inferencia que desprende *una columna*: se espera que las columnas sean escritas por escritores. Por otro lado, en (11b), *el viaje* es afortunado debido a que el verbo *llegar* establece un marco: llegar a un lugar posibilita pensar en un viaje. Sin embargo, nótese:

(12) a. Leí **una columna** ayer. # Es brillante (¿la columna o el escritor?).
b. Leí **una columna** ayer. # Él es brillante (¿quién es brillante?).

En estos casos, que una pieza léxica nos brinde la posibilidad de acceder a un referente no significa que tal referente esté activo. En (12a) la lectura preferible para el sujeto del verbo *ser* es *la columna*, pero no *el escritor*; esto sobresale con (12b) en donde al tratar de utilizar un pronombre, surge la cuestión de a qué persona se hace referencia.

Ahora bien, Lambrecht (1994) integra como un tipo de accesibilidad a aquellos referentes que se encuentren en la situación inmediata de la enunciación. Con ánimos de discernir mejor estos casos, en mi análisis optaré por entenderlos como referentes ACTIVOS POR ORIGO, entendiendo Origo como las coordenadas espacio-temporales del momento de la enunciación, incluyendo a los participantes del acto de habla (Bühler 2011). Los demostrativos plenos, aquellos que usan el gesto para señalar a lo nombrado en el contexto físico, las marcas de primera y segunda persona y los adverbios de tiempo y lugar *hoy/mañana/ayer/aquí/allá/acá/ahí* supondrán, por lo tanto, **activación constante en la MT**. Por ejemplo, atiéndase el siguiente caso:

Contexto: me encuentro en el metro y de repente ya no siento mi celular, por lo que exclamo frente a un extraño.

(13) Perd**í** mi celular.

En (13), el dispositivo referencial reducido que es la marca en el verbo de primera persona es afortunado, sin necesidad de un pronombre o un DRP previo. No es necesario que el hablante suponga que el oyente es capaz de recuperar de memoria al referente (él mismo), sino sólo que pueda interpretar la situación inmediata; lo que se supone en estos casos es que el oyente tiene en su atención al referente en cuestión, que en este caso es su interlocutor. Imaginemos que el extraño es empático con la situación y en un examen rápido del escenario, observa que el celular se encuentra en el suelo, a lo que responde apuntando:

(14) ¿no es **ese**?

El referente se activa al momento, no sólo por la mención —la forma lingüística **ese**—, sino de la función plena del demostrativo, acompañada por un gesto que señala al objeto: se activa por la percepción conjunta de la entidad física (Diessel 2012).

Otra manera de usar el Origo en la referencia es a lo que llamaré ACCESIBLE POR ORIGO, en donde operan los demostrativos determinantes en casos como:

(15) En **esta ciudad** se pierden muchos celulares.

En donde *esta ciudad* apela a la ciudad en la que se encuentre uno en el momento de la enunciación, pero es necesario el razonamiento sobre el Origo; no se nos presenta por evidencia perceptual señalada por el gesto. En este caso coincido con Lambrecht (1994, 110) quien ejemplifica que casos de DRP como *este viernes* son candidatos a este tipo de estado. De esta manera, busco distinguir entre las suposiciones de activación y accesibilidad a través

de, o el uso de deícticos reducidos o como determinantes de una frase nominal. Sin lugar a dudas, queda pendiente un trabajo que examine con mayor detalle las distintas funciones de los demostrativos y su relación con los Estados Informativos, lo cual no es un asunto resuelto (cf. Levinson 2004; Diessel 2014; Lyons 1980). Por lo pronto, sólo me interesa realizar esta diferencia a partir de lo propuesto por Lambrecht (1994).

Con respecto a la accesibilidad por texto de Lambrecht (1994), la clasifico como un subtipo de inactividad (INACTIVO TEXTUAL). En efecto, un DRR no es esperable en casos en los que, por ejemplo, se menciona un referente con un DRP en una PRIMERA ORACIÓN (POSICIÓN 1), pero se trata de hacer referencia a él con un DRR en una ORACIÓN POSTERIOR (POSICIÓN 10, por ejemplo) del mismo discurso —a menos que la lengua tenga mecanismos gramaticalizados como la cuarta persona que ilustraré en la siguiente sección. La base cognitiva de este tipo de accesibilidad es por MLP, indiferenciable con otro tipo de recuperación de memoria, pero distinguible en mi trabajo como el Registro Discursivo presentado en la §1.4 anterior. Hasta donde he podido ver en mi revisión teórica y análisis preliminar, me parece que el problema de la accesibilidad se encuentra en la anáfora asociativa. Para entender este tipo de accesibilidad, Lambrecht (1994, cap. 3) parte de la noción de marco (*frame*) desarrollada por Fillmore (1982, 111) quien sostiene lo siguiente:

> Con el término 'marco' tengo en mente un sistema de conceptos relacionados de tal manera que para entender cualquiera de ellos es necesario tener toda la estructura en donde tal concepto encaja; cuando una de las partes de dicha estructura es introducida en el texto o la conversación, todas las demás están automáticamente disponibles[33].

[33] "With the term 'frame' I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text, or into a conversation, all of the others are automatically made available" (Fillmore 1982, 111)

Sobre esto difiero, en la medida en que *disponibles* no debe entenderse como *activos en la MT*. La ACCESIBILIDAD POR ANÁFORA ASOCIATIVA se presenta, en los términos tratados en este trabajo, como la suposición de hablante sobre la capacidad de inferencia y de recuperación de información para tal inferencia.

Siguiendo la pista detallada por Noordman y Vonk (2015), entiendo que la inferencia para la anáfora asociativa está relacionada con la **coherencia discursiva**. Los estudios neurológicos y psicolingüísticos presentados por estos autores apuntan a que distintas áreas del cerebro entran en funcionamiento en la inferencia: las inmiscuidas con memoria de información semántica; las que operan la recuperación y selección de esa información; aquellas que la activan en la MT, junto con las áreas que establecen o refuerzan una relación con información en la MLP; las que buscan las relaciones pertinentes; y las que buscan las incoherencias. Como menciona Ziai, De Kuthy y Meurers (2016, 209), las anáforas asociativas presentan un reto para las tecnologías del lenguaje en cuanto a la resolución automática de referencia.

Lo investigado desde la psicología y la neurología muestra que es un mecanismo más complejo que sólo la mención de un referente en el discurso, en donde, para resolver la referencia, se involucran mecanismos y módulos relacionados con el razonamiento y la búsqueda de coherencia en el discurso. En mi análisis, en todo caso, no desarrollo una diferencia más fina de las anáforas asociativas partiendo de su base inferencial. Coincido con la visión planteada por Gerrig y O'Brien (2005) quienes sostienen que:

> […] no hay necesidad de establecer tipos de inferencias a partir de la mayor o menor tendencia a codificar por parte del lector. Las inferencias están codificadas [en la mente, memorizadas como patrones] en la medida en que la información en la memoria activa [memoria de trabajo] hace contacto con información necesaria o

> relevante de porciones inactivas del modelo discursivo y del conocimiento general del mundo[34].

Estos autores continúan diciendo que existen algunos tipos de inferencia que pueden llegar a ser codificados, como las construcciones causales, pero incluso en esos casos, la activación en MT sigue siendo un proceso pasivo. Por lo anterior, entenderé que los casos de Accesibilidad por Anáfora Asociativa no pueden predecirse por patrones gramaticales en el discurso y escapan al primer planteamiento de mi tesis: la posibilidad de predecir Estados Informativos por factores gramaticales. A pesar este escenario, estarán etiquetados en el corpus con la intención de separarlos del modelo estadístico y contrastarlos con aquellos Estados Informativos que muestren una mejor relación con las medidas propuestas (cf. Capítulo 3). Seguiré la pauta establecida por Kibrik (2011) y para evitar abordar el problema desde la coherencia del discurso, tema que está fuera de mis objetivos de investigación, sólo evaluaré la Accesibilidad de la Anáfora Asociativa si existe un referente activo del cual pueda establecer un marco para el enlace.

1.7 Conflictos de activación y ayudas referenciales

En esta sección expondré la manera en que clasifica Kibrik (2011) las ayudas referenciales a partir de conflictos de referentes en la MT, es decir, de dos o más referentes que se encuentren activos. Este autor demuestra que las lenguas integran en sus gramáticas formas que orientan al oyente para realizar una elección exitosa del referente. La mayoría de estos dispositivos

[34] "…there is no need to define categories of inferences that readers will typically encode and those that readers are less likely to encode. Inferences are encoded to the extent that information in active memory makes contact with relevant or necessary information from inactive portions of the discourse model and general world knowledge" (Gerrig y O'Brien 2005, 236).

referenciales son reducidos, y poca descripción se dedica a los plenos. Por lo que, primero, describiré brevemente la tipología de dispositivos referenciales reducidos para luego contrastar con los dispositivos referenciales plenos en español y demostrar que un análisis de la activación no puede ser tan directo debido a la manifestación de otras aristas al problema.

Al centrarse en los DR REDUCIDOS, Kibrik (2011) hace notar el conflicto en el momento en donde dos referentes activos se introducen en una oración inicial. Para ilustrar esto, nótense los siguientes ejemplos:

(16) a. Juan y Pedro tienen mucho frío.

b. Juan y María tienen mucho frío.

c. Juan y un perro tienen mucho frío.

d. Juan y el presidente tienen mucho frío.

e. Juan y los niños tienen mucho frío.

Él se tapó primero.

El caso en donde dos referentes se encuentran activos se presenta en (16). La oración inicial podría ser cualquiera de las presentadas entre (16a-e), la encadenada es la última oración *él se tapó primero*. En español, si dos referentes son activados en la primera oración, y quisiéramos recuperar un referente en la encadenada, lo ideal sería un pronombre, como en este caso *él*. En (16a), si quisiéramos indicar que el pronombre *él* hace referencia a *Juan* y no a *Pedro*, en español no contamos con maneras de codificar esa diferencia: ambos masculinos, singular y coordinados como sujetos de la oración anterior. Una posibilidad sería lo que Kibrik (2011, 328) llama referencia fresca o reciente: *Pedro*, por ser el último mencionado, contaría con una sutil diferencia que le permitiría contrastarse con *Juan*. Más

allá de que el uso de los demostrativos rescata precisamente esta diferencia, como el demostrativo *aquel* en el siguiente ejemplo de (17), es posible codificar otras diferencias.

(17) **Juan$_{i}$** y **Pedro$_{j}$** tienen mucho frío. **Aquel$_{i}$/Este$_{j}$** se tapó primero.

Por ejemplo, en el caso de (16b), el género es una propiedad semántica estable que permite seleccionar sin conflicto uno de los dos referentes activos; en el caso de (16c) podríamos decir que el rasgo HUMANO se coloca por encima de lo ANIMADO, y a su vez de lo INANIMADO como el contraste que propongo en la curiosa situación que plantea la oración siguiente:

(18) ? **Un perro$_{i}$** y **un árbol$_{j?}$** tienen mucho frío. **Él$_{i/j?}$** se tapó primero.

En el ejemplo de (16d), se muestra un caso en donde parece no haber diferencia entre si se trata de un nombre o una frase nominal definida: las dos se activan —como ya he señalado en las secciones anteriores— por el mero hecho de ser mencionadas. En este caso, pareciera que no se resuelve el conflicto de manera sencilla. En mi interpretación, *el presidente* se contrasta de *Juan*, pero podría ser por un recurso ad hoc de interpretación: se espera que el individuo con el título de presidente tenga cierta iniciativa. Otra razón sería la jerarquía implícita que invoca el marco conceptual de *el presidente*. Kibrik (2011) señala que existen lenguas en donde tal jerarquía se marca gramaticalmente, lo que ayuda en la resolución de referencia. Nótese que la dificultad para resolver la referencia se agudiza con (19) en donde se podría sostener que ambos referentes tienen el mismo rango:

(19) **El presidente$_{i}$** y **el gobernador$_{j}$** tienen mucho frío. **Él$_{¿i/j?}$** se tapó primero.

Finalmente, en (16e) ejemplifico cómo la propiedad de plural ayuda a seleccionar a uno de los dos referentes activos. En su revisión sobre los tipos de ayudas que aparecen en las

lenguas del mundo para resolver conflictos de activación Kibrik (2011) llega a una síntesis que presento en la Figura 2.

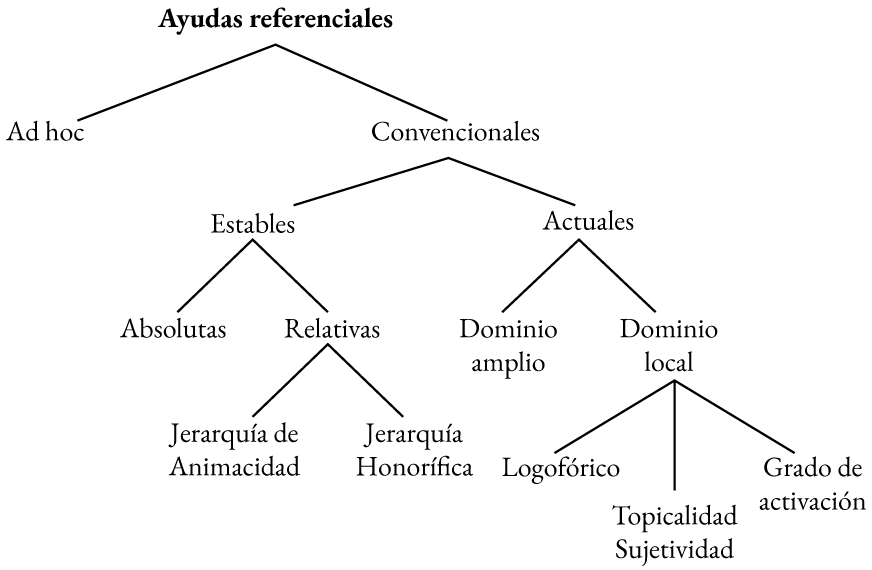


Figura 2. Clasificación de las ayudas referencias[35]

La primera diferencia retratada en el esquema está entre las ayudas referenciales ad hoc y convencionales. Las **ad hoc** apelan al conocimiento enciclopédico al momento de la enunciación, marcos o esquemas que pueden ayudar a establecer la referencia y el significado. Aparecen principalmente en lenguas en donde la marca cero es frecuente, como el japonés, el mandarín chino y el tailandés (Kibrik 2011, 292). En el caso de las **convencionales** se trata de diferencias que están codificadas en la lengua de alguna manera: se parte de propiedades **estables** de los referentes y **actuales**, entendido como el discurso *vigente*. En las estables encontramos las **absolutas**, que refieren a propiedades semánticas intrínsecas en las entidades, como la marca de género en español como vimos con (16b) o

[35] Tomado de Kibrik (2011, 332); traducción propia.

los clasificadores en otras lenguas (cf. Aikhenvald 2000). En el caso de las **relativas**, encontramos por un lado la jerarquía de animacidad que ejemplifiqué en (16c) y (18), y por otro, la jerarquía honorífica, que en español no se encuentra codificada gramaticalmente para la tercera persona pero creo que interviene como recurso ad hoc, como mostré en (16d) y (19).

En el caso de las ayudas referenciales convencionales-actuales, tenemos las de **dominio amplio.** Kibrik (2011, 308–9) ejemplifica con la lengua navajo este tipo de ayudas. En una historia con dos protagonistas que tienen las mismas propiedades semánticas y discursivas, esta lengua utiliza lo que se ha llamado *cuarta persona*: un demostrativo que se asocia a un referente particular a lo largo del discurso y se reserva para ese referente.

> In its narrative usage, **the fourth person** invariably applies to specific human (or personified) referents. If there are two or more central referents in a recounted story, at some point 'person assignment' can occur. That means that the fourth person is attributed to a certain referent as its discourse-internal but quite persistent qualification. […] The fourth person effectively operates as a **temporary proper name of the referent** (Kibrik 2011, 309; énfasis propio).

En el **dominio local** encontramos tres recursos que se gramaticalizan en las lenguas: la **logoforicidad** (*logophoricity*), que se refiere a la existencia de pronombres especiales que codifican el cambio de perspectiva de quien enuncia. Son los casos en los que el interlocutor hace énfasis en que está reportando el discurso de alguien más. En español no tenemos una forma que codifique esta propiedad, pero se utilizaría en estructuras del tipo *Juan$_i$ dijo que él$_i$*… en donde el pronombre *él* haría énfasis en que se refiere a la persona que enunció la oración, pero no quien de hecho enuncia el discurso. En español esto resultaría ambiguo y desafortunado ya que daría una construcción del tipo *Juan$_i$ dijo que yo$_i$*…. Formas especializadas para esta función discursiva las encontramos en lenguas como el angas, como se muestra en el siguiente ejemplo.

(20) Músá lə tèné { **ɗyí** / kə́}
Musa$_M$ dijo que LOG.SG.M.NPRES$_M$ / 3SG.NPRES$_i$
mét kàsúwá
ir mercado
"Musa$_M$ dijo que {él$_M$ / él$_i$} irá al mercado"
(Angas, tomado de Burquest (1986, 92 *apud* Kibrik (2011, 316)); traducción propia) [36].

En esta lengua africana de la familia chádica encontramos dos tipos de pronombres: uno general *kə́* el cuál podría hacer referencia a cualquier tercera persona singular, o el pronombre especializado *ɗyí* el cual nos indica que su antecedente es quien, en efecto, *dice* o *piensa* la cita: la fuente de la información.

En el caso de la **topicalidad o sujetividad** (*subjecthood*), existen marcas especializadas en mantener la referencia al sujeto, tópico o protagonista del discurso establecido, las cuales han sido nombradas formas de Cambio de Referencia (*switch reference*). Un ejemplo lo encontramos en las lenguas yumanas, en donde una marca en el verbo permite determinar que el sujeto de una oración subordinada es el mismo que el de la oración matriz (Langdon y Munro 1979).

(21) nya-isvar-k i:ma-k
cuando-cantar-**ss** baila-**ss**
"Cuando él$_i$ canta, él$_i$ baila"
(Mojave, tomado de Langdon y Munro (1979, 322); traducción propia).

[36] Las abreviaturas que usaré en las glosas de otras lenguas son las siguientes: ACC = acusativo; DAT = dativo; INSTR = instrumental; LOG = logofórico; NOM = nominativo; NPRES = no presente; SBJV = subjuntivo; SG = singular; SS = mismo sujeto; Subíndices: M = masculino; P = paciente.

En el ejemplo anterior de mojave, lengua yumana hablada en Arizona, Estados Unidos, la marca *-k* indica que el sujeto de *isvar* ‘cantar’, es el mismo para la segunda oración en donde el verbo matriz es *i:ma* ‘bailar’.

La última ayuda referencial la encontramos en la gramaticalización de **grados de activación**. Kibrik (2011) señala el caso del ruso y el demostrativo *on* que refiere a referentes con mayor grado de activación que *tot*, siendo esta la única diferencia. Cito su ejemplo a continuación:

(22) a.

Bylo	*by*	*kuda*	*čestnee*	*neizvestn-omu*
fueron	SBJV	mucho	más.honesto	desconocido-DAT
zloumy šlennik-u	*razbirat’sja*	*naprjamuju*	*s*	*Prokuror-om$_{p'}$*
malhechor-DAT$_M$	resolver	directamente	con	fiscal-INSTR

b.

koli	***tot***	*ego*	*čem*	*obidel*
si	ese.NOM$_P$	él.ACC$_M$	qué.INSTR	Insultado

“Podría ser mucho más honesto para el malhechor desconocido resolver las cosas directamente con el fiscal, si él lo insultó a él con algo”

(Ruso, tomado de Kibrik (2011, 328); traducción propia).

Se puede observar que *tot* hace referencia a un referente que, ni es sujeto ni tópico, entendido como tema del intercambio comunicativo. Debido a esto, aunque *prokuror-om* ‘el fiscal’ se encuentra activo debido a la mención en la oración anterior, tiene una posición menos privilegiada que los otros referentes que son argumentales y preponderantes sintáctica y semánticamente, como puede notarse por su marca *-om* ‘instrumental’.

En español no encontramos marcas especializadas para desambiguar dos o más referentes activos, a pesar de las marcas de plural y género —esto lo demostraré más adelante—. Una posibilidad está en el uso de los demostrativos para los grados de activación, como en el

ejemplo (17) en donde la posición en la oración podría establecer cierta diferencia. Sin embargo, no es una forma gramaticalizada y no es el caso que el pronombre demostrativo en español distinga en su forma, entre FN con función sintáctica de sujeto y de oblicuo, aunque sí plural y género.

Lo anterior busca explicar y clasificar los mecanismos codificados en la lengua que ayudan a discriminar dos referentes activos en los casos en donde se utilicen dispositivos referenciales reducidos. Si seguimos el mismo tono que he planteado a lo largo de este trabajo, las ayudas referenciales son recursos estructurales que tienen a disposición los hablantes, las cuales dejan ver las suposiciones de la capacidad del oyente para vincular propiedades sintácticas/semánticas/pragmáticas con el referente, y de esta manera seleccionarlo correctamente de un posible conflicto entre dos o varios referentes activos. En mi análisis resulta pertinente al momento de plantear menciones de un mismo referente: se esperaría que compartieran propiedades estables. En este punto cabría la pregunta de si estas ayudas referenciales también operan con dispositivos referenciales plenos.

Para dimensionar el problema piénsese el siguiente planteamiento: dadas dos menciones a través de dispositivos referenciales plenos de un referente en el discurso, ¿existe alguna propiedad o marca de las menciones que garantice su vínculo? En español, el número y el género parecen establecer algunas restricciones y favorecer algunas lecturas. Por ejemplo:

(23) a. El gato y el perro se pelearon. La mascota ganó sin problemas.

b. La gata y el perro se pelearon. La mascota ganó sin problemas.

c. El gato y la perra se pelearon. La mascota ganó sin problemas.

En todos los ejemplos anteriores, al igual que en los ejemplos que mostré en (16), se colocan ambas frases a examinar de manera coordinada como sujetos, para con ello explorar otras relaciones. En (23a) *la mascota* femenino singular, puede ser utilizada para hacer referencia a cualquiera de las dos entidades anteriores, ambas con los rasgos masculino y singular. En (23b) existe una preferencia por leer que *la mascota* y *la gata* son dos menciones de un mismo referente. En (23c) sucede lo mismo con *la perra* y *la mascota*. Parece que el género es una ayuda referencial en dispositivos referenciales plenos; pero esto no significa que *la mascota* y *el gato* sean imposibles de vincular. Es posible interpretar que *la mascota* se refiere a las menciones marcadas con masculino. Las marcas de género (ya sea en el nominal o en los modificadores del nominal) no establecen un criterio absoluto de vínculo. En mi trabajo no lo implemento de manera directa como factor, pero le otorgo a mi algoritmo la capacidad de observar la propiedad de género.

En contraste, la marca de número es más restrictiva:

(24) a. Las gatas y el perro se pelearon. La mascota ganó sin problemas.
b. Las gatas y las perras se pelearon. Las mascotas ganaron sin problemas.
c. El gato durmió todo el día. La mascota tiene su propio cuarto.
d. El gato durmió todo el día. Las mascotas tienen su propio cuarto.
e. Las mascotas durmieron todo el día. El gato tiene su propio cuarto.

En los ejemplos anteriores, la marca de número en la frase nominal restringe las relaciones entre las menciones. En el caso de (24a) el género no puede competir con el número: la preferencia es vincular a *la mascota* con *el perro*, a pesar de estar marcados con géneros distintos; en (24b) parece que *las mascotas* crea un nuevo conjunto conformado por las frases

coordinadas (esto mismo sucedería en el caso de presentar a las dos menciones anteriores en sus propias cláusulas); en (24c) no tenemos problemas con vincular *el gato* y *la mascota* aunque sean de género distinto, mientras que en (24d), al igual que en (24c), resulta desafortunado un vínculo entre una entidad plural y singular[37]. Pero, contrástese este último ejemplo con (24e) en donde es posible apelar a un miembro del referente mencionado por *las creaturas*. Estrictamente, no estamos hablando de dos menciones de un mismo referente, pero sí existe una relación Todo-Parte, lo que demuestra que el número tampoco está exento de problemas: no es una marca que descarte o establezca sin falla relaciones entre menciones. En todo caso, esto muestra que en español no tenemos una marca convencional que establezca las relaciones entre mismas menciones utilizando dispositivos referenciales plenos de un referente. Parece ser que, para lograr tal vínculo, para este tipo de dispositivos, se recurre al contexto y conocimiento enciclopédico de las frases en cuestión. Esto vuelve a llamar la atención sobre los casos de anáforas asociativas, en donde dos frases nominales plenas son vinculables por el conocimiento del mundo que se tiene sobre ellas. Por lo que, si se nos presenta algo como el siguiente ejemplo:

(25) Una mujer fue llevada **al hospital**. **Las heridas** fueron tratadas.

Interpretamos que *las heridas* no es una entidad completamente nueva, pero tampoco mencionada como referente de manera explícita con anterioridad. No existe un mecanismo formal exclusivo que codifique la función sobre apelar al marco para interpretar la conexión entre los referentes (en este caso, vincular que *el hospital* y *las heridas* forman parte del

[37] Parece ser que una lectura posible es interpretar que hay más gatos, pero creo que esto es parte de tratar de interpretar localmente la oración. De esto hablaré con más detalle en la sección 1.10.

mismo marco). Se podría objetar que el artículo definido hace justo esto, pero pruébese la frase encabezada con un indefinido y nos encontramos en la misma situación: *unas heridas* no aparece de la nada para presentar un referente totalmente nuevo. Se vincula, tal vez en distinto grado, pero en la misma dimensión, con *el hospital*. Como expondré más adelante, esto tiene que ver más con principios pragmáticos —en particular, con el principio pragmático de interpretación local— que con instrucciones de algún segmento de la oración.

Por lo pronto, de las ayudas referenciales analizadas por Kibrik (2011), es difícil asegurar que alguna funciona para los dispositivos referenciales plenos, excepto, la que apela al conocimiento enciclopédico o situacional que funciona en el momento de la enunciación. Sin lugar a duda, un trabajo fino que compare las ayudas referenciales utilizadas en distintos dispositivos referenciales en español es requerido. Para los objetivos de esta investigación, recurriré a la lectura plena del discurso en donde aparecen las frases para determinar su vínculo y no en principio a una marca o rasgo en particular.

Las ayudas referenciales tratadas por Kibrik (2011, caps. 9–10) ilustran diferencias entre DRR y mecanismos que resuelvan conflictos de referencia. Además de estos recursos gramaticales, también existen recursos de otros ordenes para decidir qué referente está vinculado con el dispositivo que lo menciona. Estos recursos los enuncia Kibrik como **factores de activación**. Sería deseable que un modelo que identifique los Estados Informativos pueda alimentarse tanto de la información morfosintáctica interna de la frase nominal (en donde distinguir género y número es pertinente) además de otros factores de orden sintáctico, pragmático y textual. Por tal razón, reviso de manera más detallada en la siguiente sección los factores que utiliza Kibrik en distintos experimentos para la predicción de la forma del dispositivo referencial, aunque son pocos los que puedo rescatar en el análisis que propongo. Estos

factores incluyen aspectos tan diversos como la distancia entre oraciones de un referente o la distancia retórica[38] —al inicio de párrafo, secuencia de oraciones argumentales, etc.— la animacidad y la sujetividad, entre otras propiedades. La idea de fondo de los factores es establecer una serie de propiedades que, en determinada lengua, constituyan el mejor modelo para predecir el grado de activación de un referente y, por consiguiente, el dispositivo referencial REDUCIDO preferente. En la siguiente sección hablaré sobre el acercamiento multifactorial del lingüista ruso, especificando al final cuáles factores tomo íntegros y cuáles parcialmente a partir de la forma de etiquetado de mi corpus.

1.8 Modelo multifactorial de referencia

Aunque Kibrik (2011) admite no tratar a detalle los DISPOSITIVOS REFERENCIALES PLENOS, pienso que pueden llegar a tener un correlato de análisis similar a los DISPOSITIVOS REFERENCIALES REDUCIDOS: es decir, parto de que existen propiedades lexicogramaticales que favorecen la aparición de determinados dispositivos referenciales plenos. Una escala de FN plenas puede plantearse en donde la Elección Referencial de un determinado dispositivo referencial nos permite distinguir las suposiciones del hablante sobre el Estado Informativo del referente. Esto, aunque parecido a las escalas propuestas por Givón (1983a, 17; 1983b; 2001, 417), no es del todo similar: mi atención se concentra en las tendencias y patrones de las frases nominales plenas, por lo que en principio no es de mi interés comparar estos dispositivos con DRR. Pero, además, el discurso mismo puede tener un alcance en esta

[38] Este término se refiere a las distancias planteadas por la *Rhetorical Structure Theory (RST)* que Kibrik (2011) utiliza en su trabajo, pero que yo no implementaré en mi análisis.

predicción: *algunos* de los dispositivos referenciales plenos/reducidos podrían predecirse con éxito apelando **solamente** a la estructura lexicogramatical del discurso precedente modulando la ventana de palabras considerada, que puede ir desde unas pocas palabras a todo el discurso anterior[39]. Digo *algunos* porque, como ya mencioné antes, me parece que los casos más difíciles de predecir son los de Accesibilidad por Anáfora Asociativa en donde la suposición descansa en la relación semántica entre unidades léxicas a partir de un Marco o Esquema (en términos de Fillmore (1982) y Chafe (1994)). En este momento, mi suposición principal es que, de las posibilidades de DRP, existen formas preferentes para la identificabilidad y la accesibilidad, y que tales formas pueden predecirse por patrones lexicogramaticales del discurso vigente sintetizables por un algoritmo informático. También integro la activación, aunque el trabajo de Kibrik (2011) apunta a que los DRR son los preferenciales. Esto será parte del trabajo nuclear del proyecto en lo que sigue de la investigación.

Una ventaja del trabajo de Kibrik (2011, cap. 14) como antecedente de esta investigación es que buscó implementar su propuesta en un modelo computacional para guiar la conformación de la mejor serie de factores que predijeran el DRR y el grado de activación. Para ello, planteó el uso de redes neuronales, un subtipo de técnica computacional basada en aprendizaje de máquina (*machine learning*). La diferencia entre su trabajo y el mío radica, primero, en que mi enfoque atiende FFNN plenas —en las que el método de Análisis de Semántica Latente parece tener buen resultado (Hempelmann et al. 2005)— y en que, de los factores que enlista este autor, no todos se encuentran accesibles al tipo de etiquetado que realicé en COPENOR.

[39] En términos computacionales, esto significaría variar las posibilidades de los n-gramas, sin suponer que sólo nos referimos a palabras.

En especial, en sus primeras aproximaciones utiliza un etiquetado a partir de la Teoría de la Estructura Retórica (*Rhetorical Structure Theory o RST*) y proyecta el uso de un programa/interfaz especial además de un conjunto de especialistas para un etiquetado fino de un corpus. Los resultados de este proyecto se reportan en *Referential Choice: Predictability and Its Limits* (Kibrik et al. 2016).

Kibrik (2011) señala que muchos estudios sobre la referencia establecen varios factores que ayudan a determinar el estado de activación de los referentes. Entre los cuales se encuentran: distancia oracional con el último antecedente (Givón 1983b); fronteras episódicas (Tomlin 1987); centralidad del referente en el discurso; y animacidad (Dahl y Fraurud 1996). La multiplicidad de factores involucrados es amplia, y sólo enunciar que este fenómeno representa esta complejidad no resuelve el problema de cómo determinar los pesos de cada factor y su interrelación. Kibrik (2011, 391) propone que se determine una lista inicial de factores que puedan estar abiertos a la verificación. Su planteamiento parte, primero, de establecer ciertos factores hipotéticos. Los que muestren una covariación significativa con el dispositivo referencial seleccionado son los que pasan a establecerse como factores de activación. El método para determinar la calificación para el grado de activación lo basa en el puntaje de los distintos factores que al final suma. Esta estrategia la considera básica frente a las posibilidades de las nuevas tecnologías. Admite que se deberá revisar y mejorar, no por matematizar *ars gratia artis*, sino para aproximar el método a herramientas que pueden ayudar a obtener mejores resultados.

Hoy en día son pocas las áreas de la lingüística que utilizan una visión multifactorial. Kibrik (2011) menciona que por lo general se selecciona un criterio y éste es el que se evalúa, o en el mejor de los casos, se ponen a competir varias propiedades, sin recurrir a determinar cómo

es que esas propiedades se podrían apoyar entre ellas para manifestar el fenómeno estudiado. Son los trabajos en lingüística computacional los que se basan en esta propuesta multifactorial (Kibrik et al. 2016; Recasens 2010; Strube y Wolters 2000), y agregaría que los trabajos en sociolingüística, con el uso, por ejemplo, de regresiones logísticas escalonadas (Martin Butragueño 2018)[40].

Además de que esta propuesta es pertinente por la complejidad del fenómeno estudiado, resulta necesaria para no caer en un análisis circular. Kibirk (2011) señala que analizar el grado de activación por el dispositivo referencial seleccionado reduce su poder explicativo: un DRR implica alto grado de activación porque está presente en el discurso, y está en el discurso porque el alto grado de activación del referente fue lo que oriento su selección de entre los dispositivos referenciales. Para salvar este problema, la visión multifactorial es crucial. No se asume a priori que un DRR implica un grado alto de activación, sino que una misma forma puede tener variado comportamiento dependiendo de los factores de activación que estén presentes en el discurso, y considerados en el análisis.

Los factores que Kibrik (2011) selecciona en distintos puntos de su investigación los muestro en la tabla 3. Presento los factores del primer conjunto de su primer análisis manual. El segundo conjunto es posterior al análisis manual y funciona como síntesis para realizar el experimento con redes neuronales. El tercer conjunto es el resultado de haber realizado tres ejercicios computacionales que permiten determinar cuáles de los factores mantienen una

[40] Hay que señalar que el uso de una estrategia matemática-probabilística no implica una perspectiva multifactorial, por ejemplo, el trabajo para resolución de anáforas usando teoría bayesiana (Kehler y Rohde 2018).

precisión alta en la predicción[41]. El cuarto conjunto, el más extenso, son los factores que utilizó en sus experimentos con el *RST Treebank* (Kibrik et al. 2016).

Tabla 3. Factores de activación en distintos proyectos de investigación sobre la predicción de la forma de la anáfora y el estado de activación[42]

Factor	**C1**	**C2**	**C3**	**C4**
Distancia retórica del antecedente	■	■		■
Distancia lineal del antecedente	■	■	■	■
Distancia de párrafo del antecedente	■	■	■	■
Rol sintáctico y semántico del antecedente	■		■	■
Protagonismo del referente	■	■	■	■
Animacidad del referente	■	■		■
Identidad difusa	■			
Rol sintáctico del dispositivo referencial		■	■	■
Rol sintáctico del antecedente retórico		■		■
Tipo de dispositivo referencial del antecedente retórico		■		■
Rol sintáctico del antecedente lineal		■	■	■
Tipo de dispositivo referencial del antecedente lineal		■	■	■
Género				■
Persona				■
Número				■

[41] En pocas palabras, el ejercicio consiste en que, después de entrenada la red, se van quitando y poniendo factores y la red se reentrena para determinar cuáles afectan la precisión.

[42] Los nombres de todos los factores son traducciones mías del inglés. El C1 lo tomo de Kibrik (2011, 411) el C2 de Kibrik (2011, 463); el C3 de Kibrik (2011, 466); el C4 lo tomo de Kibrik *et al.* (2016, 206).

Tipo de frase del dispositivo referencial (FN, FP, FA, etc)				
Posición del dispositivo referencial en la cadena referencial				
Tipo de descripción				
Tamaño del antecedente en palabras				
Distancia con el antecedente en palabras				
Distancia con el antecedente en Descripciones				
Distancia en oraciones con el antecedente				
Distancia en dispositivos referenciales con el último FN antecedente				

La delimitación de factores que realizaré depende de los objetivos de mi investigación. Varían con respecto a los planteados por Kibrik (2011) en que, por lo menos en los primeros ejercicios, él trata de predecir el estado de activación, cuando yo trato de predecir todos los Estados Informativos. Sólo en los experimentos del C4 se incluyen frases nominales plenas, pero el trabajo en Kibrik *et al.* (2016) se orienta a predecir la forma del dispositivo referencial, por lo que además, asume un etiquetado minucioso de las relaciones anafóricas. En mi trabajo no exploro la posibilidad de predecir la forma del dispositivo. Además, busco probar que tan sólo con las propiedades lexicogramaticales, transformadas en figuras matemáticas como los vectores, se pueden obtener resultados favorables para la predicción del Estado Informativo de una frase nominal en un discurso, por lo que no integro las relaciones anafóricas entre las frases como información que el algoritmo pueda usar en la identificación.

Podrá planteárseme por qué no realizar un etiquetado más fino, en donde también se integren las relaciones anafóricas y de correferencia. Esta es de las principales reducciones, por

razones técnicas que delinean la presente investigación. Partamos de lo siguiente: la asunción técnica principal es la voluntad para realizar el menor etiquetado del corpus y concentrar los esfuerzos en el etiquetado del Estado Informativo. Esto es debido a que **cada factor considerado implica una nueva capa de etiquetado**, lo que a su vez implica mayor tiempo, trabajo y costo (Kibrik 2011, 466–71). Debido a que mi objetivo es la automatización del proceso completo, de un texto en crudo a un texto etiquetado con Estados Informativos, se busca tener menos pasos intermedios que necesiten sub-etiquetados[43].

Por lo anterior, los factores que puedo integrar en mi trabajo por el tipo de etiquetado que plantearé (cf. §2.6) son los que muestro en la Tabla 4.

Tabla 4. Factores de activación y su inclusión en este trabajo[44]

Factor	**En mi corpus:**
Distancia retórica del antecedente	No etiquetado de RST.
Distancia lineal del antecedente	No etiquetado de Unidades Elementales del Discurso (UED).
Distancia de párrafo del antecedente	La integro parcialmente al proponer conjuntos de oraciones dependientes. Se etiquetan oraciones.
Rol sintáctico y semántico del antecedente	No etiquetado de relación anafórica, pero integrado al evaluar dispositivos referenciales anteriores.
Protagonismo del referente	No etiquetado.

[43] Sumado a esto, una hipótesis que surgió en mi revisión sobre el estado del arte computacional de este tipo de etiquetado parece indicar que el estado informativo, tanto se alimenta de la correferencia como que la ayuda. Es decir, un etiquetado automático de esta propiedad podría elevar la precisión de los identificadores automáticos de resolución de referencia.
[44] Marco en gris los considerados.

Animacidad del referente	No etiquetada.
Identidad difusa	No etiquetado.
Rol sintáctico del dispositivo referencial	Etiquetado.
Rol sintáctico del antecedente retórico	No etiquetado de RST.
Tipo de dispositivo referencial del antecedente retórico	No etiquetado de RST.
Rol sintáctico del antecedente lineal	No etiquetado de UED.
Tipo de dispositivo referencial del antecedente lineal	No etiquetado de UED.
Género del referente.	Etiquetado por Stanza pero sólo del dispositivo referencial (DR).
Persona del referente.	Etiquetado por Stanza pero sólo DR, y a partir del núcleo nominal o el determinante que encabece la frase.
Número del referente.	Etiquetado por Stanza pero sólo DR, y a partir del núcleo nominal o el determinante que encabece la frase.
Tipo de frase del dispositivo referencial (FN, FP, FA, etc)	No etiquetado, sólo etiqueto FN, aunque debido al etiquetado Stanza, podría clasificar las frases por el tipo de palabra inmediata anterior a la FN.
Posición del dispositivo referencial en la cadena referencial	No etiquetado de relación anafórica.
Tipo de descripción	No etiquetado.
Tamaño del antecedente en palabras	No contado debido a que no tengo etiquetado de relación anafórica, aunque puedo obtener el tamaño del dispositivo referencial y los dispositivos referenciales anteriores en una ventana de frases determinada.

Distancia con el antecedente en palabras	No contado debido a que no tengo etiquetado de relación anafórica, aunque puedo obtener el número de palabras al último DRR y último DRP. Tengo una medida con respecto al inicio del texto.
Distancia con el antecedente en Descripciones	No contado, igual caso que el anterior.
Distancia en oraciones con el antecedente	No contado, igual caso que el anterior, partiendo de que sí tengo etiquetadas oraciones.
Distancia en dispositivos referenciales con el último FN antecedente	No contado, igual caso que Distancia con el antecedente en palabras.

Los factores integrados será información que el algoritmo podrá utilizar para la identificación del Estado Informativo, además de la información lexicogramatical interna de la frase. Recordemos que, busco que este método computacional sea una herramienta para detectar que cierta medida —LSA y SPAN en este caso, las cuales explicaré en el Capítulo 2— puede ser tan buena para clasificar el Estado Informativo como el apelar a información de un etiquetado explícito, como el RST.

1.9 Propuestas de los estados mentales de los referentes

A manera de revisión de lo planteado, recordemos que llamo ESTADOS INFORMATIVOS al conjunto de estados abarcados por la identificabilidad, la accesibilidad y la activación. Con ello, asumo que los dispositivos referenciales, en un momento dado en el discurso, codifican la suposición del hablante sobre el estado en el que se encuentra cierto referente en la mente del oyente. A manera de llegar a la síntesis de categorías para los Estados Informativos, recupero en esta sección el recorrido de los autores y sus propuestas para nombrar las

suposiciones de los estados mentales de los referentes en general. Mi investigación partió del examen de *lo dado* como concepto en lingüística por parte de Prince (1981), investigación que la autora sintetizó en su concepto de Familiaridad Asumida y en su taxonomía, la cual muestro en el siguiente diagrama:

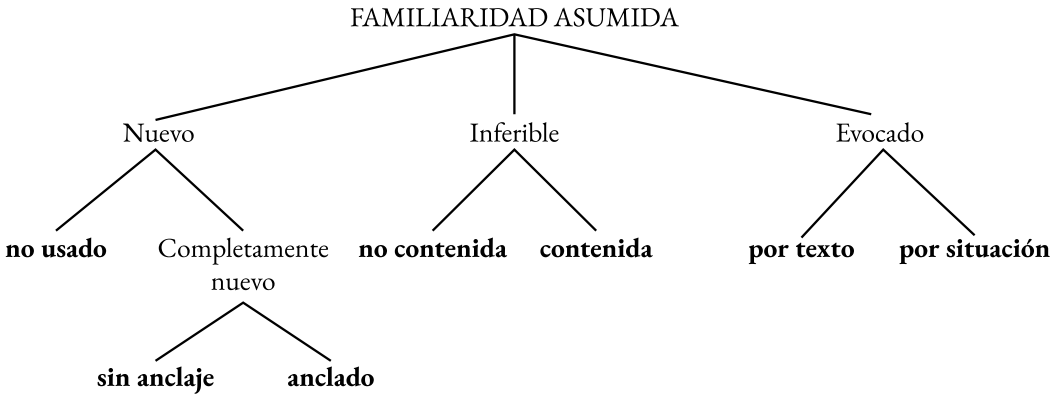


Figura 3. Clasificación de la Familiaridad Asumida[45]

De la perspectiva de Prince (1981), *texto* se entiende como una serie de instrucciones que el hablante proporciona al oyente a partir del juego de suposiciones. Siguiendo la descripción de esta autora, se dice que una entidad es NUEVA cuando "se pone en la mesa"; en algunos casos el oyente tendrá que construir el referente, por lo que será COMPLETAMENTE NUEVO (*brand-new*). En otros casos, el hablante asume que el oyente puede acceder a una plantilla que copiará al discurso vigente, por lo que tal referente es NO USADO. La diferencia entre ANCLADO y SIN ANCLAJE está en que en el primer caso el referente nuevo está relacionado con otra FN que se espera sea inferible o evocada, introducida como modificador o aposición de la primer FN, mientras que en el segundo caso la frase no es modificada por otras frases referenciales. Para los referentes INFERIBLES, Prince (1981, 236) sostiene que es el tipo más

[45] Tomado de Prince (1981, 217); traducción mía.

complejo, el cual describe como el caso en donde el "hablante asume que el oyente puede inferir un referente, vía razonamiento lógico —o, de manera más común, plausible— de las entidades discursivas evocadas en el discurso corriente o de otros inferibles"[46]. A esto último se refiere al dividir entre INFERIBLES NO CONTENIDOS y CONTENIDOS, los primeros, corresponden a la descripción anterior, los últimos los analiza a partir de estructura partitivas como *uno de los N*, en donde el referente señalado por *uno* se construye por su relación conjunto-miembro con el nominal[47]. Si el referente es EVOCADO, puede ser por dos razones: debido a que fue mencionado antes en el TEXTO o porque se encuentra en la SITUACIÓN espaciotemporal del momento de la enunciación.

Años después, Lambrecht (1994) propondrá la siguiente clasificación del estado mental de los referentes. Esta taxonomía es la base para el análisis de la dupla Tópico y Foco y su teoría sobre la estructura de la información:

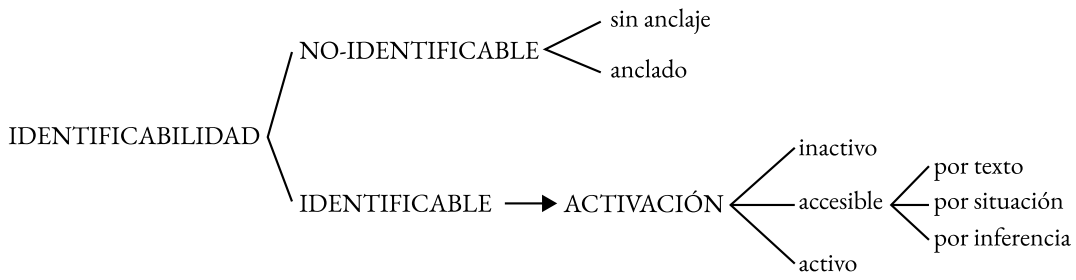


Figura 4. Sistema de identificabilidad y activación de los estados mentales de los referentes[48]

En esta propuesta, Lambrecht (1994), como ya mencioné, apela a la base cognitiva de la activación y la identificabilidad, la primera fundamentada en la consciencia, la segunda

[46] "… the speaker assumes the hearer can infer it, via logical –or, more commonly, plausible– reasoning, from discourse entities already Evoked or from other Inferrables" (Prince 1981, 236).
[47] La decisión de dividir la taxonomía de esta manera en los inferibles no queda explicada del todo por la autora, sin embargo, en mi trabajo no la considero pertinente.
[48] Tomado de Lambrecht (1994, 109); traducción propia.

fundamentada en la memoria o el conocimiento. Retoma a Prince (1981) con respecto a los términos sobre el ANCLAJE de la entidad no-identificable, así como la posibilidad de llamar a estas entidades COMPLETAMENTE NUEVAS (*brand-new*). Algo importante en su desarrollo es que sostiene que es posible analizar la identificabilidad por características lingüísticas: por patrones prosódicos o por partículas gramaticales que formalizan la definitud. Me interesa este último caso, como ya se ha adelantado en las secciones anteriores, por la naturaleza del corpus, pero también, porque no hay una correlación uno a uno sobre la identificabilidad y marcas de definitud que el mismo Lambrecht (1994) admite. Esto es de suma importancia en todo mi análisis: las formas de los determinantes son pistas, no criterios absolutos de clasificación e identificación de los Estados Informativos. Reglas del tipo "FN encabezada por *un* siempre implica que el referente no es identificable" o que "la descripción definida siempre implica activación", no son posibles.

Si el referente es identificable, significa que deberá estar en alguno de los tres estados de activación: ACTIVO, INACTIVO o ACCESIBLE. Lambrecht (1994) sugiere correlatos lexicogramaticales para identificar estas propiedades cognitivas en el texto. De nuevo, como pistas, considera que la activación usualmente se codifica como una expresión no acentuada y en los pronombres. Para los inactivos, la codificación sucede en su prominencia prosódica y, en inglés, en frases nominales léxicas definidas. En cuanto a los accesibles, Lambrecht señala que no identifica correlatos lexicogramaticales, aunque se aventura a decir que podrían encontrarse patrones indirectos en la sintaxis, pero no señala cuáles. La accesibilidad por texto y situación es similar a la de Prince (1981), mientras que a los referentes inferibles los adjunta a la categoría de accesibilidad.

Ambos autores se inspiran en las bases de Chafe (1994) que después Kibrik (2011) también tomará para desarrollar su modelo. En mi caso, como ya he señalado, me interesa una clasificación que pueda aplicarse a toda FN en el discurso, de tal forma que pueda clasificarlas a partir de su INFORMATIVIDAD, la cual defino en este trabajo como la propiedad pragmática asignada a un referente a partir de la manifestación lexicogramatical de la suposición del hablante sobre el estado mental del oyente, ya sea por su grado de activación en la Memoria de Trabajo, por su recuperación en Memoria a Largo Plazo, o por su capacidad de inferir relaciones semánticas.

Mi propuesta de categorías, inspirada en las dos clasificaciones anteriores, la presento a continuación. Apoyado en lo expuesto por Kibrik (2011), busco que tenga una base cognitiva, pero sobre todo, que sea implementable en tecnologías del lenguaje, por lo que no debe considerarse que agota la discusión teórica sobre los estados de las representaciones mentales de los referentes.

1.10 Análisis de los Estados Informativos y la referencia

Como ya adelanté, la referencia es un debate abierto en lingüística, y distintas perspectivas han tratado de darle explicación. Aunque desde las propuestas lógico-filosóficas (Frege 1984; Russell 1905; Hawkins 1978; Strawson 1950) se han desarrollado modelos para explicar la referencia, no es mi objetivo utilizar esta clase de estrategias y conceptos. En este trabajo, entiendo referencia desde una perspectiva pragmática, orientada a la idea de que la lengua está a disposición de fines comunicativos. En este sentido, las estrategias lingüísticas que ayudan a determinar la referencialidad, o su grado, son dos, de acuerdo con Alcina Caudet

(1999, sec. 2.2.4): la capacidad de recuperar a una entidad nombrada por la frase nominal con una anáfora y la dependencia del constituyente a un verbo ya sea como argumento o adjunto. El primer caso ha sido ampliamente estudiado (Givón 1983a; 1983b), y la conclusión es que, aunque existen tendencias para la recuperabilidad de un referente, ninguna estructura morfosintáctica de una frase nominal le restringe de ser recuperada con una anáfora, aunque existe una variación de grado, en donde las anáforas con dispositivos referenciales reducidos tiene un comportamiento más restringido. Por ejemplo, se puede notar a continuación en (26a y b) que un dispositivo referencial pleno como *la madera* o *ese material*, funciona de manera satisfactoria como anáfora de un referente que es modificador de una frase nominal, en este caso *la casa de* ***madera roja***.

(26) La casa de madera roja que estaba en la colonia Roma fue demolida.
Es una lástima porque…
a. la madera fue traída de Brasil (la madera roja de la casa).
b. ese material fue traído de Brasil (la madera roja de la casa).
c. esa fue traída de Brasil (¿la madera roja de la casa o la casa?).
d. **fue** traída de Brasil (¿la madera roja de la casa o la casa?).

Esta funcionalidad disminuye en el caso de un pronombre —en el ejemplo de (26c), con el pronombre demostrativo *esa*— para perder toda su funcionalidad con sólo la marca de sujeto en el verbo de (26d). En estos últimos dos ejemplos, además, existe conflicto referencial debido al género y el número que pudieran ayudar a desambiguar[49]. Para contrastar esto, hágase el ejercicio con *cristales rojos*:

[49] Se debe notar que parte del conflicto que podría suponer esta oración se resuelve al recurrir al marco que nos permite saber que es raro que una casa completa sea traída de Brasil, pero en esta misma línea, podría ponerse el ejemplo con “La capilla de metal blanco que está en Santa Rosalía”, capilla en Baja California Sur que fue traída en su totalidad desde Francia.

(27) La casa de cristales rojos que estaba en la colonia Roma fue demolida. Es una lástima porque…

a. esos fueron traídos de Brasil.

b. **fueron** traídos de Brasil.

Sin el conflicto referencial de género y número, sigue siendo difícil la recuperación con dispositivos referenciales reducidos como en (27a) pero parece rescatarse mejor con (27b). Esto ya ha sido enunciado antes, aspecto que trata Kibrik (2011) a detalle y que resumo en la sección 1.7 y no es objetivo de mi presente trabajo ahondar más en este aspecto.

Lo anterior es sólo para argumentar que **cualquier frase nominal puede introducir un referente**, y que la prueba de anáfora funciona para todos los casos. La prueba no falla en la capacidad de recuperar referentes por anáfora, pero, lo que se observa es que la forma de la anáfora varía. Es, precisamente, en la predicción de la forma de la anáfora en donde se concentran los estudios sobre jerarquía en la accesibilidad (Gundel 1996; Keenan y Comrie 1977; Ariel 1990), además de restringirse, en algunos casos, a dispositivos referenciales reducidos. No obstante, esto está fuera de mi interés en cuanto a que es un problema de resolución y predicción de ciertas de formas para las anáforas.

El segundo criterio lingüístico para determinar que una frase nominal es referencial es que su relación con el verbo de la oración en la que participa sea argumental; esto es que la frase nominal dependa del verbo. Alcina Caudet (1999) sostiene que sólo los argumentos pueden ser recuperados por anáfora, de lo que se sigue que sólo las frases nominales argumentales son referenciales. Esto no parece funcionar así. Nótese de nuevo el ejemplo de (26) con la variación que muestro en (28) a continuación.

(28) El terremoto destruyó la casa de madera roja en la colonia Roma el pasado jueves.

a. Ahí explotó un tanque (en la colonia Roma).

b. Ese será recordado por mucha gente (¿el pasado jueves?).

c. **#Tuvo** una noche muy larga (¿el pasado jueves?).

Si bien, los constituyentes que funcionan como argumentos parecen tener mayor capacidad de ser recuperados que los adjuntos por un dispositivo referencial reducido, ningún referente introducido por frase nominal, incluyendo aquellos que no dependen de una estructura verbal, como ya vimos con *madera roja* que modifica *casa* en el ejemplo (26), está gramaticalmente impedido de ser recuperado por un dispositivo referencial. Nótese que he utilizado dispositivos referenciales reducidos en los ejemplos (28 b y c). El ejemplo en (28b) el dispositivo referencial reducido ese podría recuperar tanto a *el terremoto* como a *el pasado jueves*, pero, de hecho, en este caso, *el terremoto* tiene preferencia por ser sujeto de la oración anterior. Tratar de recuperar *el pasado jueves* en este contexto con *ese* resulta anómalo. Mientras que para el ejemplo en (28c) me parece que ilustra que la lectura pretendida, en la que la marca en el verbo es un dispositivo referencial reducido que recupera *el pasado jueves*, es más anómalo que los otros casos. Nótese que, en todos los casos, un dispositivo referencial pleno como *ese día* es adecuado para recuperar a un adjunto de este tipo. Asimismo, el conflicto que podría suponer el dispositivo referencial reducido en (28a) *ahí* se resuelve al ser un demostrativo de lugar, pero finalmente, un dispositivo referencial reducido.

Sin lugar a duda, ser sujeto de la oración y ser tema de la conversación favorecen una lectura, pero como ya señalé para el caso de la anáfora, me interesan estas pruebas para dar cuenta de cuál frase nominal es referencial o no. Estas pruebas no funcionan para descartar una u

otra frase. Dan evidencia de que **todas las frases nominales pueden introducir referentes en el discurso**, aunque las estrategias para la recuperación sean distintas[50].

Esto discrepa de las posturas según las cuales existen frases nominales que no tienen referencia debido a que señalan propiedades, clases u objetos inexistentes o desconocidos (Leonetti 1990). Pero, desde una perspectiva pragmática, la referencia tiene como principal objetivo traer a la mente del oyente o interlocutor un fragmento delimitado de lo que se habla, lo que puede abarcar propiedades o clases, a través de los mecanismos propios de cada lengua; en el caso de español, a través de una frase nominal. La referencia es un ejercicio de *figuración* (crear una figura). Las entidades a las que esta figuración pueden hacer referencia no necesitan existir o ser verdaderas en los términos lógico-filosóficos, referencia contrastada por lo general con lecturas inespecíficas o que apelan a clases. En lo que sigue, preferiré utilizar el término *figura*[51] por encima del de *entidad* para evitar la connotación sobre algo que existe, aunque en mi análisis lo intercambio con el término *referente discursivo*. En este sentido, el fragmento delimitado o *la figura* se presenta en un ejercicio en donde se destaca *algo* de lo cual se puede hablar. La función principal no es el valor veritativo o la existencia de una única entidad, sino la capacidad de hablar de ese algo con nuestro interlocutor, de poder *manejar esa idea en el discurso* (Hopper y Thompson 1984). Para que funcione este ejercicio, principios pragmáticos de la comunicación deben entrar en juego. El central, ya enunciado en su momento por Grice (1989), es el de cooperación, en el que se infiere que el

[50] En lo planteado por Givón (1982) y Du Bois (1980), la *referencialidad* no se puede juzgar tan sólo por la forma de un constituyente, sino a través de un examen del discurso completo en donde aparece y la manera en que es recuperado para hablar de algo.
[51] Aunque puede relacionarse con las posturas psicológicas de figura/fondo, no es mi intención traer aquí a discusión sus paralelismos.

hablante supone las mismas capacidades generales de atención y "recorte de lo hablado" que su interlocutor. Si se enuncia:

(29) Me encontré *una chabara*

El hablante supone que el oyente puede recuperar información general que le permita entender el contenido de la oración y saber que el hablante tiene intención de destacar y delimitar *algo* que pueda ser parte de la conversación, aunque nunca en su vida haya escuchado la palabra *chabara*.

De acuerdo con Alcina Caudet (1999) otros dos principios gobiernan la recuperación del referente del contexto: el **principio de interpretación local** "que insta al oyente a no construir un contexto más amplio del necesario para llegar a una interpretación" (p. 267)[52]. Lo que supone que si se enuncia algo como:

(30) Me encontré *una chabara*. ***La tapa*** estaba suelta.

Inferimos que *la tapa* es una parte de la *chabara*, sea lo que sea, y no una tapa de repentina aparición.

El otro principio que gobierna el acto referencial es el **principio de analogía** el cual establece que la experiencia que poseemos de determinadas situaciones comunicativas nos enseña lo que debemos esperar de tal situación. Este principio es parecido a lo que se apela con los conceptos de Marco o Esquema, en la medida de que, cuando enunciamos, se espera cierto orden de aparición de entidades, así como la sucesión de ciertos eventos (cf. p. 50).

[52] Estos dos principios son derivados de los presentado en Brown y Yule (1983, 83) que a su vez parten de la máxima de relevancia enunciada por Grice (1989).

El hecho, por ejemplo, de enunciar que *la chabara* tiene *tapa* crea una serie de expectativas consecuentes, como que es algún tipo de contenedor y obedece al esquema general de su uso.

Dada la exposición anterior, estamos en posición de recuperar lo enunciado en las secciones anteriores de este capítulo con ánimos de sintetizar las distintas categorías de estados mentales de los referentes propuestas por Prince (1981), Chafe (1994) y Lambrecht (1994) para operacionalizar el análisis en un algoritmo informático.

Lo que se espera probar en esta tesis es la relación entre una medida obtenida a través de textos procesados de manera semisupervisada y el Estado Informativo. No obstante, el análisis que permite obtener el primer corpus para poder alimentar al algoritmo se realiza de manera manual. Para tal análisis, una pregunta obligada es cuáles son las características lingüísticas que permiten determinar uno u otro Estado Informativo. Trabajos como los de Kibrik (2011) sostienen que los dispositivos referenciales reducidos —pronombres, clíticos y marcas en el verbo— implican que el hablante supone un alto grado de activación del referente. Sobre esto, en lo general comparto este análisis. Por otro lado, los dispositivos referenciales plenos suponen baja activación. Son precisamente para estos dispositivos referenciales en los que planteo la necesidad de una mayor diferencia dependiendo de características morfosintácticas internas, el contexto de la frase nominal y las características morfosintácticas de las otras frases nominales del texto analizado. Por lo que el análisis que propongo utiliza como pivote heurístico el planteamiento del siguiente enunciado dada una frase nominal:

(31) “El hablante *supone* que el oyente tiene en determinado *estado mental* al referente”

El primer paso es segmentar todas las frases nominales y determinar la dependencia de los constituyentes con el verbo. Se analizan oraciones no finitas sólo en los casos en donde presenten constituyentes dependientes. Se marcará al sujeto, objeto directo y objeto indirecto cuando pueda ser reemplazado por clítico pronominal; en los casos en donde el complemento sea introducido por preposición, se colocará una etiqueta en el corpus que indique esta diferencia; aquellos casos en donde la frase nominal no dependa de un verbo y por consiguiente no sea un constituyente que exprese un rol gramatical se asignará *no aplica* (*na*).

Posteriormente, se analiza el estado informativo a partir de esta serie de criterios:

La **No-identificable [1]** como se mencionó en el apartado 1.4 se relaciona con "lo nuevo". Para este análisis se refiere a que *el hablante supone que el oyente no tiene una figura presente o creada por lo que tiene que presentar y construir una en el discurso*. Además, recordemos que No Identificable no implica inespecífico (cf. 1.2.2); es decir, una frase nominal que formaliza la suposición de no identificabilidad de un referente no implica que el referente es inexistente en la realidad. De tal manera, estas frases nominales usualmente son primeras menciones con extensas descripciones, nombres propios con apellidos y títulos asociados, pero en esta etiqueta también se integran algunos casos de frases nominales escuetas como primera mención, en los cuales sí coincide la etiqueta No Identificable con inespecífico.

Por lo que se asigna la etiqueta N**o-identificable [1]** cuando la frase nominal es **primera mención de un referente** y además:

a) tiene más de un modificador (sea izquierda o derecha, sin considerar a los determinantes) como, por ejemplo, las tres siguientes frases nominales marcadas entre corchetes:

(32) **[Las secretarías de Agricultura y Desarrollo Rural (Sader) y de Economía (SE)]**,
[el Gobierno de Sinaloa] y
[productores de tomate del país] sostuvieron una reunión para revisar
[las estrategias a seguir respecto a la negociación que los productores realizan para establecer un Acuerdo de Suspensión con Estados Unidos, mediante la integración de acciones para contribuir a que los agricultores nacionales tengan condiciones equitativas y justas en la exportación de su producto].

COPENOR-369SN[53]

O incluso tan extensas como la que inicia con *las estrategias a seguir…*

b) así también los casos en donde la frase no tiene modificadores, no son introducidas por una preposición pero no están encabezadas por un determinante, como *magistrados*, *jueces* y *regidores* en el siguiente ejemplo:

(33) A la muestra acudieron autoridades de los tres niveles de gobierno, **[magistrados]**, **[jueces]**, representantes de organizaciones civiles, directores del Gobierno Municipal y **[regidores].**

COPENOR-370CH

c) es un nombre propio con más de un modificador, en donde los apellidos de nombres propios y las aposiciones las analizo como modificadores. En el ejemplo (32) tenemos *Las secretarías de Agricultura…* y *el Gobierno de Sinaloa*, pero también se incluyen el nombre de calles y colonias que son introducidas de la siguiente manera:

[53] Con el propósito de exponer el análisis, en gran parte de los ejemplos sólo recurro a la segmentación de los constituyentes que dependen del verbo matriz analizado y no distingo las frases nominales que se encuentren dentro de cada constituyente. Nótese que esto sí lo realizo para el análisis del corpus. Un ejemplo se puede observar en la §1.11.

(34) Fueron tres los hombres ejecutados en una barbería ubicada entre **[las calles Plan de Ayala y Ramón Rayón** de **[la colonia Zaragoza]]**.

COPENOR-186CH

d) no tiene modificadores pero tiene un determinante que no forma parte del paradigma de los definidos o demostrativos.

(35) **[Un adulto]** fue golpeado por varios sujetos en el río, esto durante las primeras horas de hoy, por lo que fue enviado al hospital general Camargo.

COPENOR-162CH

Si es primera mención, pero no tiene modificadores y es encabezado por algún determinante definido se analiza como que *el hablante supone que el oyente puede recuperar al referente de conocimiento general* tal vez conocimiento compartido entre sólo ellos o conocimiento compartido por pertenecer a la misma comunidad. En el siguiente ejemplo, aunque se menciona *municipales* en otras secciones anteriores de la nota, nunca se indica de qué *comunidad* y *ciudad* se habla. Esto se infiere por la lectura del periódico, por lo que el hablante supone que el oyente puede recuperar este conocimiento de un contexto más amplio, para lo cual debe recurrir a su memoria a largo plazo.

(36) Como es prioridad en este Gobierno Municipal y por instrucción del Alcalde Armando Cabada Alvídrez, la dependencia continúa con estos operativos en diversas vialidades de **[la ciudad]** como un compromiso con **[la comunidad]**, además de implementar programas que tienen como objetivo el mejoramiento urbano.

COPENOR-061CH

En estos casos etiqueto estas frases como **Inactivo en Memoria a Largo Plazo [2]**.

Por el momento no describiré la etiqueta **Inactivo por Registro Discursivo [3]** debido a que primero atenderé los casos de **primera mención**. Las dos anteriores pertenecen a esta

situación y las dos siguientes. Existe una variación de análisis para frases de este tipo y son los casos en donde el referente es construible por algún marco dispuesto en el contexto lingüístico (como una anáfora asociativa). Estos casos los analizo en conjunto con la etiqueta **Accesible por Marco [4]**, como, por ejemplo, la siguiente nota (COPENOR- 328SO) inicia de esta manera:

(37) Con **[lesiones en nariz, mano y dorso]**, resultó Ana, luego de ser golpeada por su hermano en la colonia San Pablo, por lo que se trasladó ella misma a **[un hospital]** a recibir atención médica.

Para más adelante, terminar con la siguiente oración:

(38) **[Los médicos]** esperaban los resultados de **[las radiografías]** para el diagnóstico de las heridas

COPENOR- 328SO

No se había mencionado *los médicos* antes pero el conocimiento sobre un *hospital*, las *lesiones en nariz, mano y dorso* y el guión sobre lo que sucede en esos lugares con ese contexto vuelve accesible la figuración de un referente como *los médicos* que revisan *las radiografías* sobre *las heridas*. Aunque debe tomarse en cuenta que *las heridas* son segunda mención de una misma figura, en este caso, la que introdujo *lesiones* mencionada al inicio de la nota. Para esta etiqueta establezco que *el hablante supone que el oyente puede construir al referente por la coherencia con el guión o marco del tema dispuesto en el discurso*.

Por otro lado, también existen casos en donde se presentan frases nominales que hacen referencia a fechas con o sin modificadores y demostrativos determinantes con lo que apelan

a que *el hablante supone que el oyente puede recuperar el referente por el contexto inmediato u origo* como en el siguiente ejemplo[54]:

(39) La tarde de **[este jueves]** la alcaldesa Maru Campos acudió a las instalaciones de las oficinas de la Red por la Participación Ciudadana…

COPENOR-353CH

Estos casos los analizo bajo la categoría **Accesible por Origo [5].** También existe la posibilidad de crear una etiqueta llamada Activo por Origo, que se refiere a los pronombres de las personas que intervienen en el acto de habla y el uso de demostrativos plenos que son acompañados con gestos, no obstante, debido a que son dispositivos referenciales reducidos no los trato en esta investigación.

Todo lo anterior sucede al analizar frases nominales en las que se puede detectar que mencionan por primera vez un referente. Para aquellos casos en donde se detecta que es la **segunda mención (o más) de un referente** se siguen los siguientes pasos.

Si la frase nominal tiene un referente que se menciona más allá de la oración inmediata anterior *el hablante supone que el oyente puede recuperar al referente del registro discursivo* lo cual, se refiere a un segundo tipo de memoria, distinta a la necesitada para recuperar referentes de conocimiento general. Esta memoria está entre la memoria a largo plazo y la

[54] Un aspecto que se mencionó en la 1.6 es que la accesibilidad por origo puede ser más compleja. En el análisis que se realizó al COPENOR se descubrió que es casi nulo el uso de frases nominales encabezadas con demostrativos determinantes como primera mención para hacer referencia a alguna otra cosa que no fuera una fecha. Me parece que la distinción se encuentra entre el uso simbólico de los demostrativos y el uso contextual inmediato. Esto se podría explorar en otros trabajos. Para más información de los usos de los demostrativos se puede consultar Levinson (2004) y Diessel (1999).

memoria de trabajo (MT)[55]. Para el caso de suposición del registro discursivo lo etiqueto como **Inactivo en Registro Discursivo [3]** y son los casos de referentes no mencionados en la oración inmediata anterior. En el siguiente ejemplo, *Pulido Medrano* o *la doctora* ha sido mencionada en distintas ocasiones.

(40) … explicó **[la doctora]**. Asimismo, **mencionó** que en el caso de Los Cabos, no se cuenta con suficiente personal de verificación, motivo por el cual la cobertura no es tan extensa como en el caso de la capital de Baja California Sur. "En Los Cabos **hay** pocos verificadores, hay veces que la cobertura no **es** tan amplia como en La Paz", recalcó **[Pulido Medrano]**.

COPENOR-019BS

Para analizar el Estado Informativo de *Pulido Medrano* recurrimos a observar la oración inmediata anterior. Esta habla sobre *la cobertura*; luego un existencial *hay veces*, para después hablar sobre *Los Cabos*. La anterior a esa trata sobre la poca cantidad de personal de verificación. No es hasta el verbo *mencionar* en donde la marca en el verbo, un dispositivo reducido, la recupera de otra mención, dispositivos que, como señalé en el marco teórico de mi trabajo, no etiqueto. A esta distancia, se analiza como una entidad que forma parte del registro discursivo.

Por otro lado, si es mencionado en la **oración inmediata anterior (OIA)** entonces se encuentra en el rango de suposiciones de activación. Estas suposiciones son las que cuentan con mayor atención (Kibrik 2011), con distintas jerarquías y propiedades a evaluar para el análisis. En el caso de este trabajo sólo distinguiré tres categorías, en donde la estructura

[55] Un análisis más detallado sobre los tipos de memorias podría ayudar a crear un correlato psicológico de otro subtipo que me parece entra en función. Me refiero a aquellos casos de conocimiento que no es general pero consabido de manera íntima por parte de los interlocutores. No obstante, por el tipo de texto que trabajo en esta tesis no me parece que esta variación sea pertinente.

interna de la frase nominal no influye como en los otros casos. Las tres distinciones que realizaré son: referente activo a través de una frase nominal que formaliza (i) al sujeto de la OIA; (ii) al objeto directo o indirecto de la OIA o (iii) es una frase nominal introducida por una preposición en la OIA. Estas tres distinciones las explicaré con mayor detalle a continuación. Para estas tres, sigo el mismo enunciado de análisis: *el hablante supone que el oyente tiene en su memoria de trabajo al referente.*

Distinguiré entre si es **sujeto** de la oración inmediata anterior como en la frase *Los supuestos responsables…* presente en el ejemplo de (41) a continuación. Esto incluye a los sujetos tácitos de verbos en formas no finitas. Para el análisis, se busca el verbo inmediato anterior y se analiza su estructura. En el siguiente ejemplo, el sujeto del verbo *trasladarse* se refiere al sujeto del verbo *utilizaban* que se refiere a su vez a los *probables integrantes de una banda…*:

(41) Probables integrantes de una banda de extranjeros, acusados de despojar de dinero en efectivo a una mujer en la colonia Los Pinos, supuestamente ofreciéndole un boleto premiado del sorteo Melate, fueron detenidos cuando se encontraban hospedados en un hotel de la avenida Tecnológico y **utilizaban** un auto Volkswagen rentado con placas de Puebla, para **trasladarse** por la ciudad.
[Los supuestos responsables, un hombre y tres mujeres adultas], habrían despojado de más de 30 mil pesos a una mujer

COPENOR-371CH

Esta frase nominal la etiqueto como **Activo S [6].**

O si es objeto directo o indirecto de la oración inmediata anterior a partir de la sustitución por clítico. Para este caso utilizaré la etiqueta **Activo O [7]:**

(42) …cuando un presunto asaltante golpeó brutalmente a **[la encargada de una tienda de deportes]** en conocido centro comercial. **[La víctima]** es la señora Elvia Ochoa Cejudo…

COPENOR-052SO

En este ejemplo *la víctima* es etiquetada como Activo O debido a que su referente se encuentra activo, a través de su mención en la frase *la encargada de una tienda de deportes* que expresa el objeto directo de la oración inmediata anterior.

Para los casos en los que la frase nominal es introducida por preposición en alguna parte de la oración inmediata anterior la etiquetaré como **Activo P [8]**:

(43) …por su parte SADYFIN se impuso 6 goles a 0 a Los Cabos F. C., quedándose con el tercer lugar de **[el torneo]**. En la premiación fueron reconocidos los siguientes jugadores por su destacada participación durante **[el torneo]**…

COPENOR-170BS

En el ejemplo anterior, la primera mención de *el torneo* es introducida con la preposición *de*. Luego, en la siguiente oración, otra mención, introducida por la preposición *durante*, hace referencia a este mismo torneo, por lo que para ese momento se trata de un referente activo cuya última mención se encuentra en la oración inmediata anterior introducida por una preposición. Debe notarse que, en mi análisis, aunque etiqueto el hecho de que la segunda mención sea introducida también por preposición, es casualidad que coincida el que ambas sean introducidas de esta manera: la segunda mención pudo haber sido sujeto de la oración, pero es el estatus de la mención del antecedente el que tomo en cuenta.

De nuevo, los tres casos anteriores obedecen el mismo enunciado de análisis de Estado Informativo, y sólo etiqueto de esta manera para dar cuenta de qué tanto esta variable interviene. Esto lo destacaré en los resultados y las conclusiones.

Finalmente, distinguiré en los no-identificables una diferencia que me parece pertinente. Se trata de las frases nominales que son introducidas por preposición, sin determinante, sin modificadores y que no son nombres propios, estructura que contrasta con la señalada en los ejemplos de (33). A este conjunto le asignaré la etiqueta **No-identificable baja [0]** con lo que me refiero a que, si bien, introducen un referente discursivo, me parece que son las más difíciles de recuperar por una anáfora. Por ejemplo, en el siguiente fragmento tenemos *menores*, *salud* e *insumos*.

(44) Las muertes de **[menores]** dentro del Hospital General de Tijuana se han dado a causa de diversas complicaciones de **[salud]** y no por la falta de **[insumos]**, aclaró del director de esta institución, Clemente Zúniga Gil

COPENOR-013BC

Cabe recordar que, como en los primeros ejemplos, estos fragmentos forman parte del primer párrafo de la nota, por lo que se asumen como primeras menciones.

También distingo aquellas frases nominales que son dos o más menciones en una oración de un mismo referente, por lo que, en teoría, su Estado Informativo se encuentra activo. De tal manera, la primera es la mención y las subsecuentes son descripciones. Estas otras las etiqueto como **Identificable baja [9]**, aludiendo a su baja capacidad de activar un referente por sí solas en la oración analizada. Para este caso tenemos:

a) Aquellas frases nominales en cópulas que no son sujetos. Debe notarse que doy preferencia al orden SV para analizar al sujeto ya que en casos como el ejemplo que presento a continuación, ambos constituyentes son intercambiables:

(45) [La víctima] es **[la señora Elvia Ochoa Cejudo, de 53 años de edad]**, quien se encuentra en la clínica San José, donde es intervenida quirúrgicamente, informaron familiares cercanos.

COPENOR- 052SO

b) Las frases nominales aposicionales, como en (46), en donde también incluyo las menciones de nombres en acrónimo y siglas, como en (47). En el caso de las aposicionales, a menos de que se trate de un nombre propio, la segunda frase es la que etiqueto como Identificable baja. En el siguiente ejemplo presentó un caso en donde el segundo elemento es un nombre propio, por lo que el primer elemento *el gobernador del estado* es la Identificable baja[56].

(46) **[El gobernador del estado]**, [Carlos Mendoza Davis], instaló formalmente el Consejo Estatal de Protección Civil para la Temporada de Ciclones Tropicales del Pacífico 2019.

COPENOR-014BS

(47) En sesión ordinaria de [el Instituto Estatal Electoral] (**[IEE]**), los consejeros locales aprobaron el acuerdo donde se especifica la aplicación de sanciones a los diferentes partidos políticos que están constituidos en Sonora.

COPENOR-173SO

Con lo anterior, el conjunto de categorías luciría de la siguiente manera (Figura 5). El número que se coloca entre corchetes se utilizó como referencia en las bases de datos del trabajo computacional.

[56] Esta estrategia de análisis fue constante a lo largo de COPENOR. De los dos elementos yuxtapuestos, aquel que fuera el nombre propio era el que analizaba como núcleo, del cual dependía la aposición.

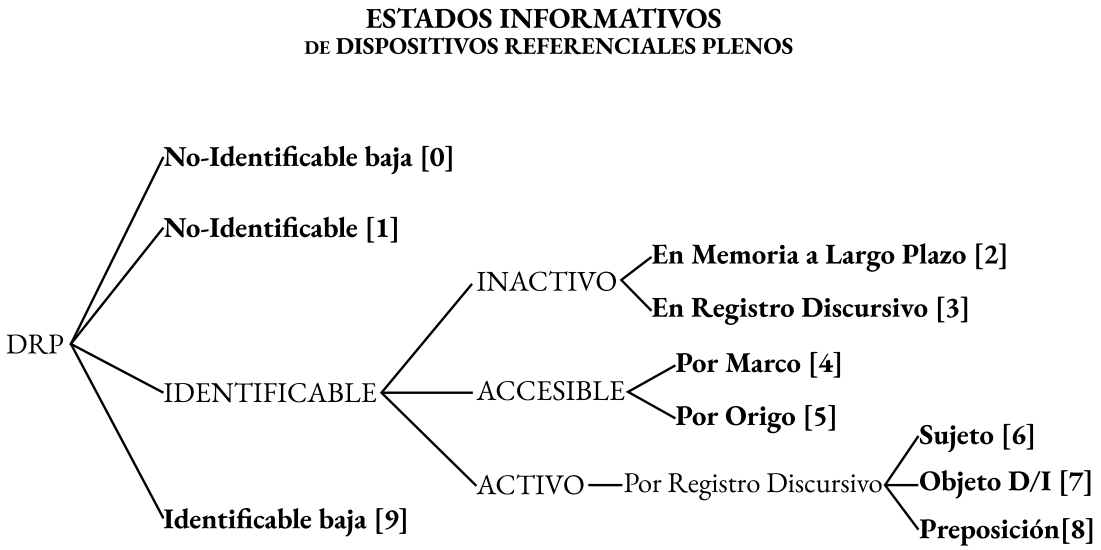


Figura 5. Categorías para el etiquetado de Estados Informativos en frases nominales plenas

1.11 Ejemplo de análisis

Para ilustrar lo desarrollado, muestro un ejemplo de análisis a continuación. En el Anexo B podrá encontrarse la nota COPENOR-253BC completa analizada en esquema XML. De ella tomaré un fragmento para ilustrar el procedimiento siguiendo los lineamientos expuestos. Partamos de la primera oración:

(OR1) 1. [La demanda de [estudiantes que [eligieron]^{VS} a [el CICESE]^{2} a través de
2. [los programas de [Verano de la Investigación Científica]^{1}]^{1}]^{1}]^{1} **[aumentó]^{VM}**
3. casi a [el doble]^{0} en relación a [el año pasado]^{5},
4. a [el [ser]^{VS} [36 estudiantes de [licenciatura]^{0}]^{8} [los seleccionados que [cuentan]^{VS}
5. con [el apoyo de [la Academia Mexicana de Ciencias]^{1} [(AMC)]^{9} y
6. [el Programa Delfín]^{1}]^{1}]^{9}]^{0}.

La primera estrategia que puede notarse en el análisis es que rompo las contracciones del tipo *al* y *del* como se observa en la L(ínea) 1, 3 y 4. Ahora, partiendo del verbo matriz *aumentó*

de la L2, podemos distinguir que la primera frase nominal argumental inicia en *la demanda* y termina en *científica*. No hay Registro Discursivo en este punto por lo que tendremos que evaluar si se apela a Memoria a Largo Plazo. Debido a que no es primera mención de un nombre propio ni un acrónimo sin modificador, no es el caso; 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 demostrativo determinante; por lo mismo que es la primera frase mencionada en la nota, queda descartado el que se vincule con algún referente que pueda ser figurado por algún marco o esquema. Por lo anterior, se analiza como **No-identificable [1]** y se coloca este número como superíndice al final de los corchetes que la abarcan. Examinemos ahora las frases nominales que no dependen de manera directa del verbo matriz. La siguiente frase inicia con un nominal escueto *estudiantes* y una relativa con el núcleo verbal *eligieron* que termina en *científica* en la L2. El mismo examen de la frase anterior lo llevamos a cabo para la frase nominal *estudiantes* y llegamos a la misma conclusión, es **[1]**. La siguiente frase nominal es *el CICESE* en L1, la cual es un acrónimo como primera mención. Por la decisión metodológica que he tomado anteriormente, esta frase nominal encabezada con artículo definido, la etiqueto como **Inactivo en Memoria a Largo Plazo [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, por lo 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* en contraste con *Juan Pérez de la Oz*. Esta pista es

la que me orienta a descartar la suposición de recuperación por MLP. Debido a que ninguna de las otras opciones para referentes no antes mencionados es analizable, la etiqueto como [1]. En este punto del análisis, 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 a el CICESE a través de los programas de Verano de la Investigación Científica
3. el CICESE
4. los programas de Verano de la Investigación Científica
5. Verano de la Investigación Científica

Continuado con el análisis, la siguiente frase nominal que se nos presenta es *el doble* en la L3. De la misma manera que algunas propiedades introducidas por preposición, así como las marcas discursivas *por [lo anterior]* o algunas frases temporales como *por [el momento]* las etiqueto como **No-identificable baja [0]**. Luego tenemos la frase *el año pasado*, su modificador nos indica que es necesario acudir al Origo para determinar el referente de esta frase nominal (el año pasado con respecto al año 2019 –año de publicación de la nota). Este referente no fue mencionado antes, por lo que la evaluación pertinente parece ser la **Accesibilidad por Origo [5]**. Aunque *el doble* tiene baja capacidad de figurar 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 reflejado 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 las analizo como

No-identificables baja [0]. La FN *36 estudiantes de licenciatura* la analizo como referente mencionado en la oración inmediata anterior con estudiantes que eligieron… –en donde el verbo *aumentar* es núcleo– pero introducida por preposición, por lo que le asigno **Activo P [8]**. La siguiente frase es *licenciatura* que en este caso al ser un nominal sin modificadores, pero modificando a otro nominal, lo marco como **No identificable baja** por lo que se marca como **[0]**.

La frase *los seleccionados que cuentan.…* es un constituyente que tiene la última mención de su referente en la misma oración, por lo que se etiqueta como **Identificable baja [9]**. En este punto, existe un referente con tres tipos de menciones:

1. estudiantes que eligieron a el CICESE a través…
2. 36 estudiantes…
3. los seleccionados que cuentan…

Dentro de esta última frase, encontramos *el apoyo de.…*. Siguiendo con el análisis propuesto, es una frase nominal con determinante y modificadores por lo que la etiqueto como **No-identificable [1]**. 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 un nombre propio extendido, y luego tenemos su acrónimo en una posición *(AMC)* que etiqueto como **Identificable baja [9]** debido a que está mención está adpuesta a la última mención del mismo referente. 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, de acuerdo con lo planteado en el modelo, sólo aquello mencionado en la última oración se encuentran activos en la MT. Es decir, aquello que forma parte de las dependencias que ejerce el verbo *contar con* que son el referente discursivo 2, 9, 10 y 11. El referente 2 sería el único que en ese momento del discurso está activo debido a la marca que se encuentra en el verbo –un dispositivo referencial reducido.

1.12 Teoría de semántica latente

En lo que sigue, quisiera exponer la pertinencia en este trabajo de la propuesta de Thomas K. Landauer (2007) sobre una teoría de semántica vectorial. Este autor ha presentado, a lo largo de investigaciones que comprenden de los años 90's del siglo pasado a la primera década de este siglo, que es posible entender la representación matemática de dimensiones como una forma de teoría semántica. El Análisis de Semántica Latente (LSA por sus siglas en inglés) no sólo es una herramienta de procesamiento de lenguaje natural que técnicamente ha mostrado resultados en distintos campos, sino que, además, puede ser usada para retratar el funcionamiento de la lengua e incluso de la mente. El principal argumento que sostiene,

desde su trabajo inicial en Landauer (1997), es que LSA ofrece una explicación para rebatir la idea de la *pobreza de estímulo* de Chomsky (1986, xxv):

> LSA does just what Chomsky thought impossible. It acquires linguistically and cognitively effective, shared, relationally embedded, representations of word meanings without any preexisting specific knowledge. And it does so by learning entirely from experience (Landauer 2002, 69).

En el apartado metodológico explico con más detalle en qué consiste esta técnica, pero por lo pronto será suficiente mencionar que la manera en que LSA captura el significado es a través de calcular las coapariciones léxicas. Después de mostrarle miles de posibles ubicaciones de una palabra, LSA logra predecir relaciones y similitudes sin necesidad de enseñarle reglas o estructuras. En este sentido, *experiencia* se entiende como un conjunto de apariciones de un ítem léxico y contextos que pueden ir de los miles a los cientos de miles. El que Landauer haya presentado evidencia de que es posible que emerjan relaciones semánticas sin necesidad de conocimiento previo, aunque sea por medio de métodos computacionales, resulta interesante pero no me concentraré en este punto. Me interesa más la propuesta en la que el significado de las palabras se entiende como:

1. Resultado de un conjunto de experiencias individuales;
2. Diferenciados a partir de las dimensiones dadas por las palabras en contexto;
3. Susceptibles a reducción dimensional de tal manera que dos conjuntos distintos de experiencias individuales iniciales puedan tener similitudes una vez reducidas.

Especialmente, me parece que en lingüística resuena la idea de una matriz de rasgos cuyo objetivo no sea dar cuenta de la naturaleza de los rasgos sino de la capacidad de la matriz de diferenciar las unidades que se presentan. Este método ha tenido sus propios problemas, como, por ejemplo, al tratar de definir los rasgos pertinentes para las diferencias de varios

campos semánticos, el análisis puede complejizarse hasta volverse ininteligible. Pero es justo este problema lo que se convierte en una virtud al momento de asistirnos por una máquina. La idea inicial de LSA es tener precisamente una matriz de alta dimensionalidad (miles de dimensiones). Los rasgos iniciales son las palabras en el contexto en donde aparece la unidad que estamos analizando. En este punto, como mencionaré en la metodología, es irrelevante el orden de palabra: lo importante es la coaparición de los elementos. En un paso posterior sucede la reducción de la dimensionalidad y la verdadera capacidad de LSA emerge. Las operaciones matemáticas para la reducción permiten crear un conjunto pequeño de rasgos abstractos, mínimos y suficientes para diferenciar el sistema que estemos analizando, pero a su vez, conservar relaciones de cercanía o similitud entre aquellas unidades que así se comporten.

Tal y como menciona Landauer (2002, 65), este tipo de semántica no detecta las diferencias entre las oraciones (48a y b):

(48) a. John hit Mary
b. Mary hit John
c. John did not hit Mary

Pero sí, el que dos entidades llamadas *John* y *Mary* están involucradas en un altercado. Además, que, entre las dos anteriores y (48c), la diferencia detectada sería mínima ya que LSA no crea inversiones con la presencia de negaciones. No obstante, aunque no se apele a un orden de palabra y haga falta reglas para otras operaciones semánticas, aun así, insiste Landauer, se obtienen resultados satisfactorios en sus distintos análisis, lo que da pista de que, aunque relevante, el orden de palabra no es definitorio; en ciertos casos incluso llega a

ser redundante. Ahora bien, debe recordarse que las investigaciones de Landauer son en inglés. Tendrá que corroborarse este comportamiento en otras lenguas.

Los rasgos de una matriz de LSA no son del tipo [±humano] o [±animado], sino rasgos abstractos cuyo objetivo es establecer la diferencia entre las unidades analizadas a partir de sus contextos de aparición. Lo crucial no es identificar qué es lo que los hace distintos, sino establecer la distinción. “In LSA, words do not have meanings on their own that define the axes, words get their meanings from their mapping” (Landauer et al. 2007, 8). Y es este punto el que me parece importante de esta propuesta, nada separado de lo que ya otros autores destacaban de un sistema lingüístico a partir de una teoría de la oposición. Por ejemplo, la propuesta de Ferdinand De Saussure se basaba en el principio de *différence* en donde se sostenía que, en la lengua, “como en todo sistema semiológico, lo que distingue a un signo es todo lo que lo constituye. La **diferencia** es lo que hace la característica, como hace el valor y la unidad” (de Saussure 2005, 145, énfasis propio).

Por su cuenta, el príncipe Trubetzkoy mencionaba que para el sistema fonológico:

> El concepto de diferenciación presupone el concepto de *contraste*, de *oposición*. Una cosa solo puede ser diferenciada de otra cosa, y ello, en la medida en que la una se pone frente a la otra o contra la otra, es decir, en la medida en que entre ellas existe una relación de contraste o de oposición (Trubetzkoy 2019, 61).

Seguido de Roman Jakobson quien, a través de ilustrar el trabajo del semiólogo americano Charles Sanders Pierce, sostuvo que la clasificación natural se lleva a cabo en dicotomías; la relación diádica básica es la oposición: “una cosa sin oposición *ipso facto* no existe” (Jakobson 1980, 35).

De acuerdo con Danesi (2009), la razón por la cual no se siguió profundizando en esta teoría de la oposición en la investigación lingüística fue por la atención que obtuvieron otras teorías

a lo largo de la segunda mitad del siglo XX. Destaca en este periodo el trabajo inicial de Noam Chomsky (1965) y la corriente que dio lugar a una lingüística cognitiva (Lakoff y Johnson 2017). Danesi (2009) menciona que, una teoría de la oposición tiene su fundamento cognitivo en una capacidad general que funciona para distintas tareas, y no sólo para el lenguaje. Esto resulta antitético para la perspectiva de Chosmky en donde un módulo especializado del lenguaje es fundamental. Por otro lado, entre la lingüística cognitiva y una teoría de la oposición, no existen realmente contradicciones ni contrargumentos directos entre ellas, por lo que se espera sean compatibles (Danesi 2009, 12).

Asistido por computadoras, una semántica vectorial sólo buscaría establecer las diferencias pertinentes entre las palabras a partir de la coaparición de otras palabras en su contexto. Esto provee al algoritmo de *experiencia* que después puede sintetizar en un proceso llamado Reducción a Valores Singulares (SVD por sus siglas en inglés) que explicaré en el Capítulo 2, § 2.3.5. En la reducción, se establecen nuevos rasgos, abstractos, que no se pueden nombrar, pero que permiten establecer la diferencia y su grado. Que el objetivo sea sólo establecer la diferencia es pertinente en una lingüística que se fundamenta en una teoría de la oposición. Con el método de Landauer, lo que se ha hecho es capturar una complejidad que, por métodos tradicionales, hubiera resultado imposible realizar. De esta manera, se puede crear un mapa que permita relacionar unidades, lo que nos recuerda que una forma válida del estudio del significado puede ser también el análisis de la relación de los signos en el sistema desde donde se enuncia.

Por lo expresado hasta este momento, pienso que vincular el principio sobre la diferencia con el método de LSA es posible y válido. No por ello, y en esto sólo hago eco de lo que Landauer ha mencionado en distintos trabajos (Landauer et al. 2007; Landauer 2002), se debe suponer

que LSA da lugar a una teoría semántica completa. Esto no es así, y el mismo autor sostiene que esto no es posible por lo menos considerando los problemas de contextualización (*grounding*) y corporeización (*embodiment*)[57]. No obstante, pienso que da lugar a maneras de volver concretas y comprobables (susceptibles a experimentación) ciertas intuiciones de cómo se maneja el significado. Esto da lugar no sólo a teorías sino a métodos que pueden ser compartidos en la comunidad para ser probados, lo que las ayuda a eclosionar de la reflexión epistemológica. Como ya he mencionado antes, el objetivo de esta tesis está orientado a probar sólo si una propiedad pragmática, como lo es la suposición del hablante sobre el estado de un referente en la mente del oyente, puede ser capturada por la metodología asociada a LSA. Y aunque este método no surge en un contexto de teoría lingüística, sí puede ayudar a fundamentar principios lingüísticos, como son la teoría de la oposición, por lo menos desde una arista que no me parece incoherente: la noción básica en que el significado es resultado de relaciones de contraste y oposición, en donde lo relevante son las relaciones entre los puntos.

> It is important to understand that in LSA, as in a map, the coordinates are arbitrary. North and south are conventionally used for the earth, but the relation of any point to any other would be just as well located by any other set of nonidentical axes (Landauer et al. 2007, 7).

Investigaciones anteriores (Hempelmann et al. 2005; Graesser et al. 2007; McCarthy et al. 2012) han encontrado que algunas diferencias que pertenecen al nivel pragmático pueden capturarse con este método, lo que extiende la aplicación de la propuesta de Landauer, la cual inició en el nivel lexicosemántico.

---

[57] De manera particular, la investigación que presento podría servir en un futuro para dar cuenta de los alcances de LSA para notar la codificación de las suposiciones sobre Estados Informativos como el Inactivo, y también aquellos que son Accesibles por la situación inmediata del momento de la enunciación.

Sin lugar a duda, esta perspectiva conlleva planteamientos que deberán ser explorados en su momento y que escapan a los presentes objetivos de investigación. No propongo una hipótesis sobre esto, pero creo que se podrían establecer en futuros trabajos de Lingüística Computacional. En todo caso, tomo lo anterior como un supuesto, que se expresaría de la siguiente manera: LSA no es sólo un método sino también una perspectiva de qué es el lenguaje; visión que descansa en el principio lingüístico de la diferencia.

En este capítulo, he buscado dejar claras cuáles categorías de análisis utilizaré en mi análisis, llamadas Estados Informativos, así como los factores que busco integrar para el análisis computacional del corpus. En esta última sección, he buscado fundamentar de manera lingüística el método de representación vectorial, el cual dará como resultado un número que, se hipotetiza, guarda una relación con los Estados Informativos. De esta manera, continúo con la exposición de los antecedentes para el análisis de información nueva/dada desde el procesamiento de lenguaje natural, para luego explicar a detalle el método LSA y las variaciones que implementaré en mi propio algoritmo, y dar pie al núcleo de mi análisis computacional.

Capítulo 2 Metodología para el etiquetado automático de Estados Informativos

First, we need to make sure that they understand what a question is.
The nature of a request for information along with a response.
—Dr. Louise Banks, *Arrival (2016)*, 42:30

2.1 Introducción

En el capítulo que sigue trato los aspectos metodológicos de esta tesis. En §2.2 empiezo con una rápida revisión sobre los antecedentes en tecnologías del lenguaje que buscan detectar la información nueva/dada en textos y los intereses que motivan investigaciones de este tipo. Luego, hablo sobre aquellos algoritmos que han buscado implementar alguna teoría lingüística en búsqueda de diversas propiedades relacionadas con lo nuevo/dado. En particular, hablo sobre la detección automática de tópico/foco, resolución de correferencia de manera automática y el etiquetado de las funciones comunicativas asociadas a la definitud. Continúo con algunos antecedentes del Análisis de Semántica Latente y la variación propuesta por Hu et al (2003). Con la intención de volver lo más transparente posible el método, en §2.3 explico cómo realizar el análisis LSA paso por paso; luego, hablo sobre algunos procesos de afinación, como el preprocesamiento del corpus en Stanza (la paquetería de Python desarrollada por la Universidad de Standford para realizar procesamiento de lenguaje natural), las variaciones de las bolsas de palabras para crear las matrices y el uso del Diccionario del español de México (§2.4). Esto me lleva a dar detalles sobre el diccionario y su capacidad de ser un inventario de sentidos que alimente la información contenida en las frases nominales (§2.4.3). Expongo de manera abreve el origen del diccionario, la creación lexicológica de las entradas para finalmente mencionar la manera de incluirlas en el

algoritmo. En este punto también muestro lo complejo de tratar de desambiguar acepciones, la polisemia capturada en el diccionario y posibles maneras de solucionarlo en mi algoritmo; aunque al final, opto por un método que se guía más por factores prácticos. En §2.5 hablo sobre la variación de LSA llamada SPAN, y al igual que el primer método, explico paso a paso cómo obtener las medidas. Durante todo este proceso, trato de mostrar poco a poco cómo se va complejizando el algoritmo. Es al final de esta sección en donde muestro un penúltimo esquema del recorrido al que se vería sometida una nota en crudo. No obstante, las notas entran a un tratamiento previo: un etiquetado manual en XML de frases nominales y oraciones, así como algunas propiedades sintácticas y de los respectivos Estados Informativos. Esto lo explico en §2.6 en la creación del Corpus Periodístico del Noroeste de México (COPENOR). En esta sección también hablo sobre la naturaleza de las notas periodísticas, la distribución de las notas por estado, la muestra inicial y el conjunto final que tomo para mi análisis. Finalmente, en §2.7 explico los conjuntos de etiquetas que construyo para realizar los experimentos, los factores que tomaré en cuenta en las pruebas probabilísticas, así como una breve exposición de cada una de ellas. Presento primero aquellas pruebas realizadas en investigaciones anteriores, contra las que compararé mis resultados, las cuales son matrices de correlación, análisis de regresión logística múltiple y análisis de varianza de una vía (ANOVA). En un segundo momento, realizo algunas pruebas para determinar la normalidad de los datos y confirmar si las pruebas paramétricas son pertinentes. Si bien, estos resultados me ayudan a la comparación con investigaciones anteriores, también incluyo un análisis de varianza de datos no paramétricos a través del método Kruskal-Wallis, que a su vez se puede implementar con el análisis de relevancia estocástica de Conover-Iman. Todo lo anterior me permitirá observar si las etiquetas propuestas son, en efecto, grupos analizables, así como el peso de las medidas como

predictores de los Estados Informativos. Finalmente, debido a que mi interés es que mi investigación fundamente la posibilidad de crear un etiquetador automático, cierro esta sección y el capítulo con la mención de los bosques aleatorios de decisiones (*random forests*) debido a que los utilizaré para probar la capacidad de las medidas como clasificadores.

2.2 Lo nuevo/dado en tecnologías del lenguaje

Supongamos que a una máquina le decimos lo siguiente a través de una interfaz:

(49) Juan golpeó a Pedro. Él nunca lo perdonó.

Entre lingüistas planteamos de manera cotidiana preguntas a los textos con el fin de evaluar, por ejemplo, relaciones de concordancia y referencia. Para que una máquina pueda realizar esta tarea, debe ser capaz de detectar un conjunto de propiedades que en ocasiones los lingüistas obvian. Por ejemplo, si quisiéramos saber ¿quién no perdonó? La computadora tendría que ser capaz no sólo de resolver la referencia del pronombre sino de establecer un marco semántico que permitiera inferir que las víctimas son las que son propensas a no perdonar un daño; e incluso, tendría que ser capaz de algo más básico, algo tal vez más obvio en esta clase de planteamientos: determinar qué es una entidad y sus correferencias (sinónimos, apodos, otros nombres). En este caso, en (49) sólo se encuentran dos pequeñas oraciones concatenadas como información base. Dado este antecedente o conocimiento previo, se le presentaría a la máquina una pregunta, por lo que, además de todo lo anterior, ésta debe ser capaz de "entender" el cuestionamiento. En realidad, solucionar este problema supera a la más simple de las elicitaciones en el campo de la lingüística. Solucionar la correferencia, establecer correctamente roles sintácticos, semánticos y marcos, son tareas

perseguidas por las tecnologías del lenguaje, las cuales se enfrentan, no a dos oraciones concatenadas, sino muchas de las veces, a corpus que constan de miles de millones de palabras.

Para la ingeniería lingüística, es relevante poder apoyarse en teorías lingüísticas que develen patrones formales de tal manera que una máquina pueda utilizarlos a su favor en la resolución de este tipo de tareas. Esta capacidad de utilizar estos patrones se conoce como Modelos de Lenguaje, entendido como la predictibilidad de que dado X ítem lingüístico le siga Y. Determinar la manera de construir la capacidad de predecir es parte de las metas de la ingeniería lingüística, y en general, de las áreas asociadas a la inteligencia artificial. Uno de los patrones lingüísticos que se podría aprovechar para lograr confeccionar esta predictibilidad es el que puede asociar a pasajes de textos la propiedad de información nueva o compartida, patrón que es el que me interesa en esta investigación. Este interés no es vanguardista, se puede rastrear hasta mediados del siglo pasado y se encontraba enmarcado en la resolución exitosa de dos grandes tareas: la **simulación de conversaciones o *chatbots*** y el **resumen automático** de textos. No es mi intención ahondar en estos estudios debido a que mi tesis no cae de manera directa dentro de alguna de esas dos áreas. Sin embargo, son antecedentes obligados que persiguen objetivos parecidos a los que he delineado en este trabajo por lo que repasaré a grandes rasgos sus alcances en la siguiente sección.

2.2.1 Chatbots y resumidores automáticos

La investigación de los chatbots busca desarrollar la técnica necesaria para hacer que una máquina genere respuestas acertadas a preguntas hechas por un ser humano. La evaluación del desarrollo de estas tecnologías no sólo se veía en la pertinencia de la respuesta dada por la máquina sino también en el uso correcto de recursos gramaticales, como concordancias, pronombres y en el seguimiento de referentes. Si bien SPAN, una de las técnicas que utilizo en la presente investigación, surge en este marco de crear un chatbot con la capacidad de asistir en la docencia —o también llamado *tutor-bot*— gran parte de las técnicas contemporáneas se basan en redes neuronales [58]. En estos algoritmos se desiste en implementar conocimiento lingüístico, lo que ha empujado a esa área a la especialización fuera de la lingüística y a demandar recursos con los que difícilmente una universidad mexicana cuenta a la fecha (c.f. Masche y Le 2018; Akma et al. 2018). Por esta razón, no profundizo en los detalles de estas tecnologías, pero, como mencioné al inicio de este documento, en un futuro no muy lejano, una *lingüística de máquina* podría explorar las maneras en que un chatbot *habla*, con el fin de detectar los patrones de estas inteligencias.

En cuanto a los resumidores automáticos, su objetivo se centra en detectar qué secciones de un texto son **relevantes**, para luego *extraerlas* o *abstraerlas*. En el primer caso se crea un texto más pequeño, y en el segundo se crea un nuevo texto. Se parte de patrones lingüísticos (etiquetado de oraciones, lematizado o *Part-of-Speech*) para guiar la calificación de la

[58] No daré detalles sobre estas tecnologías en general, pero en términos generales se puede entender una red neuronal como un sistema que pasa de A a B, atravesando *z*. Es *z* un mecanismo que aglomera un conjunto de decisiones que, en la mayoría de las veces, queda velado al investigador, resultado de haber pasado de A a B miles —a veces millones— de veces, hasta encontrar el mejor camino (c.f. Goodfellow, Bengio, y Courville 2016).

relevancia de los pasajes (Torres-Moreno 2014). A la fecha existe una amplia variedad de técnicas de cómo construir estos resumidores (Jurafsky y Martin 2009).

Al igual que con los chatbots, aunque tal vez de manera no tan preponderante, la investigación reciente para los resumidores también implementa de manera extensa redes neuronales. No obstante, muchas de las herramientas que permiten detectar qué secciones son relevantes en un texto se fueron desarrollando de manera individual y paralela, a veces no con la expresa intención de proporcionar soluciones a un resumidor sino para evaluar las capacidades de identificación de asimetrías informativas (c.f. Rajasekaran y Varalakshmi 2018).

Para crear un resumidor automático o un chatbot no es fundamental revisar alguna gramática de la lengua en cuestión. Estas tareas se suelen circunscribir al área de Ingeniería Lingüística, en donde se apremia la eficiencia de los algoritmos por encima de la adecuación de los procedimientos con análisis lingüístico. Sin embargo, existen algunas técnicas que abonan a los propósitos de resolver estas tareas y que, además, en sí mismas, son ejercicios de verificación de enunciados teóricos lingüísticos. Las primeras son aquellas relacionadas con la resolución de referencia; le sigue las que han buscado establecer relaciones semánticas entre palabras. Además de estos dos grupos, también existen trabajos orientados a apoyarse en métodos estadísticos y computacionales para verificar patrones semántico-pragmáticos. Finalmente, el trabajo asociado con las técnicas de Análisis de Semántica Latente no se ha limitado a desarrollar una herramienta matemática para encontrar relaciones entre fragmentos, sino que, además, se ha investigado si la representación obtenida a través de ese método también captura relaciones predichas desde alguna teoría lingüística. En las siguientes secciones le dedico un espacio a hablar de cada una de estas técnicas, cuyos

propósitos, ya se puede observar, se alinean a una Lingüística Computacional, entendida de la manera en que la he definido en esta tesis.

2.2.2 Resolución de referencia y relaciones semánticas

En este trabajo no ahondaré en la resolución automática de correferencia, pero no dudo que podría contribuir a mejorar sistemas para determinar el Estado Informativo de frases nominales, especialmente, en el caso de dispositivos referenciales reducidos. Sólo mencionaré que se puede notar la existencia dos conjuntos de técnicas alrededor de la resolución de referencia. Por un lado, las investigaciones que son afines a la propuesta de Heim (2008) y su teoría semántica de cambio de archivo (*File Change Semantics*), debido a la clara expresión de los mecanismos para la construcción de entidades y la manera en que se vinculan conforme avanza el discurso, con la creación de "canastas" de rasgos que ayudan a determinar que un referente es igual a otro. Por otro lado, trabajos bayesianos como los de Kehler y Rohde (2018), los cuales no toman en cuenta las propiedades semánticas de las entidades sino las características sintácticas, aunque en este caso sólo se concentra en la resolución de referentes entre un pronombre y una mención anterior. Esto mismo lo podemos encontrar para el español en Carrasco Morales (2004) para la resolución de anáfora indirecta.

Las investigaciones anteriores, en todo caso, coinciden en que es necesaria una representación computacional del discurso que permita vincular las entidades y distinguir diversas propiedades lingüísticas. Un modelo que pone en marcha esta idea, aunque no explícitamente para resolver la referencia entre menciones, es el de Kibrik et al. (2016). En este trabajo, se evalúan un conjunto de predictores que, de acuerdo con su propuesta teórica,

deberían ayudar a predecir la forma del dispositivo referencial (un pronombre o una frase nominal, por ejemplo). A excepción de este último trabajo, ya que parte de los estados de activación, las investigaciones sobre resolución de correferencia no consideran lo nuevo/dado como una propiedad lingüística. Sus intereses están orientados, como bien señala su nombre, a vincular entidades a través de sus menciones dado un discurso; a veces, no más allá de dos cláusulas concatenadas.

Una de las herramientas que ayuda a vincular dos entidades en un discurso son las ontologías semánticas. Estas se definen como "a standardized representational framework providing a set of terms for the consistent description (or "annotation" or "tagging") of data and information across disciplinary and research community boundaries" (App, Smith, y Spear 2015, xxi). Con la información etiquetada se busca encontrar relaciones subyacentes entre frases nominales que permitan identificar si un referente mencionado en un momento dado en el discurso tiene algún vínculo con algún otro referente, lo que podría interpretarse como dos menciones de una misma entidad (Nirenburg y Raskin 2004; McCarthy et al. 2012; Ziai, De Kuthy, y Meurers 2016). Aunque esta técnica forma parte de los métodos aplicados para aumentar la eficiencia de los detectores automáticos, como se mencionó en la introducción, no es parte de mi objetivo aplicarlo en esta investigación debido a los recursos que se necesitan para el español. Sin embargo, las ontologías no son la única manera de capturar relaciones semánticas. Otra solución sería tener acceso a las definiciones de un diccionario. De tal manera, si existe una relación entre, por ejemplo, la palabra *violín* y *cuerdas,* el significado plasmado en las acepciones de esas palabras y el contexto léxico nos podrían ayudar a relacionarlas. Es por ello por lo que una propuesta de esta investigación es incluir

las definiciones del Diccionario del español de México (DEM). La manera en que usaré el DEM la muestro en §2.4.3.

2.2.3 Detección automática de propiedades semántico-pragmáticas

He podido localizar por lo menos tres líneas de investigación que tienen intereses similares a los que presento en esta investigación: la detección de Tópico/Foco, de genericidad y de definitud. Todas ellas comparten de alguna manera la detección de lo nuevo/dado, aunque varían en los alcances dispuestos en sus marcos teóricos; a su vez, también consideran la frase nominal como el centro de atención y análisis. Aunque como en el caso de Tópico y Foco, también es relevante proporcionar información oracional, más allá de la frase nominal, para favorecer la realización exitosa del proceso automático.

2.2.3.1 Detección de Tópico y foco

En cuanto a la detección de tópico/foco, como ya discutí en el Capítulo 1, se debe tener mucho cuidado con estos conceptos; tópico no implica información dada ni foco información nueva. No obstante, en los trabajos de Sgall, Hajičová y Panevová (1986), Hajičová, Sgall, y Skoumalová (1995; 1993) y Mírovský et al. (2013), estas nociones discursivas son indisociables a las de tópico y foco.

> In the prototypical case, the topic (theme, “given” information) can be understood as that part of the sentence structure that is being presented by the speaker as readily available in the hearer’s memory, whereas the focus (comment, rheme) is what is being asserted about the topic (Hajičová, Sgall, y Skoumalová 1995, 81).

Su marco teórico, llamado Descripción Generativo Funcional (*Functional Generative Description*), les ha permitido implementar un algoritmo para etiquetar su noción de foco y tópico a través de un árbol de decisiones que luce de la siguiente manera:

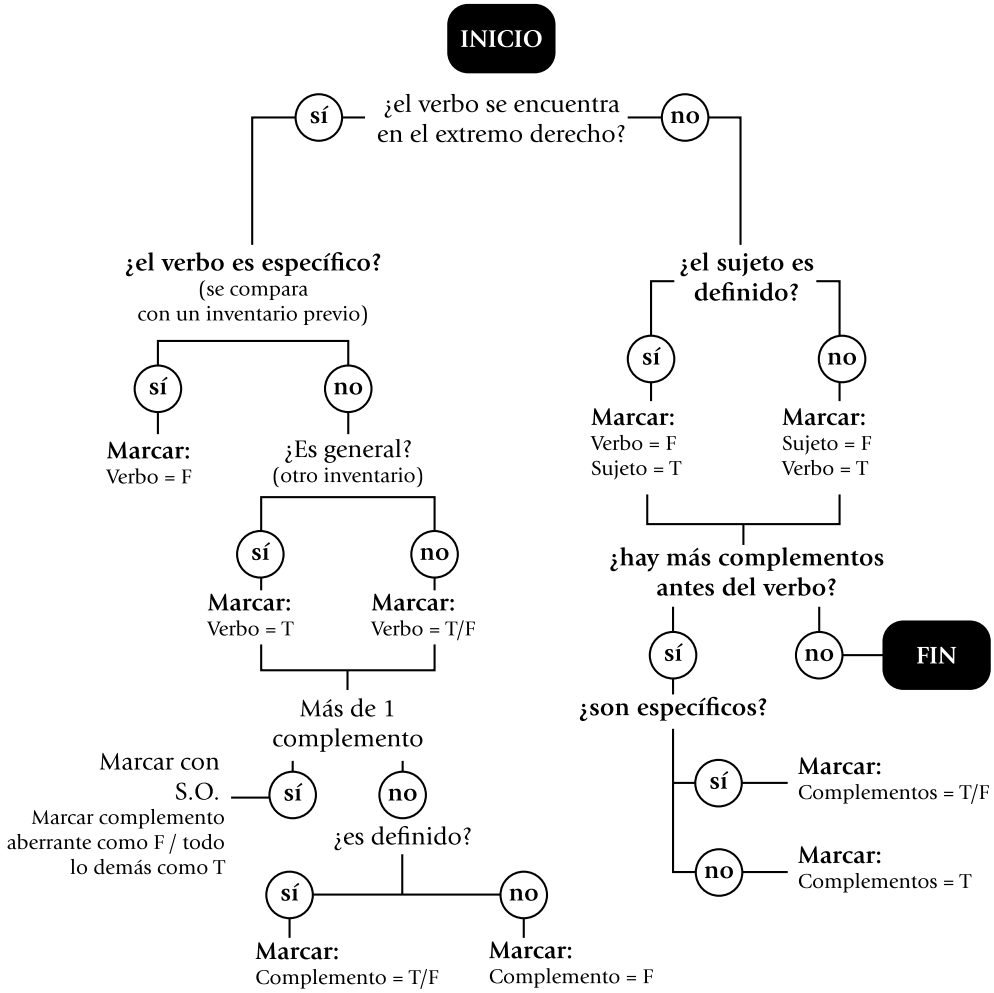


Figura 6. Árbol de decisiones para etiquetar tópico y foco de acuerdo con Mírovsky et al. (2013)[59]

Uno de los problemas de esta propuesta es que las decisiones sobre lo que es tópico y foco parten de un corpus etiquetado con su propia forma de análisis gramatical con la inclusión de un nivel de lengua llamado *tetragramático*. Esta metodología, por el momento, está

[59] Esta figura es creación propia. Su propuesta supone un etiquetado de relaciones de dependencia y de constituyentes. F = foco; T = tópico; T/F = ambiguo.

descartada de las que aplicaré a mi corpus debido a lo que implicaría el etiquetado manual preliminar. Sin embargo, como se muestra en la figura (7), presentan una técnica que resuelve el etiquetado automático de estas categorías. Son buen ejemplo de una teoría que ha probado traducirse a un algoritmo.

2.2.3.2 Detección de la Genericidad

El interés por identificar propiedades semántico-pragmáticas en frases nominales se extiende a la referencialidad, genericidad y especificidad. El trabajo de Reiter y Frank (2010), por ejemplo, es un intento por desarrollar un etiquetador automático supervisado de genericidad en frases nominales de inglés. En este contexto, es importante señalar que el concepto "supervisado" se refiere al etiquetado previo de material que se le pueda proporcionar a la máquina. Un método "no supervisado" implica sólo dar los datos crudos, con la menor intervención posible, de tal manera que todo el proceso, de inicio a fin pueda ser realizado de manera automática[60]. Reiter y Frank (2010) etiquetan de manera automática propiedades lexicogramaticales, dejando para trabajo manual el etiquetado de la propiedad "genérica" de las frases. El concepto de genericidad lo adoptan del marco teórico de Krifka et al. (1995), introducción al *The Generic Book*. Para determinar la dependencia de las variables utilizan un modelo de red bayesiana, y después, evalúan sus resultados con la medida-F. Su trabajo otorga evidencia empírica y estadísticamente válida de que la genericidad es una propiedad

[60] El ideal de procesamiento no-supervisado también es conocido como *end-to-end.*

que se ve afectada por las propiedades lexicogramaticales a nivel oración. Al integrarla, aumenta la eficacia del etiquetador automático.

2.2.3.3 Detección de Definitud

Otro trabajo, el cual persigue objetivos muy similares a los de la presente investigación, es el de Bhatia, Lin et al. (2014) y Bhatia, Simons et al. (2014). El propósito de estas investigaciones es predecir la función comunicativa asociada a la definitud en frases nominales del inglés. En su revisión teórica, se encuentran trabajos que abarcan distintos autores, desde Russel (1905) hasta Prince (1992). Dada esta revisión, sintetizan un conjunto de etiquetas, como, por ejemplo: “anáfora con dependencia de la misma cabeza”; “no anafórica copresencia”; “no anafórico único y dado para el hablante” por mencionar algunas. Después, utilizan un modelo logarítmico lineal para determinar el peso de cada una de las características morfosintácticas y bosques aleatorios para el etiquetado automático. Manejan un conjunto de propiedades llamados *precepts,* en donde se toma en cuenta toda la información morfosintáctica etiquetada en el contexto de la frase nominal, además de las propiedades dentro de la misma frase.

El etiquetado de las propiedades morfosintácticas es automático, mientras que el etiquetado de las funciones comunicativas es manual. En sus resultados hacen notar que la presencia del artículo definido es importante para el etiquetado de la definitud, pero no es única: eliminar la capacidad del etiquetador de utilizar este tipo de palabra en la clasificación no lo vuelve incapaz de detectar otras funciones comunicativas asociadas a la definitud. Como mencionan los mismos autores, esta clase de investigación ayuda a mostrar la complejidad del fenómeno sobre la definitud, y a plantear nuevas hipótesis a partir de observar qué rasgos y patrones

son los aprendidos por el clasificador, lo que a su vez puede arrojar luz en rutas opacas de gramaticalización, por lo menos, en lo que al inglés respecta.

Los trabajos anteriores funcionan como antecedente para el método de la presente investigación. En lo que sigue, describiré el Análisis de Semántica Latente, ya que es un método que no se ha utilizado con el propósito de etiquetar estados de activación o, en un entendido más amplio, Estados Informativos.

2.2.4 Antecedentes de LSA y la representación vectorial del significado

La problemática en trabajos de lingüística computacional se presenta doble: por un lado, la necesidad de contar con un conjunto de categorías más o menos estables en la comunidad científica sobre el fenómeno que se desee predecir para después realizar el análisis y etiquetado manual correspondiente de estas propiedades en corpus a mediana/gran escala; por otro lado, la selección de la técnica apropiada para crear el *clasificador*, un algoritmo informático que cataloga las entidades lingüísticas deseadas en las categorías predispuestas[61]. La distinción más grande entre las distintas variedades de clasificadores es: (i) la cantidad de datos necesaria para obtener mayor precisión; (ii) la implementación de árboles de decisiones, lo que también supone el manejo del corpus inicial (lo que impacta en el grado de "supervisión" del clasificador); y (iii) la capacidad de inspeccionar el clasificador al final del entrenamiento para extraer de él conocimiento que pueda ser nuevo para el área. De los distintos clasificadores y métodos, en esta investigación pretendo explorar, por restricciones

[61] Paralelo a los clasificadores se encuentran los agrupadores (*clustering*) que parten de datos sin categorías predispuestas (Manning y Schütze 1999, 232).

de tiempo, dos: la regresión múltiple que permite generar un modelo para clasificar y bosques aleatorios. En ambos casos, me interesa probar la capacidad de seis medidas obtenidas por dos métodos que abstraen las características lexicogramaticales de las frases: el Análisis de Semántica Latente y una variación llamada SPAN. De hecho, son estas medidas las centrales. Construir el corpus que asocie estas medidas a cada frase nominal permitirá implementar en un futuro otra clase de clasificadores.

El Análisis de Semántica Latente (*Latent Semantic Analysis*, LSA por sus siglas en inglés, aunque se le conocía inicialmente como *Latent Semantic Indexing*) (Manning y Schütze 1999; Landauer y Dumais 1997; Landauer et al. 2007) es una técnica para atender distintos problemas de procesamiento de lenguaje natural, entre los que están la recuperación de información, sistemas automáticos de tutorado, evaluación de coherencia del texto, identificación de tipos textuales y evaluación de comprensión lectora. Su relación con lo nuevo/dado se remonta a las técnicas de apoyo para la resolución de anáforas de manera automática sin la necesidad de contar con grandes cantidades de datos. El método consiste, en términos generales, en construir vectores a partir de las unidades de interés (párrafos, cláusulas, frases) y después una matriz palabra-por-unidad. Esta matriz es reducida por otro método llamado descomposición en valores singulares (*Singular Value Decomposition*, SVD) lo que nos arroja los componentes principales (*Principal Componentes*). La similitud de los vectores entre las unidades de interés se mide por su ángulo de separación, a partir del cálculo de los cosenos.

Esta herramienta se fundamenta en la **representación vectorial del significado**. En el Capítulo 1, §1.12 expliqué esta concepción, pero recuperaré a continuación algunas ideas. Primero que nada, la representación vectorial de cualquier entidad con respecto a otra supone

la localización de esas entidades en un "mapa". Lo que nos importa de esta representación del significado no es la referencialidad de la entidad sino la distinción que se hace de ésta en contraste con otras entidades en un momento determinado de un discurso. En la base, la representación vectorial es una manera de relacionar menciones *similares* alrededor de una misma entidad, y determinar en qué grado. Se debe tener cuidado con no suponer que la similitud entre palabras es lo mismo que la similitud entre constituyentes. En estos, el punto de encuentro es un referente que puede ser presentado en el discurso con frases léxicamente distintas. No ahondo en las diferencias con investigaciones que buscan similitudes léxicas, ya que están orientadas a objetivos distintos a los que persigo en este trabajo. En todo caso, la representación vectorial de textos, constituyentes y palabras ha demostrado tener un amplio uso. Si bien, era común la representación de los documentos como espacios vectoriales (Salton, Wong, y Yang 1975), es en trabajos recientes, desde la aplicación de LSA hasta el resurgimiento en la implementación de redes neuronales, en que se le vuelve a dar importancia a la representación vectorial del significado (c.f. Camacho-Collados y Pilehvar 2018).

2.2.5 De la palabra al vector

De los trabajos fundacionales al respecto se encuentran las investigaciones de Mikolov, Chen, et al. (2013) y Mikolov, Yih, et al. (2013). En estas, se mostraba que utilizando sistemas de redes neuronales se podía llegar a obtener vectores que, al aplicar operaciones algebraicas, daban como resultado posiciones que tenían sentido. Un ejemplo clásico lo ilustro en la Figura 7 a continuación. En este caso, el modelo de lenguaje producido trataba de predecir relaciones del tipo "*rey* es a *reina* como *reyes* es a *reinas*". Recordemos que son

vectores, figuras matemáticas a las cuales se les puede aplicar distintos tratamientos algebraicos. Del lado izquierdo de la Figura 7 se observa que la técnica captura una relación, en este caso podemos interpretar que es la diferencia entre masculino/femenino; a la derecha la relación capturada se interpreta como la diferencia entre lo plural-singular.

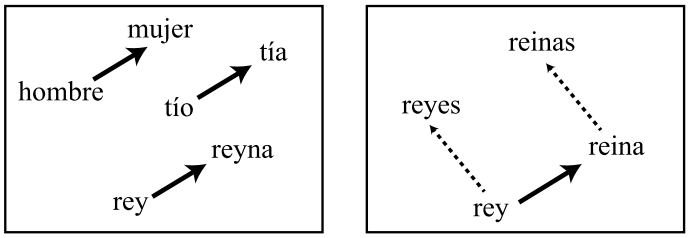


Figura 7. Captura de distintas relaciones en espacios vectoriales[62]

Lo que resultó novedoso de estos estudios fue que, si al vector de la palabra *rey* le restamos el vector de la palabra *hombre* (*rey-hombre*), el vector resultante se encontraba en una posición muy cercana a la del vector *reyna*.

Aunque este tipo de modelos de lenguaje basados en redes neuronales corresponden al estado del arte en el área, requieren de grandes cantidades de datos, muchas de las veces etiquetados, y son costosos computacionalmente. Para los modelos de Mikolov, Yih, et al. (2013), por ejemplo, la generación de los vectores de cada palabra necesitó un corpus de 267 millones de palabras/tokens. Si nos vamos al estado del arte en ingeniería lingüística, GPT-3, la inteligencia artificial desarrollada por OpenAI para generación automática de lenguaje, utilizó 499 mil millones de palabras/tokens (Brown et al. 2020). Su entrenamiento necesitaría 355 años en computadoras convencionales y su costo mínimo ronda los $4 600 000 de

[62] Tomado y traducido de Mikolov, Chen, et al. (2013, 749).

dólares[63]. Una ventaja de los métodos inspirados en bolsas de palabras y LSA es que la cantidad de datos necesarios para obtener resultados es mucho menor, con lo que se permite realizar experimentación y explorar predicciones.

La representación vectorial, ya sea construida por millones de datos y redes neuronales, o por un número de palabras menor y la reducción de matrices, se sostiene de un mismo principio: palabras/textos con significados similares tienen vectores similares o cercanos. Este mismo principio es el fundamento del Análisis de Semántica Latente. Este método, tal como su nombre lo enuncia, deja ver la estructura semántica latente de las unidades analizadas (sean palabras, frases o secciones más grandes de un texto). La relevancia de este tipo de análisis descansa en la reducción de matriz por SVD. En esta representación vectorial, se elimina información que podría ser considerada irrelevante debido a su variabilidad (ruido), por lo que se puede aprovechar mejor la estructura semántica de las unidades. Este ruido depende muchas veces de la polisemia de la palabra, o por encontrarse en una situación de homonimia; por ejemplo, para estos algoritmos, es difícil clasificar la variedad de sentidos de *banco*. No obstante, el método incluye esta reducción de las matrices, el cual crea un nuevo conjunto de representaciones vectoriales, las cuales quedan liberadas de su dependencia de la frecuencia/presencia de las palabras en los contextos analizados. De esta manera, se pueden vincular unidades que superficialmente no parecen tener palabras similares, pero sí estructuras semánticas latentes relacionadas.

[63] Cantidad interesante que retomo de Chuan Li, estudiante de doctorado, quien desarrolla este y otros temas para dimensionar los costos de este tipo de algoritmos en su página de internet: https://lambdalabs.com/blog/demystifying-gpt-3/

2.2.6 LSA, SPAN e información nueva/dada

Otra línea de investigación que impulsó el desarrollo de LSA, como ya adelante en la sección anterior, fue la búsqueda de *similitudes* entre palabras (Hempelmann et al. 2005, 945). Aunque interesante y proyectado para futuros trabajos, fue necesario apartarme drásticamente de las técnicas asociadas a esta área, entre otras razones, porque no deseo dar con un método para determinar que dos palabras son iguales o diferentes[64]. La información antecedente o dada no se reduce a la similitud del léxico. El Estado Informativo expresa las suposiciones sobre el estado mental de los referentes, no sobre la similitud léxica de una frase nominal con otra. Esto deja ver que ambas posturas implican distintos marcos teóricos. En otra investigación se podrían explorar las diferencias entre usar LSA para detectar frases que se definan como *símiles* y los Estados Informativos asociados.

La innovación en el uso de LSA surgió cuando se extendió su aplicación no sólo a la detección de similitudes entre palabras, sino también en la captura de otro tipo de relaciones semántico-pragmáticas entre frases y oraciones. El trabajo que hizo explícito este objetivo fue el de Hempelmann *et al*. (2005) "*Using LSA to Automatically Identify Givenness and Newness of Noun Phrases in Written Discourse*". Estos autores parten de las bases teóricas de Halliday (1967) y Chafe (1976) para enmarcar su definición de lo nuevo/dado y utilizan un sistema de etiquetas binario: (1) para frase nominal dada (*given*) y (0) para información que no puede ser considerada como dada pero tampoco como nueva. Este sistema, en

[64] Para realizar esta investigación también exploré LSA como técnica para encontrar similitud entre palabras del DEM. Los resultados preliminares dejaron ver la complejidad del problema y la necesidad de darle atención en su propio espacio. Aún está pendiente trabajar junto con el DEM para pulir el método. Estos primeros acercamientos se presentaron al Seminario de lexicografía del Dr. Fernando Lara en El Colegio de México en el 2019.

principio, se inspiraba en la taxonomía de Prince (1981), pero redujeron las etiquetas debido a que muchas de las otras categorías eran escasamente usadas, lo que las volvía inútiles para análisis de regresión logística. Se etiquetaron cuatro textos en inglés de cuarto de primaria, de donde se analizaron 478 frases nominales de 195 oraciones. Cada frase nominal fue, además, etiquetada de manera binaria (si/no) con otras tres propiedades: si la FN era un pronombre; si era encabezada por un artículo definido; y si compartía palabras de clase abierta con secciones anteriores del texto. Después de esto, se realizaron los cálculos de LSA entre cada frase nominal y las frases nominales antecedentes; luego se realizó un segundo cálculo llamado SPAN, una variación de LSA planteado por Hu et al. (2003). Este método, en términos generales, otorga a toda unidad de interés dos componentes: un vector ortogonal y otro paralelo, los cuales son obtenidos a partir de proyectar el vector de la unidad de interés en un hiperplano que se calcula de los vectores que constituyen el universo de lo dado en el texto revisado. La matemática de este procedimiento la desarrollo en §5.

En Hempelmann *et al*. (2005), SPAN se integró como una propiedad de las frases nominales y se realizaron tres regresiones logísticas ordinales para evaluar la capacidad de LSA y SPAN como predictores de lo nuevo/dado. La primera regresión, sin LSA ni SPAN, mostró una precisión del 66%, considerando sólo tres propiedades: pronombre, artículo de definido y palabras de clase abierta. En la segunda se incluyó LSA, y la precisión aumentó a 74%. Finalmente, al incluir SPAN, la precisión aumento a 80%. El comentario de los autores es que incluir SPAN ayuda a capturar relaciones del tipo léxicosemántico entre frases nominales en donde no hay pronombres los cuales invisibilizan la relación.

El principal antecedente de Hempelmann et al. (2005) consistió en un trabajo que, aunque citado como un borrador en el mismo año, fue publicado años después en McCarthy et al.

(2012) como "*Newness and Givenness of Information: Automated Identification in Written Discourse"*. En este trabajo también se utilizó SPAN, LSA y LSA_{MAX} el cual se refiere a medir la distancia entre no sólo una frase nominal y su antecedente sino con todas las frases nominales anteriores. También incluyen una medida de similitud basada en ontologías semánticas que implementan a través de representaciones proposicionales de significado de texto (*propositional text-meaning representations*, o TMRs). Finalmente, agregaron una medida de superposición léxica (mismas palabras de clase abierta) a partir de Coh-Metrix el cual desglosan como superposición entre unidades léxicas, constituyentes, bases léxicas, lemas y raíces (Graesser et al. 2004). Sus primeros resultados arrojaron que LSA_{MAX} y SPAN son índices útiles para identificar lo nuevo/dado respectivamente, mientras que la medida de LSA con sólo la frase nominal inmediata anterior es un índice débil, así también las ontologías semánticas y los distintos tipos de superposición léxica.

Posterior a esto, aplican diversos análisis de regresión múltiple, en donde evalúan la capacidad predictora de LSA_{MAX} y SPAN junto con otras variables dependientes para lo nuevo/dado en términos de Prince (1981). Sus resultados arrojan que SPAN, ayudado por ontologías semánticas, resulta ser un buen predictor, tanto para lo nuevo como lo dado. LSA_{MAX} no demostró superar a SPAN de manera significativa, y aunque las ontologías ayudan al predictor, su eliminación tampoco reduce significativamente la precisión de SPAN (c.f. McCarthy et al. 2012).

Las dos investigaciones anteriores corresponden al antecedente directo del tipo de técnica que deseo aplicar, además de compartir objetivos. Uno de los problemas que sobresale en ambos trabajos es que carecen de una manera de sintetizar las categorías predispuestas: suponen que el trabajo de Prince (1981) proporciona una metodología de análisis para cada

una de las etiquetas, pero no es así, como mostré en el Capítulo 1. Además, el concepto de frase nominal, del cual parten como el dispositivo referencial por excelencia, lo presentan resuelto. Este concepto varía entre distintos marcos teóricos lingüísticos y es importante especificar qué abarca para poder garantizar la reproducibilidad de los experimentos y del método. En este caso, resulta complicado comparar los resultados anteriores con los de la presente investigación precisamente porque no establecen de manera explícita los criterios que definen las frases nominales, partiendo de que sus trabajos son realizados en inglés. A pesar de lo anterior, el primer grupo de pruebas estadísticas que aplico son las presentadas en estos dos trabajos anteriores a manera de tener un primer contraste con el estado del arte de estos estudios.

Otros trabajos con objetivos más orientados a la ingeniería lingüística, pero que aplican LSA y SPAN son los de Graesser et al. (2007) y Graesser & Harter (2001) que establecen tutores automáticos que evaluan las respuestas de los alumnos; objetivo parecido al plateando en Hu et al. (2003) que motivó la creación de SPAN. Por otro lado, trabajos como los de Foltz et al. (1998) utilizan LSA para evaluar la coherencia discursiva, y de hecho, al final del trabajo de McCarthy et al. (2012) se implementó un segundo análisis con este mismo objetivo, demostrando la utilidad de la medida en otro tipo de problemas lingüísticos. Debido a que están fuera del objetivo y tema de mi investigación, no ahondo en más detalles sobre estos últimos trabajos.

2.3 Pasos para el Análisis de Semántica Latente de lo Nuevo/Dado

Dado lo anterior, a continuación, describiré el método de LSA y los pasos seguidos para obtener la medida. La manera de implementar LSA es bien conocida. Para más detalle, uno se puede remitir al *Handbook of Latent Semantic Analysis* (Landauer et al. 2007). En este caso, ejemplificaré su aplicación con las siguientes seis frases nominales:

(50) a. Mi tía Juana tiene **[un perro que come muchas croquetas]**.
b. **[El perro que me regalaron]** tiene manchas verdes.
c. **[Uno de los perros que me regalaron]** desapareció.
d. **[Un gato]** se quedó dormido en casa de mi tía Juana.
e. Muchos perros persiguieron a **[el gato]**.

La secuencia básica sigue los siguientes pasos:

Figura 8. Secuencia de pasos para LSA

2.3.1 Tokenizar

En este paso se determina la unidad mínima de análisis, lo que se entenderá como un *token*. En general, se parte de palabras gráficas separadas por espacios, por lo que aquellas palabras que se deseen segmentar como una sola unidad se juntan. Por ejemplo, una locución preposicional del tipo *al lado de* se procesaría como *a_el_lado_de*. Esto depende de cada investigación. En este paso es en donde se decide romper las contracciones del tipo *del* o mantenerlas como una sola unidad. En el tokenizado también se agregan espacios a los signos

de puntuación y se determina si se eliminan o se conservan[65]. En los casos de las oraciones en las que aparecen las FFNN en (50), sólo tenemos puntos, pero en este paso se pueden integrar punto y coma, guiones, corchetes de distinto tipo, etc. En este paso también se eliminan saltos de línea y dobles espacios, y por lo general se decide si se pasa todo a minúsculas, ya que formalmente *perro* y *Perro*, por ejemplo, serían tratadas como palabras gráficas distintas.

2.3.2 Filtrar: lista de paro

Se determinan qué palabras se eliminan para no ser tratadas en la MATRIZ DE CONTEO. En este caso, para ejemplificar, supongamos que eliminamos la preposición *en*, por lo que nuestra lista de paro sería definida por extensión de la siguiente manera:

(51) filtro_1 = [en]

Esta estructura está inspirada en la sintaxis de Python, en donde este tipo de estructura indica una lista de palabras.

2.3.3 Crear matriz de conteo

Para realizar la matriz de conteo se utiliza la paquetería de Python *scikit-learn* (Pedregosa et al., 2011). Las palabras que se integran a la matriz pueden provenir de: (i) un contexto o ventana determinada para cada frase nominal; o (ii) las palabras al interior de la frase nominal. Con estos textos se construye una Bolsa de Palabras (*Bag-of-Words*) en donde la posición de las palabras se vuelve irrelevante. En este caso tomaré como ejemplo una ventana

[65] Al respecto, es importante señalar que algunos algoritmos de etiquetado automático de propiedades lexicogramaticales necesitan de los signos de puntuación para realizar sus cálculos. Eliminar todos los signos de puntuación podría arrojar errores al momento de procesar textos muy grandes.

de 3 palabras alrededor, lo cual expresaré como VENTANA-n, en donde *n* es el número de palabras por identificar en el exterior, mientras que nombraré INTERIOR-w a las palabras internas de la frase, siendo *w* una cantidad variable de palabras. En este ejemplo tomo VENTANA-3, pero en general el parámetro que utilizo en esta investigación es VENTANA-20 con algunos criterios para determinar en dónde cortar; nótese que 20 es el número inicial, pero puede ser menor de acuerdo con los criterios dispuestos, los cuales se deciden para cada investigación (los criterios del corte de la ventana los menciono en la sección 2.4.2).

De acuerdo con una VENTANA-3 —y sin incluir las palabras en el interior—, las frases nominales en (50) tendrían el siguiente contexto:

Tabla 5. Ejemplo de frases nominales y sus contextos VENTANA-3 y filtrando ocurrencias de *en*

	Frase Nominal	**Contexto**
a.	**[un perro que come muchas croquetas]**	~~mi~~ tía juana tiene
b.	**[el perro que me regalaron]**	tiene manchas verdes
c.	**[uno de los perros que me regalaron]**	desapareció
d.	**[un gato]**	se quedó dormido ~~en casa de mi tía juana~~
e.	**[el gato]**	~~muchos~~ perros persiguieron a

Nótese en la tabla anterior que he decidido tomar *el gato* en (e) y no *muchos perros*, como se muestra en los ejemplos de (50). Esto es posible y varía con los criterios establecidos por cada investigador; incluso pudieron haber sido las dos frases, pero la segunda debió de aparecer nuevamente en su propio inciso. Observemos con mayor atención qué ha sucedido en la tabla 5. En el caso de (a) se ha eliminado *mi* debido a que sólo se contemplan las tres palabras anteriores a la frase nominal *un perro que come muchas croquetas*. Si hubiera habido palabras después de esa frase nominal, también estarían incluidas como contexto. Es

casualidad que en ninguno de estos casos tengamos seis palabras en el contexto (tres hacia enfrente y tres hacia atrás de la FN).

Dado lo anterior, la matriz de conteo estaría representada de la siguiente manera:

Tabla 6. Matriz de Conteo de 12×5

	FN_a	FN_b	FN_c	FN_d	FN_e
tía	1	0	0	0	0
juana	1	0	0	0	0
tiene	1	1	0	0	0
manchas	0	1	0	0	0
verdes	0	1	0	0	0
desapareció	0	0	1	0	0
se	0	0	0	1	0
quedó	0	0	0	1	0
dormido	0	0	0	1	0
perros	0	0	0	0	1
persiguieron	0	0	0	0	1
a	0	0	0	0	1

Cuya representación algebraica sería de la siguiente manera:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

En álgebra lineal, se asume que las columnas en una matriz como A son los vectores. No hay problema con sostener que las filas lo sean sólo es cuestión de hacerlo explícito. En este trabajo, debido a la facilidad con la cual se puede asociar una fila con una serie de palabras, utilizaré esta representación para las frases individuales, como muestro a continuación:

$$\overrightarrow{fn_a} = (1,1,1,0,0,0,0,0,0,0,0,0), donde\ fn_a \in \mathbb{R}^{12}$$

Que se lee como que el vector de fn_a, es decir *un perro que come muchas croquetas*, es miembro del conjunto de los números reales de dimensión 12.

2.3.4 Cálculo de entropía

Una vez que se tiene la matriz de conteo se realiza un cálculo de pesos para cada uno de los elementos de la matriz. Existen distintas maneras de realizar este cálculo, pero todas ellas han mostrado que una matriz de pesos arroja mejores resultados que una matriz de conteo simple al momento de aplicar técnicas de extracción de información o de similitud (Dumais 1991; Chisholm y Kolda 1999). En particular, para el caso de LSA, en Landauer et al. (2007, 38) se señala que el mejor cálculo ha sido el peso logarítmico local y global.

Antes de entrar a detalles sobre la matemática detrás del peso logarítmico, es importante señalar que el cálculo de peso se conforma de por lo menos dos componentes: el peso local (L) y el peso global (G)[66]. El primero puede calcularse por una simple transformación a binario, cálculo logarítmico, normalización aumentada de la frecuencia del término o sólo la misma frecuencia del término. Para el peso global hay más posibilidades, entre las cuales

[66] En otros cálculos de pesos, principalmente cuando se comparan documentos, se utiliza una medida de normalización (N_j) que se multiplica por $L_{i,j}$ y G_i; en este caso no se implementa directamente debido a que una vez que se factoriza esta matriz a través de SVD, pierde significatividad la normalización de, en este caso, frases nominales.

están el cálculo de entropía, frecuencia inversa del documento (IDF), IDF de frecuencia global, el cálculo Normal o inversa probabilística (Berry y Browne 2005, 35–36). Una vez obtenidos ambos pesos, el cálculo final es una multiplicación. Por lo que, si entendemos que $a_{i,j}$ es la cantidad del elemento (palabra) *i* en el documento *j* en la matriz A, su peso se obtendría a partir de la siguiente formula:

$$peso\_a_{i,j} = L_{i,j} \times G_i$$

Para calcular **el peso local** de $a_{i,j}$ de acuerdo con su entropía, se recurre al logaritmo natural del número en $a_{i,j}$, es decir, a la frecuencia del elemento en un documento particular ($tf_{i,j}$)[67], y se le suma 1. De tal manera, para aquellos casos en donde *tf = 0*, garantizamos que su logaritmo será 1:

$$L_{ij} = \log(tf_{i,j} + 1)$$

Lo anterior ayuda a no conferir un peso mayor a los términos con mayor frecuencia. Un escenario opuesto a este último, pero no por ello deseable, sería convertir la matriz a un esquema binario presencia/ausencia. Sin embargo, esto eliminaría por completo la diferencia entre un término que apareciese diez veces en un documento y sólo una vez en otro. Un terreno intermedio a esto es el logaritmo natural, lo que permite que los números pequeños crezcan de manera esperada, pero cuando empiezan a haber números más grandes, su crecimiento se reduce. Por ejemplo, el logaritmo natural de 2 es 0.69; si duplicamos el 2, es decir a 4, tendríamos que log(4) = 1.38, lo cual tiene sentido con el crecimiento de 0.69 al multiplicarlo por 2; pero si abruptamente brincamos a un número diez veces mayor, en este

[67] Debe de tomarse en cuenta que por *documentos* me refiero, en este caso, a *frases nominales*. También asumo que $a_{i,j}$ es igual a $tf_{i,j}$; he dividido estas variables para conservar la forma de las fórmulas, pero también debido a que es común que en los textos sobre estos temas se utilice el término *term frequency* de manera abreviada.

ejemplo, a 40, su logaritmo natural es 3.68, y no 13.8; resultado si, de manera intuitiva, tratamos de multiplicar el 1.38×10. Es decir, el crecimiento de los números grandes se reduce con respecto al crecimiento de los números más pequeños.

Para **el peso global** de $a_{i,j}$ de acuerdo con su entropía, se aplica la siguiente fórmula:

$$G_i = 1 + \sum_j \frac{p_{i,j} \log_2(p_{i,j})}{\log_2 n}$$

En donde *n* es el número total de documentos (cinco frases nominales en este caso) y $p_{i,j}$ se obtiene de:

$$p_{i,j} = \frac{tf_{i,j}}{gf_i}$$

En donde a su vez $tf_{i,j}$ es la frecuencia del término *i* en el documento *j* y gf_i es la frecuencia global del término *i* en todo el conjunto de documentos analizados. Esta forma de calcular el peso global logra reducir el efecto de palabras que aparecen en todos los documentos, y al mismo tiempo, toma en cuenta su distribución general. Si aplicamos el cálculo de cada peso por separado, un paso antes de realizar la multiplicación para obtener los pesos de $a_{i,j}$, tendríamos datos como los que aparecen en la Tabla 7 a continuación. Después se resuelve la multiplicación entre cada peso local por el peso global del término lo que resulta en lo que se muestra en la Tabla 8.

Tabla 7. Pesos locales y peso global de cada término

Términos	*Frases Nominales*					
	FNa	FNb	FNc	FNd	FNe	**Global**
tía	0.693147	0	0	0	0	**1**
juana	0.693147	0	0	0	0	**1**
tiene	0.693147	0.693147	0	0	0	**0.569323**
manchas	0	0.693147	0	0	0	**1**
verdes	0	0.693147	0	0	0	**1**
desapareció	0	0	0.693147	0	0	**1**
se	0	0	0	0.693147	0	**1**
quedó	0	0	0	0.693147	0	**1**
dormido	0	0	0	0.693147	0	**1**
perros	0	0	0	0	0.693147	**1**
persiguieron	0	0	0	0	0.693147	**1**
a	0	0	0	0	0.693147	**1**

Tabla 8. Matriz ponderada por entropía

Términos	*Frases Nominales*				
	FN_a	FN_b	FN_c	FN_d	FN_e
tía	0.693147	0	0	0	0
juana	0.693147	0	0	0	0
tiene	0.394624	0.394624	0	0	0
manchas	0	0.693147	0	0	0
verdes	0	0.693147	0	0	0
desapareció	0	0	0.693147	0	0
se	0	0	0	0.693147	0
quedó	0	0	0	0.693147	0
dormido	0	0	0	0.693147	0
perros	0	0	0	0	0.693147
persiguieron	0	0	0	0	0.693147
a	0	0	0	0	0.693147

En este caso, la gran mayoría de la Tabla 8 permanece igual que la Tabla 7, excepto por el término *tiene* que cambiaría, dado que: $L_{\text{tiene, FNa}} = 0.693147$; $G_{\text{tiene}} = 0.569323$, por lo que $a_{\text{tiene, FNa}} = 0.39462$. Esto también aplicaría para el caso de *tiene* en FN_b. Con lo anterior finaliza el cálculo de entropía local/global para determinar el peso de cada término. Esta tabla es la que procede al análisis para la reducción de matriz por SVD que explico a continuación.

2.3.5 Reducir matriz

Existen distintos métodos para reducir una matriz. Entre los que se mencionan en Landauer et al. (2007, 39) se encuentran la factorización por QR, la descomposición ortogonal de ULV y la descomposición semidiscreta (SDD). En particular, LSA se distingue por aplicar la reducción a valores singulares o SVD por sus siglas en inglés (*Singular Value Decomposition*). La matriz A se factoriza a las siguientes tres matrices:

$$A = U\Sigma V^T$$

La factorización es realizada a través del módulo para Python *scikit-learn* (Pedregosa et al., 2011). Muestro la aplicación de esta reducción en la Tabla 9. Para el caso de la matriz U, sólo tomo las primeras cinco dimensiones, pero esta matriz originalmente es $n \times n$.

Después de la factorización, cada uno de los componentes de V se multiplican por los vectores en Σ. En la tabla he marcado en negritas los componentes que tomo para crear la matriz reducida, que muestro en la tabla a continuación. Para este ejemplo he reducido la matriz de A a una matriz de dos dimensiones, expresado como $k = 2$. Por esta razón sólo tomo las primeras dos columnas de la matriz Σ, método que se utiliza para trazar en planos cartesianos.

Tabla 9. Factorización en las matrices U, Σ y V de la matriz ponderada en la Tabla 8

Matriz U: Vectores de términos					
tía	**0**	**0**	-0.43452	-0.5	0
juana	**0**	**0**	-0.43452	-0.5	0
tiene	**0**	**0**	-0.49476	$2.43e^{-16}$	0
manchas	**0**	**0**	-0.43452	0.5	0
verdes	**0**	**0**	-0.43452	0.5	0
desapareció	**0**	**0**	0	0	-1
se	**-0.57735**	**0**	0	0	0
quedó	**-0.57735**	**0**	0	0	0
dormido	**-0.57735**	**0**	0	0	0
perros	**0**	**-0.57735**	0	0	0
persiguieron	**0**	**-0.57735**	0	0	0
a	**0**	**-0.57735**	0	0	0
Matriz Σ: Valores Singulares					
	1.20056613	**0**	0	0	0
	0	**1.20056613**	0	0	0
	0	**0**	1.12799101	0	0
	0	**0**	0	0.98025814	0
	0	**0**	0	0	0.69314718
Matriz V: Vectores de documentos					
FN_a	**0**	**0**	-0.70711	-0.70711	0
FN_b	**0**	**0**	-0.70711	0.707107	0
FN_c	**0**	**0**	0	0	-1
FN_d	**-1**	**0**	0	0	0
FN_e	**0**	**-1**	0	0	0

Como puede observarse en la Tabla 10, para matrices pequeñas este proceso no produce resultados significativos, pero para matrices $\mathbb{R}^n$ en donde *n* puede alcanzar de cientos a cientos de miles de dimensiones, esta reducción es el paso crucial que revela la semántica latente, permitiendo asociar textos que no necesariamente comparten vocabulario.

Tabla 10. Reducción a dos dimensiones de la Tabla 8

Dimensiones	*Frases (documentos)*				
	FN_a	FN_b	FN_c	FN_d	FN_e
x	0	0	0	-1.20057	0
y	0	0	0	0	-1.20057

Como mencioné, en la tabla 10 se ha reducido a dos dimensiones, pero para los cálculos de similitud, se sugiere de 100 a 300 dimensiones (Landauer et al. 2007, 43). Algo importante a notar es que no se puede tener una matriz reducida de *n* dimensiones, siendo *n* mayor al número de elementos analizados. En este caso, no podemos obtener una matriz reducida cuya dimensionalidad sea mayor a cinco, ya que estamos analizando cinco frases nominales.

El tamaño de las dimensiones varía, como se puede suponer, por la cantidad total de vocabulario luego de aplicar el tokenizado/filtrado correspondiente además de la ventana de palabras a incluir. Para el siguiente paso en el procedimiento, utilizaré una matriz reducida a cinco ($k = 5$), como la que muestro a continuación:

Tabla 11. Reducción a cinco dimensiones de la Tabla 8

Dimensiones	*Frase (documentos)*				
	FN_a	FN_b	FN_c	FN_d	FN_e
1	0	0	0	-1.20057	0
2	0	0	0	0	-1.20057
3	-0.79761	-0.79761	0	0	0
4	-0.69315	0.693147	0	0	0
5	0	0	-0.69315	0	0

Antes de continuar con la medida de coseno, me gustaría ahondar un poco más sobre un aspecto de la reducción de matrices. Desde el texto de Landauer & Dumais (1997) "*A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge*" se había planteado que el problema central de la vectorización del significado descansaba en el cómo determinar cuántas dimensiones era pertinente reducir.

Un cambio de dimensiones es lo que puede hacer que dos elementos que se observan cercanos, en realidad estén retirados.

Por poner un ejemplo de lo anterior, compartido por todos, las tres estrellas del Cinturón de Orión se ven cercanas en un plano de dos dimensiones, pero agregar sólo una dimensión más, nos muestra que en realidad una de ellas se encuentra a una distancia considerablemente mayor con respecto a las otras dos:

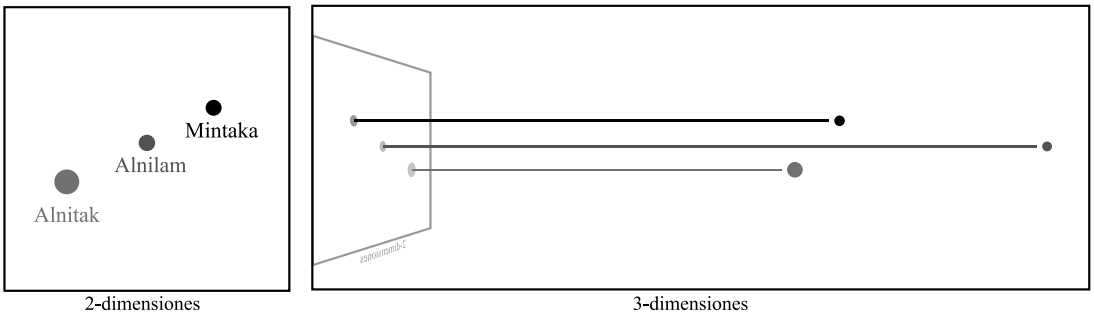


Figura 9. Representación del Cinturón de Orión[68]

Esto mismo se puede observar con la decisión que tomamos en la tabla 10 y la tabla 11. Al considerar sólo dos dimensiones, se pierde diferencia entre las $FFNN_a$, FN_b y FN_c. Pero al considerar más dimensiones, en este caso tres más, se captura la diferencia entre las frases nominales analizadas. Como he mencionado, las investigaciones han sugerido que los mejores cortes están entre las 100 y 300 dimensiones, con lo que se pueden obtener medidas significativas entre las unidades. No obstante, la cantidad precisa sigue siendo parte de la

[68] A la izquierda se encuentra la representación en dos dimensiones de las tres estrellas junto con sus nombres; a la derecha, la representación en tres dimensiones, haciendo notar que la distancia entre ellas varía. La ilustración no está a escala. Las distancias están dadas con respecto a la Tierra. No es muy difícil encontrar las distancias de estas estrellas; se pueden encontrar cálculos a partir de distintos softwares de observación y en blogs de aficionados, como https://www.space.com/3380-constellations.html o a través de aplicaciones del tipo *Star Walk 2*.

discusión en cada investigación particular. Dado lo anterior, ahora examinemos cómo medir esa diferencia.

2.3.6 Coseno: medir distancia entre vectores

Para medir la distancia entre vectores existen distintos métodos, entre los que están la medida Jaccard, Dice y JS; además de las medidas que se basan en distancia euclidiana, como la norma L1 y L2 (Jurafsky y Martin 2009, sec. 20.7). En este trabajo, siguiendo lo propuesto en Landauer et al. (2007), utilizaré la *medida de separación por* ***coseno*** o también conocida como el *coeficiente de correlación normalizado* (Manning y Schütze 1999, 541):

$$\cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \sqrt{\sum_{i=1}^{n} d_i^2}}$$

Siguiendo con nuestro ejemplo, calculamos para cada par de vectores de la tabla 11 su coeficiente, lo cual nos da como resultado la siguiente matriz de similitud:

Tabla 12. Matriz de similitud a partir del Coeficiente de correlación normalizado

	FN_a	**FN_b**	**FN_c**	**FN_d**	**FN_e**
FN_a	1	0.139463	0	0	0
FN_b	0.139463	1	0	0	0
FN_c	0	0	1	0	0
FN_d	0	0	0	1	0
FN_e	0	0	0	0	1

Lo anterior se lee como que, por ejemplo, entre la FN_a y FN_b hay un coeficiente de 0.139463, en donde 1 es similitud exacta y 0 son disímiles exactos.

2.3.6.1 LSA MAX

En este punto, el método está completo, pero en realidad lo que nos interesa es que dada una frase nominal en un punto x en nuestro texto, se evalúe la información nueva o dada. El grado con el que se aproxime este vector a los vectores anteriores nos puede ayudar a determinar alguna de estas dos propiedades. En trabajos como los de McCarthy et al. (2012) se ha determinado que no es suficiente con medir una frase con su inmediata anterior, sino todas las frases nominales anteriores con una medida que nombran LSA_{MAX} (McCarthy et al. 2012, 462). Esta medida consiste sencillamente en tomar el coeficiente más alto a partir de la comparación con las FFNN anteriores.

Tabla 13. LSA_{MAX} a partir de la matriz de similitud

	FN_a	**FN_b**	**FN_c**	**FN_d**	**FN_e**
FN_a	***s.i.***[69]				
FN_b	**0.139463**	***s.i.***			
FN_c	0	**0**	***s.i.***		
FN_d	0	0	**0**	***s.i.***	
FN_e	0	0	0	**0**	***s.i.***

Por ejemplo, si por alguna razón el coeficiente entre FN_e y FN_b hubiera sido 0.20, el valor final de FN_e no sería 0 (marcado con negritas en la tabla anterior) sino justo el de este coeficiente, el más alto, es decir, el LSA_{MAX} de FN_e sería 0.20. No obstante, el LSA_{MAX} de FN_b no sería 0.20, sino 0.139463 ya que es el número más alto a partir de ese punto hacia el inicio del mismo documento –a pesar de que, en este escenario hipotético, más adelante se

[69] Utilizo la expresión latina *solus ipse* abreviado como *s.i.* para referirme a que este valor se refiere a la medida consigo misma, la cual siempre da 1, pero para los objetivos de esta parte del método provoca ruido colocar este número en la tabla.

encontrará una FN_e con la cual obtendrá un coeficiente mayor. LSA_{MAX} es una primera variación implementada en los análisis que utilizan LSA para medir lo nuevo y lo dado. Utilizaré esta medida, y no la que mide sólo la FN inmediata anterior, dados los resultados reportados por McCarthy et al. (2012) y Hempelmann et al. (2005).

2.4. Afinar LSA

Examinemos ahora qué posibilidades hay para aumentar la precisión de LSA. El primer paso al que se suele recurrir es a la lematización de los textos, de tal manera que puedan considerarse una misma dimensión *perro* y *perros*, o *caminé* y *caminaba*. Para lograr esto, en este trabajo he implementado el etiquetador automático de propiedades lexicogramaticales Stanza (Qi et al. 2020). Con este preprocesamiento, cada unidad de análisis es primero trabajada en este programa y luego se determina el filtro de palabras, ya no por una LISTA DE PARO (§3.2). Por ejemplo, no enlistaríamos las formas *el, las, los* y *la,* sino que se filtran todas aquellas que contengan la propiedad "DET" como veremos a continuación. El segundo paso es la consideración entre vocabulario interno al documento —criterio común al momento de comparar textos— y las ventanas alrededor de la unidad de análisis seleccionada —criterio común al momento de comparar ítems léxicos—. Un aspecto innovador en esta investigación es que seleccionaré ambos criterios, pero afinados con distintos filtros. El tercer paso, también novedoso, es la inclusión de las definiciones del DEM como apoyo a vincular frases nominales con mismos significados.

2.4.1. Preprocesamiento Stanza

La rutina de Python de Stanza utiliza un modelo de lenguaje entrenado con AnCora[70], un corpus de español construido en su mayoría por notas periodísticas. Este corpus contiene distintas anotaciones gramaticales realizadas de manera manual, entre las cuales están: constituyentes y funciones sintácticas, lematización, categorías morfológicas, estructura argumental y papeles temáticos (Recasens y Martí 2010).

Decidí utilizar Stanza por encima de otras opciones por dos razones: i) necesitaba un instrumento que se adecuara al proceso completo montado en Python; ii) necesitaba etiquetas que pudieran funcionar como tokens para alimentar las bolsas de LSA. Existen distintos etiquetadores de español, por ejemplo, SpaCy (Honnibal et al. 2020), NLTK (Bird, Loper, y Klein 2009) o Freeling (Padró y Stanilovsky 2012) que tienen buenos modelos de lenguaje para distintas tareas. Debido a que la tarea que buscaba resolver era el etiquetado de partes de la oración (*POS tag*), los cuatro etiquetadores resultaron tener medidas de precisión similares. Para tomar la decisión sobre cuál etiquetador implementar, realicé un pequeño ensayo con 100 tokens (COPENOR-052SO). Los etiqueté de manera manual con EAGLES, y utilicé este sistema como intermediario para normalizar los resultados. De esta manera, comparé la precisión de los cuatros etiquetadores automáticos. En este ensayo SpaCy logró 91.65%, Freeling 77.89%, NLTK 78.79% y Stanza 94.51%. De estos etiquetadores, Freeling es el único que implica reacomodar el sistema de procesos creado (*pipeline*) para el procesamiento de las notas: en sistemas Windows se necesita tener instalado otro lenguaje de programación (a través de *Visual Studio*), u optar por Linux, lo cual implicaría, además,

[70] http://clic.ub.edu/corpus/es/ancora

cambiar de sistema operativo. Viendo los resultados de las precisiones para la tarea solicitada, el costo del cambio entre sistemas y lenguajes no se justifica. Una ventaja de Stanza y Spacy, es que ambos utilizan el *Universal Tag Set*[71]. Esto me permitió evitar un paso más en donde tradujera etiquetas como EAGLES a un sistema en donde las propiedades pudieran leerse como tokens individuales. Evidentemente hace falta un estudio particular para evaluar en su totalidad los etiquetadores actuales para español. Esto sobrepasa los objetivos de esta investigación, pero queda pendiente para futuras exploraciones. Finalmente, viendo el aprovechamiento entre Stanza y SpaCy, el primero fue elegido para realizar el proyecto.

Si partimos de la primera frase nominal como ejemplo, que reproduzco a continuación, los resultados de Stanza lucen como se muestra en la Tabla 15.

Tabla 14. Primera frase nominal segmentada de los ejemplos en 50

	Frase Nominal	**Contexto**
a.	**[un perro que come muchas croquetas]**	~~mi~~ tía juana tiene

Stanza genera un tipo de dato que puede ser transformado de manera sencilla a un diccionario (Dict_PY)[72] que después es transformado a un DataFrame_PY en Pandas (McKinney 2010), lo que a su vez permite el tratamiento en tablas. Las propiedades que se observan en la Tabla 15 provienen de un etiquetado originado por el proyecto *Universal Dependencies* (UD), un marco de trabajo para anotaciones gramaticales que incluye partes de la oración (POS tag), rasgos morfológicos y dependencias sintácticas, con una perspectiva tipológica, cuyo alcance

[71] https://universaldependencies.org/introduction.html
[72] Cuando me refiera a un tipo particular de dato manejado en Python le colocaré un guion bajo y las letras PY.

abarca alrededor de 150 lenguas[73]. Como se podrá suponer en este punto, un texto *en crudo*, sin tratamiento previo, produce resultados poco precisos. Afinar lo que se le dé a la máquina es crucial.

Tabla 15. Datos gramaticales etiquetados de manera automática por Stanza[74]

id	text	lemma	upos	feats
1	**un**	**uno**	**DET**	**Definite=Ind| Gender=Masc| Number=Sing| PronType=Art**
2	perro	perro	NOUN	Gender=Masc| Number=Sing
3	que	que	PRON	PronType=Int,Rel
4	come	come	VERB	Mood=Ind| Number=Sing| Person=3| Tense=Pres| VerbForm=Fin
5	**muchas**	**mucho**	**DET**	**Gender=Fem| NumType=Card| Number=Plur| PronType=Ind**
6	croquetas	croqueta	NOUN	Gender=Fem| Number=Plur

El primer tratamiento sugerido es una lematización y un etiquetado de los tipos de palabras. Gracias a Stanza, se tomarían los elementos de la columna *lemma* y luego se filtrarían a partir de la columna *upos*. En este ejemplo, se filtrarían los DET (en negritas). Esta modificación agregaría un paso adicional al procedimiento de LSA (un paso 0 en el esquema presentado a continuación). Finalmente, debido a que este estudio es sobre información nueva/dada y no sobre similitud, es la medida LSA_{MAX} la que nos interesa. Estas observaciones se agregan a la secuencia planteada, tal y como se muestra en la Figura 10 a continuación.

[73] https://universaldependencies.org/introduction.html
[74] En esta tabla no muestro todo lo que analiza el programa. Para más información, revisar la documentación de la paquetería en https://stanfordnlp.github.io/stanza/

Figura 10. Secuencia de pasos para LSA incluyendo afinación y LSAmax

2.4.2. Conteo interior y conteo exterior

Hasta este momento sólo he introducido el concepto de VENTANA para referirme al contexto de la frase nominal, el conteo exterior. No obstante, tomaré dos conteos para medir las frases nominales. El primero, el contexto de la frase a partir de 20 palabras a su alrededor (VENTANA-20)[75], lo construiré a partir de dos signos de puntuación: el punto (.) y el punto y coma (;). Una frase nominal como la que aparece en (52a) tendría un contexto como se resume en (52b).

(52) a. Señaló que el fenómeno es cíclico, sobre todo, cuando alguna organización delictiva pretende apoderarse del territorio de sus contrincantes. Por último, Sotomayor dijo que **[la Policía]** ya realiza operativos especiales en zonas específicas, en tanto la autoridad correspondiente ya realiza las investigaciones en torno a las muertas violentas.

b. por$_5$ último$_4$ Sotomayor$_3$ dijo$_2$ que$_1$ **[FN]** ya$_1$ realiza$_2$ operativos$_3$ especiales$_4$ en$_5$ zonas$_6$ específicas$_7$ en$_8$ tanto$_9$ la$_{10}$ autoridad$_{11}$ correspondiente$_{12}$ ya$_{13}$ realiza$_{14}$ las$_{15}$ investigaciones$_{16}$ en$_{17}$ torno$_{18}$ a$_{19}$ las$_{20}$

COPENOR-179BC

[75] Se podría ser más específico. El número de palabras que se abarcan antes de la FN no tiene que ser el mismo para el que se abarca después. En esta investigación lo generalizaré a 20 ya que mi objetivo no es evaluar las ventanas como criterio que aumente o disminuya la precisión de un identificador automático de Estados Informativos, pero se puede tener esta versatilidad de acuerdo con los objetivos de cada trabajo.

Nótese que el contexto corta un constituyente, la frase nominal las muertas violentas[76], en donde la palabra 20 es sólo el determinante las. Para estos casos, tomé la decisión de que el programa filtre este tipo de palabras, por lo que *las* quedaría fuera y tendríamos 19 palabras. Los criterios para el análisis de frase nominal en las notas periodísticas se encuentran en Anexo A. Para la ventana exterior se lematizó y filtró determinantes, preposiciones, adverbios, pronombres, nombres propios, conjunciones, interjecciones, numeración y puntuación.

Para el conteo interior (INTERIOR-*w*) considero todos los ítems léxicos de la frase nominal lematizada excepto conjunciones, interjecciones y números. También se filtran signos de puntuación en general excepto por el punto y el punto y coma. Para este conteo, todos los ítems se presentan con su categoría gramatical y sus rasgos morfológicos expuestos. Parto de que, a diferencia de los filtrados comunes para exteriores, e incluso para contenido de documentos, en este caso, es importante preservar los determinantes y las preposiciones, además que las propiedades como el género y número ayudan a relacionar frases nominales. De esta manera, una frase nominal como la policía es analizada de la siguiente manera:

(53) [el DET Definite=Def Gender=Fem Number=Sing PronType=Art policía NOUN Gender=Fem Number=Sing]

Con lo anterior estoy considerando que un token puede ser tanto un rasgo morfológico como Definite=Def como los ítems léxicos lematizados.

[76] No corrijo los errores en las notas capturadas.

Si regresamos a nuestro ejemplo en la Tabla 5 (reproducida en la tabla 16), la matriz de similitud a partir de INTERIOR-*w* y sólo considerando LSA_{MAX}, sería la que muestro en la Tabla 17.

Tabla 16. Frases nominales tomadas de la Tabla 5

	Frase Nominal
a.	**[un perro que come muchas croquetas]**
b.	**[el perro que me regalaron]**
c.	**[uno de los perros que me regalaron]**
d.	**[un gato]**
e.	**[el gato]**

Tabla 17. Triangular inferior de la matriz de similitud para resaltar LSA_{MAX}

	FN_a	**FN_b**	**FN_c**	**FN_d**	**FN_e**
FN_a	***s.i.***				
FN_b	**0.117296**	***s.i.***			
FN_c	0.167328	**0.782593**	***s.i.***		
FN_d	**0.15458**	0.000438	0.045599	***s.i.***	
FN_e	0.000319	0.143732	0.108477	**0.517069**	***s.i.***

Una manera de representar estas relaciones es utilizando dos dimensiones (k = 2) y graficándolas, como se muestra a continuación:

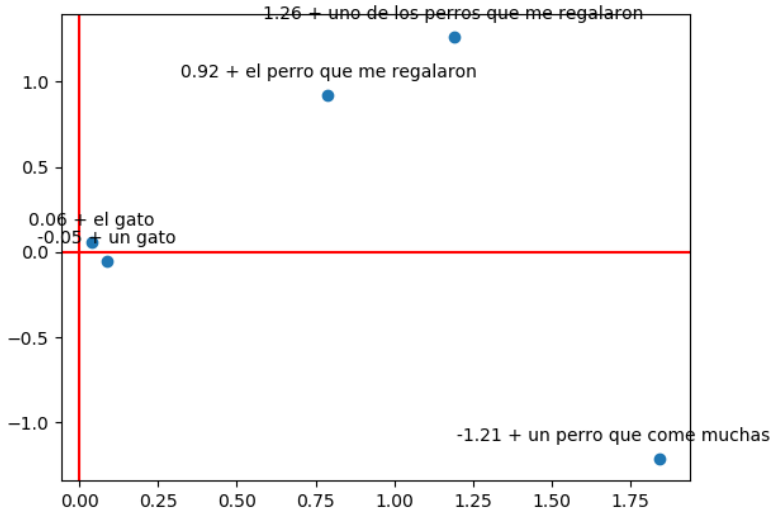


Figura 11. Representación de las frases nominales de la tabla 5 tomando k = 2

Se debe tener cuidado en cómo se interpreta esta gráfica. Lo que indica es que, al final del documento, existen cuatro puntos cercanos y uno alejado. Esto nos da señal sobre que uno de esos puntos se conserva como información "no relacionada" con nada de lo anterior.

Para el interés de esta investigación, lo relevante no es un corte final sino el coeficiente máximo o LSA_{MAX} dado en el momento en el texto en donde aparece la FN analizada. En la Tabla 18 muestro estas medidas, resultado de alimentar el programa con la ventana interior de la FN y la exterior lematizada y filtrada:

Tabla 18. LSA_{MAX} de conteo exterior e interior

	Exterior (V=3)	Interior (I=*w*)
FNa	n/a	n/a
FNb	0.139463	0.117296
FNc	1	0.782593
FNd	0	0.15458
FNe	0	0.517069

No sólo el filtrado parece más prometedor para detectar información nueva/dada, sino que el interior de la frase parece darnos mejor índice de un posible Estado Informativo dado un punto en el discurso. LSA_{MAX} exterior e interior son dos de las medidas finales que reportaré, pero hace falta describir una última inclusión al tratamiento de los textos antes de proceder a SPAN. Se trata de la sustitución de ítems léxicos por las acepciones del *Diccionario*.

2.4.3. El Diccionario del español de México como inventario de sentidos

De acuerdo con Camacho-Collados (2018), en estudios sobre la identificación latente de significado de ítems léxicos se pueden utilizar dos conjuntos de datos: (i) corpus, a través de

ventanas de palabras alrededor de los ítems léxicos o (ii) inventarios de sentidos. Para el primer caso, como expliqué en la sección anterior, propongo incluir también una ventana interior, que es en realidad el método que se utiliza en trabajos sobre extracción y recuperación de información en tecnologías del lenguaje. Para el segundo caso, los inventarios de sentidos, se trata de entradas, definiciones, o colecciones de palabras, previamente asociadas a ítems léxicos. Entre las fuentes usuales para estos inventarios se encuentran Wikipedia o WordNet. Así, el texto que genera el vector de un ítem léxico no es su ventana exterior sino un texto proveniente de alguno de estos inventarios.

En este trabajo parto del supuesto de que el Diccionario del español de México (DEM) puede ser usado como inventario de sentidos para trabajos computacionales. Esto no se ha evaluado en ningún trabajo anterior desde la fundación del diccionario, por lo que mis resultados podrán ayudar a definir usos del DEM en LSA. Antes de proseguir, daré algunos detalles generales del DEM y definiré algunos términos básicos de lexicografía a los cuales haré referencia más adelante.

2.4.3.1 Lexicografía del DEM

Los trabajos del *Diccionario* nacen en 1972, a cargo del doctor Luis Fernando Lara. Gracias al interés y apoyo del que fuese en aquel entonces director del Fondo de Cultura Económica, Don Antonio Carrillo, se logró establecer para este proyecto un fideicomiso del gobierno federal mexicano que presidía Luis Echeverría. Este proyecto, el cual ha estado resguardado desde sus inicios por El Colegio de México, buscaba crear un diccionario original, es decir, creado desde un listado propio de voces, recogidas del habla mexicana. Para ello, el equipo del doctor Lara construyó el *Corpus del español mexicano contemporáneo* (CEMC)

constituido de 1000 textos con 2000 palabras gráficas cada uno, divididos en áreas o temas como: literatura, periodismo, ciencias, técnicas, literatura popular, conversaciones grabadas, textos regionales, documentos de estudios antropológicos y jergas (Lara, Chande Ham, y García Hidalgo 1979, 30). Debido a la cantidad de información necesaria para investigar la lengua a partir de este corpus, se creó un analizador gramatical que sirvió para agilizar el trabajo. Tal fue la empresa de este analizador que le valió a la matemática Isabel García Hidalgo el Premio Dr. Arturo Rosenblueth a Sistemas de Cómputo en 1981.

Al respecto, me parece importante señalar que un antecedente en el procesamiento del lenguaje natural en México es precisamente la búsqueda del vocabulario para la creación del diccionario. Debe recordarse que, en aquel entonces, una computadora del gobierno federal apenas alcanzaba los 64 kilobytes de memoria. Los procedimientos para el análisis del corpus tomaron 8 meses y arrojaron 1 891 045 palabras gráficas, que corresponden a 64 183 tipos de palabras, las cuales el equipo del DEM revisaría una a una.

Poco tiempo después del analizador, otro producto computacional que asistió a los lexicógrafos a reducir los tiempos de análisis de contextos fue el que creó la lingüista María Pozzi. Este algoritmo informático, llamado *Horquilla*, revisaba automáticamente las concordancias para seleccionar aquellas que tuvieran patrones sintácticos distintos, y, al mismo tiempo, garantizaba que no se perdiera información en el proceso.

El primer producto de estos trabajos vio la luz en 1982 con el nombre de *Diccionario fundamental del español de México* con 2500 artículos. Esa obra ayudó a consolidar la práctica lexicográfica del equipo, y para 1986 se produjo el *Diccionario básico del español de México* en donde se presentaban 7000 artículos. Tuvieron que pasar 10 años, hasta 1996, para que el equipo publicara a el *Diccionario del español usual en México (DEUM 1)* al que

le siguió una segunda edición corregida y aumentada. De nuevo, un periodo de madurez, trabajo y análisis culminó en una de las obras magnas de El Colegio de México: el *Diccionario del español de México*, cuya segunda edición se encuentra actualmente en línea con 32000 entradas y más de 60000 acepciones (Diccionario del español de México s/f)[77].

Este recuento permite observar que construir un inventario de sentidos no es un trabajo sencillo. En los trabajos computacionales, como se mencionó, se suele colocar en un mismo tipo de documento a las enciclopedias, redes semánticas y los diccionarios, sin embargo, esto carece de precisiones que son importantes; precisiones que nos ayudarían a pulir qué le entregamos a la computadora e interpretar mejor los resultados.

Las investigaciones del equipo del DEM, desde 1972, han acumulado conocimiento de distinta índole que el doctor Lara ha publicado con el pasar de los años. He dividido estos trabajos en tres grupos de acuerdo a los intereses de mi presente investigación: (i) aquellos que constituyen un rico trabajo computacional, en donde se demuestra la utilidad de una lingüística de corpus asistida por computadora para el lexicógrafo, así como un trabajo estadístico y computacional sobre los patrones existentes en la lengua (Lara, Chande Ham, y García Hidalgo 1979; Medina Urrea 2003); (ii) la estadística propia del español fundamental mexicano, la cual ha arrojado datos interesantes; por mencionar algunos: entre 12 000 y 15 000 vocablos constituyen el español de México, el 75% de sus palabras tiende a contener entre 4 a 8 letras, y los fonemas más frecuentes son la /a/ para las vocales y la /r/ para las consonantes (Lara 2007)[78]; y (iii) la consolidación de una teoría lexicológica y un método

[77] Los datos presentados en este recuento provienen de la *Introducción al Diccionario* que se puede acceder por medio del siguiente enlace: https://dem.colmex.mx/Contenido/8

[78] Los resultados se pueden consultar en la versión en línea con el siguiente enlace: https://dem.colmex.mx/moduls/Default.aspx?id=14

lexicográfico. Para los grupos (i) y (ii), aunque interesantes y de necesaria mención, no los utilizo de una manera en la que impacten en la metodología de esta investigación. Sin embargo, el tercer grupo es crucial.

La teoría lexicográfica que retomo es la presentada en Lara (1997), Lara (2001) y Lara (2016) que proviene del trabajo de creación del DEM, pero a su vez ha ayudado a afianzar sus bases. No tengo el objetivo de crear un profundo análisis comparativo entre los tipos de inventarios de sentidos —sin embargo, sería deseable un método para realizar tal análisis— pero sí es mi objetivo remarcar, de manera general, algunas diferencias a partir de esta teoría.

Si uno de los propósitos de utilizar el DEM es establecer un primer trabajo en este tipo de investigaciones, un segundo propósito descansa en explorar los problemas que conlleva crear vectores utilizando ventanas de palabras. En muchos de los casos, las unidades léxicas son polisémicas. De acuerdo con Lara (2015, 106) la polisemia se entiende como "el fenómeno que consiste en que una palabra tenga, cuando se le considera en aislamiento, es decir, fuera de cualquier contexto, más de un significado (lo cual se muestra claramente en los diccionarios)". Esto, continua el autor, no es resultado de defectos en la lengua, todo lo contrario: "reflejan la capacidad de esta para significar cualquier nueva experiencia" (p. 106). Por lo que la polisemia no debe confundirse con ambigüedad o vaguedad, problemas que emergen de la proposición y del texto. Mientras que para una persona es fácil determinar dada una palabra qué significado es pertinente en un contexto (es decir, los casos de posible polisemia), para una computadora no resulta trivial el proceso que permita determinar cuál de los significados es el adecuado. En principio, la manera de crear el vector parte de tomar contextos de una palabra, aunque pertenezcan a significados distintos, lo que genera problemas al momento de medir los coeficientes (imagínese la palabra *banco* y los posibles

contextos de aparición que serían diluidos en un mismo vector). El problema de que en un mismo vector tengamos mezclados distintos significados, y que por lo tanto, sean indistinguibles en su aplicación computacional, se ha llamado *meaning conflation deficiency* (Camacho-Collados y Pilehvar 2018), el cual he traducido como *dilución de sentido*. La principal utilidad de un inventario de sentidos es que mantiene estas diferencias, lo que ayudaría a desambiguar las unidades léxicas. La heurística detrás de cada algoritmo para implementar estos recursos varía. El conocimiento acumulado y la complejidad de este problema, en apariencia sencillo, ha dado lugar a un campo propio dentro de las tecnologías de la información en búsqueda de una manera efectiva para desambiguar (Stevenson y Wilks 2012).

Parto de que existen tres tipos de inventarios de sentidos: enciclopedias, diccionarios y redes semánticas. Las diferencias entre los tres están en su objetivo, alcance y método de recopilación de entradas. En este punto, entenderé en un sentido amplio *entrada*, en la cual considero cada ítem recolectado por el inventario con su respectiva definición.

En el caso de una **enciclopedia**, por ejemplo, Wikipedia, su objetivo es reunir conocimiento a partir de síntesis y ampliación (la mayor cantidad de conocimiento de un elemento determinado en la menor cantidad de espacio), con la pretensión de ser universal y objetiva. En una entrada enciclopédica podemos encontrar taxonomías, recopilación de acontecimientos históricos e incluso estadística. Su concepción de la lengua es nomenclaturista: “lo que interesa del vocablo es su referencia a las cosas, las palabras son solamente nombres de cosas” (Lara 2016, 79).

Las **redes semánticas**, por ejemplo, WordNet, tienen una estructura similar a la de un *Thesaurus* (tesoro de palabras), en el que se agrupan palabras a partir de sus significados; no

obstante, en el caso de WordNet, los sentidos son agrupados en conjuntos de "sinónimos cognitivos" o *synsets*, los cuales se encuentran vinculados con otros conjuntos por relaciones lexicosemánticas. Además, la conexión entre las unidades sucede por sus sentidos, lo que ayuda a desambiguar palabras que en forma son similares.

No obstante, para un **diccionario**, se parte de distintos principios. De acuerdo con la *Teoría del diccionario monolingüe* (Lara 1997), la entrada forma parte del **artículo lexicográfico**. En esta unidad, un vocablo funciona como entrada, el cual está contenido en el lema; además, el artículo presenta una ecuación sémica que relaciona el lema con su definición. La *definición* no se construye a partir de recopilar la mayor cantidad de información sobre la entidad nombrada, sino en un ejercicio lexicográfico que tiene su impulso inicial en el acto de habla respuesta a las preguntas del tipo "¿Qué significa X?". La *definición* parte de una teoría del signo en donde se comprende que la relación entre significante y significado —entre plano de la expresión y plano del contenido, utilizando los términos de Louis Hjelmslev (1971; Lara 2001, 30)— gravita entre los estereotipos y los prototipos, y que adquiere forma a través de procesos de creación cultural. Esta concepción pone el acento en la observación rigurosa del significado como parte del signo, y en las limitaciones que conlleva: el lexicógrafo, como pensador del significado, no puede ser más que su reconstructor. Su apuesta no está en el acceso aséptico al plano del contenido, sino en dejar ver, de la manera más clara posible, su método de reconstrucción, el cual va de la documentación al análisis. Al finalizar su trabajo, el lexicógrafo se encuentra con un significado de lengua que pretende garantizar la inteligibilidad social, producto de

> … reunir, en un solo esquema, todos los datos obtenidos del análisis de los ejemplos particulares estudiados: precisiones del estereotipo, clasificaciones culturales y

> científicas del objeto significado, características detalladas de los procesos verbales y sus modos de acción, valencias actanciales sistemáticas, matices del funcionamiento semántico de los actantes, etc. (Lara 1997, 230).

En la cotidianidad, una definición de diccionario no es lo que responde un hablante a la pregunta "¿Qué significa X?". En el oficio lexicográfico se busca crear "un depósito de memoria social sintetizada en vocablos y en significados" (Lara 1997, 231) por lo que en el artículo se encontrarán "rasgos de significado que pueden no hacerse presentes en ciertos contextos, criterios de clasificación que muchos hablantes pueden ignorar, ligas culturales inadvertidas por ciertos grupos" (Lara 1997, 230). Con todo lo anterior, busco enfatizar que la *definición*, sea cual sea la manera de presentarla en un diccionario, no es una minimización o síntesis de un artículo enciclopédico u otra versión más de un Thesaurus. Se trata de un ejercicio intelectual cuyo producto, como inventario de sentidos, tiene una profundidad y alcance distinto.

Se podría objetar que los tres tipos de inventarios tienen la capacidad de participar en el acto pragmático de la pregunta y la respuesta acerca del léxico de una lengua. No obstante, y dado lo que he presentado, considero que es el diccionario el hecho orientado a resolver la necesidad de entendimiento de la sociedad (Lara 1997, 103), y no de recopilación exhaustiva, de aprehensión de esencia —caracterizada por ecuaciones sémicas con el verbo *ser* (Lara 1997, 161)— o de inventario lingüístico.

También se podría señalar que los otros dos inventarios tienen la intención de abstraerse para hablar de la lengua, pero me parece que el diccionario, incluso en este aspecto, logra diferenciarse en dos características. Primero, el ejercicio del diccionario busca ser una expresión de **entendimiento** que descansa tanto en el acto de habla respuesta como en su

fuerza ilocutiva, la cual consiste en (i) el anonimato del lexicógrafo, el cual busca desligarse como autor y se presenta como "vocero de la sociedad misma, como la manifestación lingüística de la memoria social del léxico orientada al entendimiento y por entendimiento" (Lara 1997, 104); (ii) su abstracción, propia de la institución que representa ante la sociedad y ejecutada en el ejercicio de reconstrucción del significado; y (iii) la ecuación sémica que plasma, evidencia de esta búsqueda por el entendimiento y materializada por los verbos que se usan en la correspondencia. Esto último es el segundo aspecto que me parece la diferencia crucial entre los otros inventarios.

La ecuación sémica, aproximada en este contexto de entendimiento social, establece una correspondencia entre la definición y la entrada-vocablo articuladas por un conjunto particular de verbos (Lara 2016, 85). De tal manera, si uno pregunta "¿Qué significa *banco*?" la respuesta del diccionario sería la siguiente:

(54) Institución que realiza las múltiples operaciones comerciales a que da lugar el dinero y los títulos que lo representan, como inversiones, créditos, ahorros, pagos, etc

(Diccionario del español de México s/f)

La ecuación sémica implícita en el DEM opera con el verbo *significar* de la siguiente manera (Lara 1997, 158):

(55) *banco* [significa una] institución que realiza las múltiples operaciones comerciales…

Si partimos de una oración como la siguiente:

(56) [El banco] no abre los domingos

Por la capacidad de sustitución (Lara 1997, sec. 2.2.1) que expresa la ecuación sémica, podríamos intercambiar el vocablo por la acepción, como se muestra en el siguiente ejemplo:

(57) [La institución que realiza las múltiples operaciones comerciales a que da lugar el dinero y los títulos que lo representan, como inversiones, créditos, ahorros, pagos, etc] no abre los domingos

Nótese que en (57) fue necesario que el primer sustantivo concordara con la marca de definitud para que la sustitución fuera gramatical; sería lo mismo si se quisiera colocar otro tipo de determinante o cuantificador, como *algún* / *uno de los muchos* / *cada uno de los*.

Es justo esta capacidad de sustitución, diseñada desde el origen mismo del artículo lexicográfico, lo que justifica la inclusión de una definición de diccionario en un trabajo como el que en esta investigación pretendo realizar. Con ello, la capacidad de sustitución, pretendida desde la creación del artículo, le brinda al algoritmo informático información para realizar la diferenciación o vinculación entre frases nominales. Más adelante daré detalles de cómo realicé esta sustitución en las frases nominales con ánimos de encontrar relaciones entre ellas.

En el siguiente apartado, me concentro en la estructura del artículo lexicográfico del DEM y la manera en que lo utilizaré en este trabajo; destaco que, aunque existen otras posibilidades interesantes —utilizar los ejemplos, integrar los contextos del corpus, aprovechar la polisemia, entre otras— existen limitaciones técnicas y de recursos, lo cual me lleva a dejarlas para otro momento. Para esta investigación opto por la más económica.

2.4.3.2 Los artículos del DEM y su tratamiento

Un artículo lexicográfico del DEM está compuesto de la manera en que se muestra en la Tabla 19. El artículo lo encabeza la entrada, que en este caso contiene tres vocablos en negritas. Inmediatamente después le siguen las marcas gramaticales como *s* "sustantivo" y *m* "masculino".

Tabla 19. Artículo lexicográfico de la palabra *banco*

banco[1]
s m

1 Institución que realiza las múltiples operaciones comerciales a que da lugar el dinero y los títulos que lo representan, como inversiones, créditos, ahorros, pagos, etc: *banco de depósito, banco de ahorro, banco ejidal, banco agrícola*
2 Edificio o local en el que tiene sus oficinas esta institución
3 *Banco múltiple* Organismo que concentra todas las formas de comercio con el dinero y otros valores; banca múltiple
4 Cualquier establecimiento en el que se deposita algo para ponerlo al alcance de otros individuos interesados en ello: *banco de sangre, banco de información*

banco[2]
s m
1 Asiento para una sola persona, generalmente sin respaldo
2 Mesa de trabajo, firme y resistente, que usan algunos artesanos, como los carpinteros y los herreros
3 Depósito o acumulación de arena, conchas, corales, etc que en lagos, ríos y mares da lugar a una elevación del fondo, dificultando así la navegación

banco[3]
s m Conjunto muy numeroso de peces que nadan juntos: *un banco de sardinas*

Cada uno de estos vocablos tiene una o varias acepciones, señaladas en algunos casos con números arábigos. El orden para colocar las acepciones sigue dos criterios. Los primeros

lugares las tienen las acepciones con el significado estereotípico, es decir, el significado que los hablantes le atribuyen al vocablo de manera espontánea (Diccionario del español de México s/f); el segundo criterio corresponde a identificar el significado mejor establecido en la cultura, a partir del cual se pueden inferir las otras acepciones.

En este ejemplo no se utilizan números romanos, pero en el DEM esta marca indica que existen varios significados estereotípicos. Cada número romano contiene a su vez un conjunto de acepciones. En esos casos hablamos de polisemia: distintos significados para un mismo vocablo. En los vocablos del ejemplo se puede notar que se han usado superíndices. Esto nos indica homonimia: palabras que se escriben igual, pero su origen y significado es distinto, más allá de los conjuntos de acepciones asociadas a cada vocablo.

Utilizar un diccionario para mejorar las detecciones de similitud entre textos se han reportado desde 1986 con trabajos como los de Michael Lesk y su intento por desambiguar palabras como *cone* "cono" de secuencias como *pine cone* "piña" y *ice cream cone* "cono de nieve" utilizando los traslapes entre acepciones dadas por el *Webster's 7th Collegiate*, el *Collins English Dictionary* y el *Oxford Advanced Learner's Diciontary of Current English* (Lesk 1986). Merece su propio espacio trabajar a detalle experimentos que evalúen las definiciones del DEM para las distintas tareas en donde otros diccionarios han aportado resultados. En lo que respecta a esta investigación utilizaré una heurística que no pretende resolver del todo la polisemia dado un contexto.

Partamos del hecho de que una entrada lexicográfica nos brinda varias lecturas de una misma palabra. Debido a que el texto recibe un tratamiento previo de lematización, en este aspecto, coinciden lema y vocablo; la entrada, en el caso de *banco*, nos da lugar a varios elementos que no presentan conflicto en su categoría gramatical. De esta manera, tenemos ocho

acepciones potenciales a ser sustituidas en un ejemplo como *el banco no abre los domingos* de la sección anterior. El distinguir todas las posibilidades de permutación entre las acepciones y la entrada analizada lo llamaré Permutación simple.

Si realizáramos una Permutación simple de todos los elementos lematizados contenidos en esa oración —que reproduzco a continuación— de acuerdo con el DEM, tendríamos las siguientes posibilidades que presento entre llaves para cada lema, expresadas en la segunda línea:

(58) El banco no abre los domingos

el	banco	no	abrir	el	domingo
{8}	{8}	{12}	{16}	{8}	{3}

Nótese que continúo sin hacer diferencia entre homonimia o polisemia, por lo que elimino la jerarquía predispuesta en el diccionario a través de sus números en índice, arábigos y romanos. Solo estoy integrando aquellas acepciones cuya categoría gramatical no entra en conflicto con la de la palabra analizada de acuerdo con Stanza (el lematizador que describo en §2.4.1).

Esta Permutación simple nos arrojaría la cantidad de 294 912 posibles lecturas de esa frase nominal. A cada una se le deberá construir su vector y después decidir cuál es la mejor candidata para considerar en el detector automático de Estados Informativos. Piénsese que, si se tuviera un texto de cinco frases nominales con la misma cantidad de palabras, tendríamos que generar textos distintos que nos permitan todas las permutaciones. Es decir, $294\,912^5$, lo cual no es sólo una cantidad astronómica que pronostica un tiempo de análisis superior a los ocho meses, sino que es sencillamente *intratable* (*intractable*); esto muestra

que este problema tiende a un crecimiento exponencial (cf. Garey y Johnson 1979). Además, incluso si tuviéramos una máquina que a fuerza pudiera ejecutar tal algoritmo, esto iría en contra tanto del Método computacional que planteé al inicio de mi trabajo, como de las intuiciones que nos proporciona la teoría semántica del análisis latente (§1.12) y la misma teoría del diccionario: las distintas acepciones no brotan al mismo tiempo, sino que se crean al "*precipitarse* los resultados de esos procesos metafóricos en la memoria colectiva de una comunidad lingüística, a partir de un significado principal que preside los significados reunidos en el vocablo" (Lara 2016, 122). En el caso de la homonimia es aún más claro: usualmente el contexto es mucho más transparente para determinar si, por ejemplo, se trata de un *banco*[1] para guardar dinero o un *banco*[2] para sentarse.

Partiré del supuesto de que existen distintas posibilidades de lectura de una frase nominal lo que desata distintas lecturas de un texto. De tal manera, si un texto tiene *y* frases nominales que van de FN_1 a FN_y, y la FN_i tiene n_i lexemas que van de L_1 a L_{n_i}, en donde L_j tiene $\omega_{i,j}$ acepciones, la cantidad de textos-lectura β está dada por:

$$\beta = \prod_{i=1}^{y} \prod_{j=1}^{n_i} \omega_{i,j}$$

La heurística que propongo busca reducir las acepciones $\omega_{i,j}$ a 1. Aunque aplicar un análisis de β puede ayudar a reducir el fenómeno sin perder detalle, representan tiempos extensos de

análisis computacional entre cada experimento para determinar la mejor lectura, por lo que para este trabajo, esta fórmula sólo queda como exposición del problema[79].

Trato de respetar la idea de que, en efecto, existen varias lecturas potenciales, pero nunca todas al mismo tiempo. Por lo que, sólo para el caso de este trabajo de investigación y con ánimos de realizar experimentación con mayor detalle y profundidad en otra ocasión, la reducción sería la siguiente[80]:

a) Sólo se sustituirán sustantivos y verbos, debido a que el propósito es relacionar frases nominales y sus posibles predicaciones debido a oraciones subordinadas.
b) Se seleccionará aquella acepción designada en la jerarquía lexicográfica como **I1 (uno romano, uno arábigo)**, apelando a que es la más estereotípica (Lara 2016, caps. 7–8).
c) Las acepciones dependientes de locuciones quedan descartadas.

---

[79] Si se hubiera continuado el análisis desde esta perspectiva, se nos presentarían dos casos. En el primero, $\beta = 1$, es decir, no existen casos de homónima en ningún lema buscado, por lo que ningún $\omega_{i,j} > 1$. Pero, si $\beta > 1$, se sigue un segundo proceso de selección, ya que el objetivo final es que Interior-w_{DEM} sea sólo una combinación de acepciones. De esta manera, el algoritmo continúa con cada una de las posibles lecturas hasta llegar a las medidas de LSA_{MAX}. Después, para cada texto-lectura β_k se realiza el siguiente cálculo:

$$\mu^{\text{LSA}}(\beta_k) = \frac{\frac{1}{n}\sum_{i=1}^{n} x_{k,i}}{\sqrt{\frac{\sum_{i}^{n}(x_{k,i}-\bar{x_k})^2}{n}}}$$

En donde $x_{k,i}$ es la medida de LSA_{MAX} de las frases nominales del texto β_k. La fórmula lo que expresa es la división del promedio de estas medidas entre su desviación estándar. Con ello, aquel conjunto que tenga un promedio más alto compensado por una desviación estándar más baja garantiza cierta consistencia interna, por lo que esa será la candidata a ser la que nos proporcione las medidas para el Interior-$w_{\text{DEM.}}$ Además de que necesitaríamos no sólo $\mu^{LSA}(\beta_k)$ sino $\mu^{SPAN}(\beta_k)$. Aplicar este proceso es muy costoso, razón por la cual opté por fusionar las acepciones homonímicas.

[80] Incluso el proceso de selección de la mejor acepción dado un contexto podría determinarse de manera inversa: dado un texto con estados informativos, seleccionar aquella combinación de acepciones que produzca un recorrido más cercano al plasmado en el etiquetado. Esto es otra posibilidad que se deja abierta a exploración.

d) Si existiese más de un vocablo en una entrada —en el diccionario diferenciado como homonimia— como en el caso de *banco*, se seleccionarán aquellas que correspondan al **I1** y se fusionaran todas en una misma bolsa de palabras[81].

Al método anterior lo llamo "**Selección única con homonimia**" o Selección UH, que contrasta con la Permutación simple. En esta Selección UH se pierde la capacidad del diccionario de distinguir polisemia, pero se preserva cierta distinción para detectar homonimia. Además, se gana en procesamiento y capacidad computacional. Incluir otro módulo dentro del proceso cuyo único objetivo sea analizar el contexto de la palabra para determinar la mejor acepción va más allá de los objetivos de mi investigación; una empresa de ese tipo necesita su propio espacio de desarrollo.

Se debe tener en cuenta que las acepciones son igualmente lematizadas por Stanza en un submódulo dentro del módulo de afinación del DEM. Su estructura estaría armada de la siguiente manera:

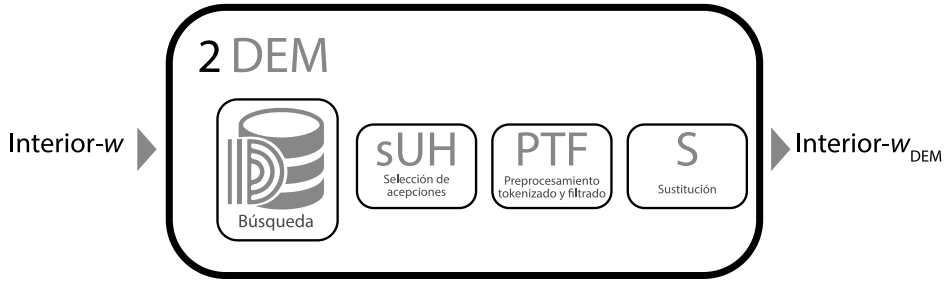


Figura 12. Módulo de integración de acepciones del DEM

[81] La estructura de un diccionario no está ordenada necesariamente para diferenciar la homonimia; puede llegar a estar integrada como parte de las acepciones en una entrada. No obstante, el DEM si especifica esta diferencia en sus entradas con números en superíndice, lo que ayuda a realizar la operación que propongo.

El módulo DEM está alimentado por un texto previamente filtrado y analizado, e integra las acepciones en la Bolsa de Palabras Interior-*w* (cf. §2.4.2). En el módulo se identifican los verbos y sustantivos a través del preprocesamiento de Stanza y se busca en el DEM sus acepciones. Se sigue el proceso de Selección UH, y a las acepciones recolectadas se le somete a un preprocesamiento, tokenizado y filtrado. En este caso, se lematiza la acepción, se eliminan las marcas gramaticales del lema y los ejemplos, y se aplica un único esquema de filtrado en donde se eliminan los pronombres, determinantes, nombres propios, conjunciones, interjecciones, numeración y puntuación.

Dado lo anterior, estamos en tiempo de realizar un ajuste al proceso mostrado en los esquemas anteriores. Hasta LSA_{max}, nuestro esquema lucía de la siguiente manera:

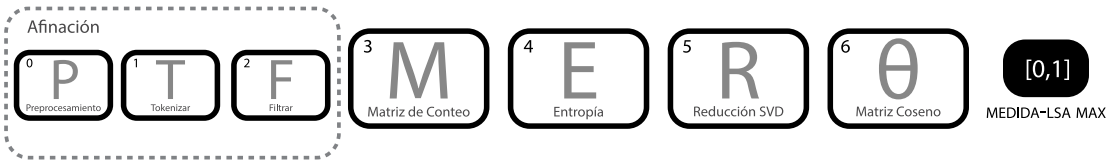


Figura 13. Secuencia de pasos para LSA incluyendo afinación y LSA_{MAX}

En este punto, uniré en un mismo módulo el preprocesamiento, el tokenizado y el filtrado, bajo el nombre de PTF. El interior del módulo DEM lo representaré como se muestra a continuación, suponiendo el proceso que ya he descrito y mostrado en la figura (12).

Figura 14. Secuencia de pasos para LSA incluyendo el DEM y las distintas salidas

Por lo que la afinación da como resultado tres Bolsas de Palabras: Ventana-*n*, Interior-*w* y su variación Interior-w_{DEM}. Cada una, en su propio recorrido de todo el proceso, arrojará sus propias medidas LSA_{MAX} para las frases nominales de la nota analizada.

Con esto se puede notar la complejidad que implica emplear variaciones en la integración del diccionario, no desdeñables para la investigación en lexicografía computacional, pero que tendrán que ser dejadas a un lado debido a los recursos con los que cuento en este trabajo. Al respecto, y antes de finalizar esta sección, me gustaría señalar tres rutas para exploración en un futuro en la utilización del DEM:

1. **Selección del texto** μ: se realizan todas las posibles combinaciones en I1 en los casos de homonimia. Luego se crean todas las posibilidades de textos y se calcula LSA, para al final, seleccionar aquel texto que tenga la medida μ más alta será el considerado al final. El cálculo de la medida μ la explico en el pie de página 78.
2. **Selección por entropía**: se evalúa cada I1 en los casos de homonimia y se selecciona aquella con la menor entropía.
3. **Permutación y entropía**: se evalúan todas las FFNN dadas por la permutación simple y se selecciona aquella con la menor entropía.

Todas las opciones anteriores integran una cadena de texto a la frase nominal, para después ingresarla en forma de Bolsa de Palabras a la matriz de conteo para calcular los vectores. El que se presente más de un texto-lectura ($\beta > 1$), implica que esa misma cantidad habrá en versiones de Interior-*w*. Como mencioné, es un problema que crece exponencialmente. De nuevo, el objetivo de la Selección UH, en este caso, es proporcionar, en la medida de lo posible, un solo texto a la matriz de conteo. Las tres rutas que propongo son sólo algunas que se podrían explorar, faltaría agregar aquellas posibilidades que manipulan las

representaciones vectoriales de cada acepción. Por ejemplo, se podrían implementar operaciones algebraicas para juntar los vectores de frases nominales con los de las acepciones o incluso, aplicar LSA y SPAN al interior del módulo del DEM para elegir la mejor acepción.

En la siguiente sección explicaré el método SPAN que se utiliza como variación de LSA para estudios de información nueva/dada. Este método alterno lo aplicaré a los vectores de la matriz reducida producto de alimentar la matriz de conteo con textos cuyas frases nominales contienen definiciones del DEM a partir de Selección única (Interior-w_{DEM}), con textos con FFNN sin acepciones (Interior-*w*) y las representaciones de las frases usando los contextos de palabras (Ventana-*n*).

2.5. SPAN

LSA es una medida que se obtiene al calcular el coseno entre dos vectores. Las estrategias para obtener mejores resultados consisten en afinar la manera de crear los vectores y en los métodos para la factorización de matrices. Esta medida produjo, en su momento, buenos resultados para buscar información dentro de los documentos o asociar palabras a partir de sus contextos de aparición. Era inevitable pensar si tareas con problemas parecidos podrían resolverse con la misma herramienta. Graesser & Harter (2001) se enfrentaron al problema de crear tutores automáticos: robots “profesores” que pudieran realizar preguntas a humanos a puro estilo mayéutico. Se necesitaba una manera de evaluar si la respuesta del alumno era parecida a la respuesta “modelo” del profesor; pero, además, si el alumno presentaba nueva información a partir de sus propias respuestas anteriores. Es decir, se evaluaba qué tanta información nueva y correcta se entregaba a la computadora con la pretensión de ayudar en

el seguimiento del estudiante y su formación. Diversos algoritmos se han utilizado con este propósito, entre los que se encuentra LSA_{MAX}, pero también tenemos los casos de las medidas de coapariciones de léxico (McCarthy et al. 2012, 463) —estrategia muy parecida a la utilizada inicialmente para desambiguar palabras del diccionario (Lesk 1986)— y las que buscaban utilizar las nociones de coherencia discursiva (Graesser et al. 2004).

En la búsqueda para solucionar la tutoría automática, Hu et al. (2003) proponen una variación algebraica a LSA llamada SPAN[82]. Como mencioné en los antecedentes, aunque esta técnica surgió en este ambiente, SPAN ha mostrado buenos resultados para comparar frases nominales y detectar información nueva/dada (Hempelmann et al. 2005), pero no ha sido evaluada como criterio para etiquetar en español las nociones pragmáticas de Estados Informativos, o sus conceptos asociados como la identificabilidad y la activación.

A continuación, describiré el cálculo realizado para obtener SPAN y señalaré en qué sección del proceso de etiquetado implementaré esta técnica. Las nociones generales de SPAN son las siguientes: en vez de obtener el coeficiente de similitud más alto entre un conjunto de vectores (LSA_{MAX}), creamos un hiperplano (también llamado subespacio) en donde se proyecta el vector "entrante". De esta manera, la pregunta no es qué tan similares son los vectores sino qué tanto es posible que un vector "exista" en el hiperplano de otro vector o de un conjunto de vectores. Esta medida es la que nos dice qué tanta información es *dada*. En términos semánticos, podría arriesgarme a decir que en el proceso se construye un horizonte de significado.

[82] Este término es dado por antonomasia. En inglés y en álgebra lineal, un sistema generador de vectores es llamado *linear span*. Como se verá, SPAN es producto de proyectar a una base, la cual genera un hiperplano.

Cuando proyectamos, se crea un nuevo vector, al cual le calculamos su vector ortogonal, es decir, un vector que, en $\mathbb{R}^3$, tiene un ángulo de 90º con respecto a todos los vectores posibles. Para calcular SPAN se toma la norma del vector nuevo y se divide entre la resta del vector dado con el vector nuevo. En esta medida el 1 indica completa asociación con lo anterior, y 0 indica total novedad.

Veamos con más detalle en qué consiste el procedimiento matemático. Una vez que tengamos la representación vectorial final de las frases nominales —a partir del filtro o conteo seleccionado y después de la reducción— utilizaremos estos vectores para proyectar, uno a la vez, al conjunto anterior. Los vectores deben ser linealmente independientes, con lo que se garantiza que cualquier conjunto de ellos crea una base a donde proyectar. Esto se resuelve con la reducción por SVD en el proceso de LSA, por lo que SPAN se inserta inmediatamente después de ese paso. Para este ejemplo, utilizaré los siguientes tres vectores de tres dimensiones linealmente independientes:

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 3 & 5 & 1 \end{pmatrix}$$

El primer vector es $\overrightarrow{v1}$ = (1, 0, 3). Como no existe ningún antecedente, no realizamos ninguna operación, por lo que su valor de SPAN es 0. El siguiente vector es $\overrightarrow{v2}$ = (2, 1, 5). En este caso, debido a que sólo hay un vector antecedente, se realiza una proyección simple, a partir de la siguiente fórmula:

$$proy_v u = \frac{u \cdot v}{|v|^2} \cdot v$$

La cual se lee como que la proyección de un vector *u* en *v* es igual al producto punto de *u* y *v* entre la magnitud del vector *v* al cuadrado por el vector *v*. El resultado anterior nos produce un nuevo vector, el vector de lo *dado*, el cual llamaremos $\overrightarrow{Dv2}$ y tiene la siguiente forma:

$$\overrightarrow{Dv2} = (1.13333333, 0.56666667, 2.83333333)$$

Para obtener el vector de lo nuevo, se resta el vector entrante, en este caso $\overrightarrow{v2}$ a $\overrightarrow{Dv2}$, con lo que garantizamos que este vector sea ortogonal a la proyección. Esta operación luciría de la siguiente manera:

$$\overrightarrow{Nv2} = \overrightarrow{v2} - \overrightarrow{Dv2}$$

Lo cual nos da como resultado el siguiente vector:

$$\overrightarrow{Nv2} = (0.3, 1.0, -0.1)$$

Finalmente, para obtener SPAN, obtenemos las normas de $\overrightarrow{Dv2}$ y $\overrightarrow{Nv2}$, y aplicamos el siguiente cálculo[83]:

$$\text{SPAN} = \frac{\overrightarrow{|Nv2|}}{|\overrightarrow{Nv2}| + |\overrightarrow{Dv2}|}$$

Esto, sustituyendo, nos da como resultado lo siguiente[84]:

$$\text{SPAN} = \frac{1.04}{1.04 + 5.37} = 0.16$$

[83] La indicación en el método sostiene que para calcular la novedad de la información "a proportion score is then taken: Span(new information) = N/(N+G). N is the component of the vector that is perpendicular to the hyperplane and G is the projection of the vector along the hyperplane" (Hempelmann et al. 2005, 944). Se podría construir otra medida a partir de a *Span(new information)* al restarle 1, con lo que se obtiene "lo dado". No obstante, mantendré en mi trabajo esta diferencia, ya que podría coincidir con los resultados de los autores: SPAN podría servir para algunos estados informativos, y LSA_{MAX} para otros.

[84] Para estos ejemplos he tomado sólo los primeros dos decimales de cada norma. En los cálculos de la investigación, tomo todo el número.

Por lo que, hasta este punto, nuestras representaciones vectoriales de las frases nominales tendrían asociados estos valores:

Tabla 20. Ejemplo de resultados previos de SPAN

	SPAN
$\overrightarrow{v1}$	0
$\overrightarrow{v2}$	0.16
$\overrightarrow{v3}$	--

Siguiendo el procedimiento, se proyecta el vector entrante $\overrightarrow{v3}$ sobre los vectores anteriores, en este caso sólo sobre $\overrightarrow{v1}$ y $\overrightarrow{v2}$, y se suman los resultados, con lo cual obtenemos el vector $\overrightarrow{Dv3}$. Esto se resume en la siguiente fórmula:

$$\overrightarrow{Dv}_k = \sum_{i=1}^{k-1} \frac{v_k \cdot v_i}{|v_i|^2} \cdot v_i$$

Lo demás sigue el proceso anterior; se continua con la resta de $\overrightarrow{v_3}$ a $\overrightarrow{Dv_3}$, y a este resultado —que en este caso será $\overrightarrow{Nv3}$— le calculamos la norma, de tal manera que podamos resolver la siguiente operación:

$$\text{SPAN} = \frac{\overrightarrow{|Nv3|}}{\overrightarrow{|Nv3|} + \overrightarrow{|Dv3|}} = \frac{46.50}{46.50 + 2.21} = 0.95$$

Nuestra tabla (20), junto con la medida SPAN de $\overrightarrow{v_3}$ y las medidas LSA_{MAX} de cada vector siguiendo el proceso que describí en la §3, luciría de la siguiente manera:

Tabla 21. Ejemplo de resultados de SPAN y LSA_{MAX}

	SPAN	LSA_{MAX}
$\overrightarrow{v_1}$	0	1
$\overrightarrow{v_2}$	0.16	0.98
$\overrightarrow{v_3}$	0.95	0.57

En este caso, los vectores no son realmente representaciones vectoriales de frases nominales; han sido sólo tres vectores de ejemplo. Algo interesante a notar es que SPAN no es complemento de LSA_{MAX}. En todo caso, este ejemplo no está basada en texto y sólo tiene como objetivo ser ilustrativo.

Por lo pronto, sólo resta señalar que si tuviéramos un hipotético $\overrightarrow{v4}$, el proceso sería el mismo:

1. Proyección a cada vector anterior.
2. Suma de las proyecciones.
3. Resta del vector entrante $\overrightarrow{v_k}$, al vector suma de proyecciones $\overrightarrow{Dv_k}$.
4. Cálculo de SPAN con las normas de $\overrightarrow{Dv_k}$, y el resultado de la resta, $\overrightarrow{Nv_k}$.

LSA_{MAX} y SPAN serán las dos medidas reportadas tanto de interior-*w*, interior-w_{DEM} y ventana-*n*. El proceso de medida culminaría con esto, para después evaluar cuál es la medida que mejor se asocia con los Estados Informativos. Por el momento, he integrado SPAN al diagrama, el cual luce de la siguiente manera:

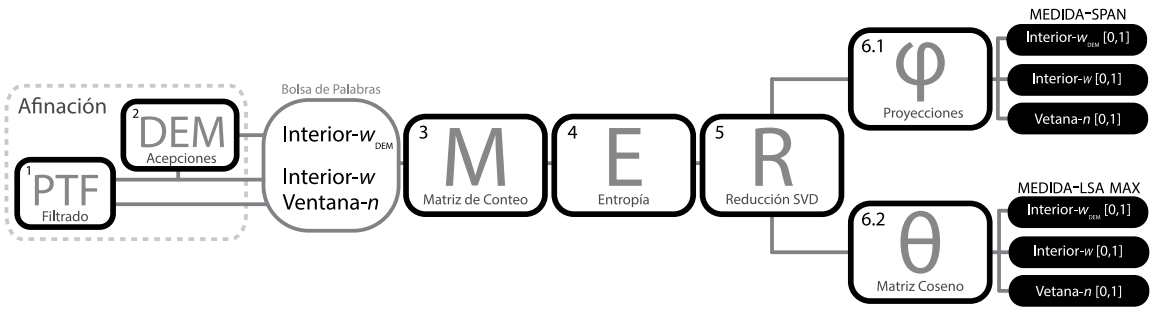


Figura 15. Secuencia de pasos para calcular LSAmax y SPAN

En lo que sigue, abordaré la forma en que construí el corpus de notas periodísticas de donde extraje las notas que forman parte del análisis final, así como sus respectivas frases nominales. Cada nota, como texto, se integra al proceso antes descrito. Pero antes de ello, las notas pasan por un proceso de etiquetado manual y, una vez se tengan identificadas las frases nominales, se construyen las tres Bolsas de Palabras. Esto lo describiré al final de la siguiente sección.

2.6. Creación de COPENOR

Los corpus para probar las medidas LSA_{MAX} y SPAN como medidas de lo nuevo/dado han consistido en diálogos del tipo pregunta-respuesta y textos de educación básica, como ya señalé en los antecedentes. En este caso, busqué variar la perspectiva de los textos y atender otra necesidad, paralela a la que trato en esta investigación. Aunque gran parte de los antecedentes en procesamiento de lenguaje natural (PLN) están en resolver tareas que ayuden a reducir los tiempos de análisis, como el caso del analizador del DEM, la explosión de información que ha sucedido en los últimos 10 años ha sido tal que ya no es un asunto de volver más eficiente el análisis: parece ser la única opción.

Por poner un ejemplo, la pandemia por coronavirus (Covid-19) que inició el 2019 —y que, al momento de escribir esta tesis, aún presenta un patrón de fallecidos y contagiados que no decrece— ha generado un alud de artículos de investigación. De acuerdo con CORD-19, una iniciativa para construir una base de datos abierta que reúne investigación sobre el tema (Wang et al. 2020), al 13 de septiembre del 2020 se han recolectado 253 454 artículos; considérese que la base inició con 44 220 artículos el 13 de marzo del 2020. Incluso con el conjunto inicial, resulta imposible que un ser humano pueda consumir esa cantidad de información. Esto tan sólo en el terreno del discurso científico especializado. Si nos enfocamos en la información que se ha divulgado por medios de comunicación, el número es incapturable. Esto no es exclusivo de los tiempos de pandemia. Este comportamiento se ha predicho y observado con la llegada de los nuevos medios de comunicación digitales y la descentralización de la creación de contenidos (Schrape 2019; Holton y Chyi 2012).

La primera característica del corpus que busco crear surge de esta necesidad en la investigación del discurso periodístico con métodos computacionales dadas estas nuevas dinámicas de producción y consumo. Los resumidores automáticos y los extractores de información de noticias no son un tema nuevo en el área de PLN. Como mencioné, muchos corpus inician con este tipo de discurso, entre otras razones, porque es sencillo recopilarlo. El corpus con el que está entrenado Stanza está construido precisamente con notas periodísticas. No obstante, las recopilaciones de los textos suceden con una perspectiva en la que las lenguas son entendidas en un sentido amplio por lo que se suponen y generalizan patrones. Se parte de un mismo y único corpus que se considera *representativo* de cada *lengua*. Además, se deja de lado el propósito original del corpus. Si bien, es posible realizar en ellos análisis del discurso, no son creados con ese fin. Esto presenta dos problemas, uno

que compete al mismo estado del arte en procesamiento de lenguaje natural, y otro, a estudios lingüísticos.

En el área de PLN, se suele utilizar un mismo conjunto de corpus para entrenar las inteligencias artificiales (IA), en particular, me refiero a los algoritmos creados desde la perspectiva del Aprendizaje de Máquina (*Machine Learning*). En parte, esto ha ayudado a crear un estándar en determinadas tareas, pero a su vez, genera sesgos. La *inteligencia* entrenada se limita a los patrones que encuentra en ese conjunto y los trata de generalizar a otros conjuntos de datos. Entrenamiento con diversos corpus ayuda a expandir el alcance de las IIAA. Además, al momento de la ejecución, se necesita probar que el algoritmo ha aprendido a generalizar de manera correcta. Para ello, tener corpus distintos, tratados por expertos, con los que la inteligencia no se haya encontrado antes, suma a la mejora del proceso. Por lo que el problema es la poca diversidad en los corpus, tanto de entramiento como de evaluación, no porque en sí mismo haya poca cantidad de corpus disponibles sino porque siempre se deseará tener acceso a más datos que ayuden a la tarea. Debido a esto, este corpus trata de noticias en español mexicano, en particular, en el dialecto hablado en el noroeste del país.

En el área de estudios de dialectología y sociolingüística, se sostiene que México puede dividirse en diversas zonas dialectales (Henríquez Ureña 1921; Lope Blanch 1970). En particular, me interesa aportar y sostenerme en la hipótesis que demarca el noroeste de México como zona dialectal. Entre los estudios que afirman esta división se encuentran Brown (1989), Mendoza Guerrero (2006; 2004), Moreno de Alba (1994) y Serrano (2000). Las investigaciones de estos autores apuntan a variaciones léxicas y fonológicas, así como a la distinción subjetiva por parte de los mismos hablantes de la zona. Por lo que considero al

noroeste de México tal y como lo divide Lope Blanch (1996), como aquella zona que comprende los estados de Baja California, Baja California Sur, Chihuahua, Durango, Sinaloa y Sonora. Se quedará pendiente la investigación que pruebe si en los medios de comunicación digitales se plasma la variación o se refuerza un español nacional. Aunque ese no es el foco de mi investigación, sí pretendo construir este corpus con esta perspectiva para abrir posibilidades a futuras investigaciones tanto en el área de la sociolingüística como en el de las ciencias de la comunicación.

Parecería que un trabajo que busca detectar información nueva demanda con naturalidad notas periodísticas, pero la progresión de información nueva y dada en un texto no es exclusiva de una nota, es parte del proceso de informar en la comunicación. En particular, en los estudios relacionados con ciencias de la comunicación, la noticiosidad (*newsworthiness*) no es una propiedad que dependa de manera exclusiva en las suposiciones de conocimiento compartido/nuevo del periodista con su audiencia. Esta propiedad, muchas de las veces producto de la experiencia del periodista, se conforma por una serie de valores intrínsecos a los hechos que suceden en el mundo, lo que los destaca de la cotidianidad y los vuelven dignos de ser noticia (O'Neil y Harcup 2009, 161). El corpus que presento está conformado de notas, no por esta intuitiva relación a presentar información nueva, sino por su fácil obtención, por tratarse de medios digitales, por su integración a un trabajo posterior de sociolingüística y a la aportación en la variación de corpus dialectales para el trabajo en PLN. Evaluar qué tanto corresponde el criterio comunicológico de la noticiosidad con el nivel lingüístico de la estructura de la información va más allá de los objetivos de este trabajo, pero sería interesante revisarlo en un futuro.

El trabajo del periodista consiste en detectar los hechos relevantes a partir de evaluar su noticiosidad, y después vaciar su síntesis en una estructura discursiva convencional que agilice tanto la producción de la nota como la lectura. Estas estructuras persiguen objetivos comunicativos distintos, lo que da lugar a los diferentes géneros periodísticos. En este corpus sólo consideraré el género informativo *noticia*, dejando de lado los géneros interpretativos, dialógicos y argumentativos (Salaverría y Cores 2005, 150). Las características generales de la noticia, establecidas desde finales del siglo XIX, son las siguientes (las tomo íntegras de Salavarría & Cores (2005, 152; negritas propias)):

a) **Título informativo.** Expone de forma escueta, con una sola frase, lo más relevante del hecho noticioso.
b) **Entrada de sumario.** El arranque del texto responde, en uno o dos párrafos, a las seis preguntas básicas de la información: quién, qué, dónde, cuándo, cómo y por qué.
c) **Pirámide invertida**. La información restante se estructura en el cuerpo del texto en orden de interés decreciente y con párrafos autónomos.
d) **Estilo impersonal.** El texto recurre a la voz impersonal y proscribe elementos que expliciten juicios de valor, tales como adjetivos calificativos.

Si bien, al momento de revisar y capturar los productos comunicativos en los medios digitales vigilé que correspondieran a este género, no todos comparten punto por punto estas características. Limitar el corpus a este género puede ayudar a circunscribir resultados de ésta y otras investigaciones que lo utilicen para estudios de medios. Lo anterior declara las motivaciones que guían la construcción del Corpus Periodístico del Noroeste de México o COPENOR.

2.6.1 Muestra y captura de las notas

El primer paso en la captura consistió en la creación de la base de medios. Este paso lo realicé en tres momentos: primero, revisé dos bases de medios de comunicación digitales: http://www.prensaescrita.com/ y http://www.abyznewslinks.com/. Después, consulté periodistas locales para evaluar si hacía falta algún medio importante. Esto resaltó la falta de un censo respaldado por alguna institución de investigación o colegio de profesionales sobre los medios de comunicación digitales e impresos activos. Salvando esto, identifiqué 125 medios digitales en la región noroeste. Finalmente, visité los sitios de los medios para corroborar que estuvieran activos y verificar que tuvieran notas con no más de una semana de antigüedad.

Este primer paso resultó en 94 medios, cuyos datos pueden ser consultados en el Anexo C. La distribución por estados y medios se puede observar a continuación, en donde se muestra que Chihuahua cuenta con la mayor cantidad de medios digitales activos (33), seguido de Sonora (21) y Baja California (17):

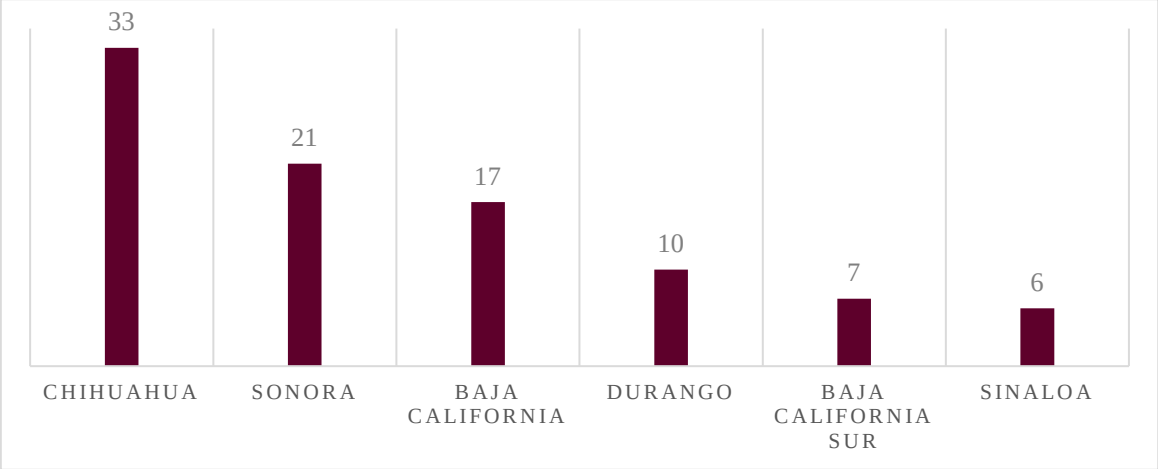


Figura 16. Gráfica de distribución de medios de comunicación activos por estado

En la captura, coloqué la ciudad si es que se mencionaba en el medio. Si no se mencionaba, coloqué el nombre de la capital del estado. Sobre esto, resalta el hecho que, en el estado de Chihuahua, la ciudad de Chihuahua tiene 15 medios activos y Ciudad Juárez 14 –compárese con Tijuana, Baja California, que tiene 11 medios activos. No todas las ciudades de cada estado cuentan con medios representados en esta primera base.

Para determinar el tamaño de la muestra a capturar –con la intención de que sea demográficamente representativa, utilicé la siguiente fórmula:[85]

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

Tomando en cuenta que un medio activo genera alrededor de 10 notas diarias, esto significa una producción de 56 400 notas en un periodo de sesenta días (dos meses) –que comprendió del 23 de mayo al 16 de julio del 2019. Pasando el cálculo con un 95% de confianza y un 5% de margen de error obtenemos 380 notas periodísticas a capturar.

La forma en que se distribuyen los medios en los estados no es regular. Para mantener esta proporción, el muestreo aleatorio que realicé es por conglomerado. El procedimiento fue el siguiente. Se contabilizaron los medios por estado, quedando cada conglomerado de la siguiente manera:

[85] 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.

Tabla 22. Suma acumulativa de los medios por estado

Estado	Rango de medios
Baja California	1-17
Baja California Sur	18-24
Chihuahua	**25-57**
Durango	58-67
Sinaloa	68-73
Sonora	74-94

Después, se obtuvo un número aleatorio entre 1 y 94 para determinar qué estado seleccionar. Por ejemplo, si llegase a salir 39, el estado a revisar sería Chihuahua (marcado con negritas en la tabla). Con el estado seleccionado, se obtiene otro número aleatorio, esta vez entre 1 a 33 (recordemos que Chihuahua tiene 33 medios activos). Este número nos dice el medio dentro de Chihuahua a seleccionar. Una vez seleccionado el medio, se selecciona la noticia a integrar al corpus. Para capturar las notas, se dio preferencia a aquellas que informaran de noticias locales firmadas por periodistas. También se integraron notas que no estuvieran firmadas por periodistas, pero que trataran de acontecimientos locales. Sólo aquellas firmadas por agencias fueron descartadas. COPENOR está codificado en Lenguaje de Marcador Extensible (XML). Asociado al documento principal se encuentra un formato de nombres únicos (XSD), proporcionado por el DEM. Los elementos capturados y sus etiquetas se muestran a continuación:

Tabla 23. Estructura del XML de la nota en COPENOR

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 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 explícita, se coloca la ciudad del medio o la capital del estado. 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 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.
<etiquetado>	Contenido textual etiquetado con frase nominal, oración, Stanza, lema y Estados Informativos.

Para determinar el muestreo por conglomerado creé una rutina en Python llamada *rndm_medios.py* (Anexo P) que realizaba todo el procedimiento y entregaba un archivo xlsx

(Excel) con medios y fechas de captura a lo largo de los dos meses programados. Este itinerario puede ser revisada en el Anexo D. Al final, las notas capturadas por cada estado siguen la distribución que muestro en la Figura 17:

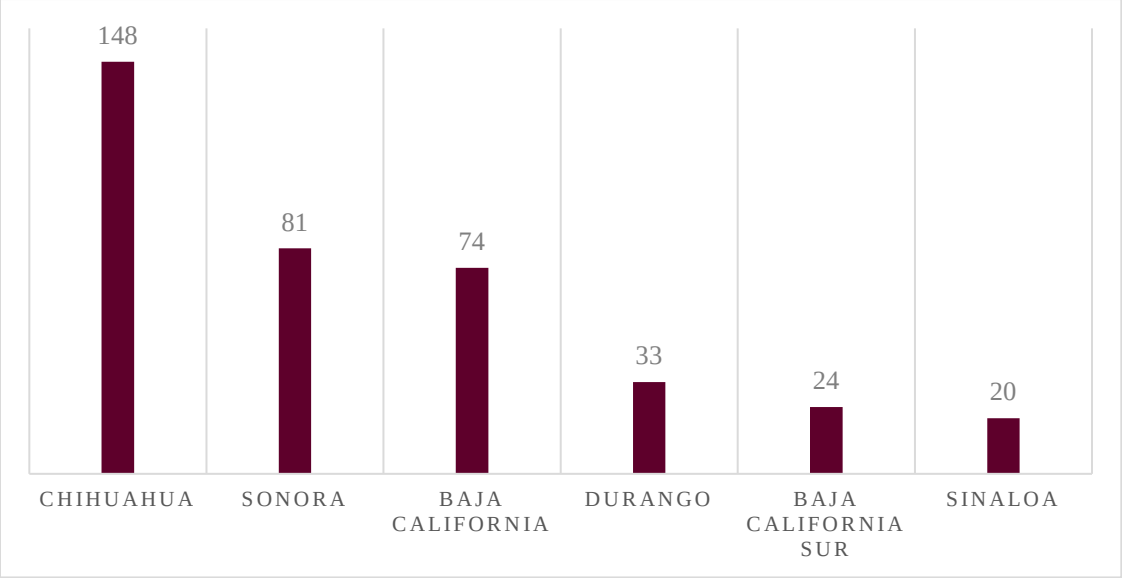


Figura 17. Distribución de notas por estado

La captura no fue un procedimiento exento de problemas. Durante los dos meses se presentaron dos tipos de contratiempos. Primero, una vez empezada la captura, resultaba que la página del medio estaba desactivada a pesar de la primera revisión. Para solucionar este problema, se escogía otro medio del estado de manera aleatoria y se marcaba el medio desafortunado con un asterisco. No se eliminaba el medio de otras entradas en el itinerario ya que podía ser que sólo estuviera temporalmente inaccesible.

Segundo, algunos medios tienen una producción de notas más baja que otros. En esos casos, como la condición fue sólo que la nota más reciente del medio no tuviera más de una semana de antigüedad, si resultaba que aparecían en el itinerario dos veces en una semana, se buscaba

una nota más allá del rango del tiempo de captura. Se encontrará que COPENOR tiene algunas notas con fechas de creación fuera de los dos meses programados.

Finalmente, otro aspecto que no resultó contraproducente, pero que debe tomarse en cuenta, es que la extensión de la nota no fue criterio de selección, por lo que existe amplia variación en esta propiedad (en el Capítulo 3 trataré algunos datos descriptivos sobre esto). Así mismo, en COPENOR también se encontrarán notas que incluyen citas directas debido a que tampoco se consideró la cita como criterio de selección, por lo que estos contenidos deberán ser ponderados si se quiere utilizar el corpus en otro tipo de análisis o cálculos estadísticos. En todo caso, dejo que Stanza etiquete las comillas, las cuales pueden ser recuperadas en rutinas de Python para extraer las citas completas.

2.6.2 Etiquetado manual de COPENOR

Posterior a la primera captura de notas se procedió al etiquetado manual de la oración, la frase nominal, de las categorías sintácticas (sintetizadas como se explicó en el Capítulo 1) y los Estados Informativos. El tomar todo el corpus de 380 notas resultó ser una tarea que se extendía más allá del tiempo estipulado para esta investigación. No obstante, este primer corpus permitió un segundo manejo para extraer suficientes notas como para superar el umbral de 2 000 frases nominales y realizar la exploración de los cálculos. De esta manera, se volvieron a seleccionar de manera aleatoria, cuidando la proporción de los estados, 38 notas (lo que representaría el 10% del corpus) con lo que se obtuvieron 2 388 frases nominales. El proceso para obtener este subconjunto fue el siguiente. Para garantizar que se mantuviera la proporción de la muestra original, se normalizó la cantidad de notas de todos los medios a partir del estado con la menor cantidad, que en este caso fue Sinaloa con 20

notas. De tal manera, este número mínimo fue de 19 notas: 7 Chihuahua, 4 Sonora, 4 Baja California, 2 Durango, 1 Baja California Sur y 1 Sinaloa. Con este primer conjunto no se alcanzaban las 2 000 frases, por lo que se tomó otro conjunto de las mismas proporciones, llegando a las 38 notas.

En cuanto al etiquetado manual de estas 38 notas, seguí los siguientes lineamientos, ejecutados en dos fases: primero, se etiquetaron las frases nominales y las oraciones en formato XML con las etiquetas <fn> y <ora> respectivamente dentro de la etiqueta <etiquetado> de cada nota en COPENOR. La fase de etiquetado manual se dividió en dos periodos: en un primer periodo que duró aproximadamente cinco meses, se probaron las estrategias para el etiquetado, lo que resultó en un primer conjunto de 30 notas etiquetadas. Este primer periodo sirvió para resaltar errores en el procedimiento y el establecimiento de un método más robusto de etiquetado. Después, siguió un segundo momento en donde se escogieron de manera aleatoria 38 notas y se etiquetaron en un lapso de tres meses aproximadamente. El sistema de lineamientos generado puede ser consultado en el Anexo A.

Algunas estrategias implementadas en XML divergen un poco de análisis sintáctico tradicional, como, por ejemplo, que las frases nominales con nombres propios como modificadores se conservan en una misma etiqueta como:

(59) <fn>la calle Eusebio Kino</fn>

O las frases nominales que refieren a ciudades como:

(60) <fn>la ciudad de Ensenada</fn>

Para las oraciones sólo se tiene en consideración que las conjunciones, preposiciones o nexos discursivos que introducen la oración se incluyen dentro de la etiqueta, como, por ejemplo:

(61) a. <ora>**Por lo anterior**, no se dejó <fn>el trabajo</fn></ora>

b. <ora>No hubo <fn>trabajo</fn></ora><ora>**pero** hubo <fn>salario</fn></ora>

c. <ora>**para** evitar <fn>cualquier incidente mayor</fn></ora>

Esto contrasta con el etiquetado de las frases nominales, en cuyo caso, aquellos elementos que las introducen quedan fuera de la etiqueta. También se consideraron los gerundios, participios y verbos en infinitivo como oración siempre y cuando pudiera analizarse algún constituyente dependiente a ellos. Por ejemplo:

(62) a. <ora>Juan se levantó molido <ora>**por correr todo el día**</ora></ora>

b. <ora>El techo <ora>**mojado por la tormenta**</ora> se cayó</ora>

c. <ora>Pedro se encontró a <ora>Juan cargando una canasta</ora></ora>

Nótese que en este ejemplo no muestro las etiquetas de <fn> con el fin de simplificar la exposición.

La primera etapa del etiquetado manual se realizó junto con un equipo de becarios de la Universidad Autónoma de Baja California de Ensenada que participaron con sus intuiciones como hablantes en la segmentación de constituyentes y en las discusiones. El medio local *Ensenada.net* abrió un espacio para poder realizar la revisión del análisis en sus instalaciones y liberar los servicios. La segunda etapa se realizó de manera individual y es la que produjo las 2 388 frases nominales. Para la asistencia en el etiquetado, se construyó una interfaz desde la línea de comando que leía el XML de COPENOR con las etiquetas de <fn> y <ora>. Este sistema, llamado *Leon-Eti* (*leoneti.py* en Anexo P), produce otro XML que puede ser leído y procesado por otra rutina que implementa Stanza.

2.6.2.1 Etiquetado de categorías sintácticas

La interfaz de Leon-Eti permitía observar las frases nominales en su contexto y asignarle una categoría gramatical. Para ello, se partió de las categorías gramaticales de Sujeto, Objeto Directo y Objeto Indirecto. El principal criterio para el análisis de roles fue la sustitución por clíticos *lo* y *le* y sus respectivas permutaciones con plural y género. Para el sujeto, se observaba la flexión en el verbo y si era necesario, permutar entre plural y singular o entre personas si es que había dudas de la función de algún constituyente. También se tomaron en cuenta las frases nominales introducidas por preposición, pero no se analizó si realizaban algún rol gramatical. Además de estos casos, también se analiza el segundo constituyente en los predicados con verbo *ser* como atributos.

2.6.2.2 Etiquetado de Estados Informativos

En el proceso creado a través de Leon-Eti se pregunta por el valor del Estado Informativo de la frase nominal analizada. Como se expuso en el Capítulo 1, estos valores van del 0 al 9. La primera pregunta que plantea el programa es si el referente ya había sido mencionado antes o no. Si bien, uno puede ingresar el número de la etiqueta desde el inicio, esta primera pregunta ayudaba a concentrar la atención en aquellas etiquetas de segundas menciones (las activas o la que apelan a la memoria del registro discursivo, por ejemplo). Además, conforme avanzaba el análisis de la nota, se guardaban las frases nominales en una lista paralela, lo que permitía explorar el registro discurso de manera rápida.

2.6.3 Etiquetado automático de COPENOR: Stanza como paso de transición

La nota etiquetada de manera manual es la que entra al preprocesamiento de Stanza, y es en este paso en donde inicia el proceso automático. Por un lado, Stanza analiza la estructura morfosintáctica y nos entrega una tabla con los valores detectados; se extraen verbos y sustantivos del interior de la frase para su búsqueda en el diccionario, se extraen las palabras del contexto, y se crean las distintas bolsas de palabras, de acuerdo con los filtros ya mencionados en la sección 2.4.2. En el Anexo P se puede encontrar una serie de carpetas que estructuran paso a paso los datos de salida, con lo que se deja ver que el corpus se va transformando en cada momento. Esto lo sintetizo de la siguiente manera, en donde tomo como guía el nombre de las carpetas digitales:

- **copenor_cero_a:** 38 notas extraídas de manera aleatoria. Etiquetado XML de frase nominal y oración.
- **copenor_cero_b:** 38 notas. Etiquetado manual de categorías sintácticas y Estados Informativos.
- **copenor_uno_stanza:** las notas se analizan en Stanza y se generan tres tipos de archivos para cada nota: un CSV, un XLSX y un XML. Se reúne el análisis de Stanza de nuevo en el archivo XML y se reúnen todas las notas en un solo archivo.
- **copenor_uno_contextos:** para cada nota en el XML se genera un archivo CSV en donde se colocan las bolsas de palabras, filtradas y lematizadas, de la ventana-*n* y el interior-*w*. En este paso, se extraen sustantivos y verbos del Interior-*w* y se buscan en el DEM. Se realiza la Selección UH (§2.4.3.2), se lematiza la cantidad de acepciones capturadas. Se filtran de acuerdo con lo estipulado, y se integra a la bolsa de palabras con el nombre clave INTERIOR-*wd*.

- **copenor_dos_salida:** para cada nota se crean archivos XLSX de los cálculos para las medidas SPAN y LSA_{MAX} de las bolsas interior-*w*, interior-*wd* y ventana-*n*. En estas tablas se da detalle de todo el proceso ya desarrollado en §2.3 a §2.5. Además, se reúnen en un solo archivo CSV todas las frases nominales y se le da un último tratamiento para la experimentación.

El recorrido final, antes de la experimentación, luce como se muestra en Figura 18 (p. 198). En ella se incluye el etiquetado manual del corpus y lo ya visto del tratamiento automático. El archivo que se extrae de **copenor_dos_salida** lo he llamado *tabla_de_salida.csv*. Este contiene las 2 388 frases nominales, sus respectivas etiquetas de Estados Informativos y las medidas SPAN y LSA_{MAX}, entre otras propiedades. En la siguiente sección detallo la estructura de esta base de datos final. De ella se extraen los factores y las medidas para realizar las pruebas estadísticas que nos indicarán la capacidad de estas para ser buenas predictoras de los Estados Informativos o si un modelo que combine factores y medidas es mejor para la clasificación.

2.7. Experimentación y pruebas estadísticas

A lo largo de este trabajo he utilizado el término *las medidas* para hacer referencia a los seis números asociados a la frase nominal después de aplicar los cálculos algebraicos con los que se obtiene SPAN y LSA_{MAX}, esto a partir de tres bolsas de palabras generadas de distinta manera. El objetivo es encontrar que, dada una medida, se pueda predecir el Estado Informativo, por lo que, de manera propia, las medidas son factores o variables independientes. No obstante, y sólo por una cuestión de dejar clara la diferencia con respecto al objetivo de esta

investigación, llamaré *factores* a aquellas propiedades que no son centrales en mi estudio pero que represento. Esto, por un lado, para vincular las deducciones mostradas en otras investigaciones con mi trabajo, en particular, aquellos factores que, de acuerdo con Kibrik (2011), parecen ser buenos predictores del estado de activación y de la forma del dispositivo referencial (§1.8). Por otro lado, lo hago para contrastar: podría ser que alguno de estos factores secundarios pudiera ser mejor predictor. En todo caso, las medidas son mi principal interés. Como introduje en el Capítulo 1, los factores integrados son sólo aquellos que puedan extraerse de manera automática con los recursos que he implementado. Dejando clara esta diferencia entre *las medidas* y *los factores*, a continuación, explico la base de datos de salida.

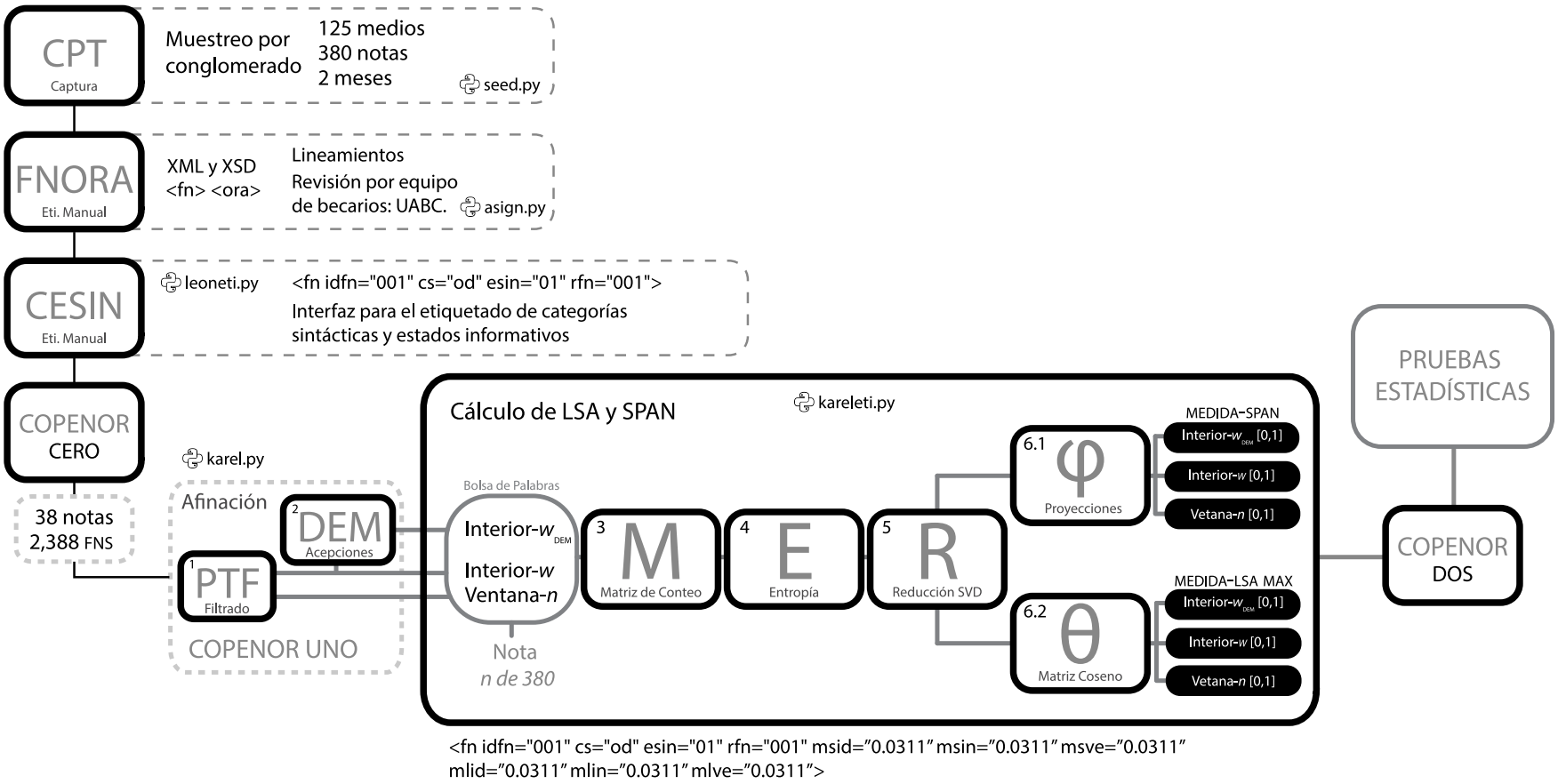


Figura 18. Recorrido final del etiquetado manual y el tratamiento automático de las notas de COPENOR

2.7.1 Variables independientes de la tabla de salida

Cada entrada de la tabla inicia con un identificador global (index), el identificador de la nota de donde proviene la frase (IDN), y el identificador de la frase dentro de esa nota (IDFN). Después se incluye la frase nominal sin signos de puntuación y luego los factores extraídos de manera automática. Estos son:

- **Definitud**: Stanza puede asignar de manera automática los valores semánticos de definitud al determinante que encabece la frase nominal. Se extrae este valor y se coloca en la base de datos asociada a esa frase nominal (1 presencia, 0 ausencia).
- **Indefinitud**: De la misma manera que el caso anterior, también se coloca en la base de datos si el primer determinante contiene el rasgo indefinido de acuerdo con el análisis automático de Stanza (1 presencia, 0 ausencia).
- **Distancia relativa en la nota:** se construyó un índice con respecto al número total de frases nominales y su posición en la nota que va de 0 (posición inicial) a 1 (última frase nominal).
- **Tamaño de la frase nominal:** de acuerdo con las palabras identificadas por Stanza.
- **Cantidad de nominales dentro de la frase nominal:** de acuerdo con el analizador de Stanza, el número de etiquetas Nominales que contiene la frase. Para este análisis se descartan los nombres propios debido a que se asume baja capacidad de este tipo de algoritmos para vincularlos.
- **Cantidad de verbos dentro de la frase nominal:** de acuerdo con el analizador de Stanza, el número de etiquetas Verbales que contiene la frase.

El único factor que no extraje de manera automática fue **el estatus gramatical de la frase** de acuerdo con el análisis de categorías gramaticales reducido propuesto. Este factor fue

etiquetado de manera manual al momento de analizar los Estados Informativos. En la base de datos, esta propiedad tiene cinco variaciones: sujeto (su), objeto directo o indirecto (ob), introducido por preposición (pr), atributo en una oración de verbo *ser* (at), y no aplica (na) que, como expuse, son los casos que no forman parte de una predicación verbal.

Con excepción de este último factor, todos los anteriores se obtienen de manera no-supervisada a partir del etiquetado de COPENOR DOS (ver Figura 18). De esta manera, tenemos un total de 13 variables independientes, de las cuales seis son las medidas y siete son factores relacionados con características morfosintácticas y discursivas. Para realizar las pruebas estadísticas, todas las variables categóricas se tradujeron a un sistema numérico.

2.7.2 Experimentación y variables dependientes

Con el fin de realizar experimentación, he creado tres conjuntos alternos al conjunto inicial de 10 etiquetas de Estados Informativos (a partir de este momento, utilizaré la abreviatura ESIN). A manera de recordatorio, se reproduce el esquema de las etiquetas propuesto en la sección 1.10:

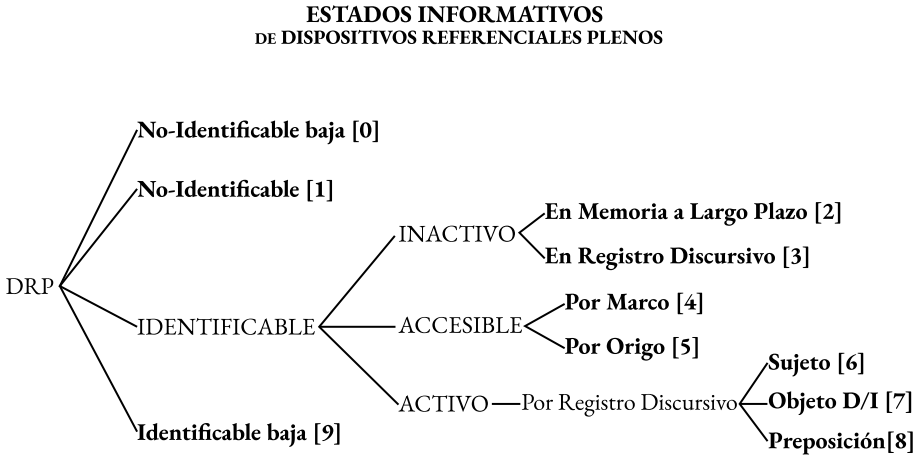


Figura 19. Etiquetas ESIN con su representación numérica

Además de estas etiquetas, confeccioné un conjunto mínimo que llamo ESIN_R1 (reducción 1) en el cual reúno en una sola etiqueta todo aquello etiquetado como Activo (las diferencias sobre el estatus gramatical de la frase nominal que expresa el antecedente de la frase analizada) así como la etiqueta [9] Identificable Baja; también en una sola etiqueta reúno los Inactivos (Inactivo por Memoria a Largo Plazo (MLP) y por Registro Discursivo (RD)); en el caso de las etiquetas Accesibles por Origo y por Marco, las reúno en una sola. Finalmente, el No Identificable y No Identificable Baja se encuentran unidas en esta reducción. Lo anterior luce de la siguiente manera:

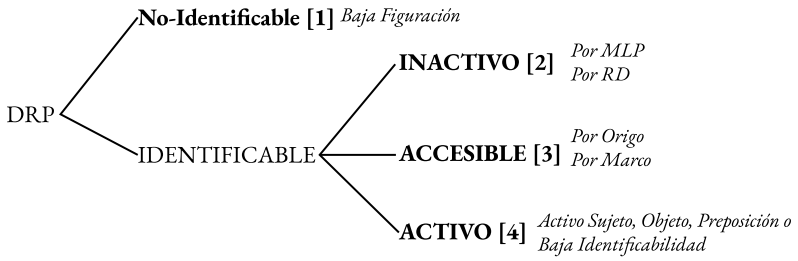


Figura 20. Etiquetas reducidas (ESIN_R1)

En un segundo filtro ESIN_R2, y sólo con la intención de comparar los resultados de los estudios anteriores (McCarthy et al. 2012; Hempelmann et al. 2005), sólo tomo en cuenta dos extremos a partir de los estudios de Prince (1981): nuevo y dado. Por esta razón, creo una segunda reducción en donde aquellas frases etiquetadas como No Identificable y No Identificable Baja las etiqueto como **Nuevo [1]**, y todo lo demás, lo etiqueto como **Dado [0]**.

Finalmente, realicé una última reducción. La ESIN_R3 está inspirada en la anterior con algunas diferencias a partir de lo revisado en el capítulo teórico. Si nos restringimos a la capacidad del algoritmo de sólo tener acceso a las experiencias de los textos, entonces todo aquello que tenga relación con el texto podría ser clasificado. Nos podríamos adelantar a

suponer que tendría problemas para aquellas etiquetas que apelan a conocimiento fuera del texto. Por esta razón, construyo este conjunto en donde las etiquetas Inactivo por Registro Discursivo, Identificable Baja y todas las etiquetas Activas las subsumo en un solo grupo que llamo sencillamente **Activo en texto [1]**; todas las demás las reúno en otro grupo que llamo **Fuera de texto [0]**. Serían aquellas situaciones en donde el hablante supone que el oyente necesita más información que sólo la nota para establecer al referente o el que no exista tal referente y se tenga que figurar en ese momento. Es el caso del Inactivo por Memoria a Largo Plazo, No identificable, No-Identificable Baja, Accesible por Origo y Marco.

En la Tabla 24 (siguiente página) se encuentra un resumen de lo anterior. En la primera columna de esta tabla coloco las 10 etiquetas, y en las subsecuentes coloco las agrupaciones con una nueva etiqueta numérica. Estas reagrupaciones tienen el objetivo de servir sólo para experimentación y contraste. Se siguen basando en las etiquetas originales y siguen funcionando bajo el análisis que propuse en el Capítulo 1. Las dos secciones anteriores exponen las características principales de la tabla de salida. Las reducciones al primer conjunto ESIN permiten asegurar que la cantidad de observaciones de cada categoría es mayor a 30 con lo que se asume que el teorema del límite central aplica (Montgomery y Runger 2014, 245). Esto posibilita la aplicación de ciertas pruebas estadísticas. En la siguiente y última sección del capítulo, expongo a qué pruebas fueron sometidas las variables independientes y estas agrupaciones de ESIN.

Tabla 24. Reducciones de etiquetas ESIN para la experimentación[86]

ESIN *(Originales)*	**ESIN_R1** *(Reducir detalle)*	**ESIN_R2** *(Prince)*	**ESIN_R3** *(Texto)*
No-identificable Baja [0]	No-Identificable [1]	Nuevo [0]	Fuera de texto [0]
No-Identificable [1]			
Inactivo por MLP [2]	Inactivo [2]	Dado [1]	
Inactivo por RD [3]			Activo por texto [1]
Accesible por Marco [4]	Accesible [3]		Fuera de texto [0]
Accesible por Origo [5]			
Activo S [6]	Activo [4]		Activo por texto [1]
Activo O [7]			
Activo P [8]			
Identificable Baja [9]			

2.7.3 Pruebas estadísticas

Para evaluar la capacidad de las reagrupaciones de ESIN de relacionarse con las medidas y los factores, realicé varias pruebas estadísticas, las cuales dividí en tres etapas. En la primera etapa apliqué pruebas con el objetivo de mostrar datos para la comparación con los trabajos de Hempelmann et al. (2005) y McCarthy et al. (2012). En la segunda etapa busqué determinar si existía normalidad en las medidas dadas las cuatro agrupaciones de ESIN. En

[86] En lo que sigue de esta sección, así como de los resultados, haré referencia a estas agrupaciones por su nombre clave: ESIN para las diez etiquetas de Estados Informativos originales; ESIN_R1 para las cuatro etiquetas de la primera reducción propuesta; ESIN_R2 para la segunda reducción a dos etiquetas, inspirada por los antecedentes; y ESIN_R3 para la tercera reducción a dos etiquetas, dada la limitación tecnológica que he explicado. Por lo que, cuando mencione el nombre clave debe leerse como "agrupaciones de etiquetas de Estados Informativos dada (o no) una reducción".

la tercera etapa realicé pruebas no paramétricas, junto con el clasificador de bosques aleatorios. Las pruebas utilizadas en el primer grupo las describo a continuación. En todas se utilizaron paqueterías en Python para su realización. Como guía para la implementación de las pruebas y los algoritmos he utilizado los manuales orientados a lingüística de Hernández Campoy y Almedia (2005) y Levshina (2015) así como apreciaciones generales estadísticas en Cohen (1977) y Mongomery y Runger (2014). En los casos del uso de las pruebas en Python, existen bibliotecas digitales en línea que permiten explorar los alcances de cada función. Hago referencia a su documentación en los casos pertinentes.

Matriz de correlación simple entre las medidas y las agrupaciones ESIN. Para reproducir el método de Hempelmann et al. (2005), se utiliza el coeficiente de correlación Pearson (Urdan 2005, 77). No obstante, cuidé la pertinencia del coeficiente de acuerdo con la escala de las variables analizadas lo cual especifico en los resultados. Este coeficiente mide la correlación lineal entre dos grupos de datos. Lo que nos dice es la fuerza con la cual dos variables están relacionadas. Puede ser -1 para una perfecta correlación negativa, o 1 para una perfecta correlación positiva. El 0 significa nula correlación. Después se realizó un segundo análisis de correlación entre los 13 predictores para destacar si existía alguna otra correlación, mayor a la de las medidas, y de esta manera incluir el predictor en el modelo de clasificación. Esto también me permite observar posibles casos de colinealidad. Para las matrices utilicé la función *corr* de los DataFrame_PY en Pandas.

Análisis de varianza de una vía (ANOVA) (Urdan 2005, 118; Hernández Campoy y Almeida 2005, 221)**.** Esta prueba la complemento con una representación visual en gráficas de cajas y bigote (*Box and Whiskers*). La hipótesis nula de esta prueba sostiene que no existe diferencia entre las medias de los grupos, lo cual también indica que las agrupaciones

propuestas son tan buenas como una segmentación aleatoria de los datos; la hipótesis alternativa es que al menos una de las medias de los grupos analizados es distinta. Buscamos que el valor *p* sea menor a 0.05. El objetivo de esta prueba es demostrar que los grupos (cualquiera de los planteados en las cuatro agrupaciones de ESIN) son significativos tomando como parámetro alguna de las medidas. En Python se utiliza de la paquetería *scipy.stats* la función *anova_lm*.

Regresión múltiple (Levshina 2015, 253). Esta clase de modelos son el preámbulo para la clasificación. Dependiendo del tipo de escala utilizada para la variable dependiente se utiliza un modelo de regresión. Para este estudio, utilicé dos tipos de regresión: Multinomial para ESIN y ESIN_R1; Logit para ESIN_R2 y ESIN_R3. En el primer caso, se toma una variable categórica y se utilizan los factores y medidas como predictores. En una primera vuelta, se utilizan las 13 variables, y luego se reajusta el modelo con aquellas que hayan tenido un mejor desempeño, además de una correlación destacable. En Python se utilizó la función *LogisticRegression* de la paquetería *sklearn*.

El análisis de varianza de una sola vía supone normalidad en los datos, aunque es robusta con datos que tengan una distribución asimétrica (Blanca et al. 2017). Si bien, nos muestran aspectos interesantes, es preferible antes evaluar la normalidad de los conjuntos. Por lo que, en la segunda etapa del estudio realizo la prueba **K^2 de D'Agostino** (D'Agostino 1971) para evaluar la normalidad de las agrupaciones ESIN dadas las medidas, además de agregar las gráficas de su distribución. Esta prueba se encuentra en la paquetería de Python *scipy.stats*.

Posterior a estas pruebas, en la tercer y última etapa, se utilizaron dos pruebas para datos no paramétricos: **Kruskal-Wallis** y un análisis de pares con **Conover-Iman** (Conover e Iman 1979). En el primer caso, la prueba nos indica si al menos una de las categorías en cada

agrupación ESIN es distinta ($p < 0.05$), en el segundo, se prueba la dominancia estocástica entre pares de categorías dada alguna de las agrupaciones de ESIN. En el fondo, lo que nos indica es que ninguna de las dos categorías analizadas en cada par domina a otra, lo que se interpreta como que pertenecen a grupos distintos. Para estas pruebas paramétricas se utilizó *posthoc_conover* de *scikit_posthocs* y *kruskal* de *scipy.stats*. Para aquellos factores que hayan resultado tener una relación más fuerte con alguna categoría ESIN, al igual que con las medidas, planteo el utilizar un modelo a partir de bosques aleatorios de decisiones para el posible clasificador. Esto con la intención de dar un primer paso hacia este objetivo y explorar clasificadores más potentes en otros estudios.

Los **bosques aleatorios** (*random forests*) se basan en árboles de clasificación (Breiman 2001; Levshina 2015, 291). Para cada nodo en el árbol se determina su capacidad de clasificar los datos a partir del parámetro que seleccionemos. Existen distintas métricas para evaluar la pureza de la clasificación. El paso crucial para pasar de un árbol a un bosque es la realización de distintos ensayos de árboles de clasificación. Para la creación de cada árbol, se pone a competir dos variables escogidas de manera aleatoria y se escoge la que en ese par tiene la menor impureza (es decir, no colocamos desde el inicio, en el nodo raíz, a la variable predictora con la mayor pureza). Estar asistidos por una computadora para intentar varios cientos de árboles construidos de manera aleatoria es crucial. Al final de este procedimiento se evalúa el modelo resultante. Al disponer de *Out-of-Bag error* (*OOB SCORE*, que corresponde al cálculo del error del tercio de los datos después de haber creado un nuevo conjunto a partir de *bootstraping*), se puede recurrir a una medida que es nombrada *accuracy* para evaluar la

precisión del modelo y omitir la validación cruzada, la cual es computacionalmente costosa[87]. El proceso para crear, depurar y evaluar el clasificador lo seguí de Amat Rodrigo (2020) y utilicé de manera fundamental la paquetería de *sklearn* de Python, en donde una de las funciones clave fue *RandomForestClassifier*. Los bosques nos permiten mostrar la preponderancia de las variables independientes del modelo a partir de dos medidas: por permutación y por la impureza. Se muestran en los resultados ambos casos del mejor modelo encontrado. Se realizaron experimentos con dos conjuntos distintos de variables independientes: uno con los factores y las medidas y otro con solo las medidas para cada agrupación ESIN.

Dado lo anterior, en el siguiente capítulo muestro los resultados de esta investigación. Sólo con ánimos de recobrar el punto de interés: busco demostrar que las medidas guardan alguna relación con los Estados Informativos y, por extensión teórica, con las nociones de los nuevo y dado. Lo expuesto en este capítulo ha buscado ser una revisión detallada de la metodología que he seguido, con el objetivo de permitir la reproducibilidad de los experimentos. La cientificidad de una disciplina no depende únicamente de realizar predicciones, sino también, de permitir comunicar, comparar y discutir las propuestas. De esta manera, aunque los resultados presentados buscan completar el objetivo propuesto en el Planteamiento, también buscan servir como un nuevo antecedente en el tema para ser discutidos, rebatidos y superados.

[87] Se debe tener en cuenta que tanto las regresiones como los bosques aleatorios en Python piden un conjunto de entrenamiento y otro conjunto de prueba. En los resultados se señalan estos grupos y los resultados de las predicciones.

Capítulo 3 Resultados de LSA y SPAN como predictores

En el siguiente capítulo doy espacio a una examinación detallada de los resultados de cada una de las pruebas, dadas las reducciones de etiquetas y la posibilidad de crear modelos para la clasificación múltiple a partir de los 13 predictores propuestos. No obstante, debido a que el principal interés son las medidas, le dedicaré mayor atención a estos predictores, dejando en un segundo plano la evaluación individual de los otros siete factores.

Antes de continuar me es importante señalar que, aunque esta sección corresponde a los resultados principales, hay resultados intermedios que ya he mostrado en otras secciones. Tres de estos me parecen centrales para poder llegar a este punto: primero, la propuesta de síntesis de las etiquetas de Estados Informativos y la manera de analizarlas (§1.10 y §1.11); segundo, el Corpus Periodístico del Noroeste de México (380 notas, 139 735 palabras, 10 038 lemas) y los lineamientos de etiquetado (§2.6, Anexo A y P); y, tercero, la programación en Python del proceso automático para la obtención de medidas dado COPENOR_cero (subconjunto de COPENOR conformado por 38 notas, 2 388 frases nominales etiquetadas, 20 914 palabras, 2 578 lemas) (§2.3 y §2.4 y Anexo P).

De esta manera, este capítulo inicia con una descripción estadística de la Tabla de Salida (descrita en la sección anterior). Inmediatamente después inicio la exposición de la correlación entre las medidas (LSA y SPAN, con las tres bolsas de palabras) y las distintas etiquetas planteadas en las agrupaciones ESIN. En una subsección distinta atiendo el tema de incluir la posición relativa de la frase nominal como un predictor y muestro que no existe una correlación mayor a las medidas con las distintas etiquetas ESIN. Le sigue la exposición de los resultados de ANOVA y luego los de la regresión logística. En esa sección, abordo

dos tipos de modelos: uno que utiliza todos los predictores, es decir, las medidas y los factores extras; otro modelo sólo utiliza aquellos predictores que tuvieron alguna correlación destacable. Los resultados hasta este punto muestran que las medidas pueden funcionar como buenos predictores de algunos Estados Informativos, superando en cierto caso a los antecedentes por un punto porcentual. No obstante, me es importante corroborar la distribución de los datos. Recordemos que la decisión de mostrar los resultados de estas pruebas y los clasificadores se debía a poder comparar los resultados de esta investigación con los de los antecedentes. De esta manera, realizo la prueba K^2 de D'Agostino para determinar la normalidad, así como los histogramas para una examinación visual. Estos resultados arrojan que la gran mayoría de las etiquetas (los conjuntos de medidas dadas las etiquetas), no tienen una distribución normal, por lo que es necesario realizar pruebas no paramétricas. De esta manera, continuo con la exposición de los resultados de las pruebas Kruskal-Wallis. Una vez confirmado que se rechaza la hipótesis nula en las seis medidas, exploro por medio del método Conover-Iman cada par de etiquetas dentro de las agrupaciones para determinar cuales etiquetas se diferencian del resto. Después, realizo una prueba de clasificación con bosques aleatorios. Propongo la construcción de los modelos a partir de tres distintos conjuntos de predictores: todos, sólo las medidas y sólo aquellos que resultaron tener una correlación relevante. Este ejercicio lo hago para cada una de las agrupaciones ESIN. Este capítulo lo cierro con una observación general del desempeño de los clasificadores (de regresión y el de bosques aleatorios) y una síntesis de los hallazgos.

3.1 Estadística descriptiva de las variables

De las 38 notas, el 50% tiene más de 57 frases nominales y entre todas suman 2 388 frases nominales etiquetadas. La nota más pequeña tiene 14 frases nominales (COPENOR-110BC) y la más extensa 183 (COPENOR-369SN). La cantidad de Estados Informativos analizados en COPENOR_CERO tiene la distribución que se observa en la Figura 21.

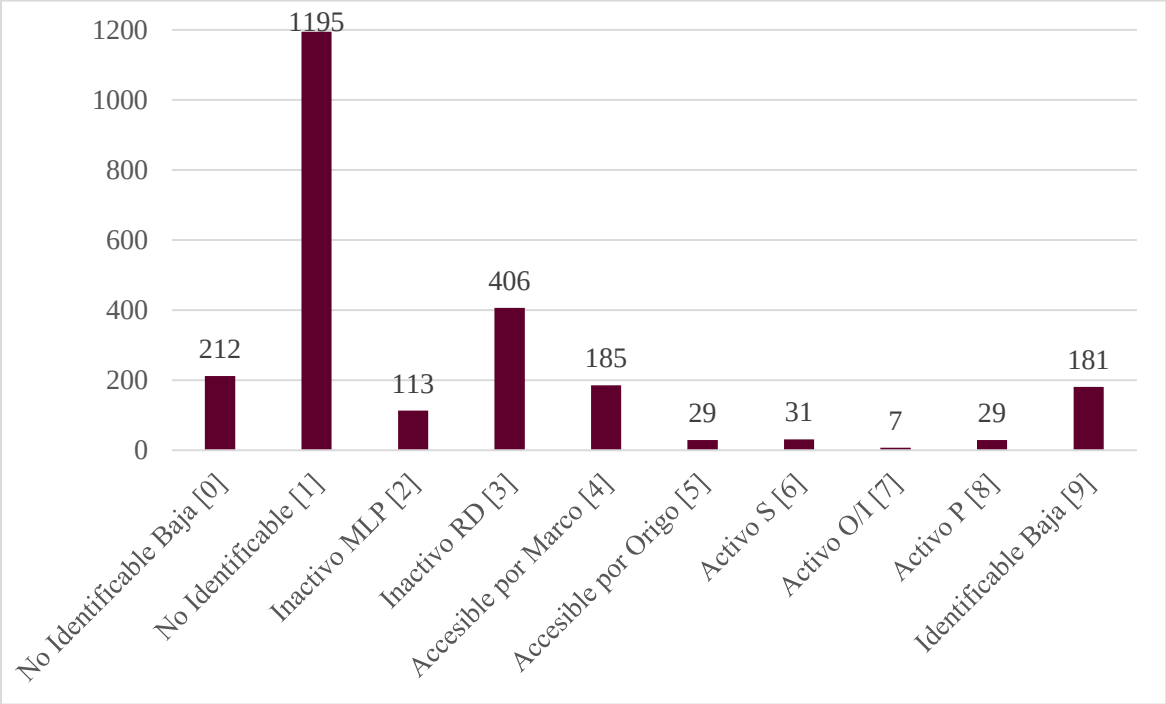


Figura 21. Ocurrencias de Estados Informativos, sin reducción (ESIN)

El Estado Informativo con la mayor cantidad de ocurrencias fue el No Identificable [1] con 1 195, mientras que el que mostró la menor cantidad de ocurrencias fue Activo de Objeto Directo/Indirecto [7] con 7. Además de esta etiqueta, otras tres etiquetas tuvieron muy bajas ocurrencias: Accesible por Origo [5] y Activo de Preposición [8] tienen 29 cada una y Activo de Sujeto [6] presentó 31. De acuerdo con lo visto en el Capítulo 1, esto es esperable: las frases nominales tienen baja probabilidad de recuperar a un referente activo, mencionado en la oración inmediata anterior; y al parecer, si sucede, estos datos muestran una preferencia

por recuperar entidades introducidas por preposiciones o en menciones que expresan el sujeto de la oración inmediata anterior. Es importante recordar que en la etiqueta Activo P [8] no distinguí el rol sintáctico, por lo que podríamos estar frente a frases nominales que funcionan como complementos circunstanciales o como argumentos en construcciones con verbos de régimen preposicional. En todo caso, como señalé en el esquema de etiquetado, el único interés de plantear estas diferencias entre los Activos era notar si existía una tendencia, principalmente en la situación en la que el referente de la frase nominal analizada haya sido mencionado en la oración inmediata anterior, en donde la frase nominal expresa además el sujeto gramatical, lo que supone un referente activo en la Memoria de Trabajo (MT). Estos datos muestran que, incluso sumando todas las etiquetas de la rama Activo, apenas se llega a 67 ocurrencias.

Para la primera reducción, ESIN_R1, la cual simplifica la presentada en ESIN, la distribución luce de la siguiente manera (Figura 22). La mayor ocurrencia la tuvo la categoría No Identificable [1] con 1 407 observaciones, le sigue la etiqueta Inactivo [2] con 519 ocurrencias.

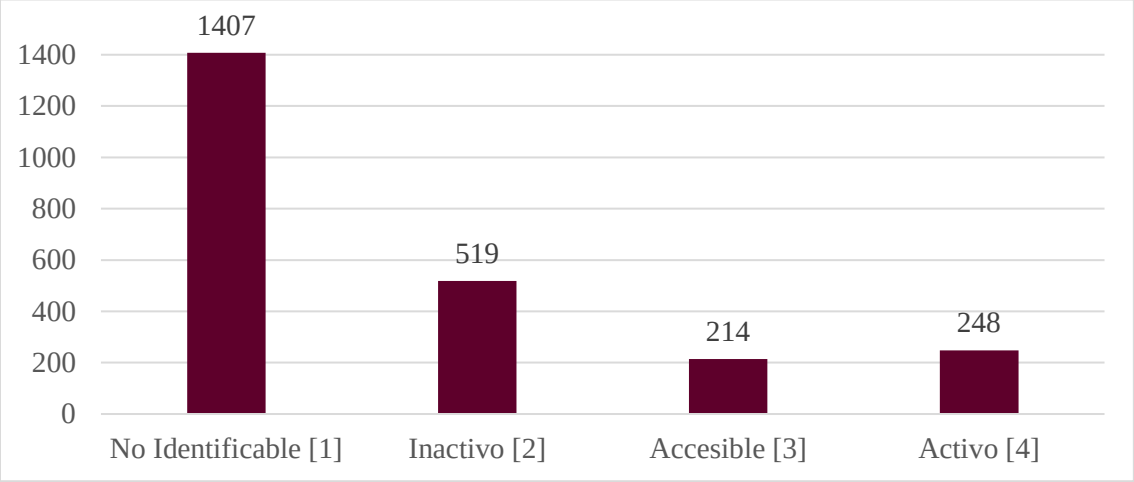


Figura 22. Ocurrencias de las etiquetas reducidas en ESIN_R1

La siguiente es Activo [4], en la cual se suman las 67 ocurrencias que mencioné en la descripción anterior más 181 que pertenecen a las Identificables Bajas; recordemos que estas suponen activación, es decir, una frase nominal que se refiere a una entidad en la MT, pero presenta baja capacidad de figurar por sí sola (es el caso de las frases nominales en oraciones copulativas o ecuativas y frases nominales aposicionales). Este grupo tuvo 248 ocurrencias. Finalmente, las Accesibles [3] tuvieron 214 ocurrencias.

Para los casos de ESIN_R2, reducción que tenía como objetivo imitar la clasificación usada en los trabajos antecedentes, y ESIN_R3, la cual es mi propuesta de reducción binaria, muestro los datos en una misma gráfica (Figura 23). La etiqueta Nuevo [0] presentó 1 407 ocurrencias, mientras que Dado [1] fue de 981. Por otro lado, para ESIN_R3, Fuera de Texto [0] presentó 1 734 ocurrencias, y Activo por texto [1] presentó 654.

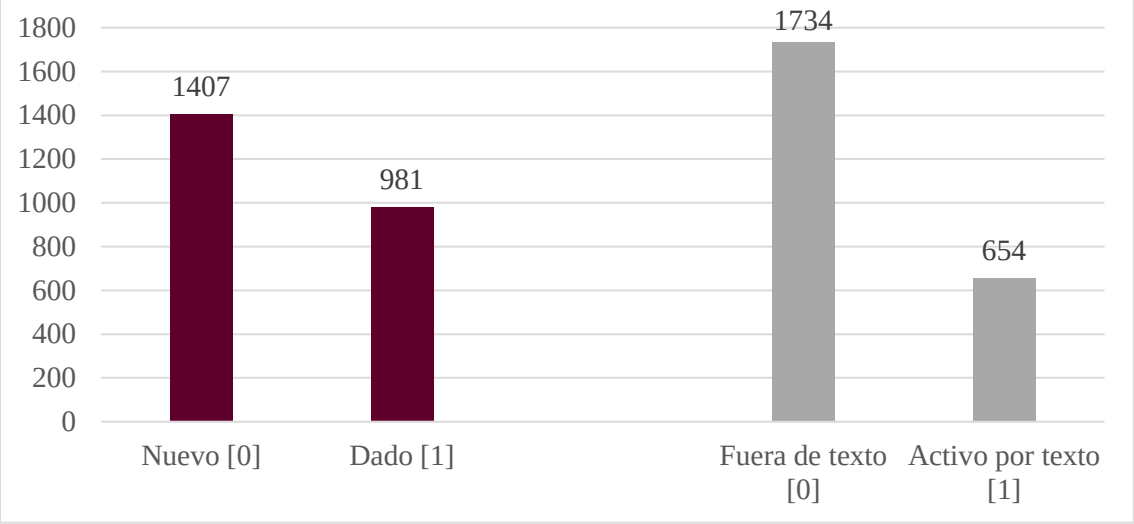


Figura 23. Ocurrencias de ESIN_R2 y ESIN_R3

En cuanto a las seis medidas, coloco su descripción estadística en la siguiente tabla, las cuales recordemos corresponden a tres ventanas distintas para LSA y SPAN.

Tabla 25. Descripción estadística de LSA_{MAX} y SPAN

Medida	Bolsas de palabras	Media	d.e. (σ)
LSA_{MAX}	Ventana-*n* (20)	.80	.21
	Interior-*w*	.37	.28
	Interior-*wd* (+ DEM)	.33	.31
SPAN	Ventana-*n* (20)	.50	.11
	Interior-*w*	.58	.18
	Interior-*wd* (+ DEM)	.63	.19

Notamos que SPAN parece tener una distribución más cercana al punto central de la medida (que va de 0 a 1 en todos los casos) con una menor desviación estándar. Por otro lado, tenemos que la medida LSA_{MAX} que utiliza la bolsa de palabras Interior-*w*, así como la que incluye el diccionario, Interior-*wd*, se comporta casi de manera opuesta a la media de la Ventana-*n*.

Sobre los factores adicionales que se incluyeron, tenemos los dos casos dicotómicos de definitud e indefinitud, los cuales tienen las siguientes ocurrencias (Tabla 26). Nótese que la suma de las presencias no llega al total de las frases nominales (*n* = 2 388). Esto es porque no todas las frases nominales tienen como primer elemento un determinante con el rasgo definido o indefinido:

Tabla 26. Descripción cuantitativa de los rasgos de Definitud e Indefinitud

	Presencia	Ausencia
Definitud	1 145	1 243
Indefinitud	138	2 250
Total	1 283	

Se observa que en las notas tenemos una gran cantidad de frases que inician con un artículo definido, las cuales representan el 47.94% de toda la muestra. Por otro lado, las frases nominales encabezadas con indefinidos representan un 5.7%. Esto nos indica que la capacidad de introducir referentes nuevos no le es exclusiva ni representativa a la indefinitud en este corpus; contrástese 138 ocurrencias de frases nominales encabezadas con indefinido y 1407 frases nominales (ESIN_R1 y ESIN_R2) cuyas estructuras expresan la suposición del hablante sobre el desconocimiento del referente o la necesidad de construir un referente en ese momento del discurso[88]. Para las categorías sintácticas que se analizaron, así como otras caracterizaciones sintácticas, se encontró la siguiente distribución:

Tabla 27. Frecuencias de categorías relacionadas con el rol sintáctico

	Frecuencias
Sujeto	409
Objeto Directo	313
Objeto Indirecto	45
Atributo	36
Preposición	568
No aplica	1 017
Total	2 388

Se nota en la Tabla 27 que gran parte de las frases nominales pertenecen a la categoría No Aplica (1 017), es decir, no son dependientes de alguna estructura verbal. Son los casos de modificadores de otras frases nominales. Recordemos que 181 frases nominales tienen la etiqueta Identificable Baja [0] (ESIN, Figura 21), en donde se incluyen frases nominales que son el complemento de oraciones copulativas y aquellas que se encuentran en aposición a

[88] Recordemos que la propiedad de indefinitud se incorporó por medio del análisis automatizado de Stanza.

otra frase nominal. De tal manera, podemos observar que, de estas 181 frases, 36 son *atributos* de esta clase de construcciones. Las 145 frases restantes representan a aquellas que funcionan como aposición, por lo que, de las 1 017 frases nominales etiquetadas como *no aplica*, 872 se encuentran en otra situación sintáctica no descrita en este trabajo, probablemente como nombres propios o modificando un núcleo nominal y encabezada por alguna preposición. Realizar esta distinción no fue parte del análisis, pero se podría investigar en un futuro.

En segundo lugar, aparecen las frases nominales que son introducidas por una preposición y dependen de una estructura verbal, ya sea como complementos circunstanciales o expresiones de alguno de los argumentos de verbos de régimen preposicional. Para estos casos se encontraron 568 ocurrencias. En tercer lugar, se encuentran las frases nominales que expresan el sujeto sintáctico de la oración en la que aparecen, con 409 ocurrencias. Enseguida, con 313 ocurrencias, se encuentran las frases nominales que expresan el Objeto Directo. Finalmente, los últimos dos lugares son para las frases nominales que expresan el Objeto Indirecto, con 45 ocurrencias, y las que expresan el atributo en oraciones ecuativas o copulativas, que como ya mencioné son 36. Aunque muestro los datos en conjunto, estas categorías relacionadas con características sintácticas las trataré en los clasificadores como variables dicotómicas.

Para el caso de las variables continuas, a excepción de las medidas que ya describí, tenemos el tamaño de la frase, su cantidad de verbos y de nominales. Esta distribución la muestro a continuación (Tabla 28).

Tabla 28. Descripción estadística del tamaño de las frases, cantidad de verbos y nominales

Variable	min	max	Media	d.s. (σ)
Tamaño de frase	1	84	5.06	5.53
Cantidad de nominales	0	17	1.34	1.36
Cantidad de verbos	0	5	0.12	0.43

Alrededor del 75% de las frases nominales tienen un tamaño de seis palabras o menos. Existe una frase nominal que se sale del patrón, la cual contiene 84 palabras. Esta frase nominal está construida como "los profesores NOMBRE + APOSICIÓN, NOMBRE + APOSICIÓN, etc." (IDFN 111, COPENOR-170BS), en la cual, además, las aposiciones tienen modificadores internos. Fuera de este caso, la tendencia en las notas se inclina a frases nominales pequeñas. Nótese que existen frases nominales que de manera paradójica tienen cero "nominales". Esto no se debe a un problema de análisis o caracterización, sino a un método de clasificación, debido a la manera en que se procesan las frases nominales. Los nombres propios no pueden buscarse en el DEM, por lo que la información que aportan a la frase es distinta en términos informáticos. Sin embargo, existen maneras computacionales para la búsqueda de entidades, lo cual requiere un inventario de sentidos distinto al trabajado y proyectado en esta investigación.

Es interesante notar que, aunque la máxima cantidad de nominales encontrados en una frase puede llegar a 17, el promedio gira alrededor de 1.34 con una desviación estándar muy pequeña de apenas 1.36. Es decir, aunque es posible tener frases nominales con cualquier cantidad de nominales, vemos que la tendencia es a uno o dos. Finalmente, la cantidad de verbos también nos dice sobre la complejidad de las frases, usualmente con oraciones subordinadas relativas. La presencia de verbos en las frases es muy baja, no llega en promedio

a 1, y una desviación estándar también muy baja de 0.43. Aunque es posible utilizar las relativas para aportar información sobre el referente descrito, este recurso parece no tener una tendencia fuerte en las notas periodísticas.

3.2 Correlación entre las medidas y las ESIN

Para la correlación, tanto Pearson como Point Biserial mostraron exactamente los mismos resultados, por lo que se muestra sólo el coeficiente de Pearson en las siguientes gráficas (Figura 24[89]). El color más claro (en amarillo) representa correlación positiva perfecta, y el color oscuro (azul marino) representa correlación negativa perfecta.

Las medidas LSA y SPAN parecen tener una correlación con algunas de las etiquetas ESIN. En particular, llama la atención la correlación entre LSA Interior-*w* e Inactivo RD [3] (*r* = 0.5**)[90]. Esto mismo se observa con la inclusión del DEM, para la medida LSA Interior-*wd* y su relación con la etiqueta Inactivo RD [3] (*r* = 0.49**). Esto empieza a delinear que el DEM no parece establecer una correlación más fuerte, aunque tampoco la debilita.

La etiqueta Inactivo por MLP [2] mostró una correlación leve con SPAN Interior-*w* (*r* = 0.24**) y No identificable [1] mostró una correlación negativa leve tanto con LSA Interior-*w* como con LSA Interior-*wd* (*r* = -0.27**).

---

[89] Cada matriz de correlación utiliza el código numérico propuesto en la Tabla 24. Para las medidas se utiliza el código siguiente: MSIN = **M**edida **S**pan **In**terior-*w*; MLIN = Medida LSA Interior-*w*; MSVN = Medida LSA Ventana-*n*; MLVN = Medida LSA Ventana-*n*; MSWD = Medida Span Interior-*wd*; MLWD = Medida LSA Interior-*wd*.

[90] ** = $p < .001$; * = $p < .05$

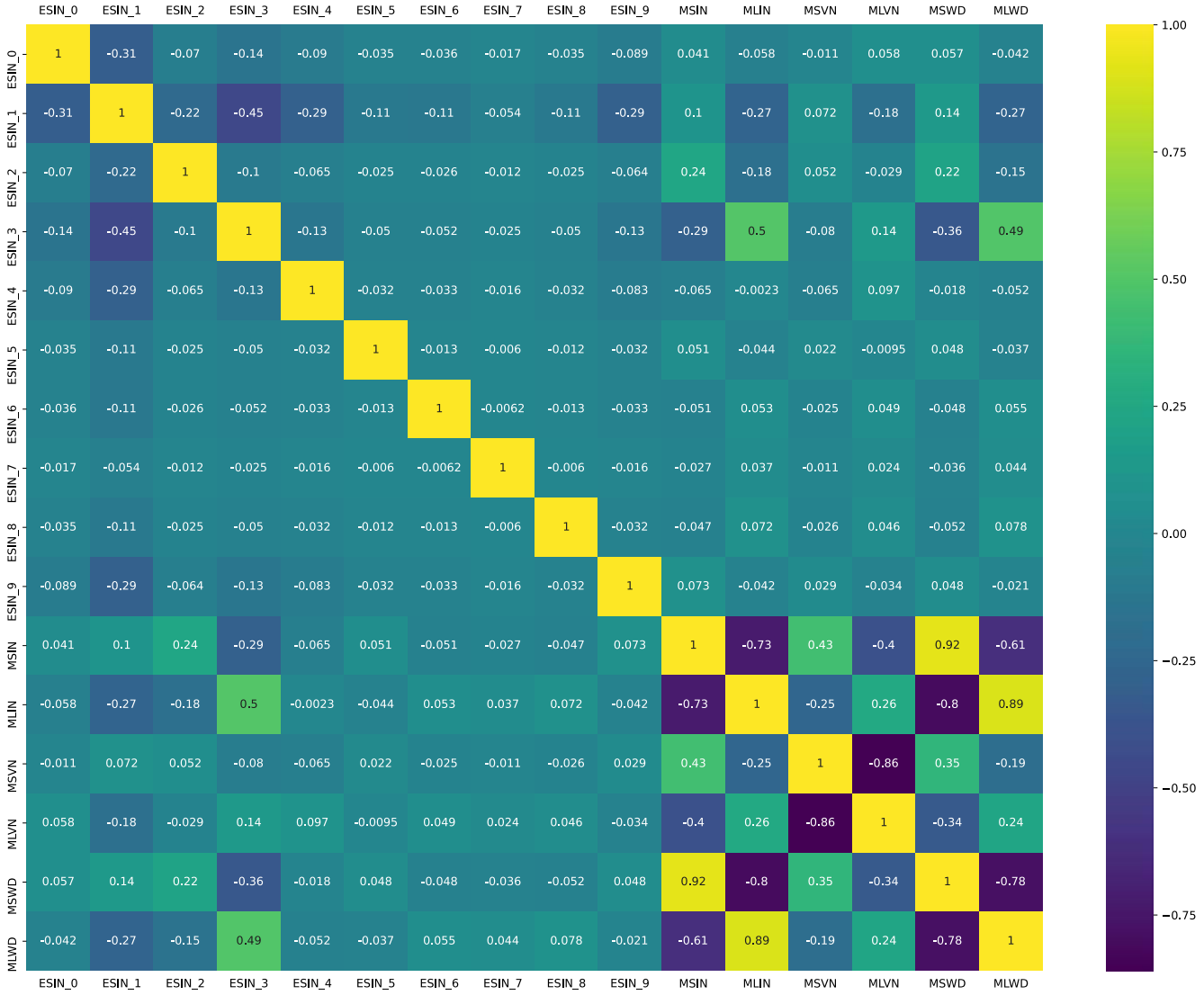


Figura 24. Matriz de coeficiente de correlación de Pearson entre ESIN y las medidas

Recordemos que para LSA, el incremento de la medida se entiende como que hay más información relacionable con la frase analizada. Por lo que, si decrece LSA esperaríamos mayor presencia de etiquetas del tipo No Identificable [1]. Este análisis se corrobora con esta prueba estadística, y aplica también con el caso del Inactivo RD [3]: a números más altos de LSA esperamos mayor presencia de alguna etiqueta que se refiera a menciones anteriores, en este caso, del Registro Discursivo (pero que no se encuentren en la Memoria de Trabajo).

Además de estas observaciones, tenemos que las medidas que usan las bolsas interiores de la frase tienen fuerte correlación entre ellas, lo que es signo de colinealidad. Esto ya nos da indicios de que, en un modelo de clasificación, sería mejor escoger dos medidas: una construida con la ventana exterior (Ventana-*n*) y otra construida con la ventana interior, ya sea la directa (Interior-*w*) o la que incluye las definiciones del diccionario (Interior-*wd*), pero no ambas.

Hasta este momento, LSA parece mostrar mejores resultados. Aquellos más relevantes de la Figura 24 los sintetizo en la Tabla 29.

Al medir las correlaciones entre la primera reducción de las ESIN (ESIN_R1) y las medidas (Figura 25), observamos que las bolsas interiores siguen predominando (Interior-*w*, Interior-*wd*), aunque al mismo tiempo, sigue siendo difícil observar una correlación entre las medidas y los Estados Informativos Activo [4] y Accesible [3]. Para este caso, llama la atención que la correlación más fuerte sucede entre Inactivo [2] y LSA Interior-*wd* ($r = 0.37^{**}$). Le sigue Inactivo [2] y LSA Interior-*w* ($r = 0.36^{**}$). Al igual que con las etiquetas sin reducir (ESIN), si se usa LSA, la diferencia entre incluir o no el diccionario es mínima.

Tabla 29. Resultados de la correlación entre SPAN, LSA y las etiquetas ESIN

Bolsa de palabras	ESIN	Pearson *r*
SPAN Interior-*w*	**Inactivo por MLP [2]**	**0.24****
	Inactivo por RD [3]	-0.29**
LSA Interior-*w*	**No Identificable [1]**	**-0.27****
	Inactivo por MLP [2]	-0.18**
	Inactivo por RD [3]	**0.5****
LSA Ventana-*n*	No identificable [1]	-0.18**
SPAN Interior-*wd* (DEM)	No Identificable [1]	0.14**
	Inactivo por MLP [2]	0.22**
	Inactivo por RD [3]	-0.36**
LSA Interior-*wd* (DEM)	**No Identificable [1]**	**-0.27****
	Inactivo por MLP [2]	-0.15**
	Inactivo por RD [3]	**0.49****

Otro detalle que surge de nuevo en este análisis es la correlación negativa de LSA con el Estado Informativo No Identificable [1] (r = -.3**), que como ya mencioné, es esperable por el tipo de medida. Este mismo patrón, pero de manera inversa, lo vemos con SPAN Interior-*w* (r = .13**) e interior-*wd* (r = .18**), lo cual también sería esperable. No obstante, tienen muy baja correlación comparadas con LSA. A continuación, muestro un resumen de las correlaciones destacables, y marco con negritas las más altas.

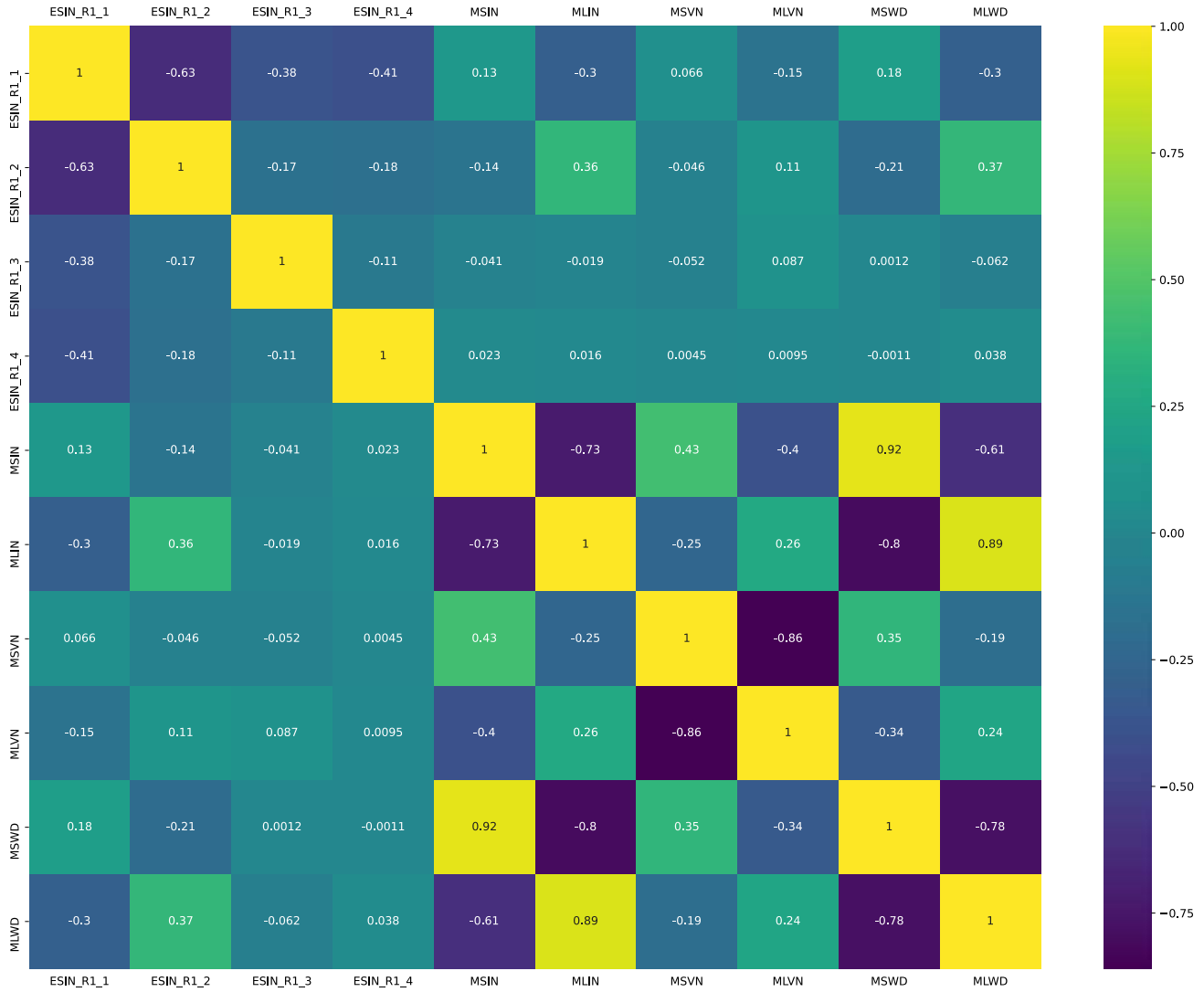


Figura 25. Matriz de coeficiente de correlación de Pearson entre ESIN_R1 y las medidas

Tabla 30. Resultados de la correlación entre SPAN, LSA y las ESIN_R1

Bolsa de palabras	**ESIN_R1**	**Pearson *r***
SPAN Interior-*w*	No Identificable [1]	0.13**
	Inactivo [2]	-0.14**
LSA Interior-*w*	**No Identificable [1]**	**-0.3****
	Inactivo [2]	**0.36****
LSA Ventana-*n*	No Identificable [1]	-0.15**
	Inactivo [2]	0.11**
SPAN Interior-*wd* (DEM)	No Identificable [1]	0.18**
	Inactivo [2]	-0.21**
LSA Interior-*wd* (DEM)	**No Identificable [1]**	**-0.3****
	Inactivo [2]	**0.37****

En el caso de la reducción de las etiquetas de Estados Informativos planteada a partir de los trabajos anteriores entre lo nuevo y lo dado (ESIN_R2), podemos observar en la matriz (Figura 26) que la correlación más fuerte es entre lo Nuevo [0] y LSA interior-*w* y LSA Interior-*wd*. En ambas bolsas se tiene que *r* = -0.3**. De manera inversa, en la etiqueta Dado [1] para LSA y en ambas bolsas interiores tenemos que *r* = 0.3**. Aunque las medidas de SPAN siguen teniendo menor fuerza de correlación, se vuelve a mostrar una mejoría a partir de incluir el DEM. Con la bolsa Interior-*w* para Nuevo [0] tenemos que *r* = 0.13** y para lo Dado [1] tenemos que *r* = -0.13**. Al incluir el diccionario, en la bolsa Interior-*wd*, la fuerza de correlación mejora: para lo Nuevo [0] tenemos que *r* = 0.18** y para lo Dado [1] tenemos que *r* = 0.18**.

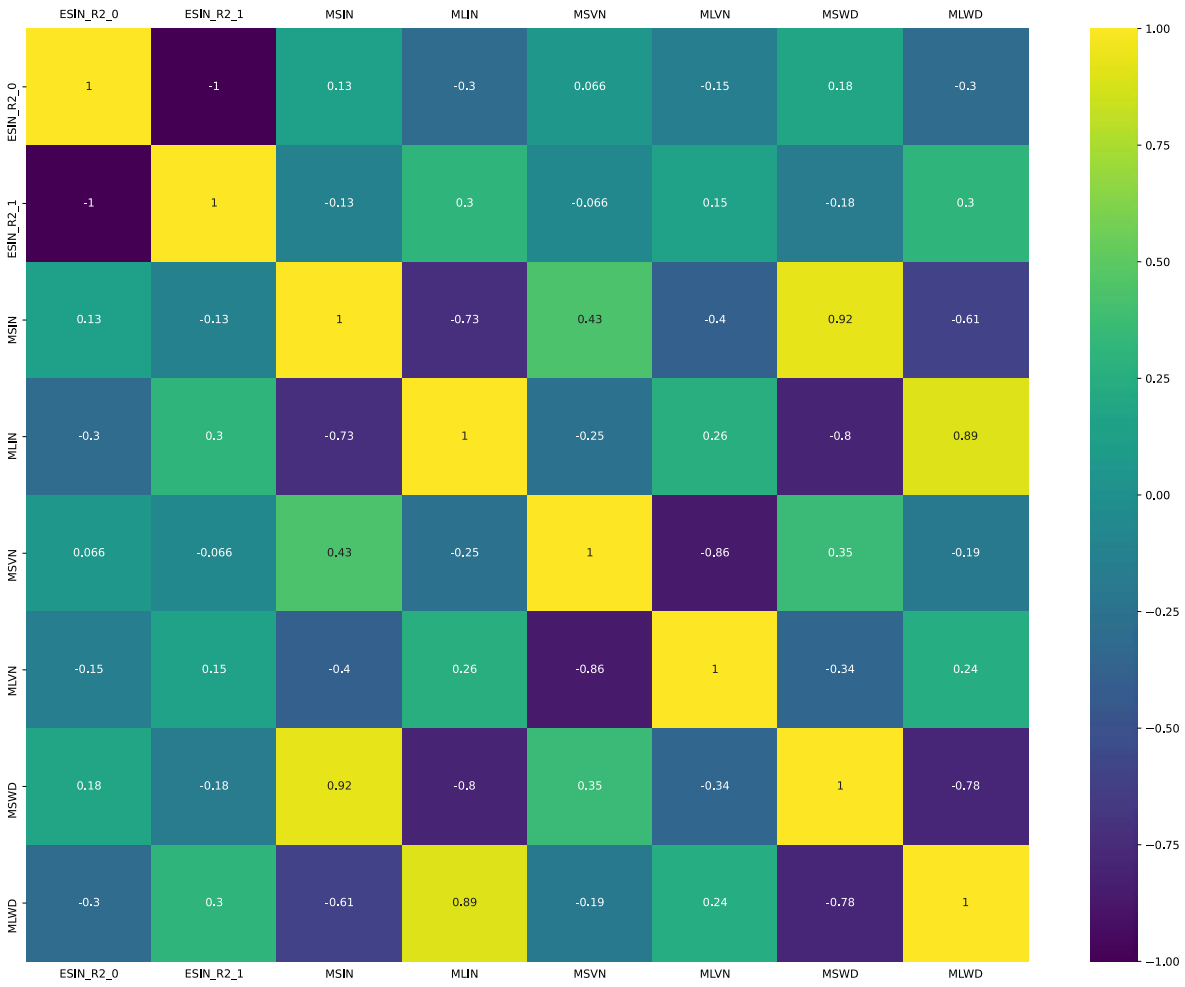


Figura 26. Matriz de coeficiente de correlación de Pearson entre ESIN_R2 y las medidas

En esta reducción cobra más sentido la correlación positiva / negativa: se espera que lo Nuevo [0] muestre una correlación negativa con LSA, y lo Dado [1], una correlación positiva, relaciones que se confirman. También se vuelve a notar que el diccionario, para LSA, no aporta mayor fuerza en la relación, pero no la disminuye.

Si comparamos con los resultados presentados en McCarthy et al. (2012), para ellos resultó que la relación entre Prince New y SPAN era de r = 0.48** y con LSA_{MAX} de r = -0.41**; para Prince Given, SPAN mostró r = 0.43** y LSA_{MAX} r = -0.31**. Esto nos permite decir que sólo uno de los resultados obtenidos en la investigación que presento es parecido a los

resultados de este antecedente, esto al usar LSA_{MAX} y relacionarlo con lo dado (LSA_{MAX} Interior-*w* e Interior-*wd*, *r* = -0.3**). Por otro lado, existe una diferencia a favor del estudio anterior para LSA_{MAX} y para SPAN, con relación a lo nuevo. Recordemos que aquel trabajo fue realizado con frases nominales en inglés y no exponen de manera clara cómo se construye cada bolsa de palabras. No obstante, los resultados tampoco parecen ser tan distintos a los obtenidos en esta ocasión. A manera de resumen de la Figura 26 y lo desarrollado hasta el momento, presento la comparación en la Tabla 31.

Tabla 31. Resultados de la correlación entre SPAN, LSA, las ESIN_R2 y los resultados de McCarthy et al. (2012)

Bolsa de palabras	ESIN_R2	Pearson R	McCarthy et al. (2012)
SPAN Interior-*w*	Nuevo [0]	0.12**	
	Dado [1]	-0.12**	
LSA Interior-*w*	**Nuevo [0]**	**-0.30****	-0.31** (LSA Prince New)
	Dado [1]	**0.30****	-0.41** (LSA Prince Given)
LSA Ventana-*n*	Nuevo [0]	-0.14**	
	Dado [1]	0.14**	
SPAN Interior-*wd* (DEM)	**Nuevo [0]**	**0.17****	0.48** (SPAN Prince New)
	Dado [1]	**-0.17****	0.43** (SPAN Prince Given)
LSA Interior-*wd* (DEM)	Nuevo [0]	-0.29**	
	Dado [1]	0.29**	

En la tabla coloco los coeficientes de McCarthy a un lado de los más altos encontrados en mi trabajo. Todos los coeficientes de SPAN en McCarthy superan a los que reporto. Sólo la correlación entre LSA Interior-*w* con lo Nuevo [0] muestra una fuerza parecida a lo reportado por McCarthy. Es importante notar que los textos de McCarthy no tenían tratamiento previo. No realizaron algún tipo de inclusión de propiedades lexicogramaticales, aspecto que sí

integro en mi investigación (§ 2.4). Quedaría pendiente para otras investigaciones observar si la inclusión de estas propiedades afecta el desempeño de las correlaciones.

Hasta este punto, la peor medida ha sido SPAN usando una bolsa de palabras construida a partir de la ventana exterior de las frases nominales (Ventana-*n*). Ni en ESIN, como en ninguna de las dos reducciones anteriores, esta medida muestra algún tipo de correlación y en algunos casos, no se alcanza a llegar siquiera a un valor *p* significativo.

Finalmente, para la correlación entre las medidas y la tercera reducción (ESIN_R3), los resultados fueron los que se muestran en la siguiente matriz (Figura 27). Esta reducción parte del supuesto de que algoritmo debería ser capaz de predecir mejor las etiquetas asociadas a suposiciones sobre recuperación de entidades nombradas en el texto (Activo por texto [1]), y fallar en aquellas exteriores al mismo (Fuera de texto [0]), en donde la suposición de desconocimiento o la necesidad de figurar una entidad como nueva se integra en esta última etiqueta. Los resultados parecen mostrar que esta clasificación, frente a la presentada en ESIN_R2, tiene una mayor fuerza de correlación con las medidas.

En este caso, SPAN adquiere mayor fuerza que en las ESIN_R2; además, de nuevo, SPAN junto con las acepciones del diccionario obtienen una correlación más fuerte. A parte de esto, también se observa que LSA sigue predominando por encima de SPAN. Un aspecto interesante es que, una vez realizada esta otra reducción, las correlaciones superan al trabajo antecedente en cuanto a LSA: de $r = -0.31^{**}$ del antecedente, pasamos a $r = -0.44^{**}$ para lo nuevo (Fuera de texto [0]) y de $r = -0.41^{**}$ pasamos a $r = 0.44^{**}$ para lo dado (Activo por texto [1]).

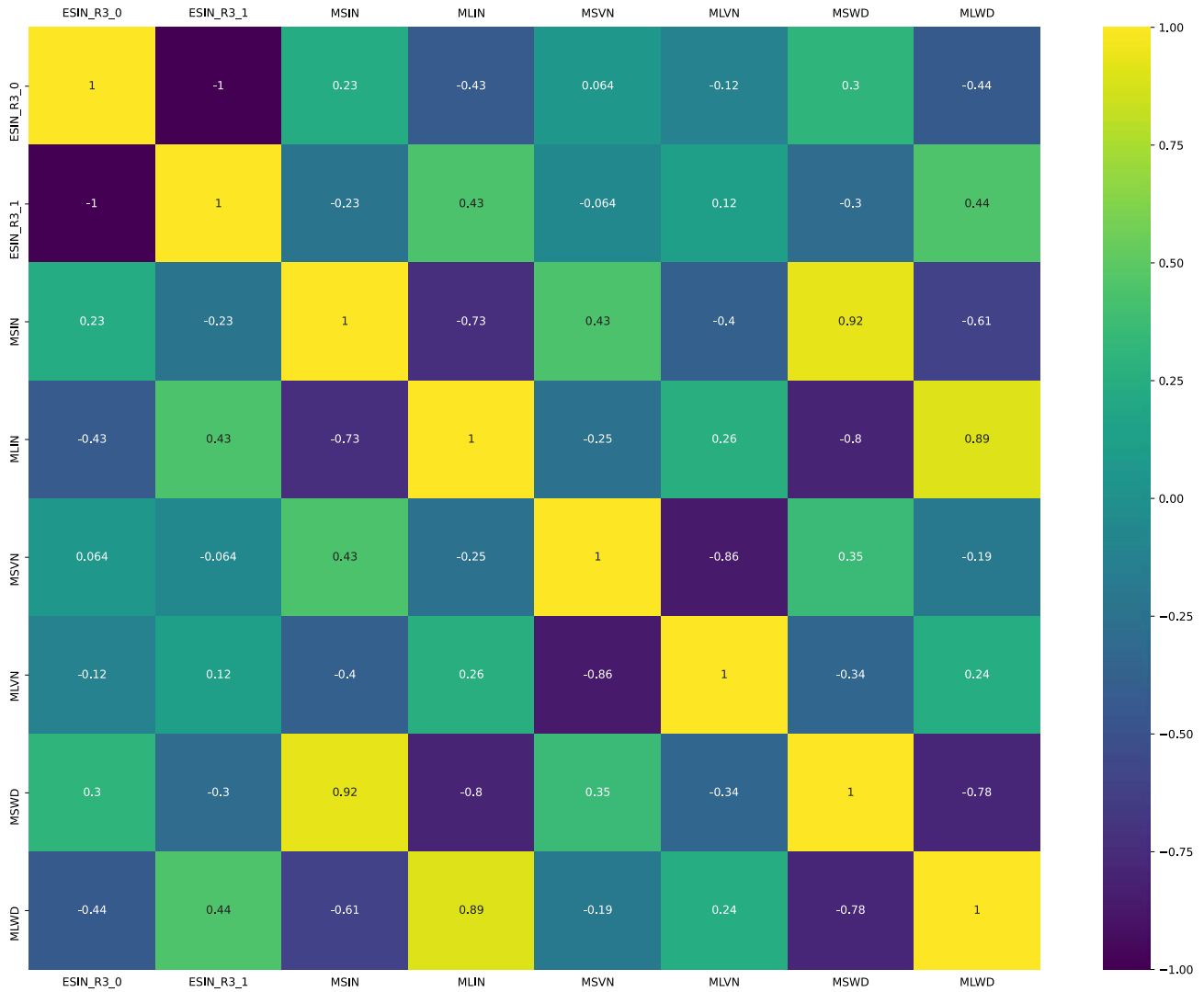


Figura 27. Matriz de coeficiente de correlación de Pearson entre ESIN_R3 y las medidas

En la siguiente tabla (Tabla 32) muestro un resumen de estos contrastes y aquellas correlaciones que superan el ±0.1. Al igual que el caso anterior, he puesto las medidas de McCarthy aun lado de las medidas más altas que reporto.

Tabla 32. Resultados de la correlación entre SPAN, LSA, las ESIN_R3 y los resultados de McCarthy et al. (2012)

Bolsa de palabras	ESIN_R3	Pearson *r*	McCarthy et al. (2012)
SPAN Interior-*w*	Fuera de texto [0]	0.23**	
	Activo por texto [1]	-0.23**	
LSA Interior-*w*	Fuera de texto [0]	-0.43**	
	Activo por texto [1]	0.43**	
LSA Ventana-*n*	Fuera de texto [0]	-0.12**	
	Activo por texto [1]	0.12**	
SPAN Interior-*wd* (DEM)	**Fuera de texto [0]**	**0.3****	0.48** (SPAN Prince New)
	Activo por texto [1]	**-0.3****	0.43** (SPANPrince Given)
LSA Interior-*wd* (DEM)	**Fuera de texto [0]**	**-0.44****	-0.31** (LSA Prince New)
	Activo por texto [1]	**0.44****	-0.41** (LSA Prince Given)

Algo que parece discordante entre los datos de McCarthy y mis hallazgos (que se contrastan también en ESIN_R2) es que, contrario a lo esperado, en este trabajo antecedente, todas las medidas obtenidas a partir de LSA tienen una correlación negativa, y todas las obtenidas a partir de SPAN tienen una correlación positiva. Es decir, en LSA esperamos una correlación positiva con lo dado (sea cual sea la etiqueta), debido a que el coseno se aproxima a 1 entre más parecido hay entre la frase analizada y las frases anteriores. Se espera que este coeficiente de similitud decrezca con frases de las cuales no se puede encontrar vínculo o algún parecido lexicogramatical (si es que esta información se le hadado para crear la bolsa de palabras). No obstante, en los resultados de McCarthy, tanto Prince New como Prince Given tienen una correlación positiva con SPAN y negativas con LSA. No he podido identificar con exactitud

a qué se deba esto, ya que incluso al homogenizar las medidas (hacer que SPAN tome como pivote el 1, y no el 0, como se explicó en §2.5) se debería seguir encontrado correlaciones "inversas" entre lo nuevo/dado, aunque, como bien mencionan los autores, no tienen que ser completamente simétricas (aspecto que, por cierto, sí sucede en los coeficientes de correlación que muestro para ESIN_R2 y ESIN_R3). Para obtener correlaciones como las que muestra McCarthy, las medidas asociadas a lo nuevo y a lo dado, crecerían a 1, lo cual, ya nos debería adelantar sobre su capacidad de diferenciar las etiquetas. Esto quedará más claro una vez realizado un análisis de varianza, lo cual nos permitirá confirmar si estos grupos están bien segmentados.

Un último aspecto que hay que señalar de los resultados mostrados es que las medidas tienen colinealidad entre ellas. Ya mencioné esto antes, pero en todos los casos, se observa que existe una correlación fuerte, ya sea positiva o negativa. En todos los casos, las únicas que parecen tener menor fuerza son las que utilizan la Ventana-*n*, especialmente SPAN, pero como se observó, esta medida tampoco muestra alguna correlación con las ESIN, ni en ninguna de sus reducciones. Esto nos permite adelantar que el clasificador sólo debería utilizar sólo una de las medidas, obtenida ya sea a partir de Interior-*w* o Interior-*wd*. También estos resultados le otorgan mayor peso a LSA que a SPAN, pero esto aún se evaluará en los clasificadores.

3.2.1 Correlación con los factores

A continuación, presentaré aquellos factores extras que se han incluido en el estudio (§ 2.7.1) y que mostraron una correlación superior al ±0.2 con alguna ESIN (ya sea alguna de las 10 etiquetas, o las reducciones) y con un valor *p* significativo ($p < .001$). El umbral de

correlación lo propongo para permitir un punto de comparación, pero como se verá, difícilmente alguno de los factores muestra una correlación más fuerte que las medidas, si es que la medida supera el ±0.1 con la etiqueta. Al inicio de esta exploración, me planteé el mostrar aquellos factores que tuvieran una correlación con las medidas superior al ±0.43, siendo este número el coeficiente más pequeño encontrado entre las medidas, sin embargo, esto no fue posible al final. Ningún factor mostró una correlación superior a ±0.43 con alguna de las medidas. Este otro umbral lo propuse debido a que la colinealidad se sospecha a partir de ±0.7. Con esto buscaba indicar y descartar factores con este comportamiento debido a que mi objetivo principal es la relación entre las medidas y las distintas etiquetas ESIN, y no el crear el clasificador último a partir de utilizar tantas propiedades como sea posible. No obstante, debido a que no se muestra una correlación fuerte entre las medidas y los factores propuestos, no se descartó ninguno por colinealidad, aunque, se descartan por otra razón: de pocos a ninguno muestran una correlación moderada con las etiquetas. En la siguiente tabla presento los factores destacables a pesar de este escenario.

Tabla 33. Resumen de correlaciones identificadas entre cada agrupación ESIN y los factores

Etiquetas ESIN (10 etiquetas)	
Etiqueta	**Correlación con factores**
No Identificable Baja [0]	**No_Definido = 0.23**** Definido = -0.23** No aplica (CS) = 0.26**
No Identificable [1]	**Tamaño de frase = 0.29**** Cantidad de nominales = 0.26** No aplica (CS) = -0.24**
Identificable Baja [9]	**No aplica (CS)= 0.24**** Atributo (CS) = 0.22**
Etiquetas ESIN_R1 (4 etiquetas). Ninguna correlación destacable y significativa.	
Etiquetas ESIN_R2 (2 etiquetas). Ninguna correlación destacable y significativa.	
Etiquetas ESIN_R3 (2 etiquetas). Ninguna correlación destacable y significativa.	

Como se ve en la Tabla 33, sólo en algunas etiquetas de los Estados Informativos sin reducción (ESIN) se detectó una correlación moderada. Primero, entre la No Identificable Baja [0] y la etiqueta No Definido que significa la ausencia del rasgo definido en oposición al conjunto de datos que sí presentaron el rasgo. Debido a que es dicotómica, se muestra también una correlación negativa con la etiqueta Definido. Esto puede entenderse como que la ausencia del rasgo Definido tiene una correlación destacable, pero leve, con aquellas frases en donde el hablante supone no identificabilidad del referente, y aparte, son introducidas por preposición, sin determinante, sin modificadores y no son nombres propios. Esto podría delinear para futuros trabajos la pregunta sobre si estas condiciones estructurales predicen mejor la poca capacidad de recuperar a un referente, y al mismo tiempo, destacar las diferencias entre un dispositivo referencial pleno y uno reducido. Se podría señalar que, precisamente por ausencia del determinante, es esperable esta correlación negativa. No obstante, el indefinido no mostró algún tipo de correlación y también depende de la forma del determinante. Esto puede deberse a la poca presencia del indefinido (n = 138) en comparación al definido (n = 1 145) en COPENOR_cero. La verificación de esto quedaría pendiente para otros trabajos, pero por lo pronto, resulta interesante destacar esta proporción en los textos periodísticos y su influencia en este etiquetado.

Luego encontramos que el tamaño de la frase y la cantidad de nominales tienen una correlación leve con la etiqueta No Identificable [1] de las ESIN (r = 0.29** para el tamaño de la frase; r = 0.26** para la cantidad de nominales). Esto nos indica que a mayor cantidad de palabras en la frase y a mayor cantidad de nominales (pero no de verbos), la correlación con esta etiqueta es más fuerte. Sobre la correlación entre estos factores se encontró que son sospechosos de presentar colinealidad, como se ve en la siguiente tabla (Tabla 34), la

correlación en todos ellos es superior al ±0.43. Debido a que los verbos tienen muy poca presencia y, además, su cantidad como factor no muestra una correlación como sí lo fue el tamaño de la frase y la cantidad de nominales, se le descarta en este estudio. No obstante, el tamaño de la frase y la cantidad de nominales tienen una correlación superior al ±0.7, lo que los vuelve aún más sospechosos de presentar colinealidad. Debido a esto, y a que el tamaño de la frase mostró una correlación un poco más alta con la etiqueta ESIN señalada, se seleccionará este factor solamente.

Tabla 34. Correlación entre la cantidad de nominales, verbos y el tamaño de la frase

Factores correlacionados		**Pearson *r***
Cantidad de Nominales	Cantidad de verbos	**0.48****
	Tamaño de la Frase	**0.73****
Tamaño de la Frase	Cantidad de verbos	**0.57****

Por último, se puede observar que la situación sintáctica puede funcionar como factor para el caso de la etiqueta No identificable Baja [0], No Identificable [1] e Identificable Baja [9]. Para la primera, se encontró que, si la frase está en una situación en donde **no aplica** el que exprese el sujeto, objeto directo o indirecto, o el que forme parte de la estructura verbal y estar encabezada por una preposición, entonces se presenta una correlación leve (r = 0.26**). En este contexto, el No aplica (CS) debe entenderse como la ausencia de aquellas caracterizaciones, por lo que una correlación positiva indica que es necesario un análisis más fino de otras posibles situaciones sintácticas que no sean las enlistadas. Por otro lado, el No identificable [1] tiene una correlación negativa con el No aplica (CS) (r = -0.24**). Si observamos la correlación de esta etiqueta con otras caracterizaciones sintácticas, no observamos alguna correlación destacable, por lo que esto significa que la caracterización propuesta (sujeto, objeto directo, objeto indirecto e introducido por preposición) requeriría

una nueva reagrupación. Aunque con coeficientes muy bajos, otras correlaciones para la etiqueta No identificable [1] fueron con la de Objeto directo (r = 0.17**) y el encontrarse en una frase nominal encabezada por preposición y en una estructura verbal (r = 0.11**). Esto deja la tarea pendiente de observar esta propiedad de manera más detallada en otros estudios[91].

Finalmente, para el Identificable Baja [9], la categoría No aplica (CS) tiene una correlación de r = 0.24**, esperable debido a que esta etiqueta incluye, por definición, a las frases nominales en aposición. De la misma manera, no sorprende la correlación con la categoría de Atributo (CS) que presenta un coeficiente de r = 0.22**. Sin embargo, lo que debería sorprender es que no tiene una correlación moderada a fuerte (superior a ±0.5 por ejemplo). Esto nos da señal de que la caracterización sintáctica no es suficiente para identificar la propiedad pragmática que se pretende predecir.

Como se observó en la Tabla 33, las únicas etiquetas que mostraron alguna correlación fueron las que pertenecen a la primera agrupación ESIN, y no todas, sólo tres de las diez. En ninguna de las reducciones propuestas para la experimentación se encontró que los factores extra propuestos tuvieran alguna correlación destacable.

Antes de continuar, presento un resumen de las medidas y los factores que tienen una correlación destacable con cada una de las etiquetas de las agrupaciones ESIN propuestas. Existen algunos casos en donde los factores mostraron una correlación más fuerte que las

[91] Cabe señalar que Du Bois (2003, 38) sostiene que el rol sintáctico preferido para hablar de entidades nuevas es el de objeto directo. Mi estudio no tiene como objetivo dilucidar si se corrobora este patrón, pero mis resultados podrían dar pie a mayor profundidad en futuras investigaciones. Nótese que sólo en este conjunto de categorías (ESIN, 10 etiquetas) es en donde el objeto directo tiene esta fuerza de correlación. En los otros conjuntos es aún más leve.

medidas, pero no superaron el ±0.2, por lo que se descartaron para este estudio. Como ya se observó, los únicos factores, más allá de las medidas, que mostraron colinealidad, fueron el conteo de nominales, de verbos y el tamaño de la frase. En la Tabla 35 sólo se encuentran aquellos que parecen ser mejores candidatos a ser predictores, tomando las medidas como prioridad.

Tabla 35. Resumen de los predictores integrados después de la correlación

Etiquetas ESIN	**Predictores**	**Pearson *r***
No identificable Baja [0]	No_Definido	0.23**
	No aplica (CS)	0.26**
No identificable [1]	LSA Interior-*w* (MLIN)	-0.27**
	Tamaño de frase	0.29**
	No aplica (CS)	-0.24**
Inactivo por MLP [2]	SPAN Interior-*w* (MSIN)	0.24**
Inactivo por RD [3]	LSA Interior-*w* (MLIN)	0.5**
Identificable Bajo [9]	No aplica (CS)	0.24**
	Atributo (CS)	0.22**
Etiquetas ESIN_R1		
No identificable [1]	LSA Interior-*wd* (MLWD)	-0.3**
Inactivo [2]	LSA Interior-*wd* (MLWD)	0.37**
Etiquetas ESIN_R2		
Nuevo [0]	LSA Interior-*w* (MLIN)	-0.3**
Dado [1]	LSA Interior-*w* (MLIN)	0.30**
Etiquetas ESIN_R3		
Fuera de texto [0]	LSA Interior-*wd* (MLWD)	-0.44**
Activo por texto [1]	LSA Interior-*wd* (MLWD)	0.44**

De los factores observados para No identificable Baja [0] del grupo ESIN, se descartó el Definido debido a la colinealidad con el No_Definido[92]. Se puede observar que en este estudio la medida de span presentó una correlación leve sólo con una etiqueta en una sola agrupación: Inactivo por MLP [2], siendo esta la mejor correlación. Existe otro factor extra que se exploró en este trabajo, pero que por su complejidad se atenderá de manera separada. Este factor es la posición relativa de la frase nominal en la nota. La razón por la cual se debe ser precavido con un factor de esta naturaleza es que, debido a que su tendencia es estrictamente lineal (aumenta conforme se progresa) podría orientar erróneamente otros factores o incluso las medidas. Resulta intuitivo pensar que una frase nominal, conforme se acerque al final de texto, presentará una alta probabilidad de formalizar la suposición sobre que el oyente puede recuperar al referente de memoria o inferirlo; es decir, no es necesaria una nueva figuración, un referente nuevo. Esto nos orillaría a proponer que tal vez la posición sea mejor predictor del Estado Informativo que las medidas propuestas. Esto se debe confrontar, a lo que le dedico la siguiente sección.

3.2.2 Correlación con la posición relativa

Obsérvese la siguiente gráfica (Figura 28) en donde la nota COPENOR-253BC[93] (*n* = 144) muestra el progreso de cada frase nominal. En el eje horizontal tenemos cada frase nominal a partir de una medida creada para su posición relativa en el texto que va de 0 a 1, en donde

[92] Aunque no para los bosques aleatorios. Debido a la manera en que se procesa la tabla final para la clasificación, se tomó al predictor dicotómico Definido, que también mostró una correlación (para más detalles ver §3.7).
[93] Esta nota se encuentra en el Anexo B para su examinación.

el 1 representa la última frase nominal y el 0 es una posición imaginaria de inicio, ninguna frase nominal es el 0 absoluto. A esta medida le llamaré Factor-P, representados por los círculos, los cuales forman una diagonal a lo largo de la gráfica, lo que muestra su progresión en el tiempo. En esta misma gráfica agrego la medida LSA Interior-*w* (superpuesta en el eje vertical, representada por cuadrados), la cual, como se observó en la sección anterior, es la que mostró mejores resultados de correlación. Debido a que también va de 0 a 1, podemos compararla con el Factor-P. Esto deja ver cierta linealidad entre ambas, es decir, que conforme crece una también la otra:

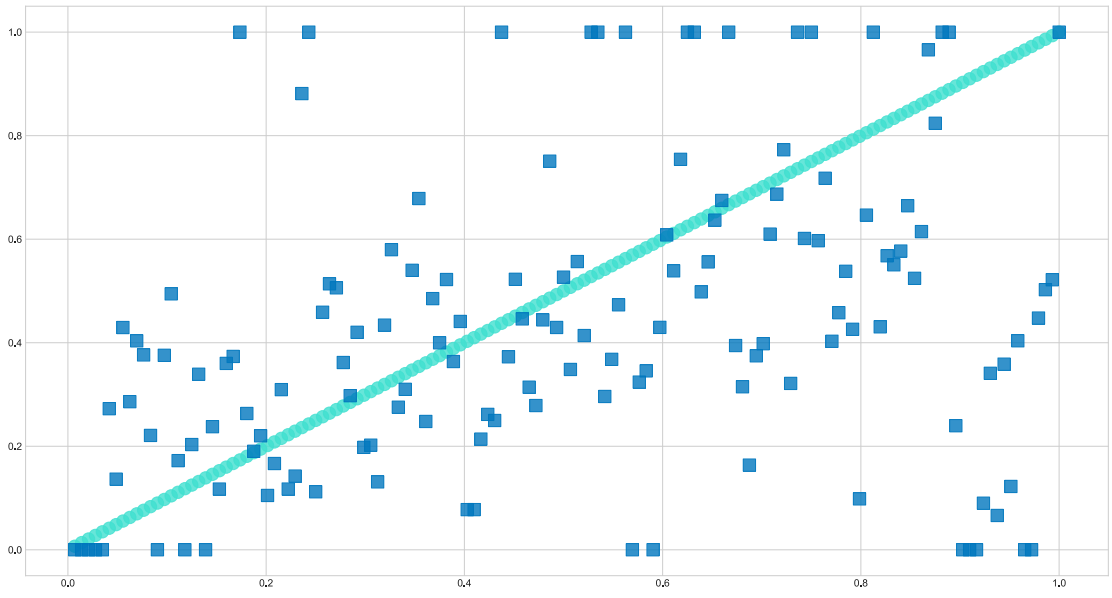


Figura 28. Gráfico de dispersión de las frases nominales en copenor-253BC

Si en el eje horizontal utilizamos el Factor-P y en el vertical esta misma medida para todas las frases nominales (*n* = 2 388) de **todas** las notas, observamos el siguiente comportamiento:

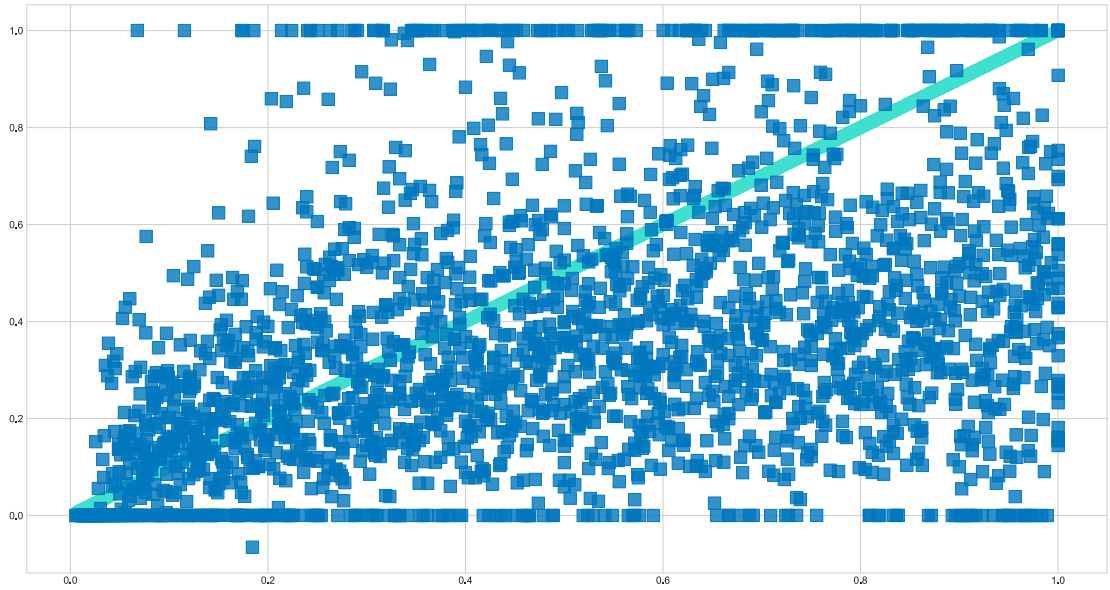


Figura 29. Gráfico de dispersión de las frases nominales en copenor_dos usando span Interior-w

En esta gráfica se observa tanta variabilidad que pareciera ser poco informativa y podría ser engañoso concluir que la medida no guarda alguna correlación con el Factor-P. Sin embargo, esto no es así. Entre las medidas y el Factor-P, la correlación más fuerte surgió con SPAN Interior-*wd* (-0.52**) y la más leve con SPAN Ventana-*n* (-0.28**). Esto lo muestro en la siguiente tabla (Tabla 36). Se puede notar que existe una correlación más fuerte que las observadas entre las ESIN y los distintos predictores.

Tabla 36. Medidas correlacionadas con el Factor-P

Medidas correlacionadas	**Pearson *r***
SPAN Interior-*w* (MSIN)	-0.49**
LSA Interior-*w* (MLIN)	0.40**
SPAN Ventana-*n* (MSVN)	-0.28**
LSA Ventana-*n* (MLVN)	0.34**
SPAN Interior-*wd* (MSWD)	-0.52**
LSA Interior-*wd* (MLWD)	0.34**

Si utilizamos el criterio de ±0.7 para sospechar colinealidad, el Factor-P y las medidas podrían convivir como predictores. No obstante, en este trabajo utilicé el ±0.43, y como se observa, es superado por varios coeficientes: entre el Factor-P y SPAN Interior-*w* (-0.49**) y SPAN Interior-*wd* (-0.52**). Además, recordemos que la correlación más alta entre las medidas y alguna de las etiquetas fue de 0.5**, entre Inactivo por RD [3] de la agrupación ESIN y LSA Interior-*w*. Debido a esto, sólo en este caso recomendaría no utilizar el Factor-P como predictor.

El siguiente paso para confrontar las medidas y el Factor-P fue el medir la correlación entre este último y las distintas agrupaciones ESIN. Podríamos intuir que a mayor Factor-P menor probabilidad de presentarnos etiquetas relacionadas con lo nuevo (Nuevo [0] de ESIN_R2, Fuera de texto [0] de ESIN_R3 o toda la secuencia de etiquetas No identificable de ESIN y ESIN_R1). Los resultados nos muestran que incluso si consideramos las medidas y el Factor-P como predictores, encontramos cerca de cero relaciones con muchas de las etiquetas. Destaca la correlación con Inactivo por RD [3] de ESIN (0.16**) pero incluso esto apenas explicaría menos del 4% de su varianza. Llama la atención que frente a una variable con un comportamiento lineal como el Factor-P, las etiquetas de la [5] a la [9] de ESIN y la [3] y [4] de las ESIN_R1 no pasen el valor de significancia mínimo requerido ($p < .05$) esto nos da por lo menos tres panoramas a explorar: primero, el que desencadena el no tener un valor *p* significativo, es que las observaciones con respecto al Factor-P son tan buenas como una observación aleatoria; segundo, esto en parte puede deberse a las pocas observaciones y a que las medidas en estas agrupaciones tienen un comportamiento más variable; tercero, llama la atención que especialmente fueran estas etiquetas debido a que ellas, y en particular las activas, probablemente se resolverían mejor usando modelos para la resolución de

anáforas entre dispositivos referenciales reducidos. Estos últimos panoramas no sólo se presentan en este caso de correlación con el Factor-P, sino también, por lo que se ha podido observar, con las medidas y las distintas agrupaciones ESIN: ninguna de estas etiquetas (las accesibles y activas) han podido ser relacionadas con las medidas en general. Un resumen de lo anterior se muestro en la siguiente tabla a continuación:

Tabla 37. Correlación entre las agrupaciones ESIN y el Factor-P

Etiquetas ESIN	**Pearson *r***
No identificable Baja [0]	-0.0049
No identificable [1]	-0.0929**
Inactivo por MLP [2]	-0.1273**
Inactivo por RD [3]	**0.1649****
Accesible por Marco [4]	0.0748**
Accesible por Origo [5]	-0.0452*
Activo S [6]	-0.0289
Activo O/I [7]	0.0388
Activo P [8]	-0.0027
Identificable Baja [9]	-0.0023
Etiquetas ESIN_R1	
No identificable [1]	-0.0973**
Inactivo [2]	0.0846**
Accesible [3]	0.0526*
Activo [4]	-0.0068
Etiquetas ESIN_R2	
Nuevo [0]	-0.0973**
Dado [1]	0.0973**
Etiquetas ESIN_R3	
Fuera de texto [0]	-0.1342**
Activo por texto [1]	0.1342**

Debido a estos resultados, podemos adelantar que la posición relativa de la frase nominal con respecto a la nota en la que aparece (Factor-P) no resultará un predictor efectivo en el modelo de clasificación.

3.3 Análisis de varianza de una sola vía

Todo lo anterior nos empieza a delinear cuáles predictores son mejores para cada etiqueta. Sin embargo, es relevante evaluar si las etiquetas en las agrupaciones pueden considerarse segmentos bien diferenciados entre ellos, por lo menos en términos estadísticos, y con respecto a las medidas propuestas. Una primera aproximación para resolver esta incógnita está en la visualización de los datos en gráficas de cajas y bigotes (también llamadas de manera sencilla *gráficas de caja* o *box plots*). A continuación, muestro estas gráficas usando los datos de LSA y SPAN y todas las bolsas trabajadas (Figura 30) usando como pivote las etiquetas ESIN. El objetivo es observar si una o varias etiquetas puede separarse de las demás en cada conjunto de medidas dadas las etiquetas ESIN.

En las gráficas presentes en la Figura 30 podemos notar que la peor medida para la clasificación sería MSVN (SPAN Ventana-*n*). Es difícil distinguir que los grupos se comportan de manera distinta. Por otro lado, es más claro observar que MLWD (LSA Interior-*wd*) separa la etiqueta 3 de otras etiquetas en la parte superior de la gráfica. Esta etiqueta, recordemos, pertenece al código numérico para Inactivo por Registro Discursivo [3] de las ESIN. Esto nos da una primera aproximación, pero podemos probar con un análisis de varianza de una vía para cada una de las bolsas.

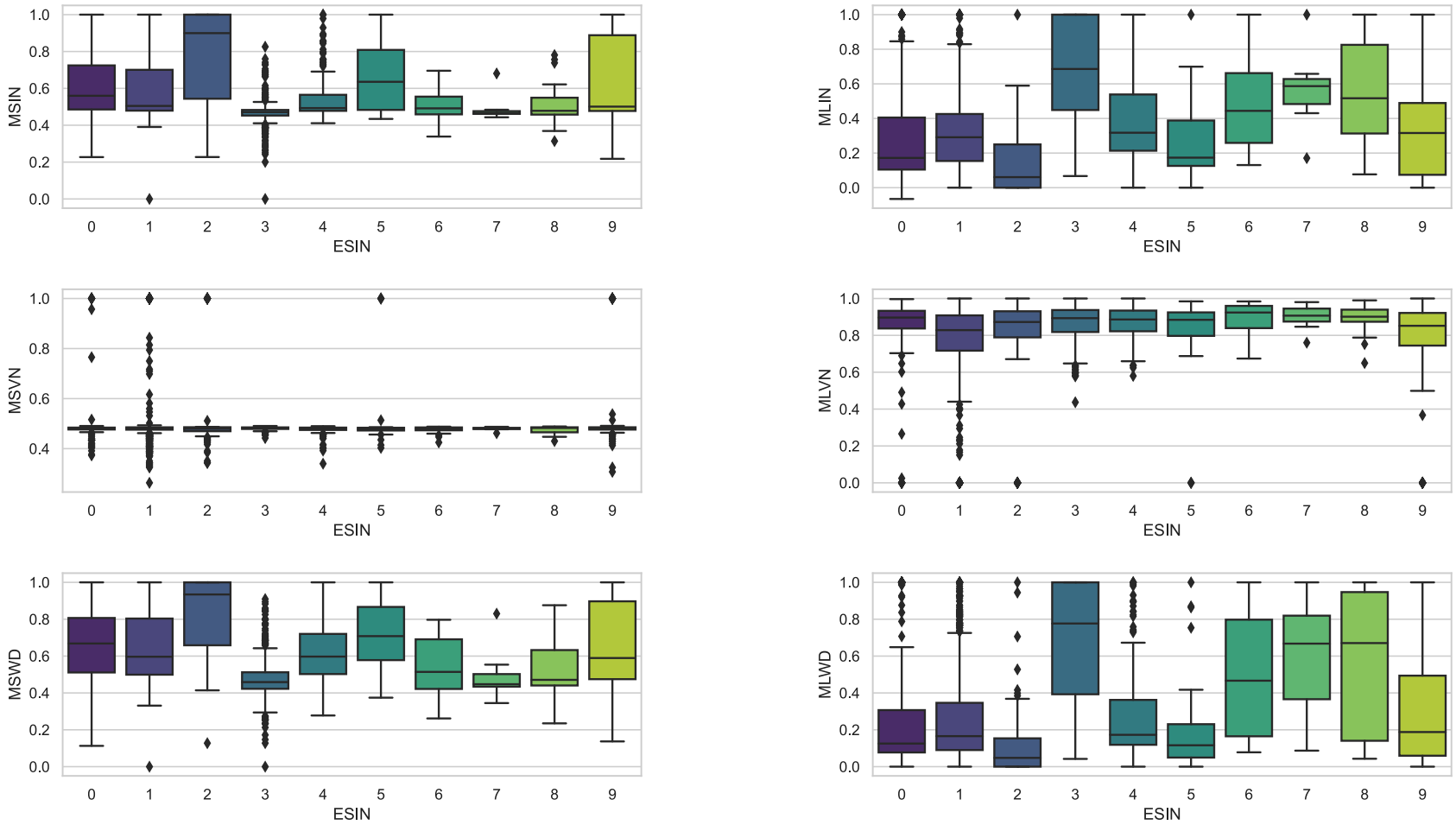


Figura 30. Gráficas de caja comparando las etiquetas ESIN con LSA y SPAN

En este caso, la hipótesis nula establece que todos los conjuntos dada una etiqueta provienen de la misma población (es decir, no hay diferencia), mientras que la hipótesis alternativa sostiene que por lo menos alguno de los conjuntos proviene de una población distinta. Los resultados los presento en la Tabla 38.

Tabla 38. ANOVA entre ESIN y medidas

Medida	**valor *p***	**estadístico *F***
MSIN	$4.3e^{-77}$**	46.18032
MLIN	$5e^{-170}$**	108.1174
MSVN	$4.97e^{-06}$**	4.590196
MLVN	$2.06e^{-22}$**	14.11798
MSWD	$1.4e^{-94}$**	57.0397
MLWD	$3.8e^{-161}$**	101.7217

Para todos los casos anteriores, los grados de libertad reportados se expresan como F(9, 2 378). En todas las bolsas, como se observa en la Tabla 38, se rechaza la hipótesis nula ($p < .001$), en consecuencia, por lo menos una de las etiquetas está siendo bien diferenciada. Podríamos desarrollar un segundo análisis para evaluar la capacidad de cada etiqueta de diferenciarse de las demás, pero esto lo dejaré para el análisis Conover-Iman (§3.6). Por lo pronto, esto nos ayuda a determinar que las medidas y la manera de agruparlas (a partir de alguna agrupación ESIN) no es producto del azar.

Para el caso de la primera reducción (ESIN_R1) sus gráficas de cajas lucen de la siguiente manera (Figura 31). Notamos en este caso que MLWD sigue mostrando una fuerte inclinación a separar la etiqueta Inactivo [3], esto también se observa en MLIN. Por otro lado, MSIN presenta muchos casos extremos (los puntos o también llamados *outsiders*).

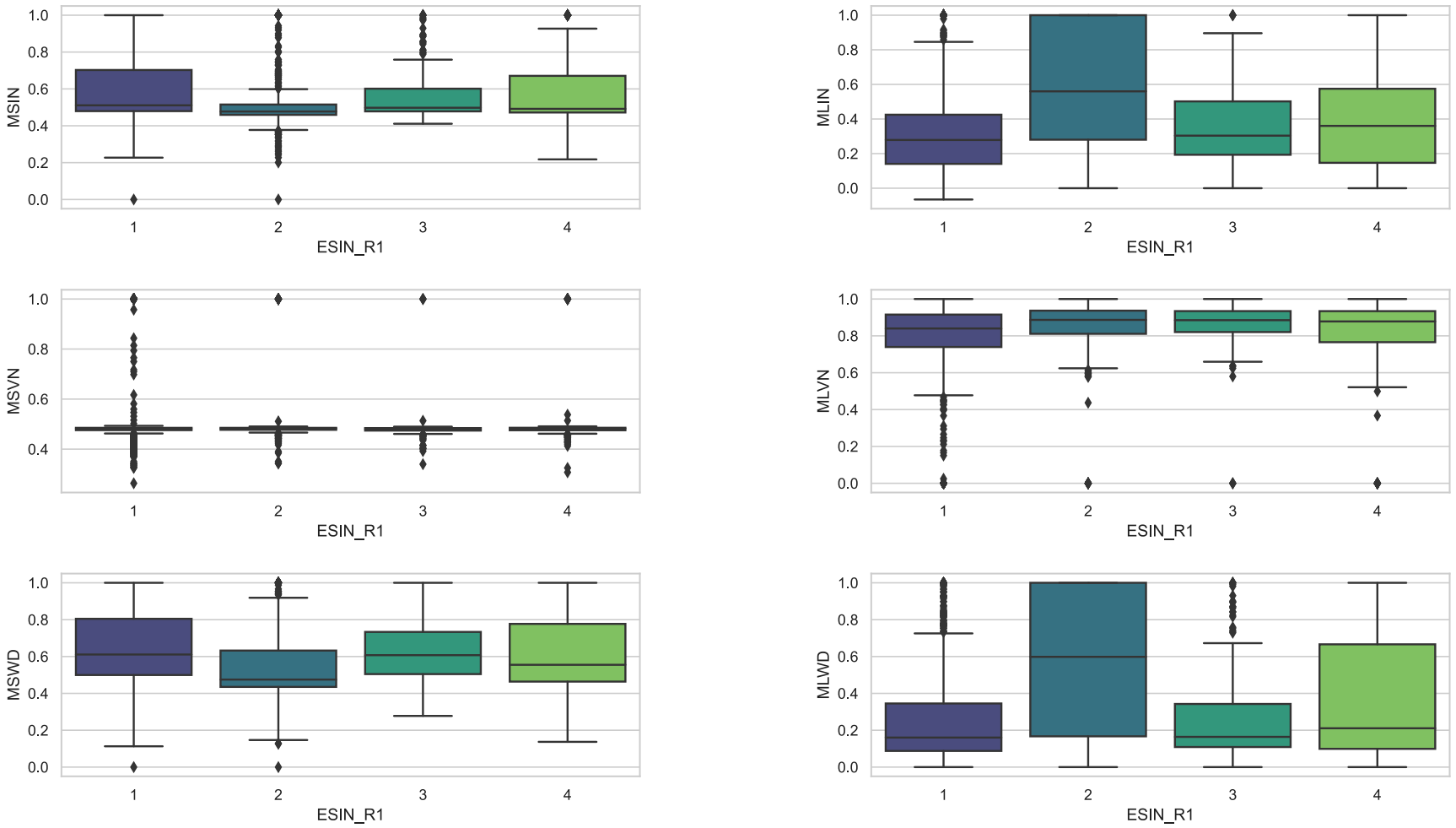


Figura 31. Gráficas de caja comparando las etiquetas ESIN_R1 con LSA y SPAN

Las dos medidas que utilizan las bolsas exteriores (Ventana-*n*) muestran una pobre diferencia entre las etiquetas. De la misma manera que el caso anterior, para corroborar esto se realiza un análisis de varianza cuyos resultados muestro en la Tabla 39.

Tabla 39. ANOVA entre ESIN_R1 y medidas

Medida	**valor *p***	**estadístico *F***
MSIN	$4.37e^{-13}$**	20.45488
MLIN	$9.29e^{-79}$**	131.3688
MSVN	0.00257*	4.766986
MLVN	$1.9e^{-13}$**	21.0354
MSWD	$2.38e^{-25}$**	40.21906
MLWD	$3.93e^{-83}$**	139.2453

De la misma manera que el caso en el análisis de varianza de ESIN, para esta agrupación casi en todas las medidas cruzadas por las etiquetas de ESIN_R1 se puede rechazar la hipótesis nula dado $p < .001$; sólo el caso de la medida SPAN Ventana-*n* (MSVN) se rechazaría por $p < .05$ que sería el mínimo aceptable. Para estos casos F(3, 2 384). Esto nos señala, en esta reducción, se presenta por lo menos una etiqueta diferenciada de las demás.

Para el caso de la reducción de acuerdo con los trabajos antecedentes (ESIN_R2), nos enfrentamos, como ya se ha visto, a una variable dependiente dicotómica. Aun así, puede observarse la manera en que se agrupan los datos dadas las medidas en las gráficas de caja mostradas en la Figura 32. En ellas se observa que, por ejemplo, MLWD muestra una gran variabilidad para clasificar Dado [1], aunque Nuevo [0] parece favorecer números pequeños. Al igual que los casos anteriores, las medidas SPAN y LSA usando la Ventana-*n* (MSVN y MLVN) no parecen diferenciar bien entre las dos etiquetas.

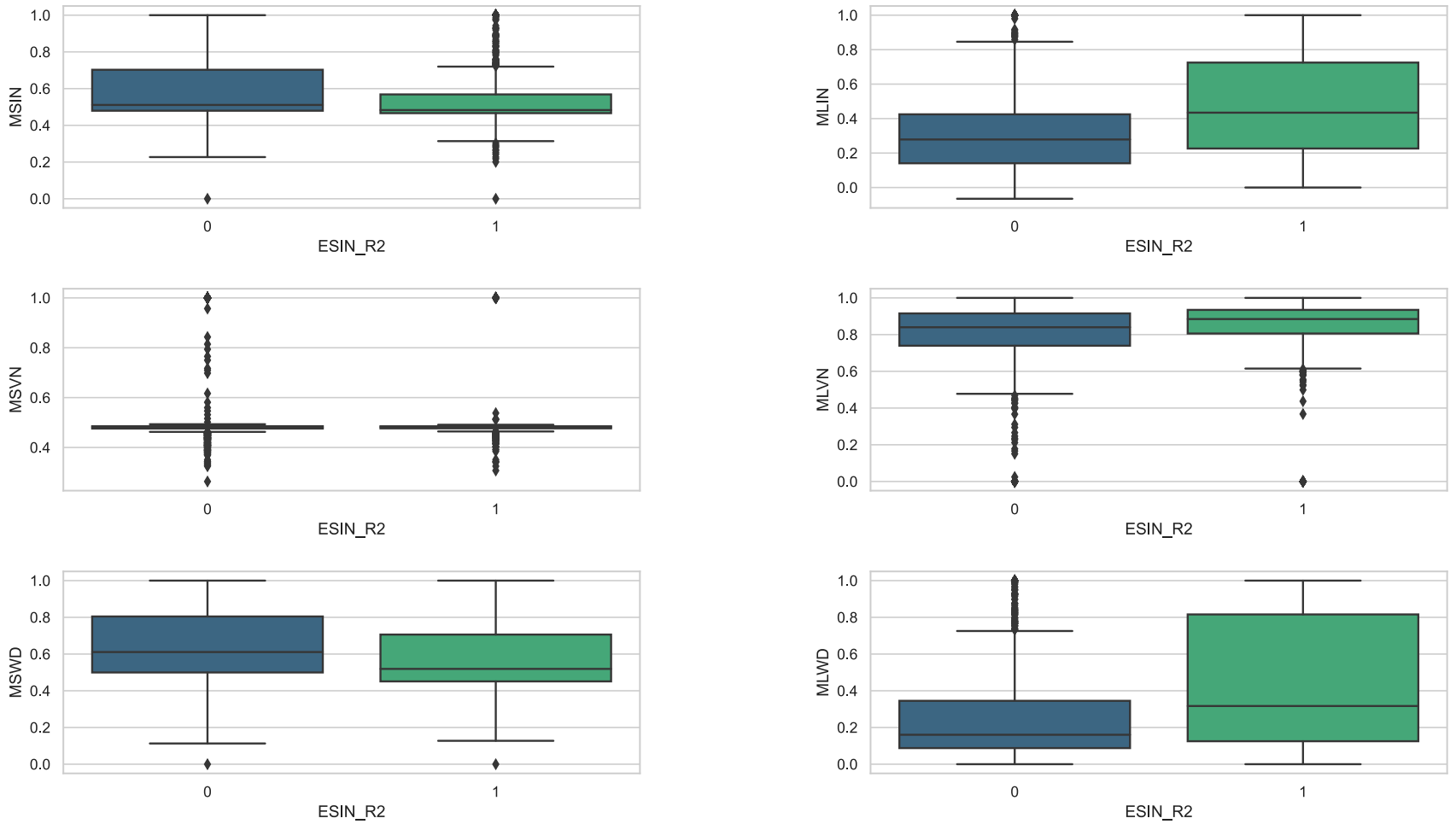


Figura 32. Gráficas de caja comparando las etiquetas ESIN_2 con LSA y SPAN

Por otro lado, MLIN muestra una mejor separación entre las etiquetas a partir de concentrar aquellas observaciones de Dado [1] en un rango medio (contrastado con MLWD, por ejemplo). MSIN y MSWD, aunque con menor dispersión, se observa menor diferencia: la etiqueta Dado [1] parece ser abarcada por la caja de lo Nuevo [0]. Esta visualización, como se ha demostrado con los cálculos, es un primer paso para determinar la capacidad de clasificación, pero no es hasta analizar las varianzas en donde tenemos una respuesta más fina a la pregunta sobre la separación de los grupos. En la Tabla 40 muestro los resultados de análisis de varianza (F(1, 2 386)) para las ESIN_R2 y las medidas. Puede observarse que, en todas, por lo menos uno de los dos grupos es distinto al de la población, es decir, en todas se rechaza la hipótesis nula. No obstante, de nuevo para MSVN, el valor *p* se evalúa en contra de $p < .05$ para considerarlo significativo.

Tabla 40. ANOVA entre ESIN_R2 y medidas

Medida	**valor *p***	**estadístico *F***
MSIN	1.99e-10**	40.82623
MLIN	4.56e-52**	242.0895
MSVN	0.001198*	10.518
MLVN	2.13e-13**	54.50323
MSWD	2.1e-18**	77.84974
MLWD	2.3e-50**	233.5039

Esta misma situación dicotómica la tenemos para la reducción que propongo, ESIN_R3. Para este caso, la visualización de las etiquetas agrupadas a partir de las medidas luce como se muestra en la Figura 33.

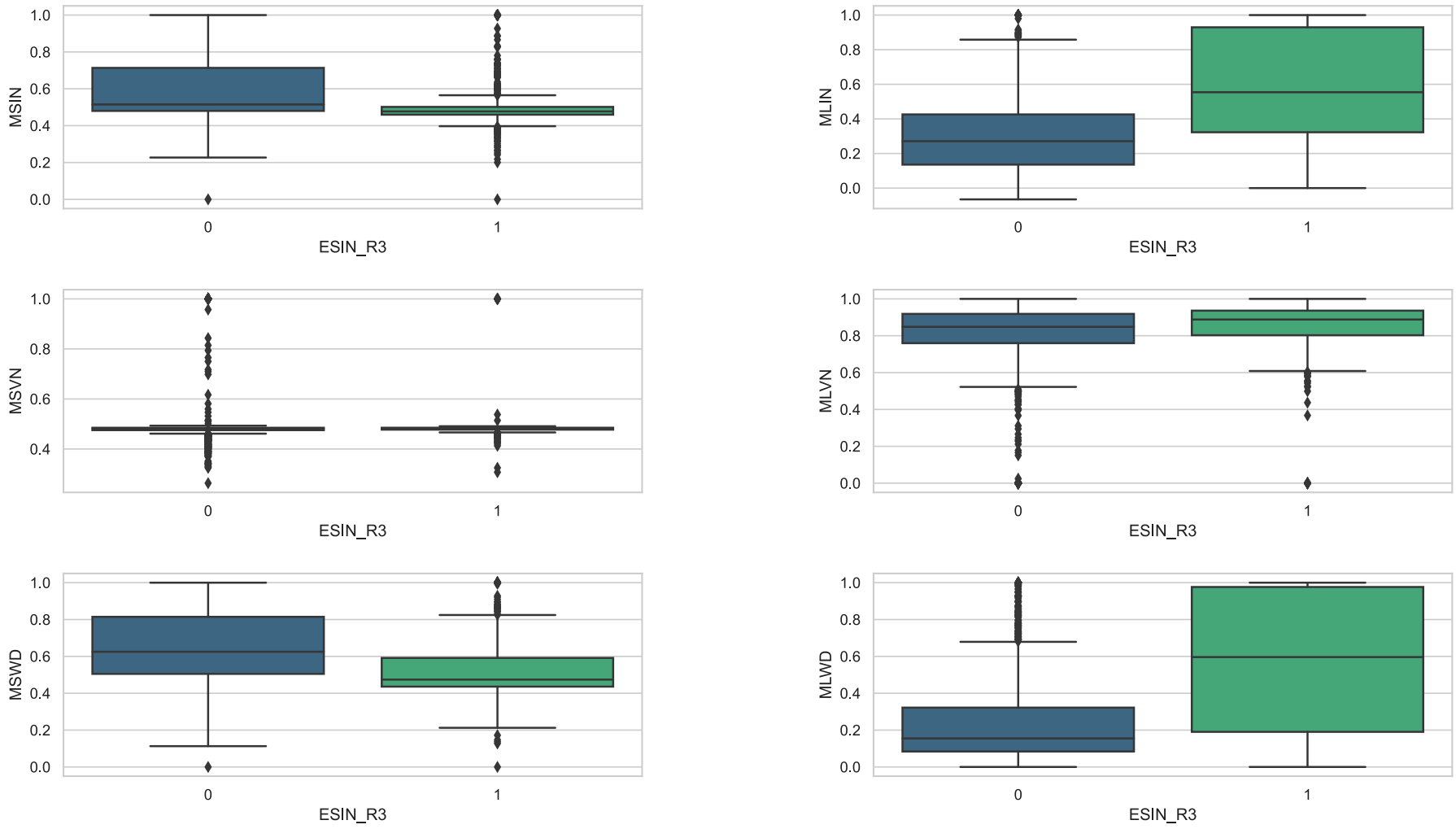


Figura 33. Gráficas de caja comparando las etiquetas ESIN_R3 con LSA y SPAN

Si uno compara las gráficas de las ESIN_R2 con las de las ESIN_R3, se puede notar que son muy similares, pero que, de hecho, ESIN_R3 parece contrastar más las diferencias en todos los casos. Por ejemplo, se observa que en MLIN la etiqueta Activo por texto [1] está más agrupada en números altos, lo que permite separarla de manera más clara de la etiqueta Fuera de texto [0], cuyos datos tienden a agruparse en números bajos (pero, nótese, no a cero). De la misma manera que en ESIN_R2, el análisis de varianza nos muestra que por lo menos uno de los dos grupos puede diferenciarse (Tabla 41) (F(1, 2 386)); y, a notar por las gráficas, parece que es Fuera de texto [0]. En todos los casos se rechaza la hipótesis nula.

Tabla 41. ANOVA entre ESIN_R3 y medidas

Medida	**valor *p***	**estadístico *F***
MSIN	$4.04e^{-30}$**	133.6657
MLIN	$1.2e^{-109}$**	550.2393
MSVN	0.001636*	9.941238
MLVN	$1.78e^{-09}$**	36.48672
MSWD	$1e^{-51}$**	240.3689
MLWD	$5.9e^{-115}$**	580.3376

3.4 Regresión múltiple

Lo desarrollado en las secciones anteriores nos permite adelantar algunas descripciones del comportamiento de las medidas y los factores como predictores:

a. No todas las medidas funcionan como predictores.
b. Sólo una medida puede funcionar como predictora so pena de colinealidad.
c. No todas las etiquetas de las agrupaciones pueden ser predichas por las medidas.

d. No todas las propiedades lexicogramaticales incluidas (factores externos) funcionarán como predictores con el mismo peso.

En particular, es LSA por medio de una bolsa interior la que mejor predice algunas etiquetas, y son las etiquetas relacionadas con Estados Informativos inactivos (Inactivo por RD [3] de ESIN; e Inactivo [2] de ESIN_R1) así como las que requieren información fuera del texto (Nuevo [0] y Fuera de texto [0], de ESIN_R2 y ESIN_R3 respectivamente) las que parecen agruparse mejor dada esta medida. Para probar esto, construiré varios modelos de regresión con el objetivo de comprender mejor el poder predictivo de las medidas para estas propiedades pragmáticas. Para la primera ronda de modelos, utilizaré todos los predictores (medidas y factores) y en una segunda ronda sólo me enfocaré en aquellos que hayan tenido una mejor correlación (de acuerdo con lo visto en §3.2); en el caso de ESIN y ESIN_R1, utilizaré un modelo de regresión logística multinomial; para ESIN_R1 y ESIN_R2 bastará la regresión logística múltiple. Los parámetros generales para crear el modelo de regresión, así como las características generales de los resultados son los siguientes:

- **Conjunto de prueba:** 33% del original, el resto conforma el conjunto de entrenamiento, creado de manera aleatoria por medio de *train_test_split* de *sklearn.*
- **Efecto de los predictores:** por medio de *f_regression* de *sklearn* utilizando el conjunto de entrenamiento. Se reporta la estadístico-F y el valor *p*.
- **Matriz de confusión:** se reportan los porcentajes de las predicciones correctas (verdaderos positivos y verdaderos negativos) e incorrectas (falsos positivos y falsos negativos) dado el conjunto de prueba.

- **Exactitud (*accuracy*):** se reporta el promedio de la exactitud de todas las etiquetas. La exactitud se define como la división entre las predicciones correctas y todas las predicciones hechas.
- **Precisión, Exhaustividad y Métrica-F_β[94]:** se reportan la cantidad de verdaderos positivos entre la suma de los verdaderos positivos y los falsos positivos (*Precisión*); la cantidad de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos (*Exhaustividad*); también se reporta una medida de precisión ponderando la precisión y la exhaustividad (*Métrica-F_β*). Particularmente, para F_β se puede modificar el valor de β para dar mayor o menor preponderancia a la Precisión, sin embargo, en este caso, se determinó que $\beta = 1$, lo que significa que no se penaliza o se mejora la Precisión.

La Métrica-F_β puede ayudar a la comparación con futuros trabajos, mientras que la Exactitud me permitirá comparar los resultados con el de los antecedentes. A continuación, expongo los resultados de los clasificadores.

3.4.1 Regresión con todos los predictores

En la primera ronda se realizó una regresión logística con todos los predictores. Los resultados para la agrupación de etiquetas ESIN (10 etiquetas) se muestran a continuación (Tabla 42). Lo primero a destacar es la relevancia de los predictores en el modelo lo cual puede explorase con el estadístico-F. He marcado con negritas los casos más altos, y con

[94] Cuando haga referencia a las métricas de reporte de resultados como Exactitud, Precisión, Exhaustividad y Métrica- F_β, utilizaré la mayúscula al inicio de la palabra.

sombra gris los predictores cuyo estadístico también presenta relevancia cercana a los más altos.

Tabla 42. Efecto de todos los predictores para la agrupación ESIN[95]

Predictores		valor *p*	estadístico *F*
MSIN		0.011254*	6.439787
MLIN		5.83e^{-08}**	29.70035
MSVN		0.035682*	4.419737
MLVN		0.000842**	11.18827
MSWD		3.5e^{-05}**	17.22174
MLWD		1.35e^{-09}**	**37.17436**
Factor-P		0.007953*	7.061817
Cantidad de Verbos (Ø)		0.367607	0.812198
Cantidad de Nominales		4.86e^{-06}**	21.0353
Cantidad de palabras (Ø)		0.108435	2.579775
NA_DEF (Ø)		0.283577	1.150642
DEF(Ø)		0.283577	1.150642
NA_IND		0.031633*	4.626416
INDEF		0.031633*	4.626416
CS	**Sujeto (Ø)**	0.540056	0.375598
	OD	1.87e^{-08}**	31.94672
	OI(Ø)	0.625167	0.238765
	AT	1.93e^{-16}**	69.1484
	PR	5.55e^{-06}**	20.77781
	NA_CS	7.68e^{-10}**	**38.305**

Se puede observar que el peso en las medidas se encuentra en MLWD y MLIN, un perfil que ya habíamos observado en las correlaciones y en el análisis de varianza. Llama la atención

[95] Marco con el signo Ø a un lado del nombre del predictor, aquellos que no tienen un valor *p* inferior a 0.05.

que el factor CS Objeto Directo tenga un peso más importante que la mayoría de las medidas, y que, de hecho, el No Aplica CS, que se refiere a las situaciones sintácticas que no se consideraron en este trabajo, tenga el mayor peso de todos los predictores. Fuera de estos casos en donde el factor CS tiene peso, lo demás corresponde a lo que ya se había alcanzado a observar en los resultados anteriores.

De esta manera, una vez entrenado el modelo y realizado las predicciones con el conjunto de prueba, obtenemos los siguientes datos (Figura 34):

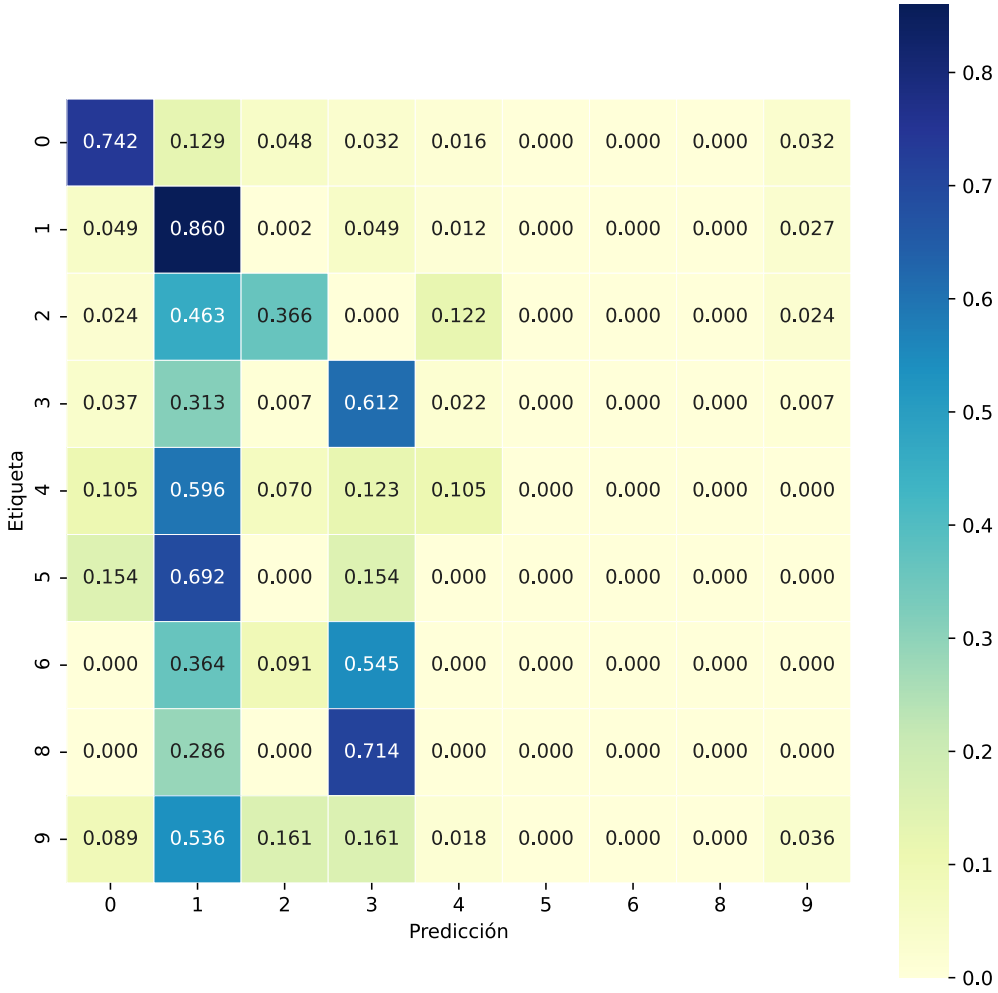


Figura 34. Matriz de confusión de ESIN

Lo anterior deja ver tres aspectos interesantes: el primero, que tiene que ver con la integridad de los datos originales, es que desapareció la etiqueta 7, es decir, Activo O/I [7]. Esto se debe probablemente a que son muy pocas observaciones (n = 7) lo que vuelve imposible hablar de esta etiqueta. De hecho, el siguiente punto tiene que ver con esta limitación en las observaciones: el modelo falla por completo en predecir las etiquetas Activo S [6] y Activo P [8]; así también para Accesible Origo [5]. Para los casos de Accesible por Marco [4] e Identificable Bajo [9], tenemos algunas predicciones correctas, pero de hecho muy pocas. El tercer aspecto es que el modelo parece identificar muy bien la etiqueta No Identificable [1] con un 86% de verdaderos positivos, sin embargo, esto es engañoso. Obsérvese que los errores cometidos en las otras etiquetas se concentran en esta columna: el predictor está sobre analizando casi todas las frases nominales como No Identificable [1]; en contraste, pareciera que el 36% que logra predecir correctamente con Inactivo MLP [2] se diferencia bien de todo el grupo; el siguiente caso sería Inactivo RD [3] con el 62% de predicciones correctas, que como se observó en los resultados de las secciones anteriores, es la etiqueta más prometedora de ser predicha. Pero, para este modelo, también se comete errores al sobre analizar: por ejemplo, 71% de las etiquetas Activo P [8] se etiquetan erróneamente como Inactivo RD [3].

La etiqueta que parece predecirse y diferenciarse bien de todas es la de No Identificable Baja [0] que, como se mencionó, son casos en donde la falta de dependencia de estructura verbal, el tamaño de la frase y la ausencia de determinantes pueden ayudar a predecir el Estado Informativo. El éxito de estas predicciones podría deberse a que el modelo aprovecha esta información: un patrón estructural regular. Podemos apuntalar esta interpretación por lo que se observó en el peso de los predictores de la tabla anterior (Tabla 42), en aquellos que no son las medidas.

Esta matriz nos muestra una Exactitud de apenas 63%, lo que está muy por debajo de los modelos en los trabajos antecedentes, los cuales lograron entre 70-80% de Exactitud. A continuación, muestro las calificaciones de acuerdo con distintas métricas (Tabla 43). Se observará que No Identificable [1] es la mejor calificada y la No identificable Baja [0] no alcanza una métrica superior, a pesar de lo visto en la matriz. Esto se debe a que para la métrica se utiliza la cantidad de predicciones hechas. En particular, se busca que el número llegue a 1, lo que significa predicción perfecta. Los ceros, como se adelantó, son casos catastróficos: el modelo es incapaz de detectar estos Estados Informativos.

Tabla 43. Métricas de evaluación del clasificador de las etiquetas ESIN[96]

Exactitud: 63%

Etiqueta ESIN	**Precisión**	**Exhaustividad**	**F_β**	**C. P.**
No identificable Baja [0]	0.54	0.74	0.62	62
No identificable [1]	**0.70**	**0.86**	**0.77**	408
Inactivo por MLP [2]	0.44	0.36	0.4	41
Inactivo por RD [3]	0.61	0.61	0.61	134
Accesible por Marco [4]	0.28	0.10	0.15	57
Accesible por Origo [5]	0	0	0	13
Activo S [6]	0	0	0	11
Activo O/I [7]	---	---	---	0
Activo P [8]	0	0	0	7
Identificable Baja [9]	0.11	0.03	0.05	56

[96] Se reportan hasta los primeros dos decimales. C. P. = Número de observaciones de cada etiqueta en el conjunto de prueba.

En lo que respecta a la primera reducción (ESIN_R1), los predictores tienen el siguiente comportamiento (Tabla 44):

Tabla 44. Efecto de todos los predictores para las etiquetas ESIN_R1

Predictores		**valor *p***	**estadístico *F***
	MSIN	0.004452*	8.112528
	MLIN	$2.1e^{-09}$**	36.29242
	MSVN	0.00778*	7.101295
	MLVN	$1.97e^{-06}$**	22.7895
	MSWD	$2.57e^{-05}$**	17.81773
	MLWD	$1.74e^{-10}$**	41.28123
	Factor-P	0.003457*	8.574911
	Cantidad de Verbos	0.499034**	0.457199
	Cantidad de Nominales	$1.12e^{-10}$**	42.16734
	Cantidad de palabras	$7.16e^{-07}$**	24.76846
	NA_DEF (Ø)	0.306315	1.047167
	DEF(Ø)	0.306315	1.047167
	NA_IND	0.000797**	11.29054
	INDEF	0.000797**	11.29054
CS	**Sujeto (Ø)**	0.686162	0.163331
	OD	$4.95e^{-11}$**	**43.80126**
	OI(Ø)	0.764903	0.08946
	AT	$7.95e^{-11}$**	42.84778
	PR	0.000159**	14.3357
	NA_CS	$1.62e^{-10}$**	41.41361

Tenemos que MLWD y MLIN son las mejores medidas de acuerdo con su peso. Pero, en resonancia con los resultados de los predictores para la agrupación ESIN simple, los predictores CS Objeto Directo, Atributo y No Aplica tienen un peso mayor en el modelo, y

para este caso, superan a las medidas por un pequeño rango. En este punto, si realizamos el contraste con los predictores que se suponen los mejores de acuerdo con las correlaciones (Tabla 35), pareciera que estos casos en particular no serían buenos para identificar sino para lograr contraste: mejorarían la capacidad de descartar frases nominales, aunque no se supiera si fueran Activo o Accesible (por ejemplo). Veamos la matriz de confusión a continuación para corroborar esto (Figura 35).

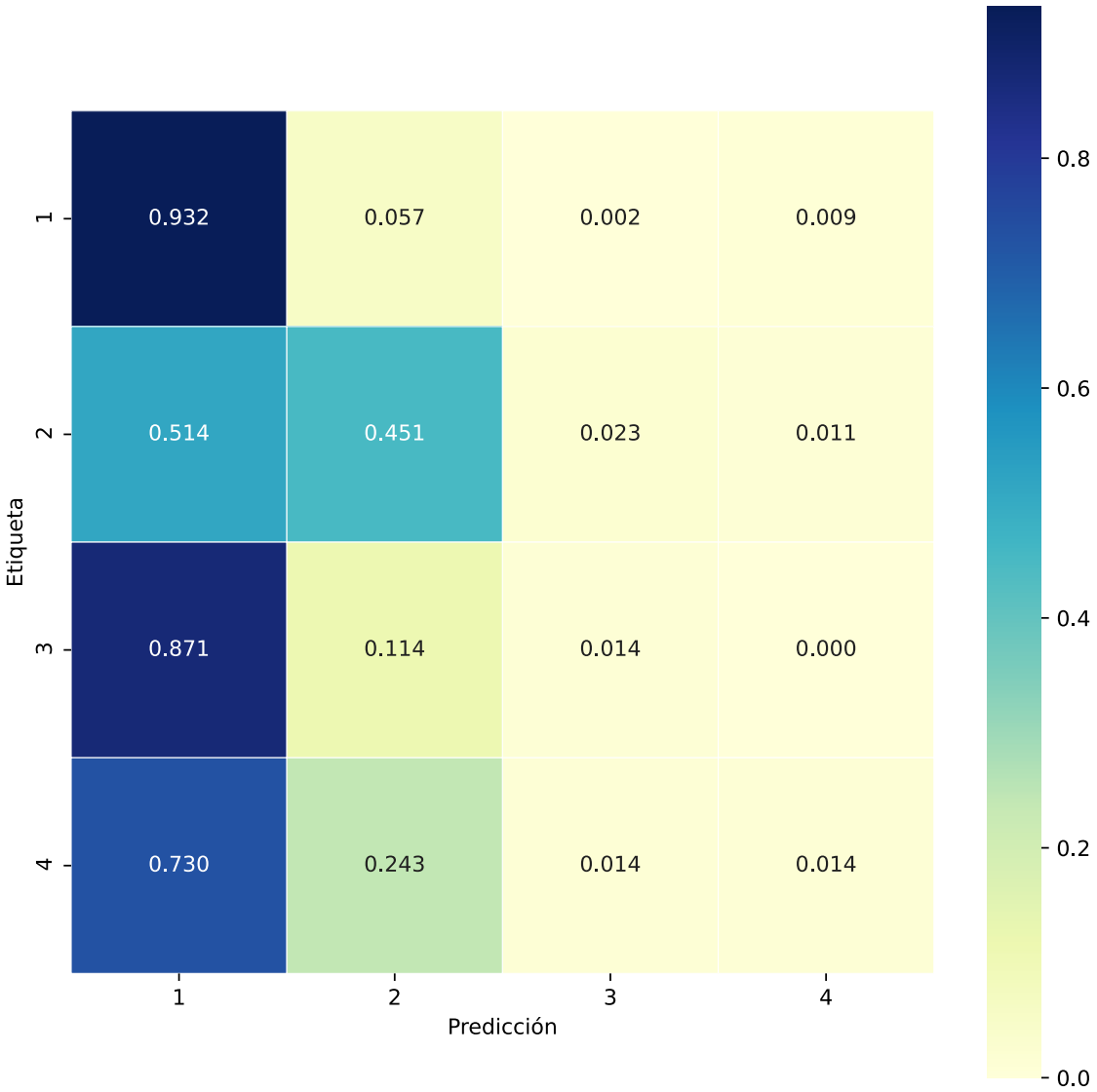


Figura 35. Matriz de confusión de ESIN_R1

Observamos que el modelo vuelve a fallar en predecir las etiquetas Accesible [3] y Activo [4], y vuelve a predecir gran parte de las etiquetas para No Identificable [1]. No obstante, el algoritmo etiqueta más del 50% de las otras etiquetas como No Identificable [1]. Esto, por otro lado, no sucede para el Inactivo [2] que tiene una menor cantidad de predicciones erróneas, aunque el total de predicciones correctas es menor al 50%. Si revisamos las métricas para la evaluación (Tabla 45), observamos que, en efecto, la medida F_β para No identificable [1] es apenas una décima mayor a la del modelo anterior para la agrupación ESIN. Aunque el modelo en general obtiene una Exactitud mayor a la del modelo anterior en 3 puntos (62 a 65%), las métricas para cada etiqueta no ofrecen un contraste tan alto, excepto por la diferencia entre la Precisión y la Exhaustividad de la No Identificable [1]: tiene mayor Exhaustividad que el modelo anterior, pero no Precisión. Podemos decir que ambos modelos se comportan casi de manera similar y la variación de las etiquetas dentro de las agrupaciones no ayudó.

Tabla 45. Métricas de evaluación del clasificador de las etiquetas ESIN_R1

Exactitud: 65%

Etiqueta ESIN	**Precisión**	**Exhaustividad**	**F_β**	**C. P.**
No identificable [1]	**0.68**	**0.93**	**0.78**	470
Inactivo [2]	0.59	0.45	0.51	175
Accesible [3]	0.14	0.01	0.02	70
Activo [4]	0.14	0.01	0.02	74

Para crear los dos modelos anteriores utilicé una regresión logística múltiple multinomial. Para los dos que siguen, utilicé una regresión logística en la que la variable dependiente es

dicotómica. Para el caso de la agrupación ESIN_R2, los predictores se comportaron de la siguiente manera (Tabla 46).

Tabla 46. Efecto de todos los predictores para las etiquetas ESIN_R2

Predictores		**valor *p***	**estadístico *F***
	MSIN	4.1e^{-08}**	30.39518
	MLIN	2.74e^{-34}**	**156.325**
	MSVN	0.001997*	9.584158
	MLVN	1.03e^{-09}**	37.72011
	MSWD	1.94e^{-14}**	59.69481
	MLWD	6.08e^{-35}**	**159.611**
	Factor-P	4.56e^{-05}**	16.71306
	Cantidad de Verbos	0.007359*	7.201665
	Cantidad de Nominales	2.28e^{-18}**	78.31328
	Cantidad de palabras	7.55e^{-17}**	71.08177
	NA_DEF	1.93e^{-05}**	18.36472
	DEF	1.93e^{-05}**	18.36472
	NA_IND	1.22e^{-06}**	23.72557
	INDEF	1.22e^{-06}**	23.72557
CS	**Sujeto**	0.063537*	3.447345
	OD	6.16e^{-11}**	43.36167
	OI(Ø)	0.180616	1.794114
	AT	0.007772*	7.103195
	PR	0.008788*	6.882411
	NA_CS	2.18e^{-05}**	18.13465

Sin lugar a duda, dada esta agrupación de etiquetas en una situación binaria, MLWD y MLIN destacan por encima de todos los predictores. Esto ya se había alcanzado a observar en las correlaciones, lo que ratifica que el diccionario aporta en la construcción de diferencias. Las otras medidas tienen pesos cercanos entre ellas, destacables por encima de otros predictores,

pero no tan altos. No obstante, se quedan detrás de dos factores: la Cantidad de Palabras dentro de la frase nominal y el CS Objeto Directo. Esto vuelve a poner el acento sobre que el modelo usa este tipo de información y no sólo las medidas, para lograr las predicciones. Veamos el desempeño de este modelo:

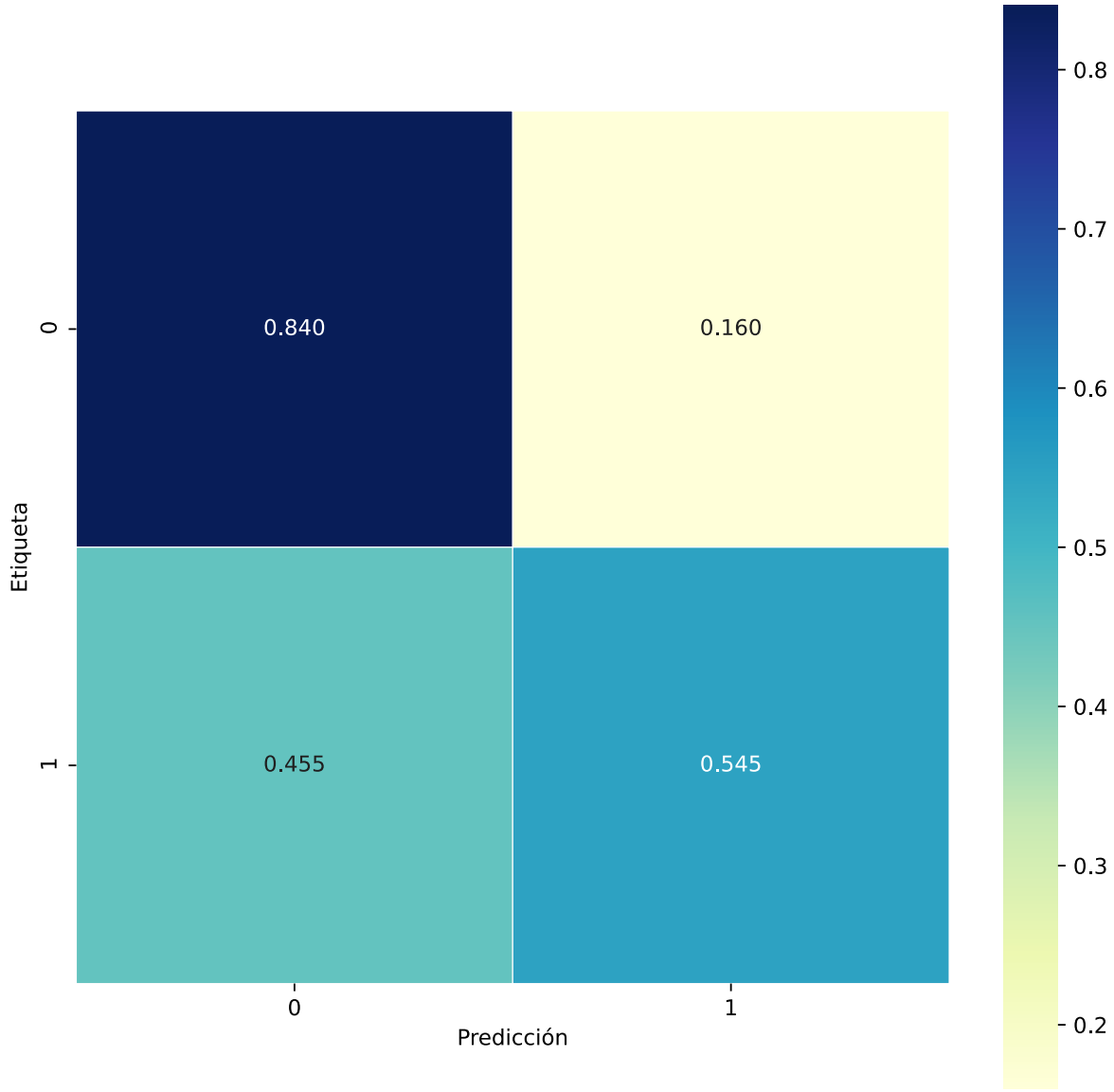


Figura 36. Matriz de confusión de ESIN_R2

En la gráfica anterior (Figura 36) se observa el patrón que hemos visto en las otras matrices de confusión: existe una mayor exhaustividad en la etiqueta Nuevo [0], mientras que para lo Dado [1] hay más errores de predicción: 45% de lo Dado [1] es etiquetado como Nuevo [0]. El modelo completo alcanzó una Exactitud del 72%, muy por encima de los otros dos modelos y cercano a lo encontrado por el trabajo antecedente, pero sin poder superarlo. Esto muestra que un modelo que contemple todas las etiquetas es ineficiente.

Si observamos el desempeño de cada etiqueta (Tabla 47), observamos que no es tan distinto con respecto a los modelos anteriores: lo Nuevo [0] tiene la misma F_β que lo No Identificable [1] de ESIN_R1 (0.78). Lo Dado [1] aumenta en comparación a las etiquetas Inactivo [2], Accesible [3] y Activo [4], pero no aporta en la Exactitud global. Fuera de este comportamiento, hay algo interesante sobre la calidad de los datos: vemos que una vez que las dos muestras tienen una cantidad de observaciones parecidas, las métricas varían, pero no drásticamente: se alcanza una Exactitud mayor en general, pero lo Nuevo [0] mantiene un mismo comportamiento.

Tabla 47. Métricas de evaluación del clasificador de las etiquetas ESIN_R2

Exactitud: 72%				
Etiqueta ESIN	**Precisión**	**Exhaustividad**	**F_β**	**C. P.**
Nuevo [0]	**0.73**	**0.84**	**0.78**	470
Dado [1]	0.7	0.54	0.61	319

El último modelo en estas condiciones fue el que tomó la agrupación ESIN_R3. Para este caso, los predictores tuvieron el comportamiento que se resumen en la siguiente tabla.

Tabla 48. Efecto de todos los predictores para las etiquetas ESIN_R3

Predictores		valor *p*	estadístico *F*
MSIN		1.37e^{-21}**	93.76703
MLIN		9.72e^{-69}**	**338.6376**
MSVN		0.000996**	10.8755
MLVN		3.33e^{-08}**	30.80552
MSWD		1.1e^{-35}**	163.3625
MLWD		9.27e^{-73}**	**361.1443**
Factor-P		2.52e^{-08}**	31.36207
Cantidad de Verbos (Ø)		0.792432	0.069273
Cantidad de Nominales		1.59e^{-06}**	23.21296
Cantidad de palabras (Ø)		0.007633*	7.135824
NA_DEF (Ø)		0.000162**	14.29994
DEF(Ø)		0.000162**	14.29994
NA_IND		0.004009*	8.303694
INDEF		0.004009*	8.303694
CS	**Sujeto**	0.000308**	13.07783
	OD	1.18e^{-05}**	19.31461
	OI(Ø)	0.279627	1.169694
	AT	3.33e^{-05}**	17.31713
	PR	4.66e^{-08}**	30.14559
	NA_CS	0.000377**	12.69544

Lo primero que sale a relucir, en contraste con lo analizado para la otra agrupación dicotómica ESIN_R2, es que el peso que alcanzan las medidas con respecto a los factores se dispara por poco más de un orden de magnitud: en primer y segundo lugar tenemos MLWD y MLIN, esto conforme a lo observado en las correlaciones. Le siguen MSWD y MSIN.

Aunque destacables entre ellos —pero para nada comparables con las medidas— se encuentran los factores CS Preposición y el Factor-P. Esto no es lo único sobresaliente de este modelo. Una vez entrenado, este modelo sobrepasa a los antecedentes con un 81% de Exactitud.

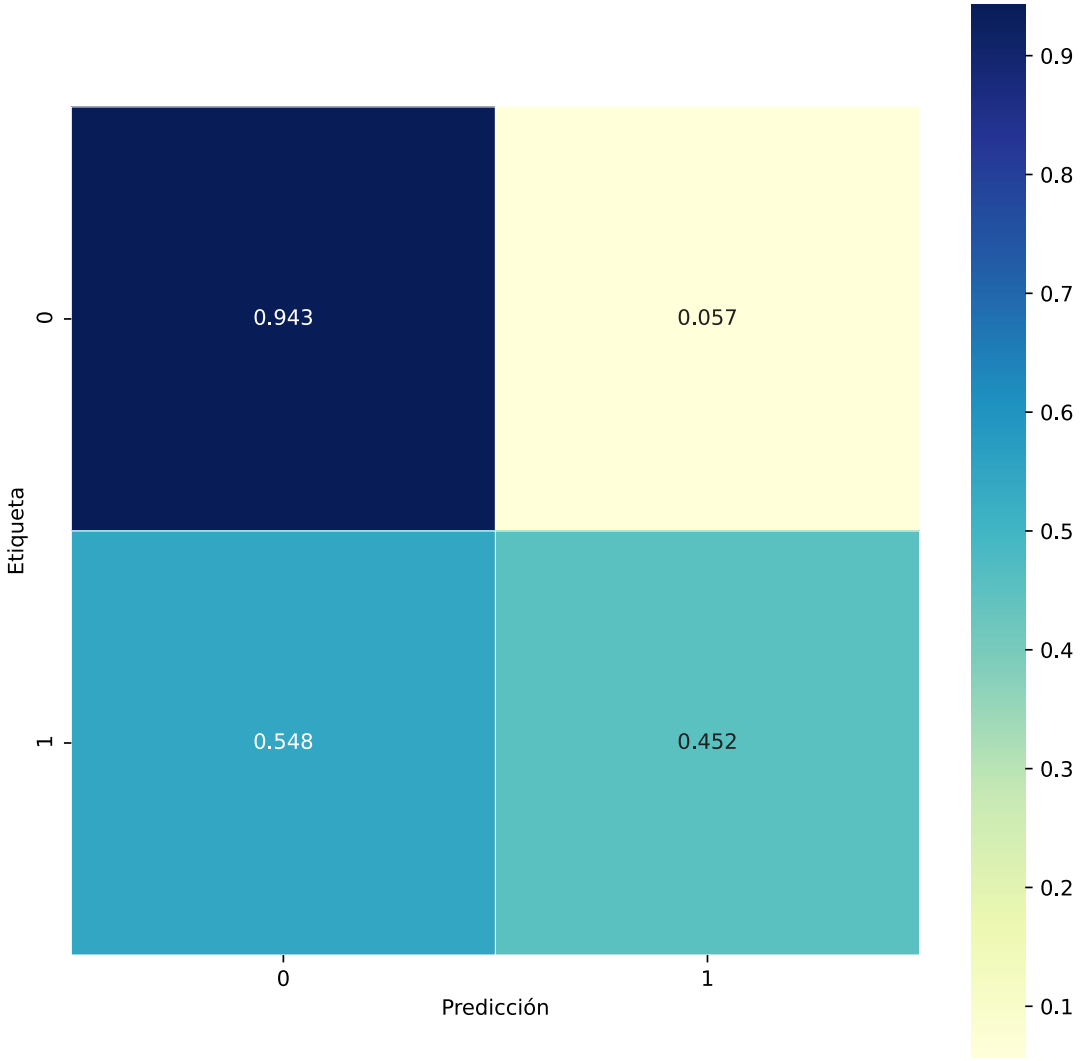


Figura 37. Matriz de confusión de ESIN_R3

En la Figura 37 observamos que pareciera no distinguirse mucho del modelo anterior, pero en realidad lo que se observa es una mayor capacidad para predecir lo Fuera de texto [0]. Tendríamos que ser cuidadosos con estos resultados: este modelo no es superior al anterior

en todos los aspectos, sobre todo, no lo es en el tamaño de las observaciones en el conjunto de prueba. También, obsérvese (Tabla 49) que, al agrupar los datos con estas etiquetas, se presenta mayor Precisión para lo Activo por texto [1] (de 0.7, en Dado [1] de ESIN_R2, pasamos a 0.74), su Exhaustividad es más pobre.

Tabla 49. Métricas de evaluación del clasificador de las etiquetas ESIN_R3

Exactitud: 81%

Etiqueta ESIN	**Precisión**	**Exhaustividad**	**F_β**	**C. P.**
Fuera de texto [0]	**0.82**	**0.94**	**0.88**	581
Activo por texto [1]	0.74	0.45	0.56	208

Lo anterior parece mostrar de manera prometedora, con sus bemoles, al modelo que agrupa las mediciones de acuerdo con ESIN_R3. A manera de contraste de estos resultados, realizo las regresiones con los predictores que mostraron una mejor correlación. Esto lo muestro a continuación.

3.4.2 Regresión con restricciones

A manera de sintetizar estos resultados, muestro a continuación (Tabla 50) el conjunto de predictores de cada modelo (reproduzco Tabla 35 omitiendo los predictores repetidos para cada etiqueta). Para este caso, omitiré las matrices de confusión y mostraré directamente las métricas. Pero antes, mostraré el peso de cada uno de los predictores para cada agrupación. Cabe recordar que el estadístico F compara con un modelo en donde el peso de los predictores es cero, por lo que no sólo resulta ilustrativo en modelos con más de un predictor, sino también en aquellos en donde sólo tenemos uno.

Tabla 50. Predictores para las regresiones logísticas restringidas

Etiquetas ESIN	**Predictores**	**Pearson *r***
No identificable Baja [0]	No_Definido	0.23**
	No aplica (CS)	0.26**
No identificable [1]	Tamaño de frase	0.29**
Inactivo por MLP [2]	SPAN Interior-*w* (MSIN)	0.24**
Inactivo por RD [3]	LSA Interior-*w* (MLIN)	0.5**
Identificable Bajo [9]	Atributo (CS)	0.22**
Etiquetas ESIN_R1		
No identificable [1]	LSA Interior-*wd* (MLWD)	-0.3**
Inactivo [2]	LSA Interior-*wd* (MLWD)	0.37**
Etiquetas ESIN_R2		
Nuevo [0]	LSA Interior-*w* (MLIN)	-0.3**
Dado [1]	LSA Interior-*w* (MLIN)	0.30**
Etiquetas ESIN_R3		
Fuera de texto [0]	LSA Interior-*wd* (MLWD)	-0.44**
Activo por texto [1]	LSA Interior-*wd* (MLWD)	0.44**

Los resultados (Tabla 51) muestran que para los predictores restringidos de la ESIN, el CS Atributo supera a las medidas. Esto confirma lo que ya habíamos observado en el modelo para ESIN con todos los predictores. Para el caso de los modelos con un solo predictor, observamos que el mejor peso está en MLWD para ESIN_R3 (Tabla 52). Los otros casos no son desdeñables, pero a falta de contraste, no podemos decir mucho de ellos. Lo único que buscamos es que no sean cero y que el valor *p* sea por lo menos menor a 0.05, en donde, por cierto, todos lo alcanzan.

Tabla 51. Efecto de todos los predictores restringidos para las etiquetas ESIN

Predictores		**valor *p***	**estadístico *F***
	MSIN	0.011254*	6.439787
	MLIN	$5.83e^{-08}$**	29.70035
	Cantidad de palabras (Ø)	0.108435	2.579775
	NA_DEF (Ø)	0.283577	1.150642
CS	**AT**	$1.93e^{-16}$**	**69.1484**
	NA_CS	$7.68e^{-10}$**	38.305

Tabla 52. Efecto de todos los predictores restringidos para ESIN_R1, R2 y R3

ESIN_R1	**valor *p***	**estadístico *F***
MLWD	$1.74e^{-10}$**	41.28123
ESIN_R2		
MLIN	$2.74e^{-34}$**	156.325
ESIN_R3		
MLWD	$9.27e^{-73}$**	361.1443

Lo anterior no resulta interpretable hasta que observemos el modelo en funcionamiento para predecir las etiquetas. A continuación, muestro en una sola tabla los resultados de estos cuatro modelos (Tabla 53). Tres puntos destacan de los resultados anteriores. Primero, la Precisión más alta la encontramos para Fuera de texto [0] en ESIN_R3. La más baja, descartando los casos con métrica cero, es Identificable Baja [9] en ESIN. Si uno observa aquellas etiquetas que no forman parte de los No identificables ([0] y [1] para ESIN y [0] para todas las demás agrupaciones), es muy raro que sean las más altas o superen a las otras etiquetas.

Tabla 53. Métricas de evaluación del clasificador restringido de todas las agrupaciones ESIN

ESIN	**Exactitud**: 64%			
	Precisión	**Exhaustividad**	**F_β**	**C. P.**
No identificable Baja [0]	0.58	0.75	0.65	62
No identificable [1]	0.67	0.91	0.77	408
Inactivo MLP [2]	0.39	0.21	0.28	41
Inactivo RD [3]	0.64	0.59	0.61	134
Accesible Marco [4]	0	0	0	57
Accesible Origo [5]	0	0	0	13
Activo S [6]	0	0	0	11
Activo O/I [7]	--	--	--	--
Activo P [8]	0	0	0	7
Identificable baja [9]	0.28	0.03	0.06	56
ESIN_R1	Exactitud: 65%			
	Precisión	**Exhaustividad**	**F_β**	**C. P.**
No identificable [1]	0.65	0.94	0.77	470
Inactivo [2]	0.63	0.40	0.49	175
Accesible [3]	0	0	0	70
Activo [4]	0	0	0	74
ESIN_R2	Exactitud: 69%			
	Precisión	**Exhaustividad**	**F_β**	**C. P.**
Nuevo [0]	0.69	0.89	0.77	470
Dado [1]	0.72	0.41	0.5	319
ESIN_R3	Exactitud: 80%			
	Precisión	**Exhaustividad**	**F_β**	**C. P.**
Fuera de texto [0]	0.81	0.94	0.87	581
Activo por texto [1]	0.73	0.40	0.52	208

El único caso que parece competir es en ESIN_R2, con lo Dado [1] con un 0.72, que supera en tres puntos a lo Nuevo [1] el cual cuenta con 0.69. No obstante, para la Exhaustividad, esto se contrasta aún más: todas las etiquetas No identificables de todas las agrupaciones superan a las otras etiquetas; además, casi ninguna de las otras etiquetas supera el 0.50. Esto ya es una señal importante de la relevancia de estas últimas etiquetas: los modelos son buenos para predecir qué frase nominal no corresponde a algo anteriormente dicho.

Segundo, pareciera que la cantidad de observaciones en el conjunto de pruebas es un gran problema en estos modelos, pero nótese No identificable Baja [0] de ESIN en contraste con Accesible por Marco [4] e Inactivo RD [3]. La etiqueta Inactivo RD [3] tiene 134 observaciones y no supera en Exhaustividad a la No identificable Baja [0] que tiene 62 observaciones. En la misma línea, Accesible por Marco [4] tiene 57 observaciones, y el modelo es incapaz de predecir esta etiqueta (en todas sus métricas tenemos ceros). La cantidad de observaciones en cada grupo es preferible para métodos paramétricos, pero en este caso, no parece influir de manera tajante.

El tercer y último punto destacable de estos resultados tiene que ver con que las agrupaciones ESIN y ESIN_R1 permiten adelantarnos a los buenos resultados que se muestran para ESIN_R3. Recordemos que esta reducción la planteé a raíz de prever una limitación técnica: aquellos Estados Informativos que apelan a conocimiento “fuera de texto” tienen mejores métricas. Esta tendencia se puede observar al contrastar, desde ESIN, las etiquetas Inactivo MLP [2] con Inactivo RD [3]: el primero apenas alcanza un 0.21 en su Métrica F_{β}, mientras que el segundo llega a 0.61. El primero apela a conocimiento que no se puede rastrear en el texto, mientras que el segundo sí.

En la siguiente gráfica (Figura 38) muestro, a manera de resumen del desarrollo anterior, el contraste entre los modelos con todos los predictores (color claro) y los modelos con predictores restringidos (color oscuro). Se utiliza la Exactitud como medida de comparación.

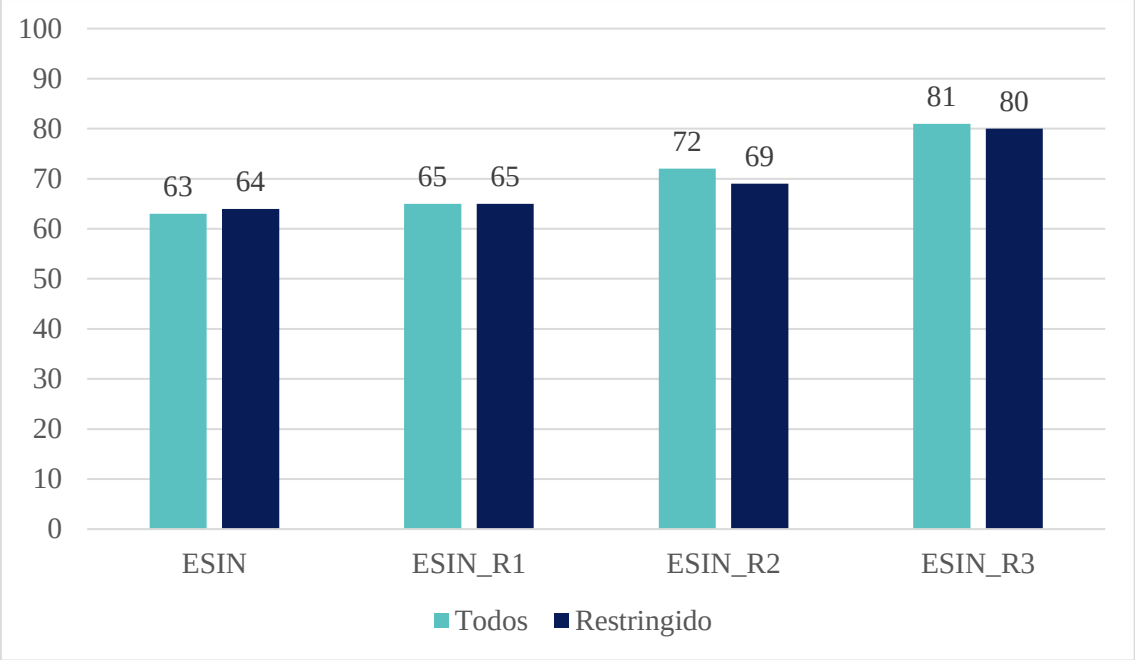


Figura 38. Exactitud (porcentual) entre los distintos modelos de regresión y los predictores

Se observa que no existe una gran diferencia entre los modelos, lo cual impacta en la cantidad de intervención que el texto necesita para obtener resultados satisfactorios. Podemos recurrir a un par de características, incluso sólo a las medidas, las cuales pueden utilizar bolsas de palabras tratadas con algún etiquetado automático de características lexicogramaticales como Stanza. El único segmentado manual necesario sería el de las frases nominales.

Las pruebas realizadas hasta este momento permiten contrastar las distintas agrupaciones, pero también —y era el principal objetivo que perseguía— permiten comparar los resultados con los antecedentes. Aunque estos resultados empiezan a encausar la respuesta al planteamiento inicial de esta investigación, me parece importante realizar un par de pruebas

más. Debido a las características de las observaciones, la siguiente prueba tiene que ver con explorar las distintas distribuciones y su relación con una distribución normal. Esto me llevará a fundamentar la razón por la cual utilizo pruebas no paramétricas en la penúltima sección, así como el clasificador de bosques aleatorios, el cual también funciona para datos que no cumplen con los requisitos de las pruebas paramétricas.

3.5 Pruebas de normalidad y gráficas

Aunque ANOVA es robusta con otras distribuciones distintas a la normal, me parece importante señalar si se sigue esta distribución. En los estudios antecedentes no se especifica este punto, aunque se supone, debido al teorema del límite central. Sin embargo, si nuestros datos no cumplen el requisito de normalidad, requisito que demandan las pruebas, sería deseable realizar pruebas no paramétricas para, por lo menos, tener una manera de contrastar los resultados.

Una forma de evaluar la normalidad de las distribuciones es observando los histogramas y comparar, a grandes rasgos, si tienen una forma de campana. Otra manera es realizar pruebas estadísticas en donde la hipótesis nula establezca que la muestra estudiada proviene de una distribución normal. Si se rechaza la hipótesis nula, entonces estamos frente a casos no paramétricos.

A continuación, expongo ambas exploraciones. En primer lugar, muestro las distribuciones en histogramas para pasar a las pruebas de cada grupo. En los siguientes casos se utilizaron las medidas agrupadas por las etiquetas, con excepción del primer caso, en donde se analizaron cada una de las medidas, es decir, seis muestras con el total de las observaciones

(*n* = 2 388). A la exploración de los histogramas le sigue una serie de tablas en donde presento los resultados de la prueba de normalidad K^2 de D'Agostino.

3.5.1 Exploración de la normalidad en las medidas

Si tomamos las seis medidas como conjuntos, cada una muestra las siguientes distribuciones (Figura 39). Algo que se puede notar en todas ellas es que tenemos picos en los extremos, lo que en algunos casos resulta ser lo único que distorsiona la distribución normal (en forma de campana). Si omitimos estos picos, la mayoría podrían lucir normales. Hay que tener en cuenta que, debido a que SPAN y LSA van de 0 a 1, no existe algo como las tendencias a ±infinito. No podemos hablar realmente de normalidad, aunque se siga una distribución de campana. No obstante, el que los datos se agrupen en una sección y desciendan poco a poco, es decir, que se agrupen alrededor de la moda, es un comportamiento deseable para estos estudios. Tomando en cuenta estos detalles, podemos observar que los casos de MLWD y MLVN tiene claras distribuciones asimétricas, en contraste con las otras cuatro.

Para disipar sospechas, realicé la prueba de normalidad K^2 de D'Agostino cuyos resultados se muestran en la Tabla 54.

Tabla 54. Resultados de la prueba de normalidad K^2 de D'Agostino a todas las medidas

Medida	**valor *p***	**estadístico K^2**
MSIN	$6.42e^{-87}$**	396.9315
MLIN	$1.86e^{-41}$**	187.5743
MSVN	0**	1854.098
MLVN	$1e^{-278}$**	1280.246
MSWD	$1.99e^{-44}$**	201.2539
MLWD	$2.43e^{-65}$**	297.5619

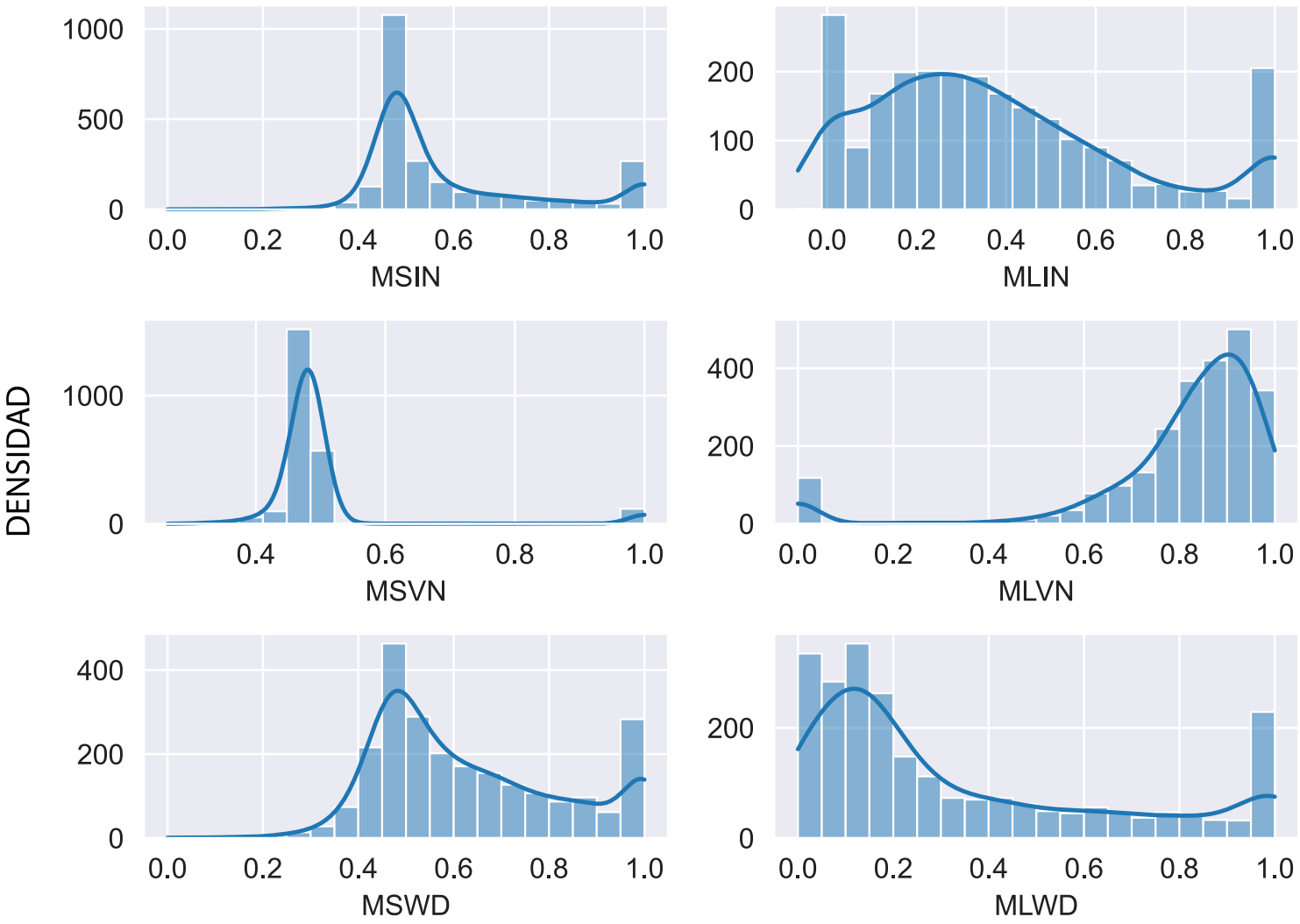


Figura 39. Histogramas de todas las medidas

El estadístico es la suma de s^2 y k^2, los cuales corresponden a los valores *z* de las dos pruebas que realiza el algoritmo: una prueba de sesgo (*skew test*) y la prueba de curtosis (*kurtosis test*); al igual que los casos anteriores, un número elevado coincide con un valor *p* bajo. Para esta prueba el valor $p < 0.05$, por lo que sólo reporto aquellos que lo superan. En todos los casos, se rechaza la hipótesis nula: no tenemos distribuciones normales.

La pista que nos otorgan los histogramas es que los extremos podrían estar influyendo en la distribución y en la prueba. Debido a esto, realizo un ensayo retirando aquellas observaciones que correspondan a ceros y unos. Esto modifica el tamaño total de la muestra, pero nos permite asegurar que estos extremos no se tocan. Los resultados de este segundo ensayo se muestran a continuación (Tabla 55), y como se puede observar, no hay un cambio tan drástico sobre el tamaño total de las observaciones (n = 2 388). En ningún caso hay un decremento de la muestra superior al 10% con respecto al original. Lo anterior nos muestra que, aun retirando los ceros y unos, las distribuciones no son normales.

Tabla 55. Resultados de la prueba de normalidad K^2 de D'Agostino a todas las medidas sin incluir las observaciones de ceros y unos

Medida	**valor *p***	**estadístico**	***n* sin 0/1**
MSIN	$3.5e^{-114}$**	522.4808	2271
MLIN	$4.71e^{-40}$**	181.1056	2194
MSVN	0**	2107.684	2271
MLVN	$4.1e^{-161}$**	738.615	2266
MSWD	$5.3e^{-30}$**	134.8199	2270
MLWD	$2.93e^{-61}$**	278.7682	2203

Todo el ejercicio lo realicé para cada una de las muestras dadas las etiquetas. Este procedimiento generó un total de 108 histogramas para cada distribución: diez por seis

medidas paras las ESIN; cuatro por seis medidas para ESIN_R1; y dos por seis medidas tanto para ESIN_R2 como ESIN_R3. Los histogramas pueden consultarse en el Anexo. Por lo pronto, muestro a continuación los resultados de la prueba de normalidad K^2 de D'Agostino a cada una de estas distribuciones (Tabla 56). Es importante notar que esta prueba no permite muestras por debajo de ocho observaciones, por lo que la etiqueta de ESIN Activo O/I [7] no muestra resultados (problema que ya observamos en las regresiones). Esto refuerza la idea de que, para analizar estos casos, podrían utilizarse otros métodos.

Por otro lado, llama la atención la distribución de Activo S [6] en MLIN, MSIN y MSWD; además Accesible por Origo [5] en MSIN y MSWD; y también Activo P [8] en MSWD. De acuerdo con este análisis, estas distribuciones tendrían una distribución normal. Coloco a continuación sus histogramas (Figura 40 y Figura 41).

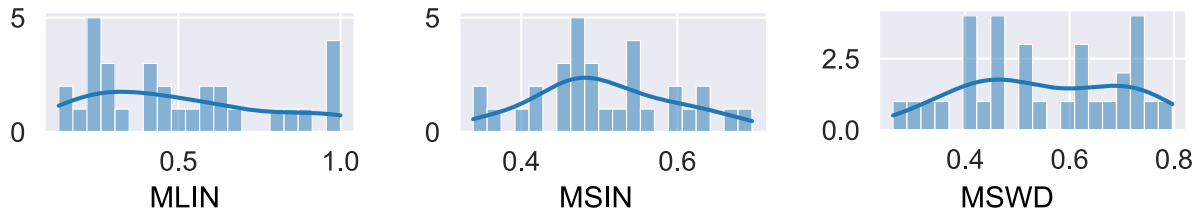


Figura 40. Histogramas de Activo S [6] en MLIN, MSIN y MSWD

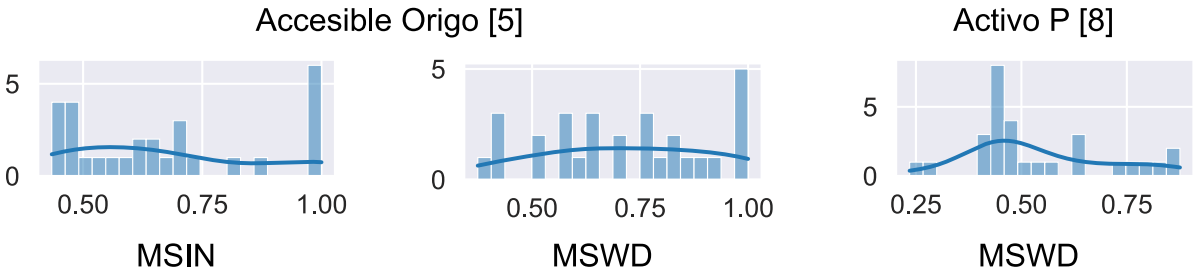


Figura 41. Histogramas de Accesible Origo [5] en MSIN y MSWD y Activo P [8] en MSWD

Tabla 56. Resultados de la prueba de normalidad K^2 de D'Agostino para todas las etiquetas de las agrupaciones ESIN

Etiquetas	**Valor *p* de las medidas**					
ESIN	**MLIN**	**MLVN**	**MLWD**	**MSIN**	**MSVN**	**MSWD**
No identificable baja [0]	1.85e^{-08}**	2.92e^{-38}**	5.91e^{-10}**	2.47e^{-05}**	1.34e^{-45}**	3.6e^{-03}*
No Identificable [1]	6.52e^{-14}**	1e^{-110}**	3.98e^{-54}**	5.05e^{-39}**	1.5e^{-169}**	5.5e^{-41}**
Inactivo MLP [2]	9.65e^{-10}**	7.48e^{-13}**	9.17e^{-21}**	5.6e^{-112}**	3.49e^{-14}**	0.00474*
Inactivo RD [3]	1.85e^{-78}**	2.2e^{-16}**	3.13e^{-56}**	1.9e^{-17}**	1.02e^{-44}**	3.05e^{-15}**
Accesible Marco [4]	0.00233*	1.59e^{-06}**	3.46e^{-12}**	2.06e^{-17}**	4.11e^{-43}**	4.18e^{-03}**
Accesible Origo [5]	0.01105*	1.6e^{-06}**	3.33e^{-04}**	**0.0854**	2.23e^{-07}**	**0.1722**
Activo S [6]	**0.1313**	0.0284*	4.55e^{-04}**	**0.8321**	6.96e^{-07}**	**0.0889**
Activo O/I [7]	--	--	--	--	--	--
Activo P [8]	0.02644*	1.7e^{-04}**	1.49e^{-08}**	9.05e^{-03}*	0.00618*	**0.3098**
Identificable baja [9]	0.00313*	1.72e^{-20}**	1.12e^{-05}**	3.26e^{-13}**	1.14e^{-30}**	8.97e^{-23}**
ESIN_R1						
No identificable [1]	2.32e^{-40}**	1.2e^{-138}**	4.66e^{-73}**	1.31e^{-43}**	4.4e^{-206}**	2.75e^{-43}**
Inactivo [2]	3.6e^{-147}**	8.28e^{-90}**	0**	1.34e^{-34}**	5.3e^{-131}**	1.75e^{-15}**
Accesible [3]	8.7e^{-04}**	1.53e^{-50}**	9.5e^{-14}**	3.7e^{-16}**	1.59e^{-70}**	1.79e^{-03}**
Activo [4]	4.96e^{-04}**	4.78e^{-33}**	5.62e^{-14}**	8.64e^{-09}**	2.74e^{-48}**	2.11e^{-09}**
ESIN_R2						
Nuevo [0]	2.32e^{-40}**	1.2e^{-138}**	4.66e^{-73}**	1.31e^{-43}**	4.4e^{-206}**	2.75e^{-43}**
Dado [1]	8.81e^{-76}**	1.1e^{-155}**	0**	1.2e^{-51}**	1.4e^{-221}**	1.52e^{-17}**
ESIN_R3						
Fuera de texto [0]	3.9e^{-45}**	4.5e^{-183}**	4.99e^{-95}**	6.69e^{-51}**	1e^{-260}**	1.72e^{-60}**
Activo por texto [1]	4.76e^{-78}**	5e^{-104}**	0**	6.72e^{-60}**	2.3e^{-177}*	7.57e^{-26}**

En todas las distribuciones anteriores tenemos el problema de contar con muy pocas observaciones. Si recuperamos los resultados previos, sobre el ANOVA y las regresiones, el etiquetado de estas categorías no muestra un desempeño sobresaliente; de hecho, en las etiquetas de la rama Activo de ESIN, los modelos de regresión fallaron completamente. Debido a esto, no pienso que sea prudente determinar que estas etiquetas puedan ser evaluadas con herramientas paramétricas. Sin embargo, me abstendré a realizar una apreciación contundente, por lo menos hasta comprar estos resultados con las pruebas no paramétricas.

Debido a los picos observados en los extremos, realicé la prueba retirando los ceros y los unos. Los resultados son similares a los anteriores (Tabla 57). Se repiten los casos de los análisis anteriores y surge que Inactivo MLP [2] y Accesible por Origo [5] también muestran una distribución normal. Los histogramas de ambos casos los muestro a continuación (Figura 42).

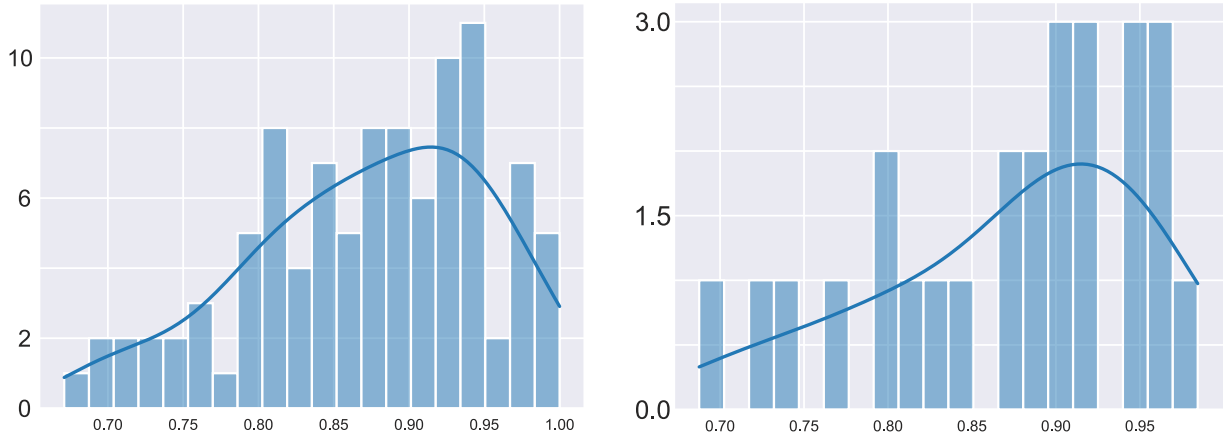


Figura 42. Histogramas de Inactivo MLP [2] (izquierda) y Accesible Origo [5] (derecha) de ESIN en MLVN

Podemos observar que, en efecto, en ambos casos tenemos distribuciones que parecerían provenir de una distribución normal. Dado los resultados de las secciones anteriores, Accesible Origo [5] es poso confiable debido a la poca cantidad de observaciones, pero no es caso para Inactivo MLP [2], el cual tiene 113 observaciones (99 si retiramos los unos y ceros). Sin embargo, esto sucede para la medida LSA a partir de la Ventana-*n*, medida que en general ha dado muy malos resultados.

Tabla 57. Resultados de la prueba de normalidad K^2 de D'Agostino para todas las etiquetas de las agrupaciones ESIN sin incluir las observaciones de ceros y unos

Etiquetas	**Valor *p* de las medidas**					
ESIN	**MLIN**	**MLVN**	**MLWD**	**MSIN**	**MSVN**	**MSWD**
No identificable baja [0]	1.16e^{-09}**	6.92e^{-45}**	2.85e^{-13}**	3.18e^{-05}**	1.48e^{-69}**	0.017425*
No Identificable [1]	8.92e^{-13}**	2.58e^{-61}**	4.33e^{-48}**	5.13e^{-55}**	2.3e^{-179}**	1.32e^{-20}**
Inactivo MLP [2]	4.98e^{-08}**	**0.085207**	1.48e^{-16}**	0	6.17e^{-17}**	0.03807*
Inactivo RD [3]	2.37e^{-46}**	2.2e^{-16}**	6.6e^{-117}**	1.9e^{-17}**	1.02e^{-44}**	3.05e^{-15}**
Accesible Marco [4]	0.00421*	1.15e^{-06}**	3.46e^{-12}**	2.06e^{-17}**	4.11e^{-43}**	0.00418*
Accesible Origo [5]	0.018006*	**0.224929**	2.67^{-04}**	**0.181264**	3.68^{-04}**	**0.38624**
Activo S [6]	**0.26066**	0.028424*	4.55^{-04}**	**0.832169**	6.96e^{-07}**	**0.08892**
Activo O/I [7]	--	--	--	--	--	--
Activo P [8]	0.06676	0.000177**	1.31e^{-06}**	0.009057*	0.006184*	**0.309819**
Identificable baja [9]	0.031876*	2.46e^{-06}**	4.55e^{-05}**	2.91e^{-06}**	1.02e^{-43}**	1.73e^{-06}**
ESIN_R1						
No identificable [1]	1.29e^{-37}**	3.07e^{-92}**	9.64e^{-67}**	2.64e^{-58}**	1.5e^{-273}**	2.06e^{-20}**
Inactivo [2]	7.35e^{-43}**	1.32e^{-18}**	0**	2.82e^{-40}**	8.3e^{-129}**	1.79e^{-17}**
Accesible [3]	0.001193*	7.74e^{-07}**	1.05e^{-13}**	1.21e^{-16}**	7.39e^{-43}**	0.002893*
Activo [4]	0.008165*	4.77e^{-11}**	1.69e^{-11}**	2.4e^{-11}**	1.23e^{-57}**	3.36e^{-05}**
ESIN_R2						
Nuevo [0]	1.29e^{-37}**	3.07e^{-92}**	9.64e^{-67}**	2.64e^{-58}**	1.5e^{-273}**	2.06e^{-20}**
Dado [1]	4.49e^{-23}**	4.1e^{-41}**	0**	9.11e^{-62}**	1.6e^{-210}**	5.8e^{-18}**
ESIN_R3						
Fuera de texto [0]	1.16e^{-40}**	2.8e^{-122}**	4.79e^{-87}**	1.44e^{-65}**	0**	6.38e^{-26}**
Activo por texto [1]	3.11e^{-25}**	8.66e^{-28}**	0**	4.04e^{-66}**	3.1e^{-182}**	3.76e^{-27}**

En general, las cantidades de observaciones entre los dos análisis (con ceros/unos y sin ellos) no varía mucho. Para comparación, muestro en la Tabla 58 las cantidades después de la eliminación. En la última columna está el tamaño original de cada muestra (descrito en §3.1). Se puede notar que los casos sospechosos de tener una distribución normal son aquellos que rozan las 30 observaciones. Es por esta razón que llama la atención el caso de Inactivo MLP [2] de ESIN, pero como mencioné, lo descarto por el desempeño general de MLVN.

Tabla 58. Tamaño de las muestras originales sin incluir las observaciones de ceros y unos

Etiquetas	**Tamaño de las muestras**						
ESIN	**MLIN**	**MLVN**	**MLWD**	**MSIN**	**MSVN**	**MSWD**	**Original**
No identificable baja [0]	196	204	192	204	204	204	212
No Identificable [1]	1111	1112	1111	1115	1115	1114	1195
Inactivo MLP [2]	100	**99**	99	100	100	100	113
Inactivo RD [3]	351	406	362	406	406	406	406
Accesible Marco [4]	184	184	185	185	185	185	185
Accesible Origo [5]	26	**26**	25	**26**	26	**26**	29
Activo S [6]	**29**	31	31	**31**	31	**31**	31
Activo O/I [7]	6	7	6	7	7	7	7
Activo P [8]	28	29	26	29	29	**29**	29
Identificable baja [9]	163	168	166	168	168	168	181
ESIN_R1							
No identificable [1]	1307	1316	1303	1319	1319	1318	1407
Inactivo [2]	451	505	461	506	506	506	519
Accesible [3]	210	210	210	211	211	211	214
Activo [4]	226	235	229	235	235	235	248
ESIN_R2							
Nuevo [0]	1307	1316	1303	1319	1319	1318	1407
Dado [1]	887	950	900	952	952	952	981
ESIN_R3							
Fuera de texto [0]	1617	1625	1612	1630	1630	1629	1734
Activo por texto [1]	577	641	591	641	641	641	654

Lo anterior señala que los casos Activos de ESIN tiene un comportamiento poco regular, con pocas observaciones, lo que los vuelve el conjunto de etiquetas con menor capacidad de ser

predichas con tan solo las medidas. Por otro lado, estos resultados muestran que no hay un impacto significativo en la gran mayoría de las distribuciones si retiramos los ceros y unos. Realizar un experimento con las muestras modificadas para los clasificadores, infiero, también sería irrelevante.

Si nos centramos en sólo aquellos casos en que se ha observado que las medidas son buenos predictores, tampoco encontramos grandes cambios. En particular, si recuperamos lo presentado en las regresiones, observamos que no existe un cambio en las muestras usadas para ESIN_R3, que fue el modelo con la mayor Exactitud. Así mismo, no encontramos variación para las distribuciones de las ramas No identificable de las distintas agrupaciones. En lo que respecta a Inactivo MLP [2] ESIN en MLVN, esta medida utilizando la ventana exterior, de hecho, no resultó buena predictora.

Lo anterior es señal que frases nominales analizadas con SPAN y LSA, agrupadas con estas medidas y las respectivas etiquetas de Estados Informativos, no tienen una distribución normal. Aunque se podrían utilizar métodos para sobrellevar este problema, como el *bootstrapping* integrado en los bosques aleatorios, es preciso señalar este comportamiento en esta clase de datos. Debido a esto, en la siguiente sección realizo pruebas no paramétricas para evaluar la pertinencia de la división entre las muestras.

3.6 Análisis de varianza: Kruskal-Wallis y Conover-Iman

Las distintas gráficas de cajas y bigotes que presenté en §3.3 siguen siendo una buena primera aproximación para evaluar si las medidas agrupan las etiquetas de manera que se distingan entre ellas. Ya vimos que existen algunas que parecen diferenciarse mejor de las otras

etiquetas. Los resultados de ANOVA señalaron que en todas las medidas se descarta la hipótesis nula dadas las agrupaciones ESIN. Para la siguiente prueba no paramétrica, Kruskal-Wallis, se sigue la misma idea: la hipótesis nula sostiene que todos los grupos provienen de la misma población y la hipótesis alternativa enuncia que por lo menos uno de los grupos es distinto. En la Tabla 59 muestro los valores *p* de la prueba.

Tabla 59. Prueba Kruskal-Wallis para todas las medidas y agrupaciones ESIN

Agrupaciones	**MLIN**	**MLVN**	**MLWD**	**MSIN**	**MSVN**	**MSWD**
ESIN	$8.4e^{-119}$**	$1.97e^{-25}$**	$4.8e^{-110}$**	$3.24e^{-87}$**	0.008141*	$2.6e^{-99}$**
ESIN_R1	$4.71e^{-52}$**	$6.42e^{-15}$**	$1.78e^{-49}$**	$2.26e^{-32}$**	0.038225*	$1.19e^{-38}$**
ESIN_R2	$9.44e^{-39}$**	$7.58e^{-16}$**	$1.78e^{-33}$**	$1.29e^{-23}$**	**0.733051**	$2.55e^{-26}$**
ESIN_R3	$2.67e^{-79}$**	$8.42e^{-10}$**	$8.38e^{-78}$**	$3.08e^{-54}$**	0.003698*	$3.9e^{-67}$**

Lo anterior es muy parecido a lo encontrado en §3.3. En ambos casos, el valor de significancia se cierra en MSVN, sólo que, para este caso en particular, ESIN_R2 es la única que no pasa la prueba: usar esa medida para agrupar lo Nuevo [0] y Dado [1] no resultó significativo. En el caso de ANOVA, dejé los resultados hasta la revisión general de las medidas. Para esta prueba no paramétrica ahondo un poco más, dada la exploración de la normalidad de cada distribución de la sección anterior. Reviso qué etiquetas se diferencian mejor de las otras para cada agrupación. Para ello, utilizo un método que describo en la siguiente sección.

3.6.1 Resultados de la prueba Conover-Iman

La prueba Conover-Iman[97] se realiza para todos los pares de etiquetas posibles dada cada agrupación ESIN. La prueba hace uso de la dominancia estocástica, lo que significa que uno de los dos grupos obtiene "mejores resultados" si se realizara una lotería con ellos. En el fondo, no nos interesa el que uno sea mejor o peor, sino que existe una diferencia en los resultados obtenidos entre los dos grupos, lo que ayuda a descartar la hipótesis nula la cual enuncia que con los dos grupos obtenemos los mismos resultados: no existe diferencia entre ellos.

Realizar Kruskal-Wallis de manera general, como he mostrado en la sección anterior, es requisito para este segundo paso en el procedimiento. Una vez que la hipótesis nula es rechazada de manera general, podemos otorgar relevancia a los resultados del método Conover-Iman. El siguiente paso consiste en revisar los pares en donde se rechace la hipótesis nula. Por lo que, si para la agrupación de etiquetas completa ESIN hablamos de 10 etiquetas a comparar entre ellas, tendríamos 100 resultados, de los cuales 10 ya sabemos que se trata de las comparaciones entre etiquetas iguales. De los 90 restantes, son relevantes 45 relaciones, ya que las otras 45 son las mismas etiquetas comparadas, pero en orden inverso. Esto lo menciono para dimensionar el problema de mostrar estos resultados. Debido a esto, no presentaré cada uno de los valores *p*, lo cuales serían 270 (45 por cada medida), sino que realizo un resumen general, en particular, para el caso de la agrupación ESIN, la cual es la

[97] https://scikit-posthocs.readthedocs.io/en/latest/generated/scikit_posthocs.posthoc_conover/

más compleja. Dejaré las matrices completas en el Anexo F por si el lector desea realizar una exploración más minuciosa de estos resultados.

En los resultados de las medidas en donde se utilizó la ventana exterior se observa la peor capacidad de lograr diferenciar las etiquetas. Aunque la prueba Kruskal-Wallis señala que sí existe por lo menos un grupo que tiene dominancia en MLVN y MSVN, gran parte de las etiquetas son indiferenciables. Para notar esto, construí el siguiente gráfico (Figura 43).

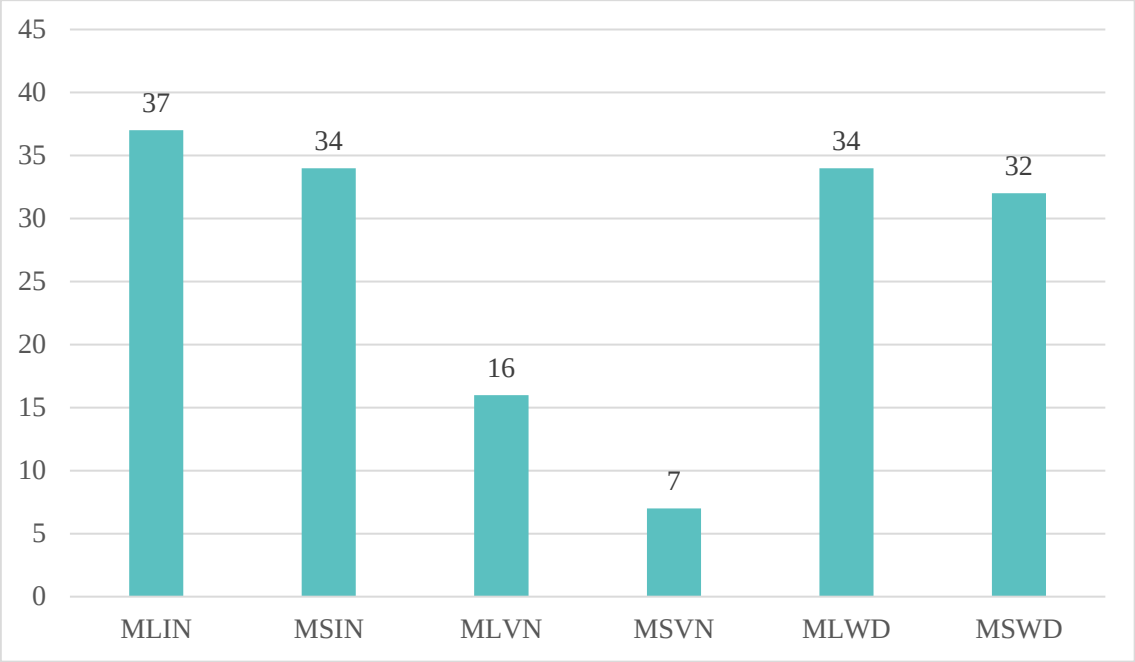


Figura 43. Cantidad de valores *p* por medida para ESIN

El número máximo de diferencias sería de 45 por medida, es decir, el escenario ideal en donde todos los valores *p* de cada par están por debajo del 0.05. Se puede observar que las que utilizaron las ventanas exteriores son las que tienen la menor cantidad de diferencias significativas. Por otro lado, las de la ventana interior parecen tener mejores resultados, con un pequeño decremento en las medidas que usan el diccionario (MLWD y MSWD). Aunque MLIN y MSIN tienen la mayor cantidad de valores *p* entre pares, se nota que en ningún caso se logra el máximo. En términos generales, todas las medidas fallan en diferenciar los pares

entre las etiquetas Activo S [6], Activo O/I [7] y Activo P [8]. Esto no resulta novedoso una vez que hemos observado los resultados anteriores, lo que coincide con el hecho de que forman el conjunto más pequeño de observaciones. Un aspecto interesante es que no falla en su diferenciación con otras etiquetas "no activas". De acuerdo con los resultados, aunque estas tres etiquetas son indiferenciables entre ellas, en relación con las otras etiquetas existe dominancia, lo que sugiere que las medidas son buenas para indicar qué frase nominal no corresponde a alguno de estos Estados Informativos. Dados los resultados para estas etiquetas hasta este momento, podemos adelantar que este comportamiento es coherente con lo propuesto por Kibrik (2011): es difícil encontrar que un dispositivo referencial pleno recupere un referente activo.

Por otro lado, MLIN en la agrupación ESIN logra diferenciar las etiquetas Inactivo MLP [2] y Accesible por Marco [4]. En las otras medidas, MSIN, MSWD y MLWD, Inactivo MLP [2] es la única etiqueta que se diferencia por completo de las demás. Otro patrón común que vemos casi en todas las medidas es la poca dominancia entre No identificable Baja [0] y No identificable [1], lo que sugiere que forman parte de un mismo grupo. No obstante, son precisamente estas etiquetas junto con Inactivo RD [3] las siguientes mejor diferenciadas. Un resumen de la relación de la cantidad de valores *p* por cada etiqueta dada la medida se puede consultar en la Tabla 60. Considérese que el valor máximo en cada celda es 9, y el total máximo de cada fila es 54. Una etiqueta con 9 en todas las celdas implicaría que es capaz de diferenciarse de todas las etiquetas dadas todas las medidas.

Tabla 60. Cantidad de valores *p* para cada etiqueta dada las medidas de ESIN

Etiqueta	MLIN	MSIN	MLVN	MSVN	MLWD	MSWD	Total
No identificable Baja [0]	8	6	3	1	6	5	29
No identificable [1]	7	6	7	1	6	5	32
Inactivo MLP [2]	9	9	5	2	9	9	**43**
Inactivo RD [3]	8	8	3	5	8	8	**40**
Accesible Marco [4]	9	8	2	1	8	6	34
Accesible Origo [5]	6	6	**0**	2	6	6	26
Activo S [6]	7	6	3	**0**	7	7	30
Activo O/I [7]	6	6	1	**0**	6	6	25
Activo P [8]	7	7	3	**0**	7	7	31
Identificable Baja [9]	7	6	5	2	5	5	30

Lo anterior confirma algunos puntos que ya se venían observando en lo que va de los resultados:

- MLIN es la mejor medida para diferenciar, colocando a LSA por encima de SPAN.
- La ventana interior es mejor que la exterior.
- El diccionario no aporta mayor capacidad de diferenciar entre las etiquetas.

La situación en la agrupación ESIN_R1 es aún más contrastante. La cantidad máxima de pares es de seis por cada medida, lo que da como resultado un total de 36 valores *p* a explorar. Las matrices pueden ser consultadas en el Anexo F. En términos generales, se observa el mismo patrón que ESIN: la etiqueta Accesible [3] y Activo [4] vuelven a ser los menos diferenciados, es el caso en MLIN y MSIN, en donde es el único par en donde no encuentra diferencia. En la misma línea, MLWD sólo falla en diferenciar un solo par, el de No identificable [1] con Accesible [3]. MSWD falla en este mismo par, y en el par Accesible [3] con Activo [4]. La peor medida fue MSVN en la cual sólo se logran diferenciar dos pares.

Le sigue MLVN que alcanza tres pares. La distribución del total de valores *p* por medida puede observarse en la Figura 44.

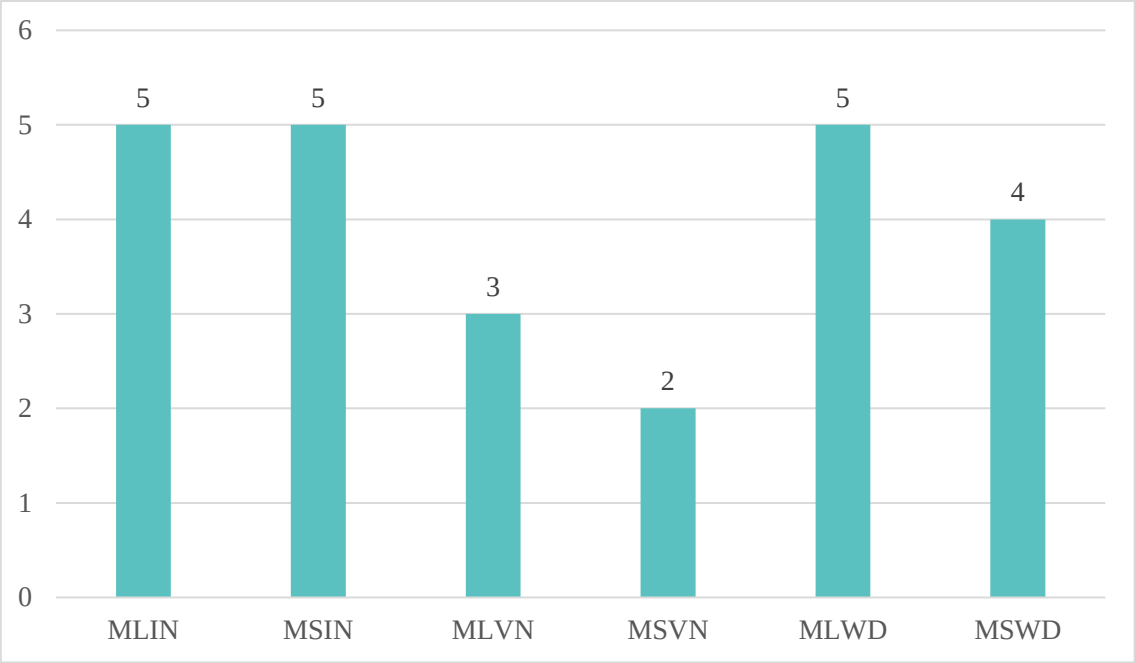


Figura 44. Cantidad de valores *p* por medida para ESIN_R1

En la Tabla muestro un resumen parecido al que mostré para la agrupación ESIN. Para el caso de las etiquetas, el máximo ideal sería un total de 18, en la que la etiqueta ha sido diferenciada de todas las otras etiquetas y en todas las medidas. Podemos observar que ninguna llega a este total, aunque para esta agrupación, la etiqueta Accesible [3] y Activo [4] no tienen un comportamiento tan distinto de las otras dos etiquetas. Esto podría ser problemático: hemos visto hasta este momento que las medidas no son buenas para todas las etiquetas, por lo que ya podemos esperar un contraste fuerte entre ellas. Parece que con esta agrupación se pierde ese contraste, por lo que tal vez la misma efectividad de las etiquetas dentro de No identificable [1] e Inactivo [2] se pierde. No obstante, las regresiones nos mostraron que esta agrupación aumenta su Exactitud con respecto a la agrupación completa ESIN. Dados estos nuevos resultados, podemos asumir que sí es un aumento en la efectividad

para diferenciar las otras etiquetas, tal vez en el mismo sentido en que se ha manejado: las medidas permiten distinguir qué no es activo.

Tabla 61. Cantidad de valores *p* para cada etiqueta dada las medidas de ESIN_R1

Etiqueta	MLIN	MSIN	MLVN	MSVN	MLWD	MSWD	Total
No identificable [1]	3	3	3	1	2	2	**14**
Inactivo [2]	3	3	1	1	3	3	**14**
Accesible [3]	2	2	1	2	2	1	10
Activo [4]	2	2	1	**0**	3	2	10

Lo anterior confirma lo que ya habíamos observado con el análisis de varianza (ANOVA), a pesar de tratarse de datos no paramétricos. Se observa que las medidas son buenas para encontrar relaciones entre frases nominales en el texto, principalmente los Estados Informativos relacionados con lo Inactivo. Por otro lado, también son buenas para saber qué no está en el texto, con las etiquetas No identificable, e incluso, en aquellos casos en donde las frases nominales son poco probables de figurar un referente que se mantenga en el texto. Pero en donde se observa mayor falla y poca capacidad de diferencia, es para las etiquetas relacionadas con estados de activación y, de manera marginal, los accesibles, en la manera en que los he entendido en esta investigación. Ya adelantaba en el marco teórico que los accesibles iban a ser difíciles de detectar por el algoritmo. Esto, debido al tipo de conocimiento y la capacidad de inferir las relaciones semánticas pertinentes —aspecto que, para tratar de solucionarlo, está muy por encima de los recursos disponibles en esta tesis— integrar las acepciones del DEM era una manera de hacerle frente a este problema, pero la manera en que se ha utilizado este recurso en este trabajo no mostró diferencia. Sobre este inventario de sentidos, los resultados muestran que las medidas aprovechan poco a nada al DEM, aunque, tampoco agregan ruido, lo cual también es destacable. Todo esto, deja ver un

patrón general que resulta interesante: las medidas son buenas a un nivel pragmático textual, pero pésimas para un nivel sintáctico inter-oracional, lo cual explicaría su mala capacidad de detectar frases nominales cuya estructura exprese la suposición de un estado activo del referente. Por lo que, si un referente es mencionado a través de una frase nominal en una oración, y luego recuperado en la siguiente oración, establecer una relación entre estas dos frases nominales escapa a las capacidades de las medidas.

Las pruebas Kruskal-Wallis y Conover-Iman no difieren mucho de lo encontrado con ANOVA, por lo que podríamos esperar que las regresiones propuestas sigan teniendo validez. Para corroborar esto y como último paso de este capítulo, muestro los resultados del clasificador a partir de bosques aleatorios, el cual permite utilizar datos que no cumplen con criterios paramétricos.

3.7 Bosques aleatorios

Si la regresión nos mostró resultados prometedores para algunas de las medidas dadas las etiquetas, veamos cómo resulta un método que no necesita de una distribución normal de los datos. Este método es el bosque aleatorio (*random forest*). Para la construcción de los modelos establecí tres escenarios: en el primero incluyo todos los trece predictores; en el segundo, incluyo sólo las seis medidas; y en el tercero incluyo sólo aquellos predictores que tuvieron una correlación destacable (tal y como mostré en §3.2).

En la ejecución del algoritmo de Python, existen dos puntos relevantes antes de correr los modelos para bosques aleatorios. Primero, se suelen modificar parámetros previos antes de ejecutar el modelo, los cuales son llamados hiperparámetros. Segundo, se puede determinar

un conjunto de hiperparámetros para que el algoritmo los combine y decida, entre las combinaciones, cuál logró el menor error. Además de esta selección, también se debe tener en cuenta la manera de evaluar y retroalimentar al modelo. Como ya mencioné, esta técnica hace uso de *Out-of-bag Error* (*OOB Error*), que corresponde a utilizar aquellos datos que quedaron fuera en la selección realizada para el entrenamiento. A diferencia de la regresión, no se determina el tamaño del conjunto de prueba y de entrenamiento por adelantado, sino que, por *bootstrapping*, se construye el conjunto de entrenamiento y aquello que queda "fuera de la bolsa" es usado para la evaluación. Ese error es el que ayuda a determinar cuál modelo tiene la mejor capacidad de predicción.

En cuanto a los hiperparámetros, se suelen modificar los siguientes (disponibles en la documentación de *RandomForestClassifier* de *sklearn*):

- **n_estimators:** el número de árboles de decisiones generados.
- **max_features:** el número de predictores observados al momento de hacer la división en el árbol.
- **max_depth:** la profundidad del árbol de decisiones.
- **criterion:** la función utilizada para medir la pureza de las divisiones en un árbol.

Para esta investigación, utilicé los siguientes conjuntos de hiperparámetros:

n_estimators: 150. Por defecto se suele utilizar 100. La ventaja con esta técnica es que no ocurre un sobreajuste al momento de agregar más árboles, sencillamente deja de crecer la capacidad predictora y el recurso computacional de los subsecuentes árboles es desaprovechado.

max_features: todas/medidas/restringidos. Este parámetro se ve directamente modificado por la cantidad de predictores que evalúo desde un inicio: **Todos**, que se refiere a los 13 predictores, en donde los CS se dividen en seis más —para convertirse en dicotómicos— lo que produce un conjunto final de 18. Sólo en este caso se modifica, además, el hiperparámetro para variar las cantidades de predictores en el modelo. Estas variaciones son 5, 7 ,9 y 18; **Medidas**, que corresponde a las seis medidas ya descritas antes; **Restringidos**, que se refiere a los predictores con mejor correlación. Para los predictores de CS, se incluyen todas las demás categorías no sólo la que haya resultado tener mejor correlación.

max_depth: Sin restricción (se expande tanto como necesite para clasificar todos los datos hasta que por lo menos en cada *hoja* haya 2 observaciones), 3, 10 y 20.

criterion: 'gini' y 'entropy'[98].

Para cada uno de los modelos, reporto la Exactitud y la Métrica $F_{\beta=1}$, en particular, para esta última, reporto tanto la métrica general como para la de cada etiqueta individual. Como es de suponerse, los resultados de la agrupación ESIN son los más complejos. También incluyo al inicio de cada sección los hiperparámetros finales. Para mantener el contraste entre los modelos, se observará que repito la Exactitud en cada tabla.

3.7.1 Bosques para ESIN

[98] Existen distintos criterios para establecer la división de un nodo en un árbol de decisiones. No daré detalles sobre esto, entre otras razones, debido a que me limito a lo que la función del algoritmo puede otorgarme, pero para mayor detalle sobre la matemática detrás de cada uno de estos criterios (o la examinación de otros criterios que no están considerados) remito al lector a Rokach y Maimon (2005).

Tabla 62. Hiperparámetros para el modelo ESIN

Predictores	Todos	Medidas	Restringidos
Exactitud	63%	56%	62%
Hiperparámetros			
n_estimators	150		
max_features	18	6	5
max_depth	10	10	10
criterion	gini	gini	entropy

Al igual que con las regresiones, para los modelos que tratan de identificar las 10 etiquetas de la primera agrupación ESIN no se supera un 65% de exactitud. Llama la atención que para lograr algún tipo de diferencia se utilicen todos los 18 predictores; el que le estén disponibles los predictores no significa que necesite usarlos. Esto contrastará con los siguientes modelos.

Tabla 63. Métrica F_β y Exactitud de los modelos que clasifican las etiquetas en ESIN

	Todos	**Medidas**	**Restringidos**	**C. P.**
Exactitud	63%	56%	62%	
No identificable Baja [0]	0.53	0.21	0.5	45
No identificable [1]	**0.79**	**0.69**	**0.79**	290
Inactivo por MLP [2]	0.38	0.22	0.46	30
Inactivo por RD [3]	**0.58**	**0.56**	**0.55**	113
Accesible por Marco [4]	0.39	0.09	0.32	52
Accesible por Origo [5]	0	0	0	10
Activo S [6]	0	0	0	3
Activo O/I [7]	0	0	0	3
Activo P [8]	0	0	0	5
Identificable Baja [9]	0.24	0.04	0.13	46
Promedio (Macro)	0.29	0.18	0.27	597
Promedio (Ponderado)	0.6	0.48	0.59	597

Para los datos presentados en la Tabla 63 conviene notar la diferencia entre el promedio visto en general (Macro) y el ponderado. Debido al desequilibrio que existe entre la cantidad de observaciones de cada etiqueta, el promedio ponderado nos sirve para contrastar. La gran mayoría de los datos pertenecen al No identificable [1], el cual logra tener una Métrica F_β de 0.69-0.79 en todos los modelos; es la etiqueta con la métrica más alta. El hecho de que tengamos 290 observaciones para este caso incrementa el promedio ponderado: aquellas etiquetas con pocas observaciones tienen una menor influencia en esta medida. El problema sobre el desequilibrio de las observaciones es algo que ya he señalado antes y que, me parece, por la naturaleza del fenómeno estudiado, no puede solucionarse con tan solo buscar obtener la misma cantidad de observaciones para cada grupo. De nuevo, las etiquetas de la rama Activo y el Accesible por Origo [5] no pueden predecirse usando los modelos propuestos: ni usando todos los predictores ni usando las medidas se obtiene algún tipo de resultado en este clasificador. En la columna de C.P. de la Tabla 63 podemos observar, además, la ínfima cantidad de observaciones de estas etiquetas en el conjunto de prueba, lo que dificulta la tarea.

Lo anterior corrobora también lo que los resultados de las pruebas paramétricas apuntaban: la etiqueta No identificable [1] y la Inactivo por Registro Discursivo [3] son las mejor predichas. El contraste más alto lo vemos al usar las medidas como los predictores, en donde notamos que no funcionan por si solas, aunque siguen siendo buenas para asociarse con la etiqueta [1] y [3], en especial con esta última en donde su Métrica F_β apenas decrece dos puntos con respecto al modelo que usa todos los predictores.

3.7.2 Bosques para ESIN_R1

Tabla 64. Hiperparámetros para el modelo ESIN_R1

Predictores	Todos	Medidas	Restringidos
Exactitud	68%	63%	60%
Hiperparámetros			
n_estimators	150		
max_features	5	6	1
max_depth	20	3	3
criterion	entropy	gini	entropy

Para el caso de la agrupación ESIN_R1, el modelo de clasificación obtuvo una mejoría en comparación al de las 10 etiquetas de la agrupación anterior. En este caso, además, la Exactitud aumenta para el modelo que usa las medidas (63%) y deja atrás al escenario de factores restringidos (60%) que, como recordaremos, sólo usa la medida LSA Interior-*w* DEM (MLWD). Al igual que con las regresiones, se vuelve a obtener un buen resultado para No identificable [1] e Inactivo [2], pero vuelve a fallar para Accesible [3] y Activo [4] (Tabla 65).

Tabla 65. Métrica F_β y Exactitud de los modelos que clasifican las etiquetas en ESIN_R1

	Todos	**Medidas**	**Restringidos**	**C. P.**
Exactitud	68%	63%	60%	
No identificable [1]	**0.81**	**0.75**	**0.74**	335
Inactivo [2]	0.60	0.51	0.40	143
Accesible [3]	0.20	0	0	62
Activo [4]	0.16	0	0	57
Promedio (Macro)	0.44	0.31	0.28	597
Promedio (Ponderado)	0.63	0.54	0.51	597

Notamos que el modelo que utiliza todos los predictores es capaz de predecir algunos casos de Accesible [3] y Activo[4]. Sabemos que el problema se encuentra con Accesible por Origo [5] y distintos tipos de Activo (dado lo visto en la Tabla anterior). Ninguna de estas métricas supera a las obtenidas de manera individual con Accesible por Marco [4] e Identificable Baja [9], lo que ayuda a sostener que la división es pertinente y nos encontramos a un fenómeno que si pueda ser atendido, aunque no por las medidas.

3.7.3 Bosques para ESIN_R2

Tabla 66. Hiperparámetros para el modelo ESIN_R2

Predictores	Todos	Medidas	Restringidos
Exactitud	77%	66%	67%
Hiperparámetros			
n_estimators	150		
max_features	9	6	1
max_depth	10	10	3
criterion	gini	entropy	gini

Para el caso de la agrupación que divide las etiquetas entre lo Nuevo [0] y Dado [1] —a manera de seguir el procedimiento del antecedente— obtenemos que la mayor Exactitud la tiene el modelo que utiliza todos los predictores (77%). Las medidas en general se quedan atrás (66%), aunque la medida LSA Interior-w (MLIN) —que es el único predictor en el modelo restringido— resulta tener un desempeño casi similar (67%). Si observamos los resultados generales (Tabla 67) podemos notar que al usar todos los predictores, lo Dado [1] (si lo entendemos como la suma de las etiquetas Activo/Inactivo/Accesible) obtienen una buena métrica en comparación con los casos anteriores. Lo Nuevo [0] se mantiene por debajo del 0.80, siendo su menor desempeño en el modelo que utiliza todas las medidas.

Tabla 67. Métrica F_β y Exactitud de los modelos que clasifican las etiquetas en ESIN_R2

	Todos	**Medidas**	**Restringidos**	**C. P.**
Exactitud	77%	66%	67%	
Nuevo [0]	0.80	0.74	0.76	335
Dado [1]	0.71	0.51	0.49	262
Promedio (Macro)	0.77	0.63	0.62	597
Promedio (Ponderado)	0.77	0.64	0.64	597

Además de que ninguno de estos modelos supera al antecedente (que llega a 80%), nos permite notar, en contraste con los casos anteriores, la pertinencia de reunir en la etiqueta Dado [1] los casos que tienen algún tipo de predicción con aquellos que, en absoluto, son identificados por un clasificador de este tipo. No obstante, al ser tan pocas observaciones, es probable que los modelos que usan la agrupación ESIN_R2 les reste peso a las observaciones extrañas, como los que pertenecen a las etiquetas del grupo Activos. Observamos para estos modelos, así como para lo que veremos de la agrupación ESIN_R3, que los promedios Macro y Ponderado empiezan a igualarse: esto se debe al escenario ideal en donde la cantidad de observaciones son equilibradas entre las etiquetas.

3.7.4 Bosques para ESIN_R3

De la misma manera que en los casos anteriores y en comparación con la regresión, al dividir los datos con la agrupación ESIN_R3 se obtienen los mejores resultados. El modelo que utiliza todos los predictores obtiene un 81% de Exactitud, lo que supera al antecedente. Pero, además, el modelo que utiliza sólo las medidas alcanza un 80%, igual que el antecedente. Esto es alentador en cuanto a la cantidad de tratamiento previo que necesita un texto para poder otorgarnos un contraste entre “fuera de texto” y “en texto” significativo. Sobre esto, se

debe mencionar que estos modelos tienen una Métrica F_β inferior para la etiqueta Activo por texto [1] (0.62) en contraste con lo Dado [1] en ESIN_R2 (0.71). Por otro lado, como se observa en la Tabla 69, es Fuera de texto [0] la etiqueta con la mejor Métrica F_β de todos los casos anteriores, para aquellos modelos en donde se usan todos los predictores o solo las medidas (0.87).

Tabla 68. Hiperparámetros para el modelo ESIN_R3

Predictores	Todos	Medidas	Restringidos
Exactitud	81%	80%	78%
Hiperparámetros			
n_estimators	150		
max_features	18	6	1
max_depth	s/n	3	10
criterion	gini	entropy	entropy

Lo anterior puede interpretarse como que las medidas sí logran capturar información para distinguir qué entidades no forman parte del texto, pero en ellas no se captura si la entidad ya fue mencionada antes directamente, inferible a través del marco establecido en la nota, o los casos en donde tenemos aposiciones.

Tabla 69. Métrica F_β y Exactitud de los modelos que clasifican las etiquetas en ESIN_R3

	Todos	**Medidas**	**Restringidos**	**C. P.**
Exactitud	81%	80%	78%	
Fuera de texto [0]	0.87	0.87	0.87	427
Activo por texto [1]	0.62	0.56	0.57	170
Promedio (Macro)	0.75	0.72	0.72	597
Promedio (Ponderado)	0.80	0.78	0.79	597

Hasta este punto, se observa que tanto las pruebas paramétricas como las no paramétricas muestran resultados muy similares. En ambos casos, los análisis de varianza son

significativos. Los clasificadores tienen un resultado muy similar, en particular para ESIN_R3.

3.8 Discusión y cierre

Una vez probado que las medidas son relevantes para las distintas agrupaciones y que, además, muestran resultados que en algunos casos superan a los trabajos antecedentes, queda comparar tanto los resultados obtenidos por métodos paramétricos como no paramétricos. Por un lado, para los clasificadores que utilizan las 10 etiquetas ESIN, observamos lo siguiente (Figura 45): la regresión con todos los predictores alcanzó un 63% de Exactitud, calificación que comparte con el clasificador de bosques aleatorios que utiliza todos los predictores.

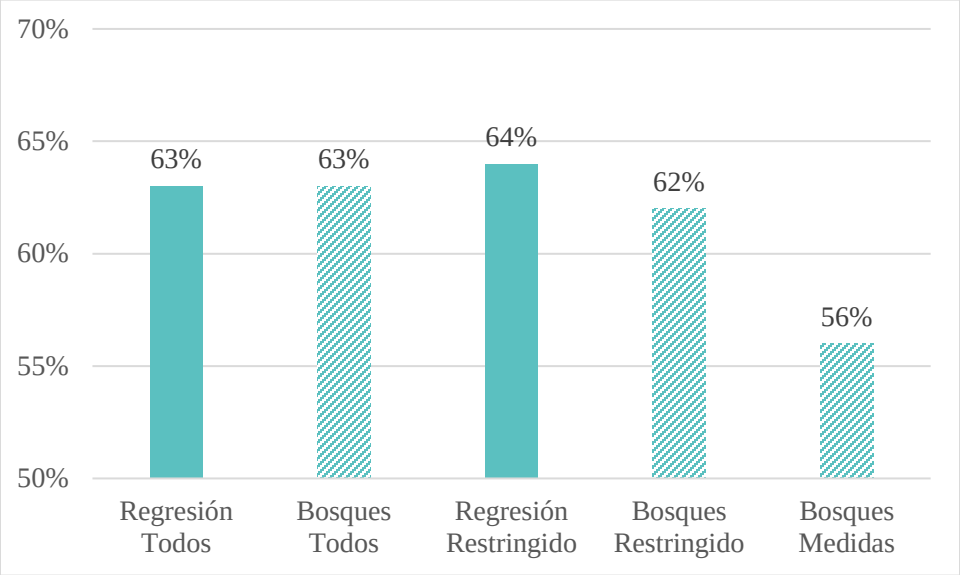


Figura 45. Comparación entre los distintos modelos para clasificación de ESIN

La línea base para los otros modelos de esta agrupación sería el 56% que se alcanza al utilizar sólo las medidas para el clasificador de bosques aleatorios, la cual es la Exactitud más baja de todos los experimentos que se reportan. La más alta Exactitud la alcanzó la regresión

logística con predictores restringidos con 64%. Sin embargo, ninguno de estos clasificadores supera a los mejores clasificadores de las otras agrupaciones. Para la agrupación reducida ESIN_R1 (Figura 46) encontramos un contraste mayor entre los distintos modelos. De la misma manera que en ESIN, el clasificador de bosques aleatorios alcanza la Exactitud más alta con 68%. Las regresiones son las que tienen la Exactitud más baja, con 65%.

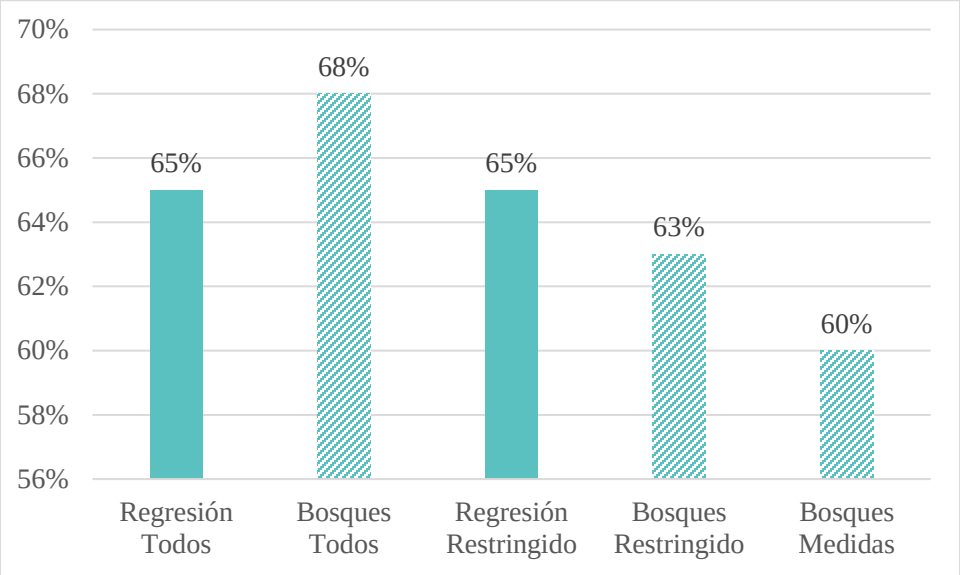


Figura 46. Comparación entre los distintos modelos para clasificación de ESIN_R1

Para la agrupación ESIN_R2 (Figura 47) observamos un incremento en la Exactitud para las regresiones, pero, en términos generales, el desempeño se mantiene casi igual que el clasificador de bosques aleatorios para ESIN_R1. Ambas coinciden en que la Exactitud más alta la alcanza este tipo de clasificador utilizando todos los predictores.

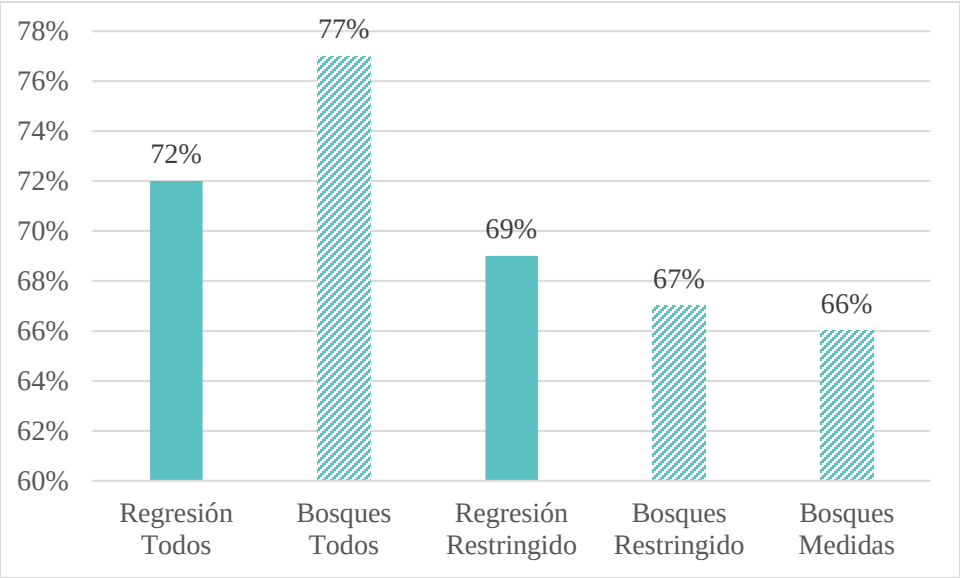


Figura 47. Comparación entre los distintos modelos para clasificación de ESIN_R2

Finalmente, como ya se venía observando en los resultados, agrupar las medidas con ESIN_R3 alcanza los mejores resultados en los clasificadores. Este es el único caso en donde no se observa una gran diferencia entre usar todos los predictores y sólo las medidas. De hecho, además de esto, también es el caso en donde las medidas alcanzan a reportar una Exactitud parecida a la de los antecedentes.

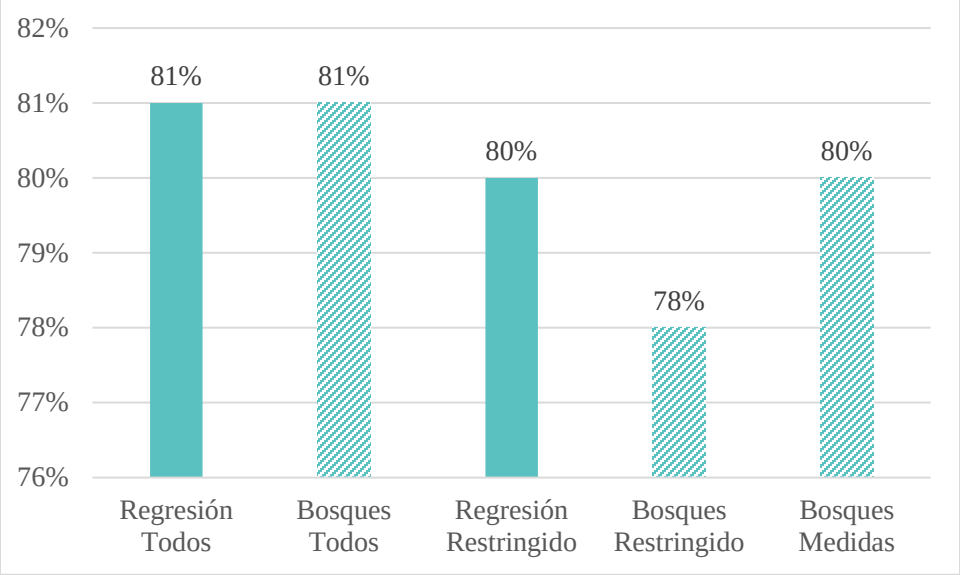


Figura 48. Comparación entre los distintos modelos para clasificación de ESIN_R3

De tal manera, los resultados los sintetizo de la siguiente manera:

- En términos generales, las medidas capturan la capacidad de distinguir qué frases nominales suponen que el oyente debe figurar un nuevo referente discursivo.
- LSA supera a SPAN, y usar la ventana interior supera a la ventana exterior. Las acepciones del Diccionario del español de México no mejoran notablemente el desempeño de los clasificadores, pero tampoco agregan ruido.
- Contrastar con los clasificadores que utilizan todos los predictores, además de establecer un punto de comparación, ha permitido hacer notar que existe información gramatical que se relaciona mejor para descartar qué no pertenece a las etiquetas relacionadas con Estados Informativos No identificable e Inactivo.
- Debido a los objetivos de la presente investigación, no ahondo en cómo los factores externos son relevantes para la clasificación y cómo podrían mejorarse, pero esto se podría atender en otras exploraciones.
- Las medidas no capturan las suposiciones sobre referentes activos a través de dispositivos referenciales plenos. Además, las observaciones de estas etiquetas son tan escasas que se vuelve difícil encontrar un patrón.
- De la primera agrupación ESIN, las medidas sólo funcionan para distinguir las etiquetas No identificable baja [0], No identificable [1], Inactivo MLP [2] e Inactivo RD [3]. Para la agrupación reducida ESIN_R1, las medidas logran diferenciar las frases etiquetadas con No identificable [1] e Inactivo [2]. En la segunda reducción ESIN_R2, dicotómica, gran parte de los aciertos los logra para identificar las frases nominales con la etiqueta Nuevo [0]. Lo Dado [1], aunque con mejores predicciones

que los casos anteriores, e incluso con un número de observaciones más equilibrada, ronda el 50% de Precisión. La tercera reducción ESIN_R3, también dicotómica, aunque no obedece estrictamente la clasificación de Estados Informativos que observamos al inicio, encuentra una mejor relación con las medidas. La etiqueta Fuera de texto [0] alcanzó una Precisión más alta que en la segunda reducción, pero la etiqueta Activo por texto [1] tiene una Precisión menor a su homólogo Dado [1].

- No es sorprendente que Accesible por Origo [5] sea, junto con los Activos, de las etiquetas peor identificadas. Las formas que requieren el Origo para recuperar un referente son escasas y dependientes de una estructura particular (demostrativo + nominal de tiempo, por ejemplo, o la primera/segunda persona, además de los demostrativos plenos que en las notas analizadas no son identificados). En cuanto a los activos, la teoría hace énfasis que, para un referente activo en la MT, un dispositivo referencial pleno es improbable. No obstante, se esperaba un desempeño más alto para Accesible por Marco [4], el cual no fue alcanzado, lo que refleja la necesidad de atender por separado el problema que plantea identificar este tipo de Estado Informativo.
- Es prometedor notar que existe un nivel (un conjunto de Estados Informativos) con los que las medidas se asocian mejor, lo que, en un sistema de clasificación completo de Estados Informativos, podría funcionar como un primer paso.

Para dimensionar el sentido de los hallazgos mostrados, permítaseme exponer dos escenarios extremos. Partiendo de lo expuesto en la introducción de esta investigación, dado un texto (T) constituido por una secuencia de frases nominales (FN), tratamos de relacionar una

medida obtenida a través de una función con un Estado Informativo de una FN en una posición particular de T.

Primer escenario: la información capturada es caótica[99]. La cantidad de palabras (e información) a la cual tiene acceso la función es tanta y tan dispar que cualquier FN en T es distinta de cualquier otra FN en T. Esto resulta en que las medidas no guardan un patrón coherente con los Estados Informativos, además, agregar más información empeora el escenario. Todos los cálculos entre las frases nominales darían lugar a medidas muy bajas de similitud (tanto el coseno como la proyección).

Segundo escenario: la información capturada es relevante en todos los aspectos. Debido a que una frase nominal comparte información con todas las frases nominales anteriores, ya sea porque la estructura lexicogramatical se repite (encontramos secuencias del tipo determinante + nominal + modificador, por ejemplo) o porque tenemos lemas repetidos (el paradigma de los definidos) una FN en T es vinculable con todas las frases nominales. Los cálculos de coseno y proyección resultan en que todas las frases nominales tienen la misma medida entre ellas.

En ambos escenarios, debido a que tenemos un mismo comportamiento para todas las frases nominales, no habría manera de usar las medidas para distinguir Estados Informativos —lo que constituiría la hipótesis nula en esta investigación—. Lo que podemos observar en los resultados es que no es el caso, existe diferencia. Hay frases cuya medida se vincula mejor con un Estado Informativo particular, por lo que, de acuerdo con lo visto en los antecedentes

[99] Me refiero por caos a la ausencia de orden y a un sistema que ha alcanzado la máxima indiferenciación.

de estas técnicas, LSA y SPAN sí capturan contrastes. De hecho, el utilizar la ventana exterior e interior de las frases para alimentar las funciones, inició como un intento de tener dos medidas que se pudieran complementar, pero que ha mostrado ser un buen ejemplo de qué es lo que sucede sí tuviéramos algún escenario extremo para todas las medidas: la ventana exterior falla en notar diferencia entre las frases debido a que la información otorgada es caótica. Esto se podría comprobar evaluando la entropía de las bolsas entre ellas, pero esto lo dejaré para otras exploraciones[100]. En esta misma línea, las acepciones del DEM podrían suponerse tan distintas al texto analizado que ocasionaría cualquiera de los dos escenarios extremos. Los resultados muestran que las acepciones del DEM no mejoran las relaciones, pero tampoco las entorpece, lo que podría interpretarse como que existe una resonancia entre los ítems léxicos y las acepciones: en donde dos ítems léxicos son distintos, la acepción vendría a redundar —pero no a amplificar— la diferencia. Esto tiene sentido con que, entre dos acepciones, sin tratamiento, no emerjan relaciones semánticas. Es probable que se necesite otra clase de procedimiento para que una etiqueta como Accesible por Marco pueda aprovechar la información que brinda el DEM.

Por lo pronto, a manera de cierre de esta sección y preámbulo del final de la investigación, puedo sostener que la representación vectorial de las frases nominales y el cálculo entre estos vectores captura distinciones pragmáticas, confirmando lo que los antecedentes habían encontrado para el inglés. La teoría expuesta para entender los Estados Informativos incluye, bajo la noción de *estados mentales de los referentes*, un conjunto de fenómenos que, en la práctica, podemos constatar, son tan distintos que implican distintos métodos para su

[100] Una posibilidad de este cálculo lo desarrollé en 2.3.4. Evaluar la pertinencia de esta manera de medir la entropía para las bolsas de palabras sería parte de un trabajo posterior.

resolución. Siguiendo el Axioma del Método Computacional, pareciera que los contrastes que emergen por la incapacidad de LSA y SPAN para reconocer ciertas etiquetas coinciden con la teoría: existen propiedades pragmáticas que, aunque nombradas bajo una misma idea, guardan relaciones distintas con el texto y con patrones lexicogramaticales. Por la naturaleza de la técnica usada, se observa el texto antecedente para evaluar el Estado Informativo de una frase nominal. Hay propiedades a las cuales les es irrelevante examinar el texto completo, como fueron las etiquetas Activo y Accesible por Origo. Para otras propiedades, no es el texto lo relevante sino las relaciones semánticas establecidas y reforzadas por el Marco global del documento. En ambos casos, una técnica como la utilizada no es pertinente. El estudio de la activación y la accesibilidad, en términos computacionales, se presenta como un ejercicio de por lo menos tres técnicas distintas, en donde una de ellas parece ser —de acuerdo con los resultados— LSA utilizando la bolsa interior de palabras e información lexicogramatical de la frase nominal.

Capítulo 4 Conclusiones

LSA y SPAN son métodos que transforman un texto en una representación vectorial para después obtener medidas de distancia entre fragmentos o entre textos mismos. En tecnologías del lenguaje, la utilidad de estas medidas se ha demostrado amplia. En particular, se han utilizado en trabajos que examinan su relación con información nueva o dada, para lo cual, se han vinculado con propuestas teóricas provenientes de la lingüística, pero no de manera determinante. No obstante, en este trabajo he dado un giro distinto. Como se ha observado, el planteamiento dio lugar a explorar la relación de estas medidas con categorías pragmáticas, sin ánimos de obedecer alguna aplicación tecnológica, aunque tampoco rechazo que exista esa posibilidad. No he tratado de construir el clasificador último de estas propiedades pragmáticas, sino de observar si existe alguna correlación entre las medidas obtenidas y los Estados Informativos propuestos. Sobre esto, cabría recordar el objetivo de la investigación:

(O_1) Identificar la propiedad pragmática del estado informativo en frases nominales a partir de herramientas de representación vectorial, para su etiquetado de manera semisupervisada.

La discusión permitió dar cuenta de lo complejo del concepto del Estado Informativo, el cual derivé de los estados de activación e identificabilidad. Siguiendo la misma lógica planteada en los trabajos antecedentes, partí de que estas propiedades pragmáticas deben tener un correlato formal. Además, está el problema de definir información nueva/dada en lingüística. Tomé la decisión de aproximarme a este problema con la intención de acotarlo, pero no de resolver. En esta tesis no se encuentra una versión unificada de lo nuevo/dado en lingüística, pero tal vez sí una versión operativa, lo que es coherente con el lineamiento que expuse al inicio de la investigación: el Axioma del Método Computacional.

Para tener una manera de aplicar estos conceptos a un algoritmo informático, decidí sólo concentrarme en las frases nominales y en el concepto de *estado mental de los referentes* que se articula en el nivel de la Estructura de la Información. Esto en principio resultó un problema porque no existe una serie de reglas claras que se le puedan dar a una máquina para identificar las frases nominales en español. Debido a esto, aquellas reglas, si se quisieran implementar, deberían ser de tipo probabilístico. Definirlas y evaluarlas era una tarea que merecía su propio espacio, tal vez otro trabajo doctoral, por lo que decidí etiquetarlas manualmente y seguir con el proceso. Además, planteé que las frases nominales, independientemente de su estructura morfosintáctica, son capaces de figurar referentes discursivos y no es hasta observar el texto, que se puede determinar si tal referente es recuperado (se habla de él) en el texto. Este trabajo no fue sobre referencia, pero es necesaria para analizar Estados Informativos. Entendiendo la referencia como un fenómeno discursivo, el contraste entre las frases nominales que son y no son referenciales no se determina a priori. Esto me permitió salvar otro obstáculo en el identificador automático: el algoritmo no tenía que determinar qué frase nominal era referencial para luego determinar su Estado Informativo. Evaluaría, de todas, sus medidas para determinar el Estado Informativo.

El marco teórico de Andrey Kibrik me permitió distinguir dos conjuntos de dispositivos referenciales: los reducidos y los plenos. Mucho trabajo existe sobre los reducidos y su ámbito de acción, siendo el estado de activación parte de lo estudiado. Pero no así para las relaciones entre dispositivos referenciales plenos, que, se asume, siempre introducirán referentes nuevos. Una vez delimitado que mi investigación sólo abarcaría dispositivos referenciales plenos, y las suposiciones sobre los Estados Informativos, le siguió la construcción de las representaciones vectoriales de estos dispositivos. Basándome en trabajos

anteriores, partí de los mismos supuestos: el vector de una frase nominal en un momento determinado en el texto se compara con los vectores de las frases nominales antecedentes, lo que, utilizando las medidas LSA y SPAN, debería permitir notar diferencias en el Estado Informativo. En función de estas ideas, planteé la primera hipótesis:

(H_1) El Estado Informativo corresponde a una medida que es resultado de aplicar una función a $\overrightarrow{FN}$, vector que se obtiene de una FN en T, su estructura morfosintáctica y un contexto.

La cual establece que existe una relación entre una frase nominal y un número, resultado de una función. Los detalles de esto ya se han dejado constatar a lo largo de esta investigación. Sumado a esta primera hipótesis, propuse una variación, en la cual integro acepciones del Diccionario del español de México, bajo el supuesto de que, las acepciones permitirían vincular frases nominales de manera más sencilla cuando sus relaciones se deben a la información léxica de las unidades que las forman. Esta hipótesis la enuncié de la siguiente manera:

(H_2) El Estado Informativo corresponde a una medida que es resultado de aplicar una función a $\overrightarrow{FN}$, vector que se obtiene de una FN en T, su estructura morfosintáctica, un contexto y **acepciones asociadas a las palabras contenidas en la FN.**

Hasta este punto, y dados los resultados mostrados en el Capítulo 3, la hipótesis nula se descarta: LSA y SPAN no son números indiferentes a los Estados Informativos. Pudimos ver que, aunque no todas las etiquetas de cada agrupación ESIN pudieron ser clasificadas, sí existe un conjunto que se asocia: el de los No Identificables. No obstante, para las acepciones, el escenario es distinto. Integrar el diccionario no resultó ni en una mejoría de los

clasificadores ni tampoco en una disminución de su desempeño. Esto, como ya lo mencioné, podría deberse a que el diccionario es redundante con la información presente en las frases nominales. En todo caso, para la segunda hipótesis, sólo se puede decir que las acepciones no modifican el resultado, por lo que la hipótesis nula se mantiene. Lo anterior me permite sostener que el objetivo de la investigación se cumplió, aunque de manera parcial. Pude identificar la propiedad pragmática del Estado Informativo en frases nominales a partir de herramientas de representación vectorial, aunque sólo para algunos casos: Inactivo y No Identificable. Por lo que las medidas pueden aportar al etiquetado de manera semisupervisada de estas propiedades. Queda pendiente para otros trabajos el evaluar cómo integrar este proceso a un etiquetador mayor, y a su vez, evaluar si las medidas, junto con otras técnicas, ayudan a la detección de lo nuevo/dado.

Como se puede observar en ambas hipótesis, se tuvo en cuenta integrar la estructura morfosintáctica y el contexto. Para ambos casos planteé dos bolsas de palabras: Interior-*w* y Ventana-*n*. Los resultados muestran que, aunque ambas bolsas no guardan correlación y podrían funcionar al mismo tiempo como predictores, la Ventana-*n* no tiene un impacto relevante en la clasificación. Quedará pendiente para otros trabajos realizar experimentos sólo con el contexto de la frase, además de aquellos que involucren probar distintas variaciones con el interior, como, por ejemplo, una versión sin información morfosintáctica. La función a la que apelan las hipótesis se refiere tanto al procedimiento para obtener el coseno más alto, a través del Análisis de Semántica Latente (LSA) y la variación realizada una vez factorizada la matriz, para luego realizar las proyecciones al hiperplano (SPAN). Los resultados muestran que LSA es la que obtuvo las mejores métricas para predecir algunos Estados Informativos.

La experimentación consistió en dos procesos: por un lado, contrastar con predictores que no fueran las medidas pero que fueran accesibles al tipo de corpus creado y relevantes en la discusión. De acuerdo con los antecedentes teóricos, y otros esfuerzos por determinar qué factores intervienen en los estados de activación, tomé en cuenta un grupo de 7 predictores extra. Los resultados muestran que las medidas por sí solas tienen un buen desempeño, pero en algunos casos, incluir estos factores, aumentan las métricas, por lo que queda pendiente demostrar para qué etiquetas les beneficia más esta clase información en los clasificadores, a veces más que las propias medidas. El segundo proceso en la experimentación consistió en crear subagrupaciones de las etiquetas de Estados Informativos originales. De tal manera, propuse tres versiones reducidas: una versión con cuatro etiquetas, y otras dos, dicotómicas. Los resultados mostraron que la Exactitud de los clasificadores aumenta en las reducciones. Esto es esperable, pero es interesante contrastar en particular las dos reducciones dicotómicas, además de notar la persistencia de algunas etiquetas de ser clasificadas. Por ejemplo, las etiquetas relacionadas con lo Activo difícilmente se pueden relacionar con las medidas. Pero, una vez reunidas con otras etiquetas, parecieran empezar a compartir patrones que les permiten ser vinculadas. No obstante, aquellas etiquetas como lo Dado [1] o lo que se encuentra Activo por texto [1] de las dos últimas reducciones no obtuvieron métricas tan altas como sus contrarias, lo Nuevo [0] y Fuera de texto [0] respectivamente, en las agrupaciones ESIN_R2 y ESIN_R3. Justamente en la tercera reducción es en donde encontramos la Exactitud más alta, superando a los trabajaos antecedentes que pretendían un objetivo similar al presentado en esta investigación.

Por último, los resultados también dejan ver que los Estados Informativos no son un fenómeno pragmático que pueda ser analizado con un sólo un método computacional. Las

suposiciones sobre identificabilidad y activación requieren distintas operaciones y métodos, en donde uno de ellos puede ser LSA. Recuérdese que el tipo de texto con el que trabajo son notas periodísticas, a veces con un número pequeño de frases nominales a comparar. Tal vez SPAN demuestre mayor puntaje en casos de textos extensos o modulando qué puede observar "del pasado" para crear el hiperplano y proyectar. Un identificador de todos los Estados Informativos propuestos probablemente requerirá no sólo estas técnicas sino también aquellas que utilizan ontologías semánticas, para inferir relaciones por Marco o Esquema, y técnicas relacionadas con resolución de anáfora en contextos inmediatos, para tratar de resolver la Activación; como sucede con las propuestas que buscan predecir que la siguiente mención de un referente activo será un dispositivo referencial reducido.

Esto va de la mano con que esta investigación muestra evidencia estadística de lo que teóricamente y a pequeña escala ya se había dicho: las frases nominales como dispositivos referenciales plenos son escasamente vinculables con referentes activos. Esto no significa que este Estado Informativo sea imposible de manifestarse en esta clase de dispositivos, pero que su escasez presenta problemas para homogenizar el proceso de identificación. Necesitan un tratamiento particular que no se cubre por el método propuesto. Los casos de activación deben ser atendidos con otros métodos, o incluso, caracterizar qué tan relevante es para alguna teoría pragmática estudiar las observaciones poco frecuentes de activación con dispositivos referenciales plenos.

4.1 Aportaciones

A parte de lo mostrado en los párrafos anteriores, la presente investigación ha dejado ver algunas áreas que, de una u otra manera, he resuelto. El primero, como ya mencioné, es un método computacional para la segmentación de las frases nominales. En este trabajo sólo pude disponer de un método manual, lo que aporta a resolver este problema, pero necesita aún de trabajo supervisado. Sería deseable contar con versiones automáticas más robustas para enfocarnos en otros problemas. Le sigue el que mi tesis es de nueva cuenta un recordatorio que algunos conceptos en lingüística no están lo suficientemente descritos como para poder implementarlos en un algoritmo informático. Es el caso de lo “nuevo” y lo “dado”. Luego, está el hecho de que el área en pragmática y estados de activación, se han concentrado principalmente en los dispositivos referenciales reducidos. Se espera que esta tesis sea un aporte a la discusión de los dispositivos referenciales plenos y la necesidad de generar jerarquías propias sobre los Estados Informativos. De hecho, el estudio comprueba que algunos estados de activación asociados a los dispositivos referenciales reducidos no pueden ser identificados en los plenos. Una jerarquía con los Estados Informativos más probables podría ser el preámbulo para un estudio pragmático de los dispositivos referenciales plenos en español.

En cuanto a la posible utilidad de estos hallazgos, como mencioné al inicio, podría utilizarse la medida en procesos automáticos para la resolución de referencia entre dispositivos referenciales plenos, en particular, llama la atención para este fin los resultados de las agrupaciones ESIN_R2 y ESIN_R3. Otra importante aportación de mi trabajo es el uso del DEM y su estructura como inventario de sentidos. En esto resulta novedoso ya que no se había utilizado el DEM para resolver tareas de este tipo en el pasado. Además, debido a que

este inventario en particular tiene un orden pensado para ciertos fines, se puede aprovechar esta estructura para orientar un analizador. No obstante, como ya mencioné, una investigación más a fondo sobre los efectos de las acepciones queda pendiente.

Un aspecto que parece prometedor, pero que es necesario seguir afinando, es la aparente imposibilidad de LSA y SPAN de identificar frases nominales con referentes activos. Pienso que una posible explicación de este comportamiento, digna de explorar en otros trabajos, está en la relación entre el ámbito local de la oración —en donde juega la activación— y el ámbito discursivo-contextual. LSA operaría en este segundo ámbito, no en el primero. Si en un futuro trabajo se decide profundizar en esta idea, los resultados que he mostrado pueden ayudar a evitar reducir la complejidad de los dispositivos referenciales plenos. No he supuesto a priori que el Estado Informativo de los referentes es no identificable o inactivo, además que en mi trabajo los determinantes no establecen categóricamente al Estado Informativo. En este sentido, mis resultados parecen indicar que observar la situación sintáctica tiene mayor relevancia para la identificación, aunque no superando las medidas.

La construcción del corpus de esta tesis doctoral, nombrado el Corpus Periodístico del Noroeste de México (COPENOR), fue un esfuerzo por aportar a la representación de otras regiones mexicanas de español. Pretendía etiquetar las primeras 380 notas, pero esto se demostró difícil por la cantidad de recursos necesarios, lo que lo dejaba fuera de los tiempos planteados para la investigación. No fue hasta realizar los primeros ejercicios, a finales del 2019 e inicios del 2020, que se redujo la cantidad de notas a analizar. Por lo que, reduje el corpus con una metodología que garantizara la distribución por estado. En una segunda vuelta de análisis construí varias herramientas paralelas para el etiquetado, pero tuvieron que ser restructuradas constantemente. Así mismo, cada nota analizada presentaba nuevos problemas

teóricos y afinaba el método de cómo analizar, manualmente, el Estado Informativo. No fue hasta la tercera vuelta que se concretaron las 38 notas y las 2 388 frases nominales.

Un aspecto que me parece importante destacar es que el tener al alcance técnicas que requieran pocos recursos es una buena señal: no todas las lenguas tienen acceso a colecciones de textos como el español o el inglés. Por poner un ejemplo, si quisiera realizar un trabajo parecido en lenguas indígenas mexicanas, 38 textos no son algo imposible de conseguir, incluso con lenguas con baja vitalidad lingüística como la lengua yumana pa ipai.

4.2. Trabajo futuro

En la revisión de los antecedentes surgió que otros métodos computacionales se han usado para encontrar correlaciones con las funciones comunicativas asociadas con la definitud y lo genérico. Una futura investigación podría ir en el mismo rumbo que la presente: construir un corpus etiquetado con funciones relacionadas con la definitud y probar si LSA o SPAN pueden funcionar para identificar alguna. Sobre esto, es interesante notar que, mientras las marcas de definitud están mucho más presentes en las frases nominales que analizo en COPENOR_CERO, la indefinitud no tanto: llama la atención que su escasez contrasta con la capacidad de LSA de clasificar lo No Identificable, asociado con menciones de nuevos referentes en el discurso. Esto no fue parte del objetivo de esta investigación, pero quedaría pendiente explorar la relación concreta de estas marcas con las medidas obtenidas a través de LSA y SPAN.

En la sección metodológica también se pudo observar que existen distintas maneras de aproximarse a evaluar las relaciones de lo nuevo y dado, no sólo LSA. Aunque hablé de

manera superficial sobre estas otras estrategias, quedaría pendiente realizar investigaciones con métodos como redes neuronales, en particular, aquellas que permiten explorar los rasgos que son relevantes para el modelo a la manera en que lo muestra Kibrik et al. (2016). También se encuentra la experimentación con técnicas asociadas a los resumidores automáticos, como, por ejemplo, *el método de puntuación de oraciones escalables*, una versión un poco distinta a la presentada en esta tesis pero con la misma medida de coseno, que consiste en maximizar la preponderancia de una oración, pero minimizar su redundancia con lo anterior, es decir, con los textos de los cuales se realiza el resumen. En todos los casos se siguen los mismos fundamentos de lo que he presentado: se buscaría si una frase/texto puede ser encontrada en otra frase/texto. Estas técnicas usan este criterio para construir el resumen, pero, en este caso, sería suficiente quedarnos con las técnicas de evaluación con los documentos o frases anteriores. Otra sugerencia ha sido el analizar una frase nominal con el resumen de lo anterior. Es decir, en vez de comparar cada frase nominal, sólo hacer un solo cálculo de coseno entre la frase y un resumen construido con alguna técnica. Estudiar y explorar los efectos de estas técnicas para las frases nominales y descubrir si en ellas se capturan aspectos lexicogramaticales es una tarea que considero pertinente para la lingüística computacional en la que enmarco esta investigación, pero que tendrá que dejarse como ideas para ser desarrolladas en otros espacios.

Además de estos otros métodos, queda pendiente seguir trabajando con las acepciones del DEM. No es necesario restringirse a LSA y SPAN, pero sí creo pertinente agotar un número mínimo de experimentos. Por ejemplo, y fue uno de los problemas al momento de incluir el diccionario en esta investigación, determinar qué acepción utilizar. Gracias a la manera en que está construido el DEM, esta decisión pudo ser dejada a la misma estructura del artículo

lexicográfico y a sus anotaciones que diferencian la polisemia y la homonimia. Aunque en sus propios términos, estas anotaciones son un recurso que no es trivial, principalmente si comparamos con otros inventarios de sentidos. No obstante, un método para desambiguar de las acepciones marcadas como homónimas era necesario. Aunque propuse una manera, resultó ser computacionalmente muy costosa (incluso con un corpus de 38 notas). Este trabajo merece su propia atención, lo cual, por su complejidad, podría resultar incluso en otro trabajo doctoral en lexicografía computacional.

Lo anterior son ideas que quedarían pendientes a estructurar. Existen dos investigaciones que no son sólo ideas, sino que cuento con las bases para profundizar en ellas en artículos independientes. Primero, realicé al inicio del 2020 experimentación con distintos etiquetadores automáticos de propiedades lexicogramaticales en Python. Comparé spaCy, NLTK, Freeling y Stanza y mi etiquetado manual. Los resultados me permitieron determinar que tienen exactitudes parecidas, pero con algunas diferencias que era importante explorar. Esto no lo reporto en su totalidad en esta tesis, debido a que no era el objetivo y me pareció pertinente planear la exposición en su propio espacio, además que necesita pasar por un etiquetado más amplio y un sistema para revisar entre pares, con lo que se pueda determinar un *límite humano*. Segundo, en distintos foros he tenido la oportunidad de exponer experimentos iniciales con las acepciones del DEM. He trabajo un pequeño grupo, comparando aquellos sospechosos de ser antónimos y sinónimos. Al mismo tiempo, probé recursos como redes semánticas en softwares como Gephi, lo cual me demostró que no es trivial la visualización de esta clase de datos. De este preámbulo, el siguiente paso consistiría en realizar pruebas con las más de 60 000 acepciones, y proponer una manera de explorar los resultados. A esos experimentos también le acompañan el trabajo con el Corpus del Español

Mexicano Contemporáneo. Me parecería interesante notar si existe alguna relación entre LSA o SPAN, las acepciones y sus respectivos contextos en el CEMC. Para este caso, sólo pude proceder a una primera limpieza del CEMC y a la búsqueda de los lemas del DEM. Para ambas investigaciones, la evaluación de los etiquetadores y los experimentos con el DEM, tengo estructurados los scripts en Python y las fuentes para realizar los artículos, por lo que no sólo es trabajo pendiente, sino que ya se encuentran en el tintero.

Existen diversas técnicas estadísticas para evaluar las distribuciones de los datos y otras pruebas para evaluar la varianza. En este trabajo traté de utilizar, sin problematizarlo, algunas pruebas que me permitieran hablar de los datos que tengo, pero reconozco que este paso necesitaría tal vez otros acercamientos. En particular, me parecería interesante buscar correspondencias de las distribuciones de propiedades pragmáticas con distribuciones de otras propiedades, tal vez sintácticas o léxicas. Esto en lingüística de corpus y cuantitativa no es nuevo, pero explorar estos campos para cumplir este objetivo quedaba lejos del objetivo de la investigación presente, entre otras razones, por la delimitación del método que proponía.

4.3 Epílogo

Con lo anterior, y a manera de síntesis puedo decir que esta investigación cumplió en gran medida con el objetivo general. Se exploraron las hipótesis, se rechazaron unas y se plantearon otras. He mostrado que algunos patrones lingüísticos pueden llegar a tener una correlación matemática, y, lo que es más, propiedades pragmáticas que muchas de las veces se nos presentan elusivas. Como mencioné al inicio de esta investigación, mi objetivo no era determinar un sistema de reglas discretas sino tendencias, y creo que esto lo he logrado con los métodos cuantitativos que he presentado. Además, esta investigación ha desprendido otros problemas a tratar, los cuales me han permitido delinear nuevas rutas de crecimiento académico y exploración.

Referencias

Abbott, Barbara. 2010. *Reference*. Oxford: Oxford University Press.

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

Aikhenvald, Alexandra Y. 2000. *Classifiers: A typology of noun categorization devices: A typology of noun categorization devices.* New York: Oxford University Press.

Akma, Nahdatul, Mohamad Hafiz, Azaliza Zainal, Muhammad Fairuz Abd Rauf, y Zuraidy Adnan. 2018. "Review of Chatbots Design Techniques". *International Journal of Computer Applications* 181: 7–10. https://doi.org/10.5120/ijca2018917606.

Alcina Caudet, Ma Amparo. 1999. "Las expresiones referenciales. Estudio semántico del sintagma nominal". Universitat de València.

Alonso, Amado. 1957. "Estilística y gramática del artículo en español." En *Estudios lingüísticos. Temas españoles*, 125–60. Madrid: Gredos.

Amat Rodrigo, Joaquín. 2020. "Random Forest con Python". Available under a Attribution 4.0 International (CC BY 4.0). 2020. https://www.cienciadedatos.net/documentos/py08_random_forest_python.html.

App, Roberto, Barry Smith, y Adnre D. Spear. 2015. *Building Ontologies with Basic Formal Ontology*. Cambridge, Massachusetts: The Mit Press.

Ariel, Mira. 1988. "Referring and Accessibility". *Journal of Linguistics* 24 (1): 65–87.

———. 1990. *Accesing Noun-Prhase Antecedents*. Londres y Nueva York: Routledge.

Awaihara, Yoshie. 2018. *Tópico y foco en japonés. Intersección entre la sintaxis y la estructura de la información.* Ciudad de México: El Colegio de México.

Baddeley, Alan. 1992. "Working Memory: The Interface between Memory and Cognition". *Journal of Cognitive Neuroscience* 4 (3): 281–88.

———. 2007. *Working memory, thought, and action*. Oxford: Oxford University Press.

Baumann, Stefan, Caren Brinckmann, Silvia Hansen-Schirra, Geert-Jan M. Kruijff, Ivana Kruijff-Korbayová, Stella Neumann, Erich Steiner, Elke Teich, y Hans Uszkoreit. 2004. “The muli project: Annotation and analysis of information structure in German and English”. En *Fourth International Conference on Language Resources and Evaluation*.

Beaugrande, Robert-Alain de, y Wolfgang Dressler Ulrich. 1997. *Introducción a la lingüística del texto*. Editado por Editorial Ariel. Barcelona.

Bello, Andrés. 2002. *Gramática de la lengua castellana destinada al uso de los americanos*. Madrid: Biblioteca Virtual Miguel de Cervantes. http://www.cervantesvirtual.com/nd/ark:/59851/bmczk5c0.

Berry, Michael W., y Murray Browne. 2005. *Understanding Search Engines*. 2a ed. Philadelphia: Society for Industrial and Applied Mathematics.

Bhatia, Archna, Chu-Cheng Lin, Nathan Schneider, Yulia Tsvetkov, Fatima Talib Al-Raisi, Laleh Roostapour, Jordan Bender, et al. 2014. “Automatic Classification of Communicative Functions of Definiteness”. En *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1059–70. Dublin, Ireland.

Bhatia, Archna, Mandy Simons, Lori Levin, Yulia Tsvetkov, Chris Dyer, y Jordan Bender. 2014. “A Unified Annotation Scheme for the Semantic/Pragmatic Components of Definiteness”. En *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, 910–916. Reykjavik, Iceland.

Bird, Steven, Edward Loper, y Ewan Klein. 2009. *Natural Language Processing with Python*. O’Reilly Media Inc.

Blanca, María J., Rafael Alarcón, Jaume Arnau, Roser Bono, y Rebecca Bendayan. 2017. “Non-normal data: Is ANOVA still a valid option?” *Psicothema* 29 (4): 552–57. https://doi.org/10.7334/psicothema2016.383.

Bois, John W. Du. 1980. “Beyond Definiteness: The Trace of Identity in Discourse”. En *The Pear stories: cognitive, cultural, and linguistic aspects of narrative production*, editado por W. L. Chafe, Ablex Pub., 203–74. Norwood, N. J.

———. 2003. “Argument structure: Grammar in use”. En *Preferred Argument Structure: Grammar as architecture for function*, editado por John W. Du Bois, Lorraine E. Kumpf, y William J. Ashby, 11–60. https://doi.org/10.1075/sidag.14.04dub.

Breiman, Leo. 2001. “Random Forests”. *Machine Learning* 45: 5–32.

Brown, Dolores. 1989. “El habla juvenil de Sonora, México: la fonética de 32 jóvenes”. *Nueva Revista de Filología Hispánica* 37 (1): 43–82. https://doi.org/https://doi.org/10.24201/nrfh.v37i1.730.

Brown, Gillian, y George Yule. 1983. *Discourse Analysis*. Cambridge: Cambridge University Press.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models are Few-Shot Learners”. arXiv:2005.14165 [cs.CL]. 2020. https://arxiv.org/abs/2005.14165.

Brucart, José M. A. 1999. “La estructura del sintagma nominal: las oraciones del relativo.” En *Gramática descriptiva de la lengua española. Vol. 1. Sintaxis básica de las clases de palabras.*, editado por Ignacio Bosque y Violeta Demonte, 395–522. Madrid: Espasa Calpe.

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

Burquest, Donald Arden. 1986. “The pronoun system of some Chadic languages”. En *Pronominal systems*, editado por Ursula Wiesemann, 71–101. Tübingen: Gunter Narr Verlag.

Camacho-Collados, Jose, y Mohammad Taher Pilehvar. 2018. “From Word To Sense Embeddings: A Survey on Vector Representations of Meaning”. *Journal of Artificial Intelligence Research* 63 (diciembre): 743–88. https://doi.org/10.1613/jair.1.11259.

Carrasco Morales, Raúl. 2004. “Resolución automática de la anáfora indirecta en el Español”. Instituto Politécnico Nacional.

Chafe, W. L. 1976. “Givenness, contrastiveness, definitiness, subjects, topics, and point of view”. En *Subject and Topic*, editado por C. Li, 25–55. New York: Academic Press.

———. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing.* Chicago: University of Chicago Press.

Chisholm, Erica, y Tamara G. Kolda. 1999. “New terms weighting formulas for the vector space method in information retrieval”. Oak Ridge, Tennessee: Oak Ridge National Laboratory. http://www.kolda.net/publication/ornl-tm-13756.pdf.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, Massachusetts: MIT Press.

———. 1986. *Knowledge of language. Its Nature, Origin, and Use.* New York: Praeger.

Cohen, Jacob. 1977. *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Company Company, Concepción, ed. 2006a. *Sintaxis histórica de la lengua española. Segunda parte: la frase nominal Vol. 1*. Ciudad de México: Fondo de Cultura Económica. Universidad Nacional Autónoma de México.

———. , ed. 2006b. *Sintaxis histórica de la lengua española. Segunda parte: la frase nominal Vol. 2.* Ciudad de México: Fondo de Cultura Económica. Universidad Nacional Autónoma de México.

Conover, W. J., y R. L. Iman. 1979. “On multiple-comparisons procedures”. https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-07677-MS.

D’Agostino, Ralph B. 1971. “An Omnibus Test of Normality for Moderate and Large Size Samples”. *Biometrika* 58 (2): 341. https://doi.org/10.2307/2334522.

Dahl, Östen, y Kari Fraurud. 1996. “Animacy in grammar in discourse”. En *Reference and Referent Accesibility*, editado por Thorstein Fretheim y Jeanette K. Gundel, 47–64. Amsterdam/Philapdelphia: John Benjamins Publishing Company.

Danesi, Marcel. 2009. “Opposition theory and the interconnectedness of language, culture, and cognition”. *Sign Systems Studies* 37 (1/2): 11–41.

Diccionario del español de México. s/f. “Diccionario del español de México”. El Colegio de México. Consultado el 1 de mayo de 2020. http://dem.colmex.mx.

Diessel, Holger. 1999. *Demonstratives: Form, function and grammaticalization.* Amsterdam: John Benjamins Publishing.

———. 2012. “Deixis and demonstratives”. En *An international handbook of natural language meaning*, editado por C Maienborn, K von Heusinger, y P Portner, 3:2407–31. Berlin: Walter de Gruyter.

———. 2014. “Demonstratives, Frames of Reference, and Semantic Universals of Space”. *Language and Linguistics Compass* 8 (3): 116–32.

Dijk, Teun A. Van. 1980. *Estructuras y funciones del discurso: una introducción intedisciplinaria a la lingüística del texto y a los estudios del discurso*. Ciudad de México: Siglo Veintiuno.

Dumais, Susan T. 1991. “Improving the retrieval of information from external sources”. *Behavior Research Methods, Instruments, & Computers* 23 (2): 229–36.

Endriss, Cornelia, y Ralf Klabunde. 2000. “Planning word-order dependent focus assignments”. En *Proceedings of the first international conference on Natural language generation - INLG ’00*, 14:156–62. Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1118253.1118275.

Fillmore, Charles J. 1982. “Frame semantics”. En *Linguistics in the morning calm.*, editado por Linguistic Society of Korea, 111–38. Seoul: Hanshin Pub. Co.

Foltz, P. W., W. Kintsch, y Thomas. K. Landauer. 1998. “The measurement of textual coherence with latent semantic analysis”. *Discourse Processes* 25: 285–307.

Frege, Gottlob. 1984. “Sobre sentido y referencia”. En *Estudios sobre semántica*. Barcelona, España: Ediciones Orbis.

García Fajardo, Josefina. 1994. “Hacia el universo del discurso desde la semántica formal. El articulo definido”. En *Segundo Encuentro de Lingüistas y Filólogos de España y México.*, editado por A. Alegría, B. Garza, y J. A. Pascual, 221–29. Universidad de Salamanca.

———. 2016. *Semántica de la oración : instrumentos para su análisis*. Ciudad de M{é}xico: El Colegio de México, Centro de Estudios Lingüísticos y Literarios.

Garey, Michael R., y David S. Johnson. 1979. *Computers and Intractability. A guide to the Theory fo NP-Completeness*. New York: W. H. Freeman & Co.

Gerrig, Richard J., y Edard J. O’Brien. 2005. “The Scope of Memory-Based Processing”. *Discourse Processes* 39 (2 y 3): 225–42.

Giuffrè, Mauro. 2017. *Text Linguistics and Classical Studies. Dressler and de Beaugrande’s Procedural Approach*. Cham, Switzerland: Springer.

Givón, T. 1982. “Logic vs. pragmatics, with human language as the referee: Toward an empirically viable epistemology”. *Journal of Pragmatics* 6 (2): 81–133. https://doi.org/https://doi.org/10.1016/0378-2166(82)90026-1.

———. 1983a. “Topic continuity in discourse: An introduction”. En *Topic continuity in discourse: A quantitative cross-language study.*, 5–41. Amesterdam: Benjamins.

———. 1983b. “Topic continuity in spoken English”. En *Topic continuity in discourse: A quantitative cross-language study.*, 345–63. Amsterdam: Benjamins.

———. 2001. *Syntax. An Introduction. Vol. 1.* Amsterdam/Philapdelphia: John Benjamins Publishing Company.

Goodfellow, Ian, Yoshua Bengio, y Aaron Courville. 2016. *Deep Learning*. Cambridge, Massachusetts: MIT Press.

Graesser, Arthur C., y Derek Harter. 2001. “Teaching Tactics and Dialog in AutoTutor”. *International Journal of Artificial Intelligence in Education* 12.

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse, y Zhiqiang Cai. 2004. "Coh-Metrix: Analysis of text on cohesion and language". *Behavior Research Methods, Instruments, & Computers* 36: 193–202.

Graesser, Arthur C., Phanni Penumatsa, M. Ventura, Zhiqiang Cai, y Xiangen Hu. 2007. "Using LSA in AutoTutor: Leaning through mixed initiative dialogue in natural language." En *Handbook of Latent Semantic Analysis*, editado por Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, y Walter Kintsch, 243–62. New York: Lawrence Erlbaum Associates, Inc.

Grice, Paul. 1989. *Studies in the Way of Words*. Cambridge, Massachusetts: Harvard University Press.

Gundel, Jeanette K. 1996. "Relevance Theory Meets the Givenness Hierarchy. An Account of Inferrables". En *Reference and Referent Accesibility*, editado por Thorstein Fretheim y Jeanette K. Gundel, 141–53. Amsterdam/Philapdelphia: John Benjamins Publishing Company.

Gutiérrez Bravo, Rodrigo. 2008. "La identificación de los tópicos y los focos". *Nueva Revista de Filología Hispánica* 56 (2): 363–401.

Hajičová, Eva, Petr Sgall, y Hana Skoumalová. 1993. "Identifying topic and focus by an automatic procedure". En *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics* -, 178–82. Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/976744.976766.

———. 1995. "An automatic procedure for topic-focus identification". *Computational Linguistics* 21 (1): 81–94.

Halliday, Michael Alexander Kirkwood. 1967. "Notes on Transitivity and Theme in English: Part 2". *Journal of Linguistics* 3 (2): 199–244. http://www.jstor.org/stable/4174965.

Hausser, Roland. 2014. *Foundations of Computational Linguistics*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41431-2.

Hawkins, John A. 1978. *Definiteness and Indefiniteness: A Study in Reference and Grammaticality Prediction*. Atlantic Highlands, N.J.: Humanities Press.

Heim, Irene. 2008. “File Change Semantics and the Familiarity Theory of Definiteness”. En *Formal Semantics*, 223–48. Oxford, UK: Blackwell Publishers Ltd. https://doi.org/10.1002/9780470758335.ch9.

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*, 941–46. https://escholarship.org/uc/item/87n81224.

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

Hernández Campoy, Juan Manuel, y Manuel Almeida. 2005. *Metodología de la investigación sociolingüística*. Granada, España: Editorial Comares.

Hjelmslev, Louis. 1971. *Prolegómenos a una teoría del lenguaje*. Madrid: Gredos.

Holton, Avery E., y Hsiang Iris Chyi. 2012. “News and the Overloaded Consumer: Factors Influencing Information Overload Among News Consumers”. *Cyberpsychology, Behavior, and Social Networking* 15 (11): 619–24.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, y Adriane Boyd. 2020. “spaCy: Industrial-strength Natural Language Processing in Python”. 2020. https://doi.org/10.5281/zenodo.1212303.

Hopper, Paul J., y Sandra A. Thompson. 1984. “The discourse basis for lexical categories in universal grammar”. *Language* 60 (4): 703–52.

Hu, Xiangen, Zhiqiang Cai, Max Louwerse, Andrew Olney, Phanni Penumatsa, y Arthur C. Graesser. 2003. “A revised algorithm for latent semantic analysis”. En *IJCAI International Joint Conference on Artificial Intelligence*, 1489–1491.

Jakobson, Roman. 1980. *The framework of language*. Ann Arbor: Michigan Slavic Publications.

Jensen, O. 2006. “Maintenance of multiple working memory items by temporal segmentation.” *Neuroscience* 139: 237–49.

Jones, Karen Sparck. 1994. “Natural Language Processing: A Historical Review”. En *Current Issues in Computational Linguistics: In Honour of Don Walker*, 3–16. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-0-585-35958-8_1.

Jurafsky, Daniel, y James H. Martin. 2009. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2nd ed. Upper Saddle River, N.J.: Pearson Prentice Hall.

Kay, Martin. 2003. “Introduction”. En *The Oxford Handbook of Computational Linguistics*, editado por Rurslan Mitkov, 1a ed., xviii–xx. Oxford: Oxford University Press.

Keenan, Edward L., y Bernard Comrie. 1977. “Noung Phrase Accesibility and Universal Grammar”. *Linguistic Inquiry* 8 (1): 63–99.

Kehler, Andrew, y Hannah Rohde. 2018. “Prominence and coherence in a Bayesian theory of pronoun interpretation”. *Journal of Pragmatics* 154 (mayo): 63–78. https://doi.org/10.1016/j.pragma.2018.04.006.

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* 7 (septiembre): 204–24. https://doi.org/10.3389/fpsyg.2016.01429.

Krifka, Manfred. 2008. “Basic notions of information structure”. *Acta Linguistica Hungarica* 55 (3–4): 243–76. https://doi.org/10.1556/ALing.55.2008.3-4.2.

Krifka, Manfred, Francis Jeffry Pelletier, Gregory N. Carlson, Alice ter Meulen, Gennaro Chierchia, y Godehard Link. 1995. “Genericity: An Introduction”. En *The Generic Book*, editado por Gregory Norman Carlson y Francis Jeffry Pelletier. Chicago: The Chicago Press.

Kruijff-Korbayová, Ivana, y Geert-Jan M. Kruijff. 2004. “Discourse-Level Annotation for Investigating Information Structure”. En *Association for Computational Linguistics Annual Meeting 2004 Workshop on Discourse Annotation*, 941–46. Barcelona.

Kruijff-Korbayová, Ivana, y Mark Steedman. 2003. “Discourse and Infromation Structure”. *Journal of Logic, Language and Information* 12: 249–59.

Lakoff, George. 1990. “The Invariance Hypothesis: is abstract reason based on image-schemas?” *Cognitive Linguistics* 1 (1): 39–74. https://doi.org/10.1515/cogl.1990.1.1.39.

Lakoff, George, y Mark Johnson. 2017. *Metáforas de la vida cotidiana*. Barcelona: Cátedra.

Lambrecht, Knud. 1986. “Pragmatically motivated syntax: presentational cleft constructions in spoken French.” En *Proceedings of the Twenty-Second Meeting of the Chicago Linguistic Society. Papers from the Paressesion on Pragmatics and Grammatical Theory.*, 115–26.

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

Landauer, Thomas. K., y S. T. Dumais. 1997. “A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge”. *Psychological review* 104 (2): 211–40.

Landauer, Thomas K. 2002. “On the computational basis of learning and cognition: arguments from LSA”. En *The psichology of learning and motivation. Advances in Research and Theory. Volume 41.*, editado por Brian H. Ross, 41–84. Amsterdam: Elsevier.

Landauer, Thomas K., Danielle S. McNamara, Simon Dennis, y Walter Kintsch, eds. 2007. *Handbook of Latent Semantic Analysis*. New York: Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9780203936399.

Langdon, Margaret, y Pamela Munro. 1979. “Subject and (Switch-)reference in yuman”. *Folia Linguistica* 13 (3–4): 321–44.

Lara. 2015. *Curso de lexicología*. Ciudad de México: El Colegio de México.

Lara, Luis Fernando. 1997. *Teoría del diccionario monolingüe*. Ciudad de México: El Colegio de México.

———. 2001. *Ensayos de teoría semántica: lengua natural y lenguajes científicos*. El Colegio de México.

———. , ed. 2007. *Resultados numéricos del vocabulario fundamental del español de México*. Ciudad de México: El Colegio de México, Centro de Estudios Lingüísticos y Literarios, Diccionario del Español de México.

———. 2016. *Teoría semántica y método lexicográfico*. Ciudad de México: El Colegio de México.

Lara, Luis Fernando, Roberto Chande Ham, y Ma. Isabel García Hidalgo. 1979. *Investigaciones lingüísticas en lexicografía*. Ciudad de México: El Colegio de México.

Leonetti, Manuel. 1990. *El artículo y la referencia*. Madrid: Taurus.

———. 1999. "El artículo". En *Gramática descriptiva de la lengua española. Vol. 1. Sintaxis básica de las clases de palabras.*, 787–890.

Lesk, Michael. 1986. "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone." En *Proceedings of ACM SIGDOC Conference*, 24–26. Toronto.

Levelt, Willem. J. M. 1989. *Speaking: From intention to articulation*. Cambridge: MIT Press.

Levinson, Stephen C. 2004. "Deixis". En *The handbook of pragmatics*, editado por Laurence R Horn y Gregory Ward, 97–121. Blackwell.

Levshina, Natalia. 2015. *How to do Linguistics with R. Data exploration and statistical analysis*. Amsterdam/Philapdelphia: John Benjamins Publishing.

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

———. 1996. “México”. En *Manual de dialectología hispánica. El español de América*, editado por Manuel Alvar, 81–89. Barcelona: Ariel.

Lyons, John. 1980. *Semantica*. Vol. 1–2. Barcelona: Teide.

Manning, Christopher D., y Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The Mit Press.

Martin Butragueño, Pedro. 2018. “La expresión del sujeto pronominal en la Ciudad de México: explorando la variación lingüística con efectos estadísticos fijos y con efectos mixtos”. 2018. http://goo.gl/dsKcp1.

Masche, Julia, y Nguyen-Thinh Le. 2018. “A Review of Technologies for Conversational Systems”. En *Advanced Computational Methods for Knowledge Engineering*, editado por Nguyen-Thinh Le, Tien van Do, Ngoc Thanh Nguyen, y Hoai An Le Thi, 212–25. Cham: Springer International Publishing.

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 *Applied Natural Language Processing*, editado por Philip M. McCarthy y Chutima Boonthum-Denecke, 457–78. Pennsylvania: IGI Global. https://doi.org/10.4018/978-1-60960-741-8.ch027.

McKinney, Wes. 2010. “Data structures for statistical computing in python”. En *Proceedings of the 9th Python in Science Conference*, 51–56.

Medina Urrea, Alfonso. 2003. “Investigación cuantitativa de afijos y clíticos del español de México: glutinometría en el corpus del español mexicano contemporáneo”. El Colegio de México.

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

———. 2006. “El español del noroeste mexicano: un acercamiento desde adentro.” En *Estudios sociolingüísticos del español de España y América*, editado por Ana Ma. Cestero Mancera, Isabel Molina Martos, y Florentino Paredes García, 159–67. Madrid: Arco Libros.

Mikolov, Tomas, Kai Chen, Greg Corrado, y Jeffrey Dean. 2013. “Efficient estimation of word representations in vector space.” En *Proceedings of the International Conference on Learning Representations (ICLR 2013)*. https://arxiv.org/abs/1301.3781.

Mikolov, Tomas, Net-tau Yih, y Geoffrey Zweig. 2013. “Linguistic Regularities in Continuous SpaceWord Representations”. En *Proceedings of NAACL-HLT 2013*, 746–51. Atlanta: Association for Computational Linguistics.

Mírovský, Jiří, Kateřina Rysová, Magdaléna Rysová, y Eva Hajičová. 2013. “(Pre-)Annotation of Topic-Focus Articulation in Prague Czech-English Dependency Treebank”. En *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 55–63. Nagoya, Japan: Asian Federation of Natural Language Processing. http://aclweb.org/anthology/I13-1007.

Montgomery, Douglas C., y George C. Runger. 2014. *Applied statistics and probability for engineers*. Estados Unidos: Wiley.

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

Nadkarni, Prakash M, Lucila Ohno-Machado, y Wendy W Chapman. 2011. “Natural language processing: an introduction”. *Journal of the American Medical Informatics Association* 18 (5): 544–51. https://doi.org/10.1136/amiajnl-2011-000464.

Nirenburg, Sergei, y Victor Raskin. 2004. *Ontological Semantics*. Cambridge, Massachusetts: The Mit Press.

Noordman, Leo G.M., y Wietske Vonk. 2015. “Inferences in Discourse, Psychology of”. En *International Encyclopedia of the Social & Behavioral Sciences*, 37–44. Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.57012-3.

Nyberg, L. 1996. “Classifying Human Long-term Memory: Evidence from Converging Dissociations”. *European Journal of Cognitive Psychology* 8 (2): 163–84.

O’Neil, D., y T. Harcup. 2009. “News values and selectivity”. En *The handbook of journalism studies*, editado por K. Wahl-Jorgensen y T. Hanitzsch, 161–74. New York: Routledge.

Padró, Lluís, y Evgeny Stanilovsky. 2012. “FreeLing 3.0: Towards Wider Multilinguality”. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*. Istanbul. http://nlp.lsi.upc.edu/publications/papers/padro12.pdf.

Paducheva [Padučeva], Elena V. 1985. *Vyskazyvanie i ego sootnesennost’ s dejstvitel’-nost’ju [Enunciación y su correlación con la realidad]*. Moscow: Nauka.

Picallo, M. Carme. 1999. “La estructura del sintagma nominal: las nominalizaciones y otros sustantivos con complementos argumentales.” En *Gramática descriptiva de la lengua española. Vol. 1. Sintaxis básica de las clases de palabras.*, editado por Ignacio Bosque y Violeta Demonte, 363–93. Madrid: Espasa Calpe.

Pozas Loyo, Julia. 2016. *El artículo definido. Origen y gramaticalización.* Ciudad de México: El Colegio de México.

Prince, Ellen F. 1978. “A comparison of wh-clefts and it-clefts in discourse.” *Language* 54: 883–906.

———. 1981. “Towards taxonomy of Given-New Information”. En *Radical Pragmatics*, editado por Peter Cole, 223–55. New York: Academic Press.

———. 1992. “The ZPG letter: Subjects, definiteness, and information-status”. En *Discourse description: Diverse analyses ofa fundraising text*, editado por S. Thompson y W. Mann, 295–325. Amsterdam, Paises Bajos: John Benjamins Publishing.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, y Christopher D. Manning. 2020. “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.” En *Association for Computational Linguistics (ACL) System Demonstrations. 2020.* https://arxiv.org/pdf/2003.07082.pdf.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, y Ilya Sutskever. 2018. “Language Models are Unsupervised Multitask Learners”. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Rajasekaran, Abirami, y R. Varalakshmi. 2018. “Review on automatic text summarization”. *International Journal of Engineering & Technology* 7 (3.3): 456. https://doi.org/10.14419/ijet.v7i2.33.14210.

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

Recasens, Marta. 2010. “Coreference: Theory, Annotation, Resolution and Evaluation.” Universidad de Barcelona.

Recasens, Marta, y M. Antònia Martí. 2010. “AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan”. *Language Resources and Evaluation* 44 (4): 315–45. https://doi.org/10.1007/s10579-009-9108-x.

Recio Diego, Álvaro. 2015. “La estructura argumental del sintagma nominal en español”. Universidad de Salamanca. Facultad de Filología.

Reiter, Nils, y Anette Frank. 2010. “Identifiying Generic Noun Phrases”. En *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 40–49. Uppsala, Suecia: Association for Computational Linguistics.

Rigau, Gemma. 1999. “La estructura del sintagma nominal: los modificadores del nombre.” En *Gramática descriptiva de la lengua española. Vol. 1. Sintaxis básica de las clases de palabras.*, editado por Ignaco Bosque y Violeta Demonte, 311–62. Madrid: Espasa Calpe.

Rijkhoff, Jan. 2004. *The Noun Phrase*. Oxford: Oxford University Press.

Rivero, María-Luisa. 1975. “Referential Properties of Spanish Noun Phrases.” *Language* 51 (1): 32–48.

Rokach, L., y O. Maimon. 2005. “Top-down induction of decision trees classifiers-a survey”. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews.* 35 (4): 476–87. https://doi.org/https://doi.org/10.1109%2FTSMCC.2004.843247.

Russell, Bertrand. 1905. “On denoting”. *Mind* 14: 479–93.

Salaverría, Ramón, y Rafael Cores. 2005. “Géneros periodísticos en los cibermedios hispanos”. En *Cibermedios. El impacto de internet en los medios de comunicación en España*, editado por Ramón Salaverría y Rafael Cores, 145–85. Sevilla: Comunicación Social Ediciones y Publicaciones.

Salton, G., A. Wong, y C. S. Yang. 1975. “A vector space model for automatic indexing.” *Communications of the ACM* 18 (11): 613–20.

Saussure, F de. 2005. *Curso de lingüística general*. 24a ed. Biblioteca de obras maestras del pensamiento. Buenos Aires: Losada. https://books.google.com.mx/books?id=l3_lAAAACAAJ.

Schrape, Jan-Felix. 2019. “The Promise of Technological Decentralization. A Brief Reconstruction”. *Society* 56: 31–37.

Schwarzschild, Roger. 1999. “GIVENness, AvoidF and other constraints on the placement of accent”. *Natural Language Semantics* 7 (2): 141–77.

Searle, John R. 1980. “Minds, brains, and programs”. *The behavorial and brain sciences* 3: 417–57.

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

Sgall, Petr, Eva Hajičová, y Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Dordrecht: Reidel.

Stevenson, Mark, y Yorick Wilks. 2012. “Word-Sense Disambiguation”. En *The Oxford Handbook of Computational Linguistics*, editado por Ruslan Mitkov, 1a ed., 254–65. Oxford: Oxford University Press.

Strawson, Peter. 1950. “On referring”. *Mind* 59: 320–44.

Strube, Michael, y Maria Wolters. 2000. “A Probabilistic genre-independent model of pronominalization”. En *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 18–25. Seattle.

Sukthanker, Rhea, Soujanya Poria, Erik Cambria, y Ramkumar Thirunavukarasu. 2018. “Anaphora and Coreference Resolution: A Review”.

Tomasello, Michael. 2008. *Origins of human communication*. Cambridge: MIT Press.

Tomlin, Rusell S. 1987. “Linguistic reflextions of cognitive events”. En *Coherence and Grounding in Discourse*, editado por Rusell S. Tomlin, 455–79. Amsterdam/Philapdelphia: John Benjamins Publishing Company.

———. 1995. “Focal attention, voice, and word order. An experimental, cross-linguistic study.” En *Word order in discourse*, editado por Pamela Downing y Michael Noonan, 517–54. Amsterdam/Philapdelphia: John Benjamins Publishing Company.

Torres-Moreno, Juan-Manuel. 2014. *Automatic Text Summarization*. New York: John Wiley & Sons.

Trubetzkoy, Nikolái. S. 2019. *Principios de fonología*. Ciudad de México: El Colegio de México.

Turing, Alan M. 1950. “Computing Machinery and Intelligence”. *Mind* LIX (236): 433–60. https://doi.org/10.1093/mind/LIX.236.433.

Urdan, Timothy C. 2005. *Statistic in Plain English*. New Jersey: Lawrecen Erlbaum Associates, Publishers Inc.

Wang, Lucy Lu, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, et al. 2020. “CORD-19: The COVID-19 Open Research Dataset”, abril. http://arxiv.org/abs/2004.10706.

Woodward, Todd S., Tara A. Cairo, Christian C. Ruff, Yoshio Takane, Michael A. Hunter, y Elton T. C. Ngan. 2006. “Functional connectivity reveals load dependent neural systems underlying encoding and maintenance in verbal working memory”. *Neuroscience* 139 (1): 317–25.

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*, 209–18. Berlin, Germany: Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-2026.

Ziai, Ramon, y Detmar Meurers. 2018. “Automatic Focus Annotation: Bringing Formal Pragmatics Alive in Analyzing the Information Sutrcutre of Authentic Data”. En *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2018*, 117–28. New Orleans, Louisiana: ACL.

Anexo A. Lineamientos para etiquetado de Oración y Frase Nominal

A continuación, se exponen los lineamientos para etiquetado de Frase Nominal y Oración en COPENOR. Ténganse en cuenta los siguientes cuatro aspectos técnicos del etiquetado y notación: (1) las etiquetas XML inician con una llave simple entre corchetes (p.e. <fn>) y deben terminar con la misma llave pero con una diagonal "/" (p.e. </fn>); (2) aquellos fragmentos que inicien con asterisco (*) indican etiquetados incorrectos; (3) como se mencionó antes, todas las contracciones del tipo *al* o *del* se separan para incluir en la etiqueta el determinante, pero separar la preposición; y (4) las etiquetas XML rodean **sin espacios** al fragmento etiquetado.

I. Nombres propios adposicionales de calles y avenidas, empresas y personas

Se analiza el cargo, lugar o el nombre de la calle junto con el nombre propio:

(1) <fn>la calle Privada Bilbao</fn>

(2) <fn>el presidente Armando Ayala</fn>

(3) <fn>la empresa Ecogas</fn>

(4) <fn>la clínica San José</fn>

Por lo que los siguientes casos son etiquetados incorrectos:

(5) *<fn>la calle <fn>Privada Bilbao</fn></fn>

(6) *<fn>el presidente <fn>Armando Ayala</fn></fn>

Los nombres propios de lugar, como las calles, pero también como ciudades y estados, se etiquetan también junto con su nominal:

(7) <fn>El estado de Baja California</fn>

(8) <fn>La ciudad de Ensenada</fn>

Pero deben considerarse como FN separadas cuando tienen referencia distinta, por ejemplo:

(9) <fn>la capital de <fn>Baja California</fn></fn>

En este caso, *la capital de Baja California* se refiere a una entidad distinta (Mexicali) que el nombre *Baja California*.

Los **apellidos de personas** se incluyen dentro de la FN, lo que vuelve incorrecto etiquetar el conjunto de apellidos por separado o incluso individualmente:

(10) <fn>José Guadalupe González Fernández</fn>

(11) *<fn>José Guadalupe <fn>González Fernández</fn></fn>

(12) *<fn>José <fn>Guadalupe</fn> <fn>González</fn><fn>Fernández</fn></fn>

Una FN que contenga adposiciones y un nombre propio tomará como núcleo el nombre propio. De esta manera, en una FN como:

(13) <fn><fn>La jefa de el Gobierno Estatal</fn>, Gloria Valenti</fn>

Se etiqueta *la jefa de el Gobierno Estatal*, lo que indica que esta pieza está subordinada a *Gloria Valenti*.

A veces no es sencillo detectar un nombre propio que contiene preposiciones, por ejemplo, en el siguiente caso:

(14) <fn>agentes de <fn>la Secretaría de Seguridad Pública Municipal</fn></fn>

Podría analizarse que *Seguridad Pública Municipal* es otro nombre propio, subordinado a *Secretaría de….* Para el caso de COPENOR, se etiqueta junto, con la suposición de que *Seguridad Pública Municipal* no es el conjunto a donde pertenece *la Secretaría* sino todo ello construye un solo referente. Nótese el siguiente caso:

(15) <fn>el Estado Mayor de <fn>la Secretaría de la Defensa Nacional</fn></fn>

En este caso, este análisis supone que la *Secretaría de la Defensa Nacional* sí es un referente en donde un miembro es *el Estado Mayor*, entre otros.

II. Frases y Núcleos Coordinados

Para el caso de las Frases Nominales con **núcleos coordinados de nombres propios**, se etiquetarán los núcleos en casos como el siguiente:

(16) <fn>las calles <fn>Guachinango</fn> y <fn>Pez Martillo</fn></fn>

Pero en casos de nombres comunes se etiquetarán en un solo grupo:

(17) <fn>las descuidadas zonas y espacios</fn>

En el caso de **FN coordinadas**, se etiquetan ambas por separado. La marca de la conjunción se etiqueta fuera de las etiquetas de FN:

(18) <fn>Personal de <fn>la empresa Ecogas</fn></fn> e <fn>integrantes de <fn>el H. Cuerpo de Bomberos</fn></fn>

III. Lineamientos generales de FN

En términos generales, se etiqueta un FN después de preposición –excepto los casos expuestos.

(19) <fn>la crema de <fn>mantequilla</fn></fn>

El determinante es pista de una FN, incluso con núcleos verbales, por ejemplo:

(20) a <fn>el trabajar por <fn>las tardes</fn></fn>

IV. Fechas y cantidades

Todos los constituyentes que expresan fechas se etiquetan sin jerarquía interna, como en los siguientes ejemplos. Nótese que las contracciones se separan:

(21) <fn>el primero de julio</fn>

(22) <fn>30 de marzo de el 2019</fn>

En el caso de cantidades, se etiqueta sólo la FN que encabeza la cantidad, sin jerarquía interna:

(23) <fn>2 kilos de cemento</fn>

(24) <fn>Tres millones de pesos</fn>

(25) <fn>2 millones 247 mil 457 pesos</fn>

(26) <fn>53 años de edad</fn>

(27) <fn>una distancia de <fn>2 mil 995 metros cuadrados</fn></fn>

V. Etiquetado de Oración

En general, una oración se detecta por un núcleo verbal. Ejemplos sencillos sin <fn> son los siguientes:

(28) <ora>el perro **duerme** furiosamente</ora>

(29) <ora>el perro **come** croquetas</ora>

(30) <ora>el perro le **regala** unas flores al gato de manera inadvertida y sorpresiva</ora>

Las oraciones subordinadas adjetivales y adverbiales de infinitivo, participio y gerundio se etiquetan como oraciones siempre y cuando se detecte su estructura verbal. Por ejemplo:

(31) a. <ora>el agente atrapó al ladrón <ora>**golpeando** el vidrio</ora></ora>
b. <ora>el agente atrapó al ladrón <ora>**golpeado** por el dueño del local</ora></ora>
c. <ora>el agente atrapó al ladrón <ora>por **golpear** el vidrio</ora></ora>

Las oraciones, a diferencia de las FFNN, sí incluyen en la etiqueta las preposiciones y conjunciones.

(32) <ora>**para** evitar <fn>cualquier incidente mayor</fn></ora>

(33) <ora>**y** en <fn>la calle Sin Nombre de <fn>la Che Guevara</fn></fn> se invirtieron <fn>799 mil 620 pesos</fn></ora>

Todas las relativas del tipo *que* introducen oración, incluso dentro de FN como en el último ejemplo (no se etiquetan todas las FFNN para realizar la exposición de manera más clara):

(34) <ora>**que** a el <ora>decirle a Juan <ora>**que** dejara en paz a la víctima</ora></ora>, éste se molestó</ora>

(35) <ora><fn>el <ora>que no lo invitaras a la fiesta</ora></fn> lo hizo enojar</ora>

VI. Pronombres relativos y oraciones

Los casos de pronombres relativos del tipo *quien/es* o *cual/es* deben pertenecer a una <fn> que es parte de una oración subordinada de otra oración matriz. Para este ejemplo utilizaré el formato en XML para ilustrar la jerarquía y no etiquetaré las FFNN de la oración matriz:

```
<ora>el agente tiene mucho trabajo
    <ora>
        <fn>el cual</fn> tiene a <fn>su familia</fn> en <fn>Ensenada</fn>
    </ora>
</ora>
```

Nótese, por otro lado, la siguiente FN que tiene un pronombre relativo pero que no es analizado como FN:

```
<fn>los cuerpos de emergencia
    <ora>quienes arribaron a
        <fn>el lugar de <fn>los hechos</fn></fn>, en
        <fn>una toma domiciliaria de
            <fn>la colonia en <fn>mención</fn></fn>
        </fn>
    </ora>
</fn>
```

Como ya se expuso en el punto III, sólo se etiquetan frases nominales cuando están introducidas por determinante, lo que aplica también para los verbos en infinitivo y los pronombres relativos. En el ejemplo anterior quienes no se etiqueta como FN –pero esto no significa que no se considere un dispositivo referencial reducido; como se expuso en el apartado teórico, todo DDR, por definición, implica que su referente está Activo.

Las FFNN que formen parte de conjunciones compuestas que funcionen como nexos no se analizan. De esta manera, construcciones del tipo *por su parte*, *por su lado*, *por lo cual* se analizan dentro de la etiqueta <ora> que introducen:

```
<ora>por lo cual se invita
    <ora>a no utilizar <fn>la vialidad de
    <fn>el río</fn></fn></ora>,
    <ora>así como evitar <fn>el paso por <fn>arroyos y vados de
    <fn>la ciudad</fn>
            </fn>
        </fn>
    </ora>
</ora>
```

Debido a la manera en que está etiquetado el corpus, se podrá acceder a estas conjunciones compuestas al buscar las etiquetas <ora> y buscar el primer verbo que aparece para después seleccionar todo lo que se encuentra entre la etiqueta <ora> y el verbo, descartando <fn>.

VII. SIGNOS DE PUNTUACIÓN Y CÓDIGOS HTML

Todo signo de puntuación queda fuera de la última etiqueta colocada, por ejemplo:

(36) <ora>quienes se han fijado <ora>para lograr <fn>el desarrollo económico y productivo de <fn>el sector pecuario de <fn>la región</fn></fn>**,**
específicamente, de <fn>los criadores</fn>**,**
<fn>engordadores</fn> y <fn>comercializadores de <fn>ovinos</fn></fn>**.**

COMILLAS Y CITAS

La única excepción a la puntación tiene que ver con las comillas. En las notas periodísticas, las comillas significan, en muchos casos, citas. Si es así, van dentro de la etiqueta <ora>. Una oración que introduce una cita indirecta, como del tipo *la directora dijo que* a veces no aparece de una manera que sea sencillo analizar con las etiquetas. Por ejemplo, nótese el siguiente caso:

(37) <ora>
<fn>Esta situación</fn>, **<ora>agregó <fn>la directora de el Registro</fn></ora>**, se presenta principalmente en <fn>las comunidades
<ora>que se ubican en <fn>las regiones más alejadas de <fn>la entidad</fn></fn></ora></fn>,
<ora>pues debido a <fn>las distancias
<ora>que se tienen
<ora>que recorrer
<ora>para llegar a <fn>la Oficialía de el Registro más cercana</fn>
</ora></ora></ora></fn>,
<fn>muchas familias</fn> deciden esperar para registrar <fn>el nacimiento de <fn>las niñas</fn> y <fn>niños</fn></fn>
</ora>

En el ejemplo anterior, *agregó la directora de el Registro* es la oración que introduce *esta situación … se presenta* pero por la estructura en la que está presentada, se coloca dentro de la oración que funciona como cita indirecta.

LOCUCIONES

En general, las locuciones preposicionales no se etiquetarán, aunque tengan estructura de frase nominal. Por ejemplo, en el siguiente caso tenemos el *mismo tiempo* introducida por la preposición *a*. En este caso particular no se etiqueta como FN, aunque si se encontrara etiquetada, se marcaría como **no identificable baja**:

```
<ora>
    A el mismo tiempo, manifestó <fn>Brenda Rosas Gamboa</fn>
    <ora>que
        <ora>aunque se han registrado <fn>casos de <fn>personas adultas
                    <ora>que acuden a solicitar <fn>su registro</fn>
                        <ora>porque necesitan
                            <fn>su acta de <fn>nacimiento</fn>
                            </fn>
                        </ora>
                    </ora></fn></fn>
        </ora>,
        es más frecuente
        <ora>que se presente <fn>esta situación</fn> en
            <fn>el caso de <fn>jóvenes y niños</fn>
            </fn>
        </ora>
    </ora></ora>.
```

(Última modificación: 17/03/2021)

Anexo B. Etiquetado completo de COPENOR-253BC

```
<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>http://sintesistv.com.mx/gano-ensenada-conflicto-territorial-con-rosarito/</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 a través de los programas de Verano de
      la Investigación Científica aumentó casi al doble en relación al año pasado, al ser 36 estudiantes de
      licenciatura los seleccionados que cuentan con el apoyo de la Academia Mexicana de Ciencias (AMC) y el
      Programa Delfín.

                  Los estudiantes provienen de 20 universidades y tecnológicos de México, entre ellos la
      UNAM, la Universidad de Guanajuato, ITESO y las universidades autónomas de Nayarit, Sinaloa y Baja
      California Sur, por mencionar algunas. Ellos pasarán alrededor de 6 semanas en los laboratorios de
      todas las divisiones del CICESE y tendrán la oportunidad de participar en prácticas de campo y
      actividades deportivas y culturales. En este verano participan 20 mujeres y 16 hombres.

                  La AMC y el Programa Delfín apoyan con una beca económica a los estudiantes seleccionados
      en las convocatorias de sus XXIX y XXIV Verano de la Investigación Científica y Tecnológica,
      respectivamente. Incluso participan varios estudiantes que tienen la oportunidad de gestionar su
      estancia con recursos propios o patrocinios de otro tipo. El objetivo de estos programas es fomentar y
      reforzar la vocación científica de las nuevas generaciones.

                  Las estancias de verano son un escalón previo al posgrado, ya que uno de los requisitos de
      estas convocatorias es cursar la etapa terminal de una licenciatura en ciencia o ingeniería. Por ello,
      la Dra. Rufina Hernández Martínez, encargada de la Dirección de Estudios de Posgrado del CICESE, se
      encargó de dar la bienvenida a los estudiantes. Además les compartió la historia de este centro de
      investigación y la diversidad de disciplinas que lo componen.
```

Cabe mencionar que la Dra. Hernández también participó en el Verano de la Investigación en los inicios de su carrera académica. Así como ella, muchos investigadores que tuvieron esta oportunidad son los que ahora invitan a las nuevas generaciones a participar y recomiendan al CICESE como una de las mejores opciones.

Para brindarles perspectivas distintas sobre la experiencia que representan estos programas, Eleyra Sena Lozoya, egresada de la maestría en Ciencias de la Tierra, y Marco A. Domínguez Bureos, actual estudiante del mismo posgrado, coincidieron en sus testimonios: realizar una estancia de verano en el CICESE fue para ambos un parteaguas al tomar la decisión de estudiar un posgrado en este centro de investigación.

“Tuve la oportunidad de convivir con investigadores, técnicos y administrativos que brindaron apoyo en mi enseñanza […] el investigador a mi cargo estaba muy involucrado en que aprendiéramos”, mencionó Eleyra Sena, “Además de preocuparse por transmitir conocimiento científico, el CICESE también se preocupa mucho por el bienestar de los estudiantes”, dijo.

En su momento, Marco Antonio Domínguez les compartió: “Un investigador me hizo saber que el lugar en el que debes estudiar combina tres cosas: apoyo de tu familia, estabilidad emocional y hacer lo que más te gusta. Creo que el CICESE combina muy bien estas cualidades. Nunca terminaré de agradecerle, a su personal y a la sociedad misma, que es la que nos da la oportunidad de hacer esto, la que nos brinda el chance de estar en el Verano de Investigación y en el posgrado del CICESE”, finalizó.

Para complementar la bienvenida, los estudiantes de posgrado José Ricardo Santillán, Danna Lyn Arellano, Gabriela Reséndiz y Georgina Rojo de Anda, compartieron los proyectos de tesis que actualmente desarrollan en los posgrados de Óptica, Ciencias de la Vida, Ecología Marina y Oceanografía Física, respectivamente, con el afán de expresar la calidad de investigación que es posible desarrollar en el 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>a[ADP_AdpType=Prep]a través[NOUN_]través de[ADP_AdpType=Prep]de <fn cs="pr"
esin="1_NO_IDENT" idfn="4">los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
programas[NOUN_Gender=Masc|Number=Plur]programa de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
idfn="5">Verano[PROPN_]Verano de[ADP_AdpType=Prep]de
la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el Investigación[PROPN_]Investigación
Científica[PROPN_]Científica
</fn></fn></ora></fn></fn>aumentó[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin]aumentar
casi[ADV_]casi a[ADP_AdpType=Prep]a <fn cs="na" esin="0_NO_IDENT_BAJA"
idfn="6">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
doble[NOUN_Gender=Masc|Number=Sing]doble </fn>en[ADP_AdpType=Prep]en relación[NOUN_]relación
a[ADP_AdpType=Prep]a <fn cs="pr" esin="5_ACCESIBLE_ORIGO"
idfn="7">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el año[NOUN_AdvType=Tim]año
pasado[ADJ_Gender=Masc|Number=Sing|VerbForm=Part]pasado </fn>,[PUNCT_PunctType=Comm], a[ADP_AdpType=Prep]a
<fn cs="pr" esin="0_NO_IDENT_BAJA" idfn="8">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
ser[NOUN_Gender=Masc|Number=Sing]ser <fn cs="su" esin="8_ACTIVO_P" idfn="9"
refe="1">36[NUM_NumForm=Digit|NumType=Card]36 estudiantes[NOUN_Number=Plur]estudiante de[ADP_AdpType=Prep]de
licenciatura[NOUN_Gender=Fem|Number=Sing]licenciatura </fn><fn cs="at" esin="9_IDENT_BAJA"
idfn="10">los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
seleccionados[ADJ_Gender=Masc|Number=Plur|VerbForm=Part]seleccionado <ora idora="3"
tp="vs">que[PRON_PronType=Int,Rel]que
cuentan[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]contar con[ADP_AdpType=Prep]con <fn
cs="pr" esin="1_NO_IDENT" idfn="11">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
apoyo[NOUN_Gender=Masc|Number=Sing]apoyo de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
idfn="12">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el Academia[PROPN_]Academia
Mexicana[PROPN_]Mexicana de[ADP_AdpType=Prep]de Ciencias[PROPN_]Ciencias </fn><fn cs="na"
esin="9_IDENT_BAJA" idfn="13" refe="10">([PUNCT_PunctSide=Ini|PunctType=Brck]( AMC[PROPN_]AMC
)[PUNCT_PunctSide=Fin|PunctType=Brck]) </fn>y[CCONJ_]y <fn cs="na" esin="1_NO_IDENT"
idfn="14">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el Programa[PROPN_]Programa
Delfín[PROPN_]Delfín </fn></fn></ora></fn></fn></ora>.[PUNCT_PunctType=Peri]. <ora idora="4" tp="vm"><fn
cs="su" esin="6_ACTIVO_S" idfn="15">Los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
estudiantes[NOUN_Number=Plur]estudiante
</fn>provienen[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]provenir de[ADP_AdpType=Prep]de
<fn cs="pr" esin="1_NO_IDENT" idfn="16">20[NUM_NumForm=Digit|NumType=Card]20
universidades[NOUN_Gender=Fem|Number=Plur]universidad y[CCONJ_]y
tecnológicos[NOUN_Gender=Masc|Number=Plur]tecnológico de[ADP_AdpType=Prep]de <fn cs="na"
```

```
esin="2_INACTIVO_MLP" idfn="17">México[PROPN_]México </fn></fn>,[PUNCT_PunctType=Comm],
entre[ADP_AdpType=Prep]entre ellos[PRON_Case=Acc,Nom|Gender=Masc|Number=Plur|Person=3|PronType=Prs]él <fn
cs="na" esin="2_INACTIVO_MLP" idfn="18">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
UNAM[PROPN_]UNAM </fn>,[PUNCT_PunctType=Comm], <fn cs="na" esin="1_NO_IDENT"
idfn="19">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el Universidad[PROPN_]Universidad
de[ADP_AdpType=Prep]de Guanajuato[PROPN_]Guanajuato </fn>,[PUNCT_PunctType=Comm], <fn cs="na"
esin="2_INACTIVO_MLP" idfn="20">ITESO[PROPN_]ITESO </fn>y[CCONJ_]y <fn cs="na" esin="1_NO_IDENT"
idfn="21">las[DET_Definite=Def|Gender=Fem|Number=Plur|PronType=Art]el
universidades[NOUN_Gender=Fem|Number=Plur]universidad autónomas[ADJ_Gender=Fem|Number=Plur]autónomo
de[ADP_AdpType=Prep]de Nayarit[PROPN_]Nayarit ,[PUNCT_PunctType=Comm], Sinaloa[PROPN_]Sinaloa y[CCONJ_]y
Baja[PROPN_]Baja California[PROPN_]California Sur[PROPN_]Sur </fn></ora>,[PUNCT_PunctType=Comm], <ora
idora="5" tp="vm">por[ADP_AdpType=Prep]por mencionar[VERB_VerbForm=Inf]mencionar
algunas[PRON_Gender=Fem|Number=Plur|PronType=Ind]alguno </ora>.[PUNCT_PunctType=Peri]. <ora idora="6"
tp="vm">Ellos[PRON_Case=Acc,Nom|Gender=Masc|Number=Plur|Person=3|PronType=Prs]él
pasarán[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Fut|VerbForm=Fin]pasar alrededor[ADV_]alrededor
de[ADP_AdpType=Prep]de <fn cs="pr" esin="0_NO_IDENT_BAJA" idfn="22">6[NUM_NumForm=Digit|NumType=Card]6
semanas[NOUN_Gender=Fem|Number=Plur]semana </fn>en[ADP_AdpType=Prep]en <fn cs="pr" esin="1_NO_IDENT"
idfn="23">los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
laboratorios[NOUN_Gender=Masc|Number=Plur]laboratorio de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
idfn="24">todas[DET_Gender=Fem|Number=Plur|PronType=Tot]todo
las[DET_Definite=Def|Gender=Fem|Number=Plur|PronType=Art]el divisiones[NOUN_Gender=Fem|Number=Plur]división
de[ADP_AdpType=Prep]de <fn cs="na" esin="3_INACTIVO_RD" idfn="25"
refe="2">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el CICESE[PROPN_]CICESE
</fn></fn></fn></ora><ora idora="7" tp="vm">y[CCONJ_]y
tendrán[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Fut|VerbForm=Fin]tener <fn cs="od" esin="1_NO_IDENT"
idfn="26">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
oportunidad[NOUN_Gender=Fem|Number=Sing]oportunidad de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
idfn="27">participar[VERB_VerbForm=Inf]participar en[ADP_AdpType=Prep]en <fn cs="na" esin="1_NO_IDENT"
idfn="28">prácticas[NOUN_Gender=Fem|Number=Plur]práctica de[ADP_AdpType=Prep]de <fn cs="na"
esin="0_NO_IDENT_BAJA" idfn="29">campo[NOUN_Gender=Masc|Number=Sing]campo </fn></fn>y[CCONJ_]y <fn cs="na"
esin="1_NO_IDENT" idfn="30">actividades[NOUN_Gender=Fem|Number=Plur]actividad
deportivas[ADJ_Gender=Fem|Number=Plur]deportivo y[CCONJ_]y culturales[ADJ_Number=Plur]cultural
</fn></fn></fn></ora>.[PUNCT_PunctType=Peri]. <ora idora="8" tp="vm">En[ADP_AdpType=Prep]en <fn cs="pr"
esin="3_INACTIVO_RD" idfn="31" refe="4">este[DET_Gender=Masc|Number=Sing|PronType=Dem]este
verano[NOUN_Gender=Masc|Number=Sing]verano
</fn>participan[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]participar <fn cs="su"
```

```
esin="1_NO_IDENT" idfn="32">20[NUM_NumForm=Digit|NumType=Card]20 mujeres[NOUN_Gender=Fem|Number=Plur]mujer
</fn>y[CCONJ_]y <fn cs="su" esin="1_NO_IDENT" idfn="33">16[NUM_NumForm=Digit|NumType=Card]16
hombres[NOUN_Gender=Masc|Number=Plur]hombre </fn></ora>.[PUNCT_PunctType=Peri]. <ora idora="9" tp="vm"><fn
cs="su" esin="3_INACTIVO_RD" idfn="34" refe="10">La[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
AMC[PROPN_]AMC </fn>y[CCONJ_]y <fn cs="su" esin="3_INACTIVO_RD" idfn="35"
refe="11">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el Programa[PROPN_]Programa
Delfín[PROPN_]Delfín </fn>apoyan[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]apoyar
con[ADP_AdpType=Prep]con <fn cs="pr" esin="1_NO_IDENT"
idfn="36">una[DET_Definite=Ind|Gender=Fem|Number=Sing|PronType=Art]uno beca[NOUN_Gender=Fem|Number=Sing]beca
económica[ADJ_Gender=Fem|Number=Sing]económico </fn>a[ADP_AdpType=Prep]a <fn cs="od" esin="1_NO_IDENT"
idfn="37">los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
estudiantes[NOUN_Number=Plur]estudiante <ora idora="10"
tp="vs">seleccionados[ADJ_Gender=Masc|Number=Plur|VerbForm=Part]seleccionado en[ADP_AdpType=Prep]en <fn
cs="pr" esin="1_NO_IDENT" idfn="38">las[DET_Definite=Def|Gender=Fem|Number=Plur|PronType=Art]el
convocatorias[NOUN_Gender=Fem|Number=Plur]convocatoria de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
idfn="39">sus[DET_Number=Plur|Person=3|Poss=Yes|PronType=Prs]su XXIX[PROPN_]XXIX y[CCONJ_]y XXIV[PROPN_]XXIV
Verano[PROPN_]Verano de[ADP_AdpType=Prep]de la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
Investigación[PROPN_]Investigación Científica[PROPN_]Científica y[CCONJ_]y Tecnológica[PROPN_]Tecnológica
</fn></fn>,[PUNCT_PunctType=Comm], respectivamente[ADV_]respectivamente
</ora></fn></ora>.[PUNCT_PunctType=Peri]. <ora idora="11" tp="vm">Incluso[ADV_]incluso
participan[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]participar <fn cs="su"
esin="1_NO_IDENT" idfn="40">varios[DET_Gender=Masc|Number=Plur|PronType=Ind]varios
estudiantes[NOUN_Number=Plur]estudiante <ora idora="12" tp="vs">que[PRON_PronType=Int,Rel]que
tienen[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]tener <fn cs="od" esin="1_NO_IDENT"
idfn="41">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
oportunidad[NOUN_Gender=Fem|Number=Sing]oportunidad <ora idora="13" tp="vs">de[ADP_AdpType=Prep]de
gestionar[VERB_VerbForm=Inf]gestionar <fn cs="od" esin="4_ACCESIBLE_MARCO"
idfn="42">su[DET_Number=Sing|Person=3|Poss=Yes|PronType=Prs]su estancia[NOUN_Gender=Fem|Number=Sing]estancia
</fn>con[ADP_AdpType=Prep]con <fn cs="pr" esin="1_NO_IDENT"
idfn="43">recursos[NOUN_Gender=Masc|Number=Plur]recurso propios[ADJ_Gender=Masc|Number=Plur]propio
</fn>o[CCONJ_]o <fn cs="pr" esin="1_NO_IDENT" idfn="44">patrocinios[NOUN_Gender=Masc|Number=Plur]patrocinio
de[ADP_AdpType=Prep]de <fn cs="na" esin="0_NO_IDENT_BAJA"
idfn="45">otro[DET_Gender=Masc|Number=Sing|PronType=Ind]otro tipo[NOUN_Gender=Masc|Number=Sing]tipo
</fn></fn></ora></fn></ora></fn></ora>.[PUNCT_PunctType=Peri]. <ora idora="14" tp="vm"><fn cs="su"
esin="1_NO_IDENT" idfn="46">El[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
objetivo[NOUN_Gender=Masc|Number=Sing]objetivo de[ADP_AdpType=Prep]de <fn cs="na" esin="3_INACTIVO_RD"
```

```
idfn="47">estos[DET_Gender=Masc|Number=Plur|PronType=Dem]este
programas[NOUN_Gender=Masc|Number=Plur]programa
</fn></fn>es[AUX_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin]ser <ora idora="15"
tp="vs">fomentar[VERB_VerbForm=Inf]fomentar y[CCONJ_]y reforzar[VERB_VerbForm=Inf]reforzar <fn cs="od"
esin="1_NO_IDENT" idfn="48">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
vocación[NOUN_Gender=Fem|Number=Sing]vocación científica[ADJ_Gender=Fem|Number=Sing]científico
de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
idfn="49">las[DET_Definite=Def|Gender=Fem|Number=Plur|PronType=Art]el
nuevas[ADJ_Gender=Fem|Number=Plur]nuevo generaciones[NOUN_Gender=Fem|Number=Plur]generación
</fn></fn></ora></ora>.[PUNCT_PunctType=Peri]. <ora idora="16" tp="vm"><fn cs="su" esin="4_ACCESIBLE_MARCO"
idfn="50">Las[DET_Definite=Def|Gender=Fem|Number=Plur|PronType=Art]el
estancias[NOUN_Gender=Fem|Number=Plur]estancia de[ADP_AdpType=Prep]de <fn cs="na" esin="0_NO_IDENT_BAJA"
idfn="51">verano[NOUN_Gender=Masc|Number=Sing]verano
</fn></fn>son[AUX_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]ser <fn cs="at" esin="1_NO_IDENT"
idfn="52">un[DET_Definite=Ind|Gender=Masc|Number=Sing|PronType=Art]uno
escalón[NOUN_Gender=Masc|Number=Sing]escalón previo[ADJ_Gender=Masc|Number=Sing]previo a[ADP_AdpType=Prep]a
<fn cs="na" esin="4_ACCESIBLE_MARCO" idfn="53">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
posgrado[NOUN_Gender=Masc|Number=Sing]posgrado </fn></fn></ora>,[PUNCT_PunctType=Comm], <ora idora="17"
tp="vm">ya[ADV_]ya que[SCONJ_]que <fn cs="su" esin="1_NO_IDENT"
idfn="54">uno[PRON_Gender=Masc|Number=Sing|PronType=Ind]uno de[ADP_AdpType=Prep]de
los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
requisitos[NOUN_Gender=Masc|Number=Plur]requisito de[ADP_AdpType=Prep]de <fn cs="na" esin="3_INACTIVO_RD"
idfn="55" refe="31">estas[DET_Gender=Fem|Number=Plur|PronType=Dem]este
convocatorias[NOUN_Gender=Fem|Number=Plur]convocatoria
</fn></fn>es[AUX_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin]ser <ora idora="18"
tp="vs">cursar[VERB_VerbForm=Inf]cursar <fn cs="od" esin="1_NO_IDENT"
idfn="56">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el etapa[NOUN_Gender=Fem|Number=Sing]etapa
terminal[ADJ_Number=Sing]terminal de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
idfn="57">una[DET_Definite=Ind|Gender=Fem|Number=Sing|PronType=Art]uno
licenciatura[NOUN_Gender=Fem|Number=Sing]licenciatura en[ADP_AdpType=Prep]en <fn cs="na"
esin="0_NO_IDENT_BAJA" idfn="58">ciencia[NOUN_Gender=Fem|Number=Sing]ciencia </fn>o[CCONJ_]o <fn cs="na"
esin="0_NO_IDENT_BAJA" idfn="59">ingeniería[NOUN_Gender=Fem|Number=Sing]ingeniería
</fn></fn></fn></ora></ora>.[PUNCT_PunctType=Peri]. <ora idora="19" tp="vm">Por[ADP_AdpType=Prep]por
ello[PRON_Case=Acc,Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs]él ,[PUNCT_PunctType=Comm], <fn cs="su"
esin="1_NO_IDENT" idfn="60">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el Dra[PROPN_]Dra
.[PUNCT_PunctType=Peri]. Rufina[PROPN_]Rufina Hernández[PROPN_]Hernández Martínez[PROPN_]Martínez
```

```
,[PUNCT_PunctType=Comm], <ora idora="20"
tp="vs">encargada[ADJ_Gender=Fem|Number=Sing|VerbForm=Part]encargado de[ADP_AdpType=Prep]de <fn cs="pr"
esin="1_NO_IDENT" idfn="61">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
Dirección[PROPN_]Dirección de[ADP_AdpType=Prep]de Estudios[PROPN_]Estudios de[ADP_AdpType=Prep]de
Posgrado[PROPN_]Posgrado del[ADP_AdpType=Prepprон]del CICESE[PROPN_]CICESE
</fn></ora></fn>,[PUNCT_PunctType=Comm],
se[PRON_Case=Acc,Dat|Person=3|PrepCase=Npr|PronType=Prs|Reflex=Yes]él
encargó[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin]encargar de[ADP_AdpType=Prep]de
dar[VERB_VerbForm=Inf]dar <fn cs="od" esin="1_NO_IDENT"
idfn="62">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
bienvenida[NOUN_Gender=Fem|Number=Sing]bienvenida </fn>a[ADP_AdpType=Prep]a <fn cs="oi" esin="3_INACTIVO_RD"
idfn="63" refe="12">los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
estudiantes[NOUN_Number=Plur]estudiante </fn></ora>.[PUNCT_PunctType=Peri]. <ora idora="21"
tp="vm">Además[ADV_]además les[PRON_Case=Dat|Number=Plur|Person=3|PronType=Prs]él
compartió[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin]compartir <fn cs="od" esin="1_NO_IDENT"
idfn="64">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
historia[NOUN_Gender=Fem|Number=Sing]historia de[ADP_AdpType=Prep]de <fn cs="na" esin="8_ACTIVO_P" idfn="65"
refe="2">este[DET_Gender=Masc|Number=Sing|PronType=Dem]este centro[NOUN_Gender=Masc|Number=Sing]centro
de[ADP_AdpType=Prep]de <fn cs="na" esin="0_NO_IDENT_BAJA"
idfn="66">investigación[NOUN_Gender=Fem|Number=Sing]investigación </fn></fn></fn>y[CCONJ_]y <fn cs="od"
esin="1_NO_IDENT" idfn="67">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
diversidad[NOUN_Gender=Fem|Number=Sing]diversidad de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
idfn="68">disciplinas[NOUN_Gender=Fem|Number=Plur]disciplina <ora idora="22"
tp="vs">que[PRON_PronType=Int,Rel]que
lo[PRON_Case=Acc|Gender=Masc|Number=Sing|Person=3|PrepCase=Npr|PronType=Prs]él
componen[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]componer
</ora></fn></fn></ora>.[PUNCT_PunctType=Peri]. <ora idora="23"
tp="vm">Cabe[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin]caber
mencionar[VERB_VerbForm=Inf]mencionar <ora idora="24" tp="vs">que[SCONJ_]que <fn cs="su" esin="6_ACTIVO_S"
idfn="69">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el Dra[PROPN_]Dra .[PUNCT_PunctType=Peri].
Hernández[PROPN_]Hernández </fn>también[ADV_]también
participó[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin]participar en[ADP_AdpType=Prep]en <fn
cs="pr" esin="3_INACTIVO_RD" idfn="70">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
Verano[PROPN_]Verano de[ADP_AdpType=Prep]de la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
Investigación[PROPN_]Investigación </fn>en[ADP_AdpType=Prep]en <fn cs="pr" esin="1_NO_IDENT"
idfn="71">los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
```

```
inicios[NOUN_Gender=Masc|Number=Plur]inicio de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
idfn="72">su[DET_Number=Sing|Person=3|Poss=Yes|PronType=Prs]su carrera[NOUN_Gender=Fem|Number=Sing]carrera
académica[ADJ_Gender=Fem|Number=Sing]académico </fn></fn></ora></ora>.[PUNCT_PunctType=Peri]. <ora
idora="25" tp="vm">Así[ADV_]así como[SCONJ_]como
ella[PRON_Case=Acc,Nom|Gender=Fem|Number=Sing|Person=3|PronType=Prs]él ,[PUNCT_PunctType=Comm], <fn cs="su"
esin="1_NO_IDENT" idfn="73">muchos[DET_Gender=Masc|NumType=Card|Number=Plur|PronType=Ind]mucho
investigadores[NOUN_Gender=Masc|Number=Plur]investigador <ora idora="26"
tp="vs">que[PRON_PronType=Int,Rel]que
tuvieron[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin]tener <fn cs="od"
esin="4_ACCESIBLE_MARCO" idfn="74">esta[DET_Gender=Fem|Number=Sing|PronType=Dem]este
oportunidad[NOUN_Gender=Fem|Number=Sing]oportunidad
</fn></ora></fn>son[AUX_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]ser <fn cs="at"
esin="9_IDENT_BAJA" idfn="75">los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el <ora idora="27"
tp="vs">que[PRON_PronType=Int,Rel]que ahora[ADV_]ahora
invitan[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]invitar a[ADP_AdpType=Prep]a <fn cs="od"
esin="1_NO_IDENT" idfn="76">las[DET_Definite=Def|Gender=Fem|Number=Plur|PronType=Art]el
nuevas[ADJ_Gender=Fem|Number=Plur]nuevo generaciones[NOUN_Gender=Fem|Number=Plur]generación
</fn>a[ADP_AdpType=Prep]a participar[VERB_VerbForm=Inf]participar </ora><ora idora="28" tp="vs">y[CCONJ_]y
recomiendan[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]recomendar a[ADP_AdpType=Prep]a <fn
cs="od" esin="3_INACTIVO_RD" idfn="77">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
CICESE[PROPN_]CICESE </fn>como[SCONJ_]como <fn cs="na" esin="9_IDENT_BAJA"
idfn="78">una[PRON_Gender=Fem|Number=Sing|PronType=Ind]uno de[ADP_AdpType=Prep]de
las[DET_Definite=Def|Gender=Fem|Number=Plur|PronType=Art]el mejores[ADJ_Degree=Cmp|Number=Plur]mejor
opciones[NOUN_Gender=Fem|Number=Plur]opción </fn></ora></fn></ora>.[PUNCT_PunctType=Peri]. <ora idora="29"
tp="vm"><ora idora="30" tp="vs">Para[ADP_AdpType=Prep]para brindar[VERB_VerbForm=Inf]brindar
les[PRON_Case=Dat|Number=Plur|Person=3|PronType=Prs]él <fn cs="od" esin="1_NO_IDENT"
idfn="79">perspectivas[NOUN_Gender=Fem|Number=Plur]perspectiva distintas[ADJ_Gender=Fem|Number=Plur]distinto
sobre[ADP_AdpType=Prep]sobre <fn cs="na" esin="1_NO_IDENT"
idfn="80">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
experiencia[NOUN_Gender=Fem|Number=Sing]experiencia <ora idora="31" tp="vs">que[PRON_PronType=Int,Rel]que
representan[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]representar <fn cs="su"
esin="3_INACTIVO_RD" idfn="81">estos[DET_Gender=Masc|Number=Plur|PronType=Dem]este
programas[NOUN_Gender=Masc|Number=Plur]programa </fn></ora></fn></fn></ora>,[PUNCT_PunctType=Comm], <fn
cs="su" esin="1_NO_IDENT" idfn="82">Eleyra[PROPN_]Eleyra Sena[PROPN_]Sena Lozoya[PROPN_]Lozoya
</fn>,[PUNCT_PunctType=Comm], <fn cs="na" esin="9_IDENT_BAJA"
idfn="83">egresada[NOUN_Gender=Fem|Number=Sing]egresada de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
```

```
idfn="84">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
maestría[NOUN_Gender=Fem|Number=Sing]maestría en[ADP_AdpType=Prep]en Ciencias[PROPN_]Ciencias
de[ADP_AdpType=Prep]de la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el Tierra[PROPN_]Tierra
</fn></fn>,[PUNCT_PunctType=Comm], y[CCONJ_]y <fn cs="su" esin="1_NO_IDENT" idfn="85">Marco[PROPN_]Marco
A[PROPN_]A .[PUNCT_PunctType=Peri]. Domínguez[PROPN_]Domínguez Bureos[PROPN_]Bureos
</fn>,[PUNCT_PunctType=Comm], <fn cs="na" esin="9_IDENT_BAJA" idfn="86">actual[ADJ_Number=Sing]actual
estudiante[NOUN_Number=Sing]estudiante de[ADP_AdpType=Prep]de <fn cs="na" esin="8_ACTIVO_P" idfn="87"
refe="74">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
mismo[DET_Gender=Masc|Number=Sing|PronType=Dem]mismo posgrado[NOUN_Gender=Masc|Number=Sing]posgrado
</fn></fn>,[PUNCT_PunctType=Comm],
coincidieron[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin]coincidir en[ADP_AdpType=Prep]en <fn
cs="pr" esin="4_ACCESIBLE_MARCO" idfn="88">sus[DET_Number=Plur|Person=3|Poss=Yes|PronType=Prs]su
testimonios[NOUN_Gender=Masc|Number=Plur]testimonio </fn>:[PUNCT_PunctType=Colo]: <ora idora="32"
tp="vs">realizar[VERB_VerbForm=Inf]realizar <fn cs="od" esin="1_NO_IDENT"
idfn="89">una[DET_Definite=Ind|Gender=Fem|Number=Sing|PronType=Art]uno
estancia[NOUN_Gender=Fem|Number=Sing]estancia de[ADP_AdpType=Prep]de <fn cs="na" esin="0_NO_IDENT_BAJA"
idfn="90">verano[NOUN_Gender=Masc|Number=Sing]verano </fn>en[ADP_AdpType=Prep]en <fn cs="pr"
esin="3_INACTIVO_RD" idfn="91">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
CICESE[PROPN_]CICESE </fn></fn></ora>fue[AUX_Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin]ser
para[ADP_AdpType=Prep]para ambos[NUM_Gender=Masc|NumType=Card|Number=Plur]ambos <fn cs="at"
esin="1_NO_IDENT" idfn="92">un[DET_Definite=Ind|Gender=Masc|Number=Sing|PronType=Art]uno
parteaguas[NOUN_Gender=Masc|Number=Sing]parteagua a[ADP_AdpType=Prep]a <fn cs="na" esin="1_NO_IDENT"
idfn="93">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el tomar[VERB_VerbForm=Inf]tomar <fn
cs="od" esin="1_NO_IDENT" idfn="94">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
decisión[NOUN_Gender=Fem|Number=Sing]decisión <ora idora="33" tp="vs">de[ADP_AdpType=Prep]de
estudiar[VERB_VerbForm=Inf]estudiar <fn cs="od" esin="1_NO_IDENT"
idfn="95">un[DET_Definite=Ind|Gender=Masc|Number=Sing|PronType=Art]uno
posgrado[NOUN_Gender=Masc|Number=Sing]posgrado </fn>en[ADP_AdpType=Prep]en <fn cs="pr" esin="8_ACTIVO_P"
idfn="96" refe="80">este[DET_Gender=Masc|Number=Sing|PronType=Dem]este
centro[NOUN_Gender=Masc|Number=Sing]centro de[ADP_AdpType=Prep]de
investigación[NOUN_Gender=Fem|Number=Sing]investigación
</fn></ora></fn></fn></fn></ora>.[PUNCT_PunctType=Peri]. “[PUNCT_PunctType=Quot]“ <ora idora="34"
tp="vm"><ora idora="35" tp="vs">Tuve[VERB_Mood=Ind|Number=Sing|Person=1|Tense=Past|VerbForm=Fin]tener <fn
cs="od" esin="1_NO_IDENT" idfn="97">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
oportunidad[NOUN_Gender=Fem|Number=Sing]oportunidad <ora idora="36" tp="vs">de[ADP_AdpType=Prep]de
convivir[VERB_VerbForm=Inf]convivir con[ADP_AdpType=Prep]con <fn cs="pr" esin="1_NO_IDENT"
```

```
idfn="98">investigadores[NOUN_Gender=Masc|Number=Plur]investigador </fn>,[PUNCT_PunctType=Comm], <fn cs="pr"
esin="1_NO_IDENT" idfn="99">técnicos[NOUN_Gender=Masc|Number=Plur]técnico </fn>y[CCONJ_]y <fn cs="pr"
esin="1_NO_IDENT" idfn="100">administrativos[NOUN_Gender=Masc|Number=Plur]administrativo <ora idora="37"
tp="vs">que[PRON_PronType=Int,Rel]que
brindaron[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin]brindar
apoyo[NOUN_Gender=Masc|Number=Sing]apoyo en[ADP_AdpType=Prep]en <fn cs="pr" esin="4_ACCESIBLE_MARCO"
idfn="101">mi[DET_Number=Sing|Number[psor]=Sing|Person=1|Poss=Yes|PronType=Prs]mi
enseñanza[NOUN_Gender=Fem|Number=Sing]enseñanza </fn></ora></fn></ora></fn></ora>[[PUNCT_PunctType=Dash][
…[PROPN_]… ][PUNCT_PunctType=Dash]] <ora idora="38" tp="vs"><fn cs="su" esin="1_NO_IDENT"
idfn="102">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
investigador[NOUN_Gender=Masc|Number=Sing]investigador a[ADP_AdpType=Prep]a <fn cs="na"
esin="0_NO_IDENT_BAJA" idfn="103">mi[DET_Number=Sing|Number[psor]=Sing|Person=1|Poss=Yes|PronType=Prs]mi
cargo[NOUN_Gender=Masc|Number=Sing]cargo
</fn></fn>estaba[AUX_Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin]estar muy[ADV_]mucho
involucrado[ADJ_Gender=Masc|Number=Sing|VerbForm=Part]involucrado <ora idora="39"
tp="vs">en[ADP_AdpType=Prep]en que[SCONJ_]que
aprendiéramos[VERB_Mood=Sub|Number=Plur|Person=1|Tense=Imp|VerbForm=Fin]aprendiérar
</ora></ora>”[PUNCT_PunctType=Quot]” ,[PUNCT_PunctType=Comm],
mencionó[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin]mencionar <fn cs="su"
esin="3_INACTIVO_RD" idfn="104" refe="72">Eleyra[PROPN_]Eleyra Sena[PROPN_]Sena
</fn></ora>,[PUNCT_PunctType=Comm], “[PUNCT_PunctType=Comm]“ <ora idora="40" tp="vm"><ora idora="41"
tp="vs"><ora idora="42" tp="vs">Además[ADV_AdpType=Prep]además de[ADP_AdpType=Prep]de
preocupar[VERB_VerbForm=Inf]preocupar se[PRON_Case=Acc,Dat|Person=3|PrepCase=Npr|PronType=Prs|Reflex=Yes]él
<ora idora="43" tp="vs">por[ADP_AdpType=Prep]por transmitir[VERB_VerbForm=Inf]transmitir <fn cs="od"
esin="1_NO_IDENT" idfn="105">conocimiento[NOUN_Gender=Masc|Number=Sing]conocimiento
científico[ADJ_Gender=Masc|Number=Sing]científico </fn></ora></ora>,[PUNCT_PunctType=Comm], <fn cs="su"
esin="3_INACTIVO_RD" idfn="106" refe="80">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
CICESE[PROPN_]CICESE </fn>también[ADV_]también
se[PRON_Case=Acc,Dat|Person=3|PrepCase=Npr|PronType=Prs|Reflex=Yes]él
preocupa[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin]preocupar mucho[ADV_]mucho
por[ADP_AdpType=Prep]por <fn cs="pr" esin="1_NO_IDENT"
idfn="107">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
bienestar[NOUN_Gender=Masc|Number=Sing]bienestar de[ADP_AdpType=Prep]de <fn cs="na" esin="0_NO_IDENT_BAJA"
idfn="108">los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
estudiantes[NOUN_Number=Plur]estudiante </fn></fn></ora>”[PUNCT_PunctType=Quot]” ,[PUNCT_PunctType=Comm],
dijo[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin]decir </ora>.[PUNCT_PunctType=Peri]. <ora
```

```
idora="44" tp="vm">En[ADP_AdpType=Prep]en su[DET_Number=Sing|Person=3|Poss=Yes|PronType=Prs]su
momento[NOUN_Gender=Masc|Number=Sing]momento ,[PUNCT_PunctType=Comm], <fn cs="su" esin="3_INACTIVO_RD"
idfn="109" refe="75">Marco[PROPN_]Marco Antonio[PROPN_]Antonio Domínguez[PROPN_]Domínguez
</fn>les[PRON_Case=Dat|Number=Plur|Person=3|PronType=Prs]él
compartió[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin]compartir :[PUNCT_PunctType=Colo]:
“[PUNCT_PunctType=Dash]“ <ora idora="45" tp="vs"><fn cs="su" esin="1_NO_IDENT"
idfn="110">Un[DET_Definite=Ind|Gender=Masc|Number=Sing|PronType=Art]uno
investigador[NOUN_Gender=Masc|Number=Sing]investigador
</fn>me[PRON_Case=Acc,Dat|Number=Sing|Person=1|PrepCase=Npr|PronType=Prs]yo
hizo[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin]hacer saber[VERB_VerbForm=Inf]saber <ora
idora="46" tp="vs">que[SCONJ_]que <fn cs="su" esin="1_NO_IDENT"
idfn="111">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
lugar[NOUN_Gender=Masc|Number=Sing]lugar en[ADP_AdpType=Prep]en <fn cs="na" esin="0_NO_IDENT_BAJA"
idfn="112">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el <ora idora="47"
tp="vs">que[PRON_PronType=Int,Rel]que debes[AUX_Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin]deber
estudiar[VERB_VerbForm=Inf]estudiar
</ora></fn></fn>combina[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin]combinar <fn cs="od"
esin="1_NO_IDENT" idfn="113">tres[NUM_NumType=Card|Number=Plur]tres cosas[NOUN_Gender=Fem|Number=Plur]cosa
:[PUNCT_PunctType=Colo]: <fn cs="na" esin="1_NO_IDENT" idfn="114">apoyo[NOUN_Gender=Masc|Number=Sing]apoyo
de[ADP_AdpType=Prep]de tu[DET_Number=Sing|Number[psor]=Sing|Person=2|Poss=Yes|PronType=Prs]tu
familia[NOUN_Gender=Fem|Number=Sing]familia </fn>,[PUNCT_PunctType=Comm], <fn cs="na" esin="1_NO_IDENT"
idfn="115">estabilidad[NOUN_Gender=Fem|Number=Sing]estabilidad emocional[ADJ_Number=Sing]emocional
</fn>y[CCONJ_]y <ora idora="48" tp="vs">hacer[VERB_VerbForm=Inf]hacer <fn cs="od" esin="0_NO_IDENT_BAJA"
idfn="116">lo[PRON_Case=Acc|Definite=Def|Gender=Masc|Number=Sing|Person=3|PrepCase=Npr|PronType=Prs]él <ora
idora="49" tp="vs">que[PRON_PronType=Int,Rel]que más[ADV_Degree=Cmp]más
te[PRON_Case=Acc,Dat|Number=Sing|Person=2|PrepCase=Npr|PronType=Prs]tú
gusta[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin]gustar
</ora></fn></ora></fn></ora></ora></ora>.[PUNCT_PunctType=Peri]. <ora idora="50" tp="vm"><ora idora="51"
tp="vs">Creo[VERB_Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin]creer <ora idora="52"
tp="vs">que[SCONJ_]que <fn cs="su" esin="3_INACTIVO_RD" idfn="117"
refe="80">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el CICESE[PROPN_]CICESE
</fn>combina[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin]combinar muy[ADV_]mucho
bien[ADV_]bien <fn cs="od" esin="7_ACTIVO_O" idfn="118"
refe="98">estas[DET_Gender=Fem|Number=Plur|PronType=Dem]este cualidades[NOUN_Gender=Fem|Number=Plur]cualidad
</fn></ora></ora>.[PUNCT_PunctType=Peri]. <ora idora="53" tp="vs">Nunca[ADV_]nunca
terminaré[VERB_Mood=Ind|Number=Sing|Person=1|Tense=Fut|VerbForm=Fin]terminar de[ADP_AdpType=Prep]de
```

```
agradecer[VERB_VerbForm=Inf]agradecer le[PRON_Case=Dat|Number=Sing|Person=3|PronType=Prs]él
,[PUNCT_PunctType=Comm], a[ADP_AdpType=Prep]a <fn cs="oi" esin="4_ACCESIBLE_MARCO"
idfn="119">su[DET_Number=Sing|Person=3|Poss=Yes|PronType=Prs]su
personal[NOUN_Gender=Masc|Number=Sing]personal </fn>y[CCONJ_]y a[ADP_AdpType=Prep]a <fn cs="oi"
esin="1_NO_IDENT" idfn="120">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
sociedad[NOUN_Gender=Fem|Number=Sing]sociedad misma[ADJ_Gender=Fem|Number=Sing]mismo
,[PUNCT_PunctType=Comm], <ora idora="54" tp="vs">que[PRON_PronType=Int,Rel]que
es[AUX_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin]ser <fn cs="at" esin="9_IDENT_BAJA"
idfn="121">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el <ora idora="55"
tp="vs">que[PRON_PronType=Int,Rel]que
nos[PRON_Case=Acc,Dat|Number=Plur|Person=1|PrepCase=Npr|PronType=Prs]yo
da[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin]dar <fn cs="od" esin="1_NO_IDENT"
idfn="122">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
oportunidad[NOUN_Gender=Fem|Number=Sing]oportunidad <ora idora="56" tp="vs">de[ADP_AdpType=Prep]de
hacer[VERB_VerbForm=Inf]hacer esto[PRON_Number=Sing|PronType=Dem]este
</ora></fn></ora></fn>,[PUNCT_PunctType=Comm], <fn cs="at" esin="9_IDENT_BAJA"
idfn="123">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el <ora idora="57"
tp="vs">que[PRON_PronType=Int,Rel]que
nos[PRON_Case=Acc,Dat|Number=Plur|Person=1|PrepCase=Npr|PronType=Prs]yo
brinda[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin]brindar <fn cs="od" esin="1_NO_IDENT"
idfn="124">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
chance[NOUN_Gender=Masc|Number=Sing]chance <ora idora="58" tp="vs">de[ADP_AdpType=Prep]de
estar[VERB_VerbForm=Inf]estar en[ADP_AdpType=Prep]en <fn cs="pr" esin="3_INACTIVO_RD"
idfn="125">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el Verano[PROPN_]Verano
de[ADP_AdpType=Prep]de Investigación[PROPN_]Investigación </fn>y[CCONJ_]y en[ADP_AdpType=Prep]en <fn cs="pr"
esin="1_NO_IDENT" idfn="126">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
posgrado[NOUN_Gender=Masc|Number=Sing]posgrado de[ADP_AdpType=Prep]de <fn cs="na" esin="3_INACTIVO_RD"
idfn="127">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el CICESE[PROPN_]CICESE
</fn></fn></ora></fn></ora></fn></ora></fn></ora>”[PUNCT_PunctType=Quot]” ,[PUNCT_PunctType=Comm],
finalizó[VERB_Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin]finalizar </ora>.[PUNCT_PunctType=Peri].
<ora idora="59" tp="vm"><ora idora="60" tp="vs">Para[ADP_AdpType=Prep]para
complementar[VERB_VerbForm=Inf]complementar <fn cs="od" esin="3_INACTIVO_RD" idfn="128"
refe="54">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
bienvenida[NOUN_Gender=Fem|Number=Sing]bienvenida </fn></ora>,[PUNCT_PunctType=Comm], <fn cs="su"
esin="1_NO_IDENT" idfn="129">los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
estudiantes[NOUN_Number=Plur]estudiante de[ADP_AdpType=Prep]de
```

```
posgrado[NOUN_Gender=Masc|Number=Sing]posgrado <fn cs="na" esin="9_IDENT_BAJA" idfn="130">José[PROPN_]José
Ricardo[PROPN_]Ricardo Santillán[PROPN_]Santillán </fn>,[PUNCT_PunctType=Comm], <fn cs="na"
esin="9_IDENT_BAJA" idfn="131">Danna[PROPN_]Danna Lyn[PROPN_]Lyn Arellano[PROPN_]Arellano
</fn>,[PUNCT_PunctType=Comm], <fn cs="na" esin="9_IDENT_BAJA" idfn="132">Gabriela[PROPN_]Gabriela
Reséndiz[PROPN_]Reséndiz </fn>y[CCONJ_]y <fn cs="na" esin="9_IDENT_BAJA" idfn="133">Georgina[PROPN_]Georgina
Rojo[PROPN_]Rojo de[ADP_AdpType=Prep]de Anda[PROPN_]Anda </fn></fn>,[PUNCT_PunctType=Comm],
compartieron[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin]compartir <fn cs="od"
esin="1_NO_IDENT" idfn="134">los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
proyectos[NOUN_Gender=Masc|Number=Plur]proyecto de[ADP_AdpType=Prep]de <fn cs="na" esin="0_NO_IDENT_BAJA"
idfn="135">tesis[NOUN_Gender=Fem]tesis </fn><ora idora="61" tp="vs">que[PRON_PronType=Int,Rel]que
actualmente[ADV_]actualmente
desarrollan[VERB_Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin]desarrollar en[ADP_AdpType=Prep]en
<fn cs="pr" esin="1_NO_IDENT" idfn="136">los[DET_Definite=Def|Gender=Masc|Number=Plur|PronType=Art]el
posgrados[NOUN_Gender=Masc|Number=Plur]posgrado de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
idfn="137">Óptica[NOUN_Gender=Fem|Number=Sing]Óptica </fn>,[PUNCT_PunctType=Comm], <fn cs="na"
esin="1_NO_IDENT" idfn="138">Ciencias[PROPN_]Ciencias de[ADP_AdpType=Prep]de
la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el Vida[PROPN_]Vida </fn>,[PUNCT_PunctType=Comm],
<fn cs="na" esin="1_NO_IDENT" idfn="139">Ecología[PROPN_]Ecología Marina[PROPN_]Marina </fn>y[CCONJ_]y <fn
cs="na" esin="1_NO_IDENT" idfn="140">Oceanografía[PROPN_]Oceanografía Física[PROPN_]Física
</fn></fn></ora></fn>,[PUNCT_PunctType=Comm], respectivamente[ADV_]respectivamente ,[PUNCT_PunctType=Comm],
con[ADP_AdpType=Prep]con <fn cs="pr" esin="1_NO_IDENT"
idfn="141">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el
afán[NOUN_Gender=Masc|Number=Sing]afán <ora idora="62" tp="vs">de[ADP_AdpType=Prep]de
expresar[VERB_VerbForm=Inf]expresar <fn cs="od" esin="1_NO_IDENT"
idfn="142">la[DET_Definite=Def|Gender=Fem|Number=Sing|PronType=Art]el
calidad[NOUN_Gender=Fem|Number=Sing]calidad de[ADP_AdpType=Prep]de <fn cs="na" esin="1_NO_IDENT"
idfn="143">investigación[NOUN_Gender=Fem|Number=Sing]investigación <ora idora="63"
tp="vs">que[PRON_PronType=Int,Rel]que es[AUX_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin]ser
posible[ADJ_Number=Sing]posible <ora idora="64" tp="vs">desarrollar[VERB_VerbForm=Inf]desarrollar
en[ADP_AdpType=Prep]en <fn cs="pr" esin="3_INACTIVO_RD"
idfn="144">el[DET_Definite=Def|Gender=Masc|Number=Sing|PronType=Art]el CICESE[PROPN_]CICESE
</fn></ora></ora></fn></fn></ora></fn></ora>.[PUNCT_PunctType=Peri].

      </etiquetado>

</nota>
```

Anexo C. Medios de comunicación en COPENOR

ID	ESTADO	CIUDAD	MEDIO	URL
M001	Baja California	Ensenada	El Vigía	https://www.elvigia.net/
M002	Baja California	Ensenada	Ensenada.net	http://www.ensenada.net/
M003	Baja California	Mexicali	Enlace Informativo	https://enlaceinformativo.net/
M004	Baja California	Mexicali	El Imparcial	https://www.elimparcial.com/mexicali/
M005	Baja California	Mexicali	La Voz de la Frontera	https://www.lavozdelafrontera.com.mx/
M006	Baja California	Mexicali	Monitor Económico	http://monitoreconomico.org/noticias/
M007	Baja California	Tijuana	Agencia Fronteriza de Noticias	http://www.afntijuana.info/
M008	Baja California	Tijuana	El Mexicano	http://www.el-mexicano.com.mx/inicio.htm
M009	Baja California	Tijuana	El Sol de Tijuana	https://www.elsoldetijuana.com.mx/
M010	Baja California	Tijuana	Monitor BC	http://www.monitorbc.info/
M011	Baja California	Tijuana	Tijuana Press	https://tijuanapress.com/
M012	Baja California	Tijuana	Uniradio Informa	https://www.uniradioinforma.com/noticias/bajacalifornia
M013	Baja California	Tijuana	Televisa Californias	http://xewt12.com/category/noticias/
M014	Baja California	Tijuana	Periódico Baja California	http://www.periodicobajacalifornia.info/
M015	Baja California	Tijuana	Síntesis	http://sintesistv.com.mx/noticias/
M016	Baja California	Tijuana	La Jornada BC	http://jornadabc.mx/
M017	Baja California	Tijuana	Zeta	https://zetatijuana.com/
M018	Baja California Sur	La Paz	BCS Noticias	http://www.bcsnoticias.mx/
M019	Baja California Sur	La Paz	El Peninsular	http://peninsulardigital.com/
M020	Baja California Sur	La Paz	El Sudcalifornia no	https://www.elsudcalifornia no.com.mx/
M021	Baja California Sur	La Paz	Raíces	http://revistaraicesbcs.com/
M022	Baja California Sur	La Paz	Cabovision	http://cabovision.tv/index.php/noticias-m
M023	Baja California Sur	San José del Cabo	El Independiente	https://www.diarioelindependiente.mx/
M024	Baja California Sur	San José del Cabo	Tribuna de los Cabos	https://www.tribunadeloscabos.com.mx/
M025	Chihuahua	Camargo	Impacto	http://impactonoticias.com.mx/
M026	Chihuahua	Chihuahua	Acento Noticias	https://www.acento.com.mx/

M027	Chihuahua	Chihuahua	Al Contacto	https://www.alcontacto.com.mx/
M028	Chihuahua	Chihuahua	Cambio 16	http://www.cambio16.gob.mx/
M029	Chihuahua	Chihuahua	El Ágora	http://www.elagora.com.mx/
M030	Chihuahua	Chihuahua	El Diario	https://www.eldiariodechihuahua.mx/
M031	Chihuahua	Chihuahua	El Digital	https://www.eldigital.com.mx/
M032	Chihuahua	Chihuahua	El Heraldo de Chihuahua	https://www.elheraldodechihuahua.com.mx/
M033	Chihuahua	Chihuahua	El Pueblo	http://elpueblo.com/
M034	Chihuahua	Chihuahua	Entre Líneas	http://entrelineas.com.mx/category/local/
M035	Chihuahua	Chihuahua	La Crónica de Chihuahua	http://www.cronicadechihuahua.com/
M036	Chihuahua	Chihuahua	La Crónica de Hoy Chihuahua	http://www.omnia.com.mx/
M037	Chihuahua	Chihuahua	La Opción	http://laopcion.com.mx/
M038	Chihuahua	Chihuahua	La Parada Digital	https://www.laparadadigital.com/
M039	Chihuahua	Chihuahua	Segundo a Segundo	http://segundoasegundo.com/chihuahua/
M040	Chihuahua	Chihuahua	Tiempo	http://tiempo.com.mx/seccion/local/
M041	Chihuahua	Ciudad Juárez	Canal 44	https://canal44.com/
M042	Chihuahua	Ciudad Juárez	El Fronterizo	http://www.elfronterizo.com.mx/categoria/local
M043	Chihuahua	Ciudad Juárez	El Mexicano	https://www.periodicoelmexicano.com.mx/
M044	Chihuahua	Ciudad Juárez	Frontenet	http://www.frontenet.com/_blog/cat/local
M045	Chihuahua	Ciudad Juárez	Hoy	http://www.juarezhoy.com.mx/index.php/juarez
M046	Chihuahua	Ciudad Juárez	Juárez a Diario	https://www.juarezadiario.com/secciones/juarez/
M047	Chihuahua	Ciudad Juárez	Juárez Noticias	http://juareznoticias.com/
M048	Chihuahua	Ciudad Juárez	La Polaka	https://www.lapolaka.com/category/juarez/
M049	Chihuahua	Ciudad Juárez	La Red Noticias	https://larednoticias.com/
M050	Chihuahua	Ciudad Juárez	Net Noticias	https://netnoticias.mx/
M051	Chihuahua	Ciudad Juárez	Norte de Ciudad Juárez	https://nortedigital.mx/
M052	Chihuahua	Ciudad Juárez	Puente Libre	http://puentelibre.mx/
M053	Chihuahua	Ciudad Juárez	Televisa Juárez	http://televisajuarez.tv/seccion/noticias-locales
M054	Chihuahua	Ciudad Juárez	Vivir en Juárez	http://vivirenjuarez.com.mx/
M055	Chihuahua	Hidalgo del Parral	El Monitor	http://www.elmonitorparral.com/

M056	Chihuahua	Hidalgo del Parral	El Sol de Parral	https://www.elsoldeparral.com.mx/
M057	Chihuahua	Nuevo Casas Grandes	Akro Noticias	https://www.akronoticias.com/category/principal
M058	Durango	Durango	Canal 10	http://canal10.com.mx/#/
M059	Durango	Durango	Contacto Hoy	https://contactohoy.com.mx/
M060	Durango	Durango	Contexto de Durango	https://contextodedurango.com.mx/noticias/
M061	Durango	Durango	Durango al Día	http://www.durangoaldia.com/
M062	Durango	Durango	El Siglo de Durango	https://www.elsiglodedurango.com.mx/
M063	Durango	Durango	El Sol de Durango	www.elsoldedurango.com.mx
M064	Durango	Durango	La Voz de Durango	http://lavozdgo.com/
M065	Durango	Durango	La Neta	http://lanetadurango.com/
M066	Durango	Durango	Órale, qué chiquito!	http://durango.notigram.com/
M067	Durango	Durango	Victoria de Durango	http://periodicovictoria.mx/
M068	Sinaloa	Culiacán	El Debate	https://www.debate.com.mx/seccion/culiacan/
M069	Sinaloa	Culiacán	El Sol de Sinaloa	https://www.elsoldesinaloa.com.mx/
M070	Sinaloa	Culiacán	Noroeste	https://www.noroeste.com.mx/
M071	Sinaloa	Culiacán	Viva Voz	http://www.vivavoz.com.mx/portal/
M072	Sinaloa	Culiacán	Viva la Noticia	https://vivalanoticia.com/category/sinaloa-centro/
M073	Sinaloa	Mazatlán	El Sol de Mazatlan	https://www.elsoldemazatlan.com.mx/
M074	Sonora	Caborca	Caborca Noticias	http://www.caborcanoticias.com/noticias/
M075	Sonora	Ciudad Obregón	Diario del Yaqui	https://diariodelyaqui.mx/
M076	Sonora	Ciudad Obregón	Tribuna	https://www.tribuna.com.mx/seccion/sonora/
M077	Sonora	Ciudad Obregón	Medios Obson	https://www.mediosobson.com/
M078	Sonora	Guaymas	El Autónomo	http://www.elautonomo.mx/index.php/regionales
M079	Sonora	Guaymas	El Vigía	http://187.243.249.242/
M080	Sonora	Hermosillo	Crítica	https://www.critica.com.mx/
M081	Sonora	Hermosillo	Dossier	http://www.dossierpolitico.com/seccion.php?categoria=1
M082	Sonora	Hermosillo	El Imparcial	https://www.elimparcial.com/sonora/seccion/hermosillo/
M083	Sonora	Hermosillo	El Sol de Hermosillo	https://www.elsoldehermosillo.com.mx/
M084	Sonora	Hermosillo	Entorno Informativo	http://www.entornoinformativo.com.mx/

M085	Sonora	Hermosillo	Expreso	https://www.expreso.com.mx/
M086	Sonora	Hermosillo	Uniradio Noticias	https://www.uniradionoticias.com/noticias/sonora
M087	Sonora	Hermosillo	Televisa Sonora	http://televisasonora.tv/seccion/noticias-locales
M088	Sonora	Navojoa	La Verdad	http://www.diariolaverdad.mx/
M089	Sonora	Nogales	El Diario de Sonora	http://www.eldiariodesonora.com.mx/?r=1&Ancho=1088
M090	Sonora	Nogales	El Observador Mexico	https://elobservadormexico.com/category/estatal/
M091	Sonora	Nogales	Nuevo Día	https://www.nuevodia.mx/secciones.php?cat=62
M092	Sonora	Puerto Peñasco	Número Uno	http://www.numerounoonline.com/new/index.php/noticias/sonora
M093	Sonora	San Luis Río Colorado	Tribuna de San Luis	https://www.tribunadesanluis.com.mx/tags/temas/sanluisriocolorado
M094	Sonora	San Luis Río Colorado	Noticias	https://www.diarionoticias.info/

Anexo D. Itinerario de captura de las notas

FECHA CAPTURA	IDN	ESTADO	ID	CIUDAD	MEDIO
2019-05-23	001CH	**Chihuahua**	M049	Ciudad Juárez	La Red Noticias
	002BC	**Baja California**	M017	Tijuana	Zeta
	003CH	**Chihuahua**	M048	Ciudad Juárez	La Polaka
	004CH	**Chihuahua**	M031	Chihuahua	El Digital
	005DU	**Durango**	M059	Durango	Contacto Hoy
	006DU	**Durango**	M065	Durango	La Neta
	007CH	**Chihuahua**	M057	Nuevo Casas Grandes	Akro Noticias
2019-05-24	008SN	**Sinaloa**	M069	Culiacán	El Sol de Sinaloa
	009CH	**Chihuahua**	M034	Chihuahua	Entre Líneas
	010CH	**Chihuahua**	M052	Ciudad Juárez	Puente Libre
	011SO	**Sonora**	M079	Guaymas	El Vigía
	012DU	**Durango**	M064	Durango	La Voz de Durango
	013BC	**Baja California**	M012	Tijuana	Uniradio Informa
	014BS	**Baja California Sur**	M023	San José del Cabo	El Independiente
2019-05-25	015BC	**Baja California**	M008	Tijuana	El Mexicano
	016SN	**Sinaloa**	M070	Culiacán	Noroeste
	017CH	**Chihuahua**	M036	Chihuahua	La Crónica de Hoy Chihuahua
	018SO	**Sonora**	M083	Hermosillo	El Sol de Hermosillo
	019BS	**Baja California Sur**	M018	La Paz	BCS Noticias
	020BC	**Baja California**	M012	Tijuana	Uniradio Informa
	021CH	**Chihuahua**	M033	Chihuahua	El Pueblo
2019-05-26	022SO	**Sonora**	M080	Hermosillo	Crítica
	023SO	**Sonora**	M084	Hermosillo	Entorno Informativo
	024BS	**Baja California Sur**	M022	La Paz	Cabovision
	025BC	**Baja California**	M004	Mexicali	El Imparcial
	026BC	**Baja California**	M007	Tijuana	Agencia Fronteriza de Noticias
	027BC	**Baja California**	M016	Tijuana	La Jornada BC
	028CH	**Chihuahua**	M048	Ciudad Juárez	La Polaka
2019-05-27	029SO	**Sonora**	M091	Nogales	Nuevo Día
	030DU	**Durango**	M060	Durango	Contexto de Durango
	031DU	**Durango**	M066	Durango	Órale, qué chiquito!
	032BC	**Baja California**	M008	Tijuana	El Mexicano
	033DU	**Durango**	M063	Durango	El Sol de Durango
	034CH	**Chihuahua**	M033	Chihuahua	El Pueblo
	035SN	**Sinaloa**	M072	Culiacán	Viva la Noticia
2019-05-28	036SO	**Sonora**	M090	Nogales	El Observador México
	037CH	**Chihuahua**	M049	Ciudad Juárez	La Red Noticias

	038CH	**Chihuahua**	M038	Chihuahua	La Parada Digital
	039BS	**Baja California Sur**	M019	La Paz	El Peninsular
	040CH	**Chihuahua**	M036	Chihuahua	La Crónica de Hoy Chihuahua
	041CH	**Chihuahua**	M031	Chihuahua	El Digital
	042BC	**Baja California**	M002	Ensenada	Ensenada.net
2019-05-29	043CH	**Chihuahua**	M036	Chihuahua	La Crónica de Hoy Chihuahua
	044CH	**Chihuahua**	M052	Ciudad Juárez	Puente Libre
	045CH	**Chihuahua**	M055	Hidalgo del Parral	El Monitor
	046SO	**Sonora**	M081	Hermosillo	Dossier
	047BC	**Baja California**	M014	Tijuana	Periódico Baja California
	048BC	**Baja California**	M011	Tijuana	Tijuana Press
	049SO	**Sonora**	M088	Navojoa	La Verdad
2019-05-30	050SO	**Sonora**	M090	Nogales	El Observador México
	051SO	**Sonora**	M093	San Luis Río Colorado	Tribuna de San Luis
	052SO	**Sonora**	M085	Hermosillo	Expreso
	053CH	**Chihuahua**	M029	Chihuahua	El Ágora
	054CH	**Chihuahua**	M039	Chihuahua	Segundo a Segundo
	055BC	**Baja California**	M017	Tijuana	Zeta
	056DU	**Durango**	M059	Durango	Contacto Hoy
2019-05-31	057BC	**Baja California**	M009	Tijuana	El Sol de Tijuana
	058SO	**Sonora**	M093	San Luis Río Colorado	Tribuna de San Luis
	059BC	**Baja California**	M008	Tijuana	El Mexicano
	060SO	**Sonora**	M077	Ciudad Obregón	Medios Obson
	061CH	**Chihuahua**	M047	Ciudad Juárez	Juárez Noticias
	062SO	**Sonora**	M091	Nogales	Nuevo Día
	063BC	**Baja California**	M007	Tijuana	Agencia Fronteriza de Noticias
2019-06-01	064BS	**Baja California Sur**	M022	La Paz	Cabovision
	065CH	**Chihuahua**	M030	Chihuahua	El Diario
	066CH	**Chihuahua**	M037	Chihuahua	La Opción
	067CH	**Chihuahua**	M048	Ciudad Juárez	La Polaka
	068BC	**Baja California**	M011	Tijuana	Tijuana Press
	069SN	**Sinaloa**	M068	Culiacán	El Debate
	070BC	**Baja California**	M011	Tijuana	Tijuana Press
2019-06-02	071SO	**Sonora**	M077	Ciudad Obregón	Medios Obson
	072CH	**Chihuahua**	M030	Chihuahua	El Diario
	073SO	**Sonora**	M082	Hermosillo	El Imparcial
	074SO	**Sonora**	M079	Guaymas	El Vigía
	075SN	**Sinaloa**	M071	Culiacán	Viva Voz
	076CH	**Chihuahua**	M031	Chihuahua	El Digital
	077SO	**Sonora**	M081	Hermosillo	Dossier
2019-06-03	078CH	**Chihuahua**	M038	Chihuahua	La Parada Digital
	079DU	**Durango**	M060	Durango	Contexto de Durango

	080BC	**Baja California**	M002	Ensenada	Ensenada.net
	081DU	**Durango**	M062	Durango	El Siglo de Durango
	082SO	**Sonora**	M093	San Luis Río Colorado	Tribuna de San Luis
	083CH	**Chihuahua**	M030	Chihuahua	El Diario
	084CH	**Chihuahua**	M044	Ciudad Juárez	Frontenet
2019-06-04	085BC	**Baja California**	M006	Mexicali	Monitor Económico
	086CH	**Chihuahua**	M047	Ciudad Juárez	Juárez Noticias
	087CH	**Chihuahua**	M039	Chihuahua	Segundo a Segundo
	088SN	**Sinaloa**	M072	Culiacán	Viva la Noticia
	089CH	**Chihuahua**	M041	Ciudad Juárez	Canal 44
	090CH	**Chihuahua**	M042	Ciudad Juárez	El Fronterizo
	091CH	**Chihuahua**	M046	Ciudad Juárez	Juárez a Diario
2019-06-05	092CH	**Chihuahua**	M042	Ciudad Juárez	El Fronterizo
	093SO	**Sonora**	M082	Hermosillo	El Imparcial
	094CH	**Chihuahua**	M035	Chihuahua	La Crónica de Chihuahua
	095SN	**Sinaloa**	M068	Culiacán	El Debate
	096CH	**Chihuahua**	M044	Ciudad Juárez	Frontenet
	097CH	**Chihuahua**	M028	Chihuahua	Cambio 16
	098CH	**Chihuahua**	M041	Ciudad Juárez	Canal 44
2019-06-06	099CH	**Chihuahua**	M047	Ciudad Juárez	Juárez Noticias
	100CH	**Chihuahua**	M043	Ciudad Juárez	El Mexicano
	101SO	**Sonora**	M086	Hermosillo	Uniradio Noticias
	102CH	**Chihuahua**	M047	Ciudad Juárez	Juárez Noticias
	103BC	**Baja California**	M003	Mexicali	Enlace Informativo
	104SO	**Sonora**	M076	Ciudad Obregón	Tribuna
	105SN	**Sinaloa**	M068	Culiacán	El Debate
2019-06-07	106DU	**Durango**	M067	Durango	Victoria de Durango
	107CH	**Chihuahua**	M046	Ciudad Juárez	Juárez a Diario
	108SN	**Sinaloa**	M072	Culiacán	Viva la Noticia
	109SO	**Sonora**	M083	Hermosillo	El Sol de Hermosillo
	110BC	**Baja California**	M007	Tijuana	Agencia Fronteriza de Noticias
	111DU	**Durango**	M064	Durango	La Voz de Durango
	112DU	**Durango**	M067	Durango	Victoria de Durango
2019-06-08	113CH	**Chihuahua**	M025	Camargo	Impacto
	114SO	**Sonora**	M089	Nogales	El Diario de Sonora
	115BC	**Baja California**	M014	Tijuana	Periódico Baja California
	116CH	**Chihuahua**	M056	Hidalgo del Parral	El Sol de Parral
	117SN	**Sinaloa**	M072	Culiacán	Viva la Noticia
	118DU	**Durango**	M066	Durango	Órale, qué chiquito!
	119CH	**Chihuahua**	M037	Chihuahua	La Opción
2019-06-09	120SN	**Sinaloa**	M071	Culiacán	Viva Voz
	121DU	**Durango**	M065	Durango	La Neta
	122SO	**Sonora**	M074	Caborca	Caborca Noticias

	123CH	**Chihuahua**	M028	Chihuahua	Cambio 16
	124BC	**Baja California**	M009	Tijuana	El Sol de Tijuana
	125BS	**Baja California Sur**	M021	La Paz	Raíces
	126SO	**Sonora**	M085	Hermosillo	Expreso
2019-06-10	127BC	**Baja California**	M007	Tijuana	Agencia Fronteriza de Noticias
	128SN	**Sinaloa**	M073	Mazatlán	El Sol de Mazatlan
	129SO	**Sonora**	M076	Ciudad Obregón	Tribuna
	130CH	**Chihuahua**	M034	Chihuahua	Entre Líneas
	131BC	**Baja California**	M015	Tijuana	Síntesis
	132BC	**Baja California**	M017	Tijuana	Zeta
	133SO	**Sonora**	M077	Ciudad Obregón	Medios Obson
2019-06-11	134SO	**Sonora**	M077	Ciudad Obregón	Medios Obson
	135BC	**Baja California**	M009	Tijuana	El Sol de Tijuana
	136CH	**Chihuahua**	M025	Camargo	Impacto
	137CH	**Chihuahua**	M040	Chihuahua	Tiempo
	138CH	**Chihuahua**	M046	Ciudad Juárez	Juárez a Diario
	139BS	**Baja California Sur**	M023	San José del Cabo	El Independiente
	140BC	**Baja California**	M010	Tijuana	Monitor BC
2019-06-12	141CH	**Chihuahua**	M035	Chihuahua	La Crónica de Chihuahua
	142CH	**Chihuahua**	M028	Chihuahua	Cambio 16
	143SO	**Sonora**	M086	Hermosillo	Uniradio Noticias
	144SO	**Sonora**	M089	Nogales	El Diario de Sonora
	145CH	**Chihuahua**	M048	Ciudad Juárez	La Polaka
	146CH	**Chihuahua**	M049	Ciudad Juárez	La Red Noticias
	147SO	**Sonora**	M089	Nogales	El Diario de Sonora
2019-06-13	148CH	**Chihuahua**	M052	Ciudad Juárez	Puente Libre
	149CH	**Chihuahua**	M047	Ciudad Juárez	Juárez Noticias
	150SO	**Sonora**	M088	Navojoa	La Verdad
	151SO	**Sonora**	M084	Hermosillo	Entorno Informativo
	152BC	**Baja California**	M005	Mexicali	La Voz de la Frontera
	153CH	**Chihuahua**	M028	Chihuahua	Cambio 16
	154BS	**Baja California Sur**	M022	La Paz	Cabovision
2019-06-14	155BC	**Baja California**	M012	Tijuana	Uniradio Informa
	156SN	**Sinaloa**	M068	Culiacán	El Debate
	157SO	**Sonora**	M086	Hermosillo	Uniradio Noticias
	158CH	**Chihuahua**	M044	Ciudad Juárez	Frontenet
	159SO	**Sonora**	M094	San Luis Río Colorado	Noticias
	160BC	**Baja California**	M013	Tijuana	Televisa Californias
	161CH	**Chihuahua**	M025	Camargo	Impacto
2019-06-15	162CH	**Chihuahua**	M025	Camargo	Impacto
	163SO	**Sonora**	M086	Hermosillo	Uniradio Noticias

	164CH	**Chihuahua**	M045	Ciudad Juárez	Hoy
	165BC	**Baja California**	M007	Tijuana	Agencia Fronteriza de Noticias
	166CH	**Chihuahua**	M055	Hidalgo del Parral	El Monitor
	167CH	**Chihuahua**	M055	Hidalgo del Parral	El Monitor
	168CH	**Chihuahua**	M056	Hidalgo del Parral	El Sol de Parral
2019-06-16	169SO	**Sonora**	M093	San Luis Río Colorado	Tribuna de San Luis
	170BS	**Baja California Sur**	M021	La Paz	Raíces
	171SO	**Sonora**	M081	Hermosillo	Dossier
	172CH	**Chihuahua**	M026	Chihuahua	Acento Noticias
	173SO	**Sonora**	M081	Hermosillo	Dossier
	174CH	**Chihuahua**	M031	Chihuahua	El Digital
	175BS	**Baja California Sur**	M022	La Paz	Cabovision
2019-06-17	176BC	**Baja California**	M014	Tijuana	Periódico Baja California
	177BC	**Baja California**	M006	Mexicali	Monitor Económico
	178CH	**Chihuahua**	M035	Chihuahua	La Crónica de Chihuahua
	179BC	**Baja California**	M012	Tijuana	Uniradio Informa
	180CH	**Chihuahua**	M042	Ciudad Juárez	El Fronterizo
	181DU	**Durango**	M067	Durango	Victoria de Durango
	182CH	**Chihuahua**	M033	Chihuahua	El Pueblo
2019-06-18	183CH	**Chihuahua**	M057	Nuevo Casas Grandes	Akro Noticias
	184BC	**Baja California**	M014	Tijuana	Periódico Baja California
	185BC	**Baja California**	M007	Tijuana	Agencia Fronteriza de Noticias
	186CH	**Chihuahua**	M045	Ciudad Juárez	Hoy
	187CH	**Chihuahua**	M035	Chihuahua	La Crónica de Chihuahua
	188SO	**Sonora**	M090	Nogales	El Observador México
	189SO	**Sonora**	M077	Ciudad Obregón	Medios Obson
2019-06-19	190SO	**Sonora**	M080	Hermosillo	Crítica
	191CH	**Chihuahua**	M047	Ciudad Juárez	Juárez Noticias
	192BS	**Baja California Sur**	M019	La Paz	El Peninsular
	193CH	**Chihuahua**	M050	Ciudad Juárez	Net Noticias
	194BC	**Baja California**	M016	Tijuana	La Jornada BC
	195CH	**Chihuahua**	M026	Chihuahua	Acento Noticias
	196SO	**Sonora**	M084	Hermosillo	Entorno Informativo
2019-06-20	197BC	**Baja California**	M003	Mexicali	Enlace Informativo
	198BC	**Baja California**	M017	Tijuana	Zeta
	199CH	**Chihuahua**	M052	Ciudad Juárez	Puente Libre
	200SO	**Sonora**	M080	Hermosillo	Crítica
	201SO	**Sonora**	M075	Ciudad Obregón	Diario del Yaqui
	202SO	**Sonora**	M088	Navojoa	La Verdad
	203BC	**Baja California**	M010	Tijuana	Monitor BC
2019-06-21	204SO	**Sonora**	M086	Hermosillo	Uniradio Noticias

	205BC	**Baja California**	M014	Tijuana	Periódico Baja California
	206BC	**Baja California**	M012	Tijuana	Uniradio Informa
	207SO	**Sonora**	M089	Nogales	El Diario de Sonora
	208CH	**Chihuahua**	M038	Chihuahua	La Parada Digital
	209CH	**Chihuahua**	M037	Chihuahua	La Opción
	210SO	**Sonora**	M076	Ciudad Obregón	Tribuna
2019-06-22	211CH	**Chihuahua**	M047	Ciudad Juárez	Juárez Noticias
	212SO	**Sonora**	M075	Ciudad Obregón	Diario del Yaqui
	213DU	**Durango**	M065	Durango	La Neta
	214SO	**Sonora**	M074	Caborca	Caborca Noticias
	215CH	**Chihuahua**	M029	Chihuahua	El Ágora
	216SO	**Sonora**	M077	Ciudad Obregón	Medios Obson
	217CH	**Chihuahua**	M041	Ciudad Juárez	Canal 44
2019-06-23	218SN	**Sinaloa**	M073	Mazatlán	El Sol de Mazatlan
	219CH	**Chihuahua**	M052	Ciudad Juárez	Puente Libre
	220SO	**Sonora**	M075	Ciudad Obregón	Diario del Yaqui
	221CH	**Chihuahua**	M047	Ciudad Juárez	Juárez Noticias
	222CH	**Chihuahua**	M026	Chihuahua	Acento Noticias
	223CH	**Chihuahua**	M037	Chihuahua	La Opción
	224BC	**Baja California**	M015	Tijuana	Síntesis
2019-06-24	225CH	**Chihuahua**	M034	Chihuahua	Entre Líneas
	226SO	**Sonora**	M089	Nogales	El Diario de Sonora
	227CH	**Chihuahua**	M051	Ciudad Juárez	Norte de Ciudad Juárez
	228SO	**Sonora**	M078	Guaymas	El Autónomo
	229CH	**Chihuahua**	M052	Ciudad Juárez	Puente Libre
	230CH	**Chihuahua**	M032	Chihuahua	El Heraldo de Chihuahua
	231SO	**Sonora**	M093	San Luis Río Colorado	Tribuna de San Luis
2019-06-25	232CH	**Chihuahua**	M048	Ciudad Juárez	La Polaka
	233CH	**Chihuahua**	M033	Chihuahua	El Pueblo
	234SN	**Sinaloa**	M068	Culiacán	El Debate
	235SO	**Sonora**	M083	Hermosillo	El Sol de Hermosillo
	236BS	**Baja California Sur**	M023	San José del Cabo	El Independiente
	237CH	**Chihuahua**	M038	Chihuahua	La Parada Digital
	238CH	**Chihuahua**	M039	Chihuahua	Segundo a Segundo
2019-06-26	239BC	**Baja California**	M013	Tijuana	Televisa Californias
	240BS	**Baja California Sur**	M022	La Paz	Cabovision
	241CH	**Chihuahua**	M033	Chihuahua	El Pueblo
	242CH	**Chihuahua**	M034	Chihuahua	Entre Líneas
	243CH	**Chihuahua**	M034	Chihuahua	Entre Líneas
	244SO	**Sonora**	M088	Navojoa	La Verdad
	245SO	**Sonora**	M078	Guaymas	El Autónomo
2019-06-27	246BC	**Baja California**	M002	Ensenada	Ensenada.net

	247BS	**Baja California Sur**	M022	La Paz	Cabovision
	248BS	**Baja California Sur**	M022	La Paz	Cabovision
	249CH	**Chihuahua**	M040	Chihuahua	Tiempo
	250SO	**Sonora**	M079	Guaymas	El Vigía
	251DU	**Durango**	M062	Durango	El Siglo de Durango
	252BC	**Baja California**	M015	Tijuana	Síntesis
2019-06-28	253BC	**Baja California**	M002	Ensenada	Ensenada.net
	254BS	**Baja California Sur**	M018	La Paz	BCS Noticias
	255CH	**Chihuahua**	M036	Chihuahua	La Crónica de Hoy Chihuahua
	256CH	**Chihuahua**	M025	Camargo	Impacto
	257CH	**Chihuahua**	M043	Ciudad Juárez	El Mexicano
	258CH	**Chihuahua**	M057	Nuevo Casas Grandes	Akro Noticias
	259CH	**Chihuahua**	M050	Ciudad Juárez	Net Noticias
2019-06-29	260CH	**Chihuahua**	M055	Hidalgo del Parral	El Monitor
	261CH	**Chihuahua**	M030	Chihuahua	El Diario
	262CH	**Chihuahua**	M049	Ciudad Juárez	La Red Noticias
	263SO	**Sonora**	M088	Navojoa	La Verdad
	264BS	**Baja California Sur**	M023	San José del Cabo	El Independiente
	265BC	**Baja California**	M004	Mexicali	El Imparcial
	266DU	**Durango**	M059	Durango	Contacto Hoy
2019-06-30	267BC	**Baja California**	M008	Tijuana	El Mexicano
	268BC	**Baja California**	M006	Mexicali	Monitor Económico
	269BC	**Baja California**	M015	Tijuana	Síntesis
	270SN	**Sinaloa**	M070	Culiacán	Noroeste
	271CH	**Chihuahua**	M039	Chihuahua	Segundo a Segundo
	272BS	**Baja California Sur**	M020	La Paz	El Sudcaliforniano
	273CH	**Chihuahua**	M038	Chihuahua	La Parada Digital
2019-07-01	274CH	**Chihuahua**	M056	Hidalgo del Parral	El Sol de Parral
	275SO	**Sonora**	M094	San Luis Río Colorado	Noticias
	276CH	**Chihuahua**	M057	Nuevo Casas Grandes	Akro Noticias
	277BC	**Baja California**	M013	Tijuana	Televisa Californias
	278CH	**Chihuahua**	M037	Chihuahua	La Opción
	279DU	**Durango**	M062	Durango	El Siglo de Durango
	280BC	**Baja California**	M015	Tijuana	Síntesis
2019-07-02	281SN	**Sinaloa**	M068	Culiacán	El Debate
	282DU	**Durango**	M061	Durango	Durango al Día
	283BC	**Baja California**	M007	Tijuana	Agencia Fronteriza de Noticias
	284DU	**Durango**	M062	Durango	El Siglo de Durango
	285BS	**Baja California Sur**	M023	San José del Cabo	El Independiente
	286BC	**Baja California**	M017	Tijuana	Zeta

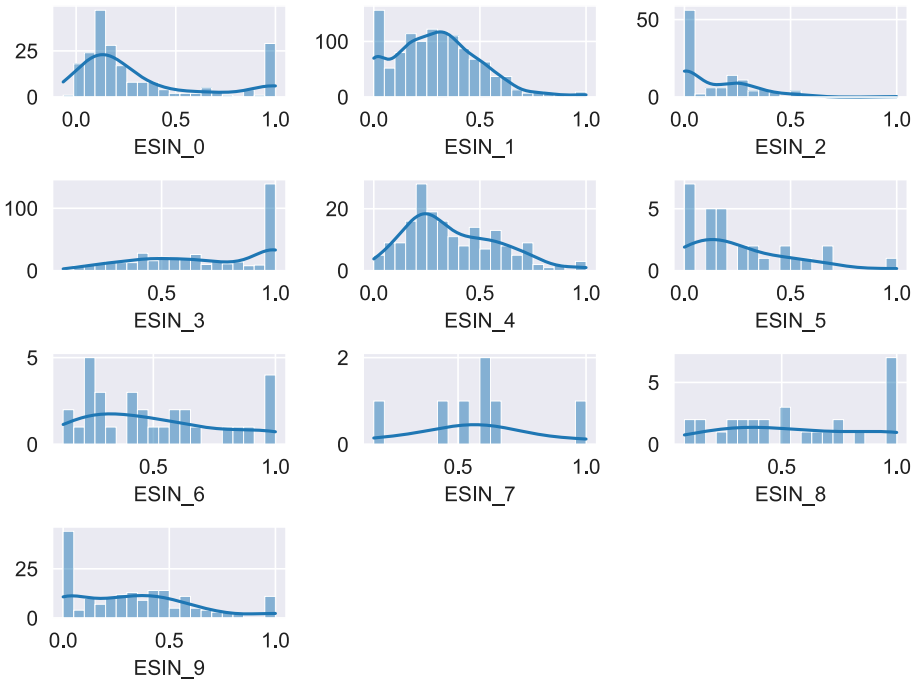
	287SO	**Sonora**	M075	Ciudad Obregón	Diario del Yaqui
2019-07-03	288CH	**Chihuahua**	M036	Chihuahua	La Crónica de Hoy Chihuahua
	289CH	**Chihuahua**	M056	Hidalgo del Parral	El Sol de Parral
	290CH	**Chihuahua**	M027	Chihuahua	Al Contacto
	291CH	**Chihuahua**	M036	Chihuahua	La Crónica de Hoy Chihuahua
	292BC	**Baja California**	M005	Mexicali	La Voz de la Frontera
	293CH	**Chihuahua**	M051	Ciudad Juárez	Norte de Ciudad Juárez
	294SN	**Sinaloa**	M072	Culiacán	Viva la Noticia
2019-07-04	295CH	**Chihuahua**	M028	Chihuahua	Cambio 16
	296BC	**Baja California**	M006	Mexicali	Monitor Económico
	297SO	**Sonora**	M090	Nogales	El Observador México
	298DU	**Durango**	M060	Durango	Contexto de Durango
	299DU	**Durango**	M062	Durango	El Siglo de Durango
	300CH	**Chihuahua**	M032	Chihuahua	El Heraldo de Chihuahua
	301BC	**Baja California**	M005	Mexicali	La Voz de la Frontera
2019-07-05	302BC	**Baja California**	M017	Tijuana	Zeta
	303CH	**Chihuahua**	M049	Ciudad Juárez	La Red Noticias
	304BC	**Baja California**	M005	Mexicali	La Voz de la Frontera
	305BC	**Baja California**	M016	Tijuana	La Jornada BC
	306DU	**Durango**	M065	Durango	La Neta
	307SO	**Sonora**	M086	Hermosillo	Uniradio Noticias
	308SO	**Sonora**	M086	Hermosillo	Uniradio Noticias
2019-07-06	309DU	**Durango**	M064	Durango	La Voz de Durango
	310SN	**Sinaloa**	M068	Culiacán	El Debate
	311BC	**Baja California**	M013	Tijuana	Televisa Californias
	312BC	**Baja California**	M012	Tijuana	Uniradio Informa
	313SO	**Sonora**	M084	Hermosillo	Entorno Informativo
	314CH	**Chihuahua**	M041	Ciudad Juárez	Canal 44
	315SN	**Sinaloa**	M073	Mazatlán	El Sol de Mazatlan
2019-07-07	316SO	**Sonora**	M077	Ciudad Obregón	Medios Obson
	317SO	**Sonora**	M090	Nogales	El Observador México
	318CH	**Chihuahua**	M047	Ciudad Juárez	Juárez Noticias
	319DU	**Durango**	M065	Durango	La Neta
	320SO	**Sonora**	M078	Guaymas	El Autónomo
	321CH	**Chihuahua**	M028	Chihuahua	Cambio 16
	322CH	**Chihuahua**	M036	Chihuahua	La Crónica de Hoy Chihuahua
2019-07-08	323DU	**Durango**	M059	Durango	Contacto Hoy
	324BC	**Baja California**	M016	Tijuana	La Jornada BC
	325BC	**Baja California**	M002	Ensenada	Ensenada.net
	326CH	**Chihuahua**	M043	Ciudad Juárez	El Mexicano
	327DU	**Durango**	M062	Durango	El Siglo de Durango
	328SO	**Sonora**	M077	Ciudad Obregón	Medios Obson
	329BC	**Baja California**	M013	Tijuana	Televisa Californias

2019-07-09	330BC	**Baja California**	M003	Mexicali	Enlace Informativo
	331CH	**Chihuahua**	M044	Ciudad Juárez	Frontenet
	332SO	**Sonora**	M088	Navojoa	La Verdad
	333DU	**Durango**	M064	Durango	La Voz de Durango
	334CH	**Chihuahua**	M042	Ciudad Juárez	El Fronterizo
	335CH	**Chihuahua**	M030	Chihuahua	El Diario
	336SO	**Sonora**	M085	Hermosillo	Expreso
2019-07-10	337CH	**Chihuahua**	M030	Chihuahua	El Diario
	338CH	**Chihuahua**	M032	Chihuahua	El Heraldo de Chihuahua
	339SN	**Sinaloa**	M071	Culiacán	Viva Voz
	340SO	**Sonora**	M080	Hermosillo	Crítica
	341CH	**Chihuahua**	M038	Chihuahua	La Parada Digital
	342CH	**Chihuahua**	M042	Ciudad Juárez	El Fronterizo
	343CH	**Chihuahua**	M057	Nuevo Casas Grandes	Akro Noticias
2019-07-11	344SO	**Sonora**	M090	Nogales	El Observador México
	345SO	**Sonora**	M078	Guaymas	El Autónomo
	346CH	**Chihuahua**	M030	Chihuahua	El Diario
	347BC	**Baja California**	M006	Mexicali	Monitor Económico
	348SO	**Sonora**	M076	Ciudad Obregón	Tribuna
	349CH	**Chihuahua**	M035	Chihuahua	La Crónica de Chihuahua
	350SO	**Sonora**	M076	Ciudad Obregón	Tribuna
2019-07-12	351CH	**Chihuahua**	M054	Ciudad Juárez	Vivir en Juárez
	352SN	**Sinaloa**	M072	Culiacán	Viva la Noticia
	353CH	**Chihuahua**	M039	Chihuahua	Segundo a Segundo
	354BC	**Baja California**	M013	Tijuana	Televisa Californias
	355BS	**Baja California Sur**	M022	La Paz	Cabovision
	356CH	**Chihuahua**	M038	Chihuahua	La Parada Digital
	357BC	**Baja California**	M016	Tijuana	La Jornada BC
2019-07-13	358DU	**Durango**	M061	Durango	Durango al Día
	359CH	**Chihuahua**	M035	Chihuahua	La Crónica de Chihuahua
	360CH	**Chihuahua**	M050	Ciudad Juárez	Net Noticias
	361CH	**Chihuahua**	M047	Ciudad Juárez	Juárez Noticias
	362SN	**Sinaloa**	M070	Culiacán	Noroeste
	363SO	**Sonora**	M084	Hermosillo	Entorno Informativo
	364CH	**Chihuahua**	M050	Ciudad Juárez	Net Noticias
2019-07-14	365CH	**Chihuahua**	M030	Chihuahua	El Diario
	366BC	**Baja California**	M004	Mexicali	El Imparcial
	367CH	**Chihuahua**	M035	Chihuahua	La Crónica de Chihuahua
	368BC	**Baja California**	M008	Tijuana	El Mexicano
	369SN	**Sinaloa**	M072	Culiacán	Viva la Noticia
	370CH	**Chihuahua**	M047	Ciudad Juárez	Juárez Noticias
	371CH	**Chihuahua**	M041	Ciudad Juárez	Canal 44

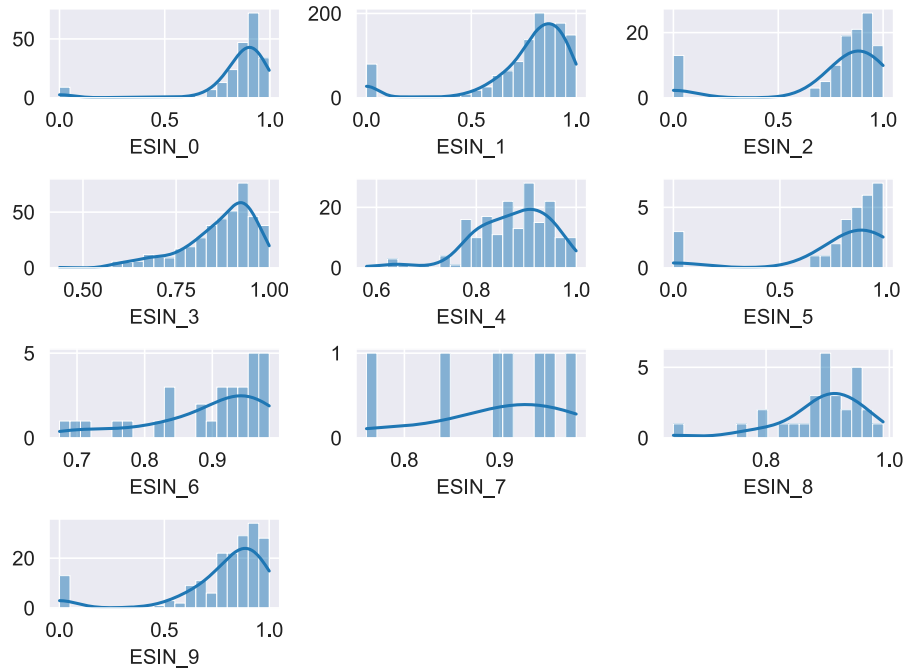
2019-07-15	372SO	**Sonora**	M078	Guaymas	El Autónomo
	373DU	**Durango**	M061	Durango	Durango al Día
	374CH	**Chihuahua**	M031	Chihuahua	El Digital
	375CH	**Chihuahua**	M034	Chihuahua	Entre Líneas
	376CH	**Chihuahua**	M031	Chihuahua	El Digital
	377DU	**Durango**	M061	Durango	Durango al Día
	378CH	**Chihuahua**	M043	Ciudad Juárez	El Mexicano
2019-07-16	379DU	**Durango**	M060	Durango	Contexto de Durango
	380BC	**Baja California**	M007	Tijuana	Agencia Fronteriza de Noticias

Anexo E. Histogramas de las distribuciones de las agrupaciones ESIN

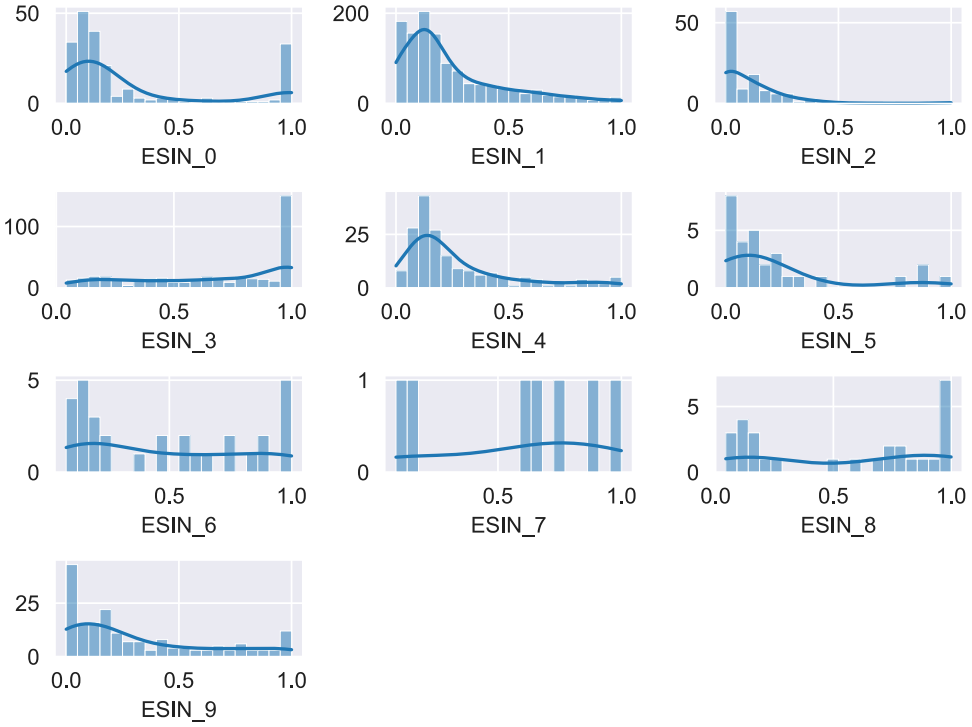
Histograma 1. Agrupación ESIN y Medida LSA con bolsa Interior-*w*



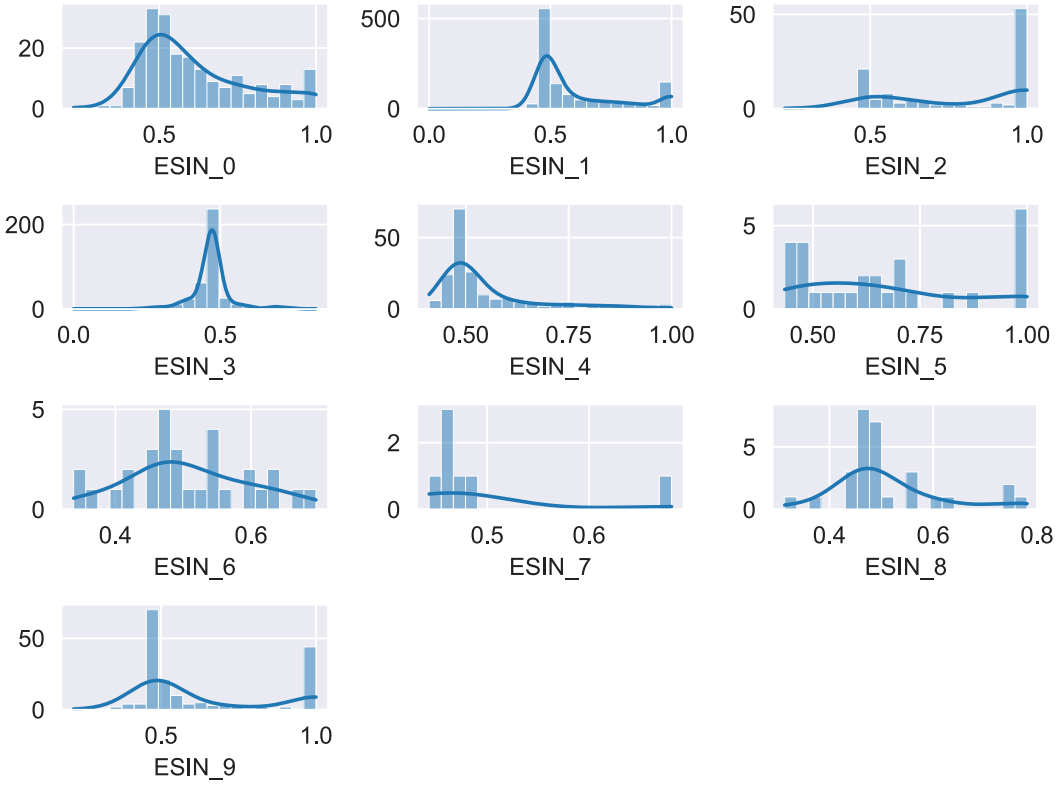
Histograma 2. Agrupación ESIN y Medida LSA con bolsa Ventana-*n*



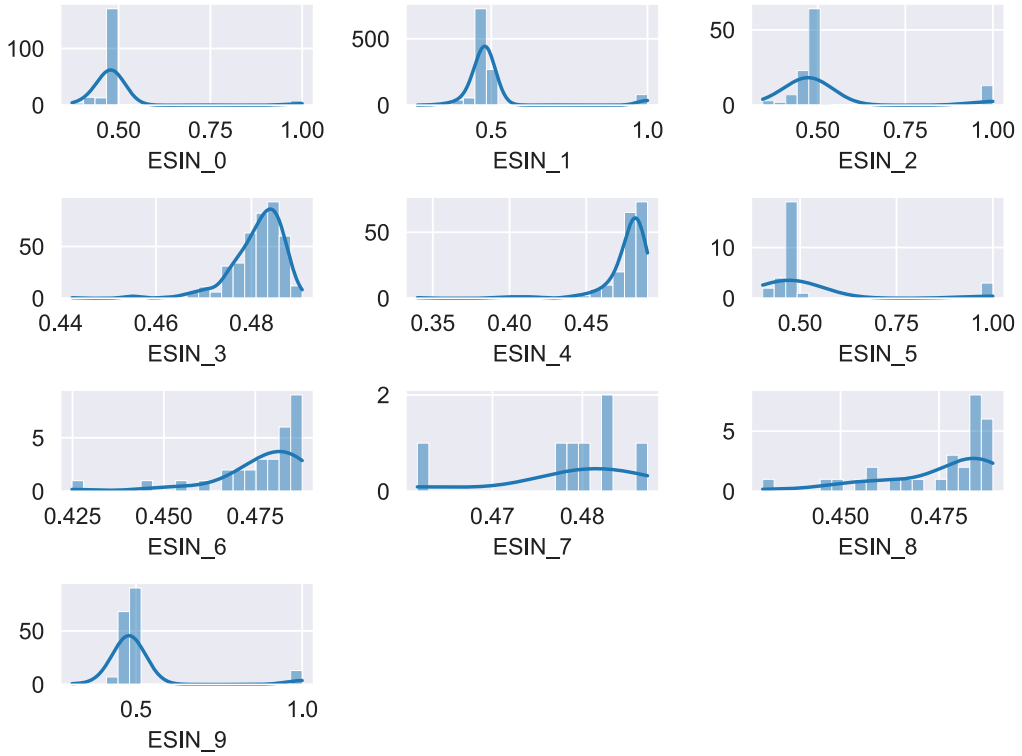
Histograma 3. Agrupación ESIN y Medida LSA con bolsa Interior-*w* DEM



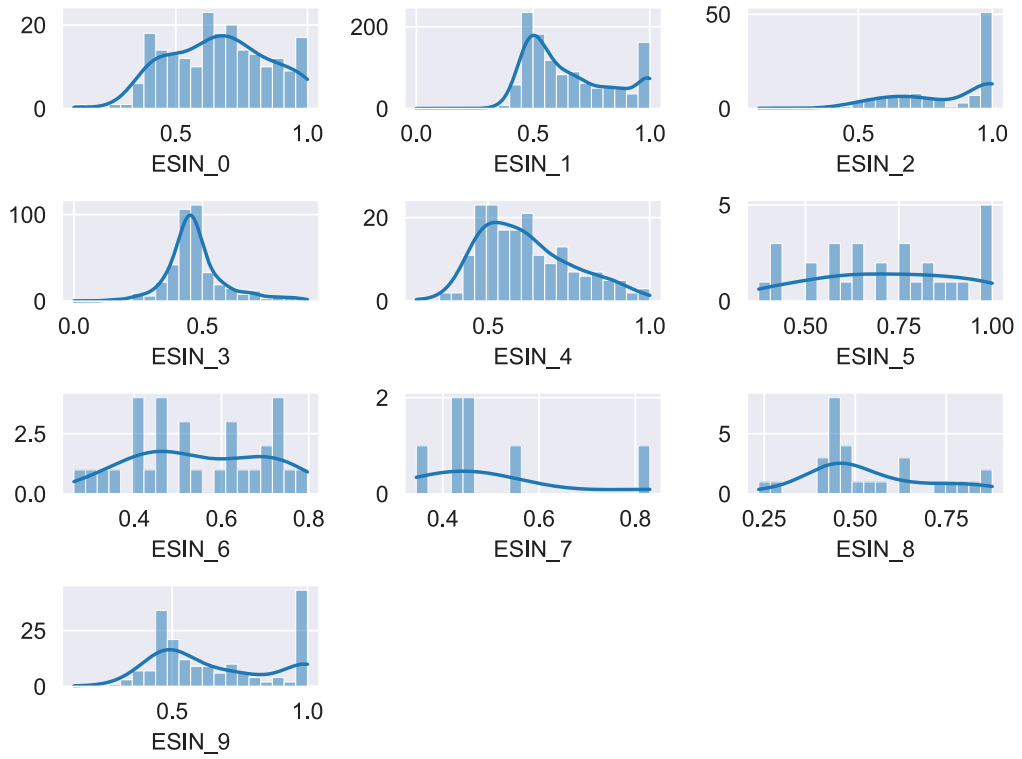
Histograma 4. Agrupación ESIN y Medida SPAN con bolsa Interior-*w*



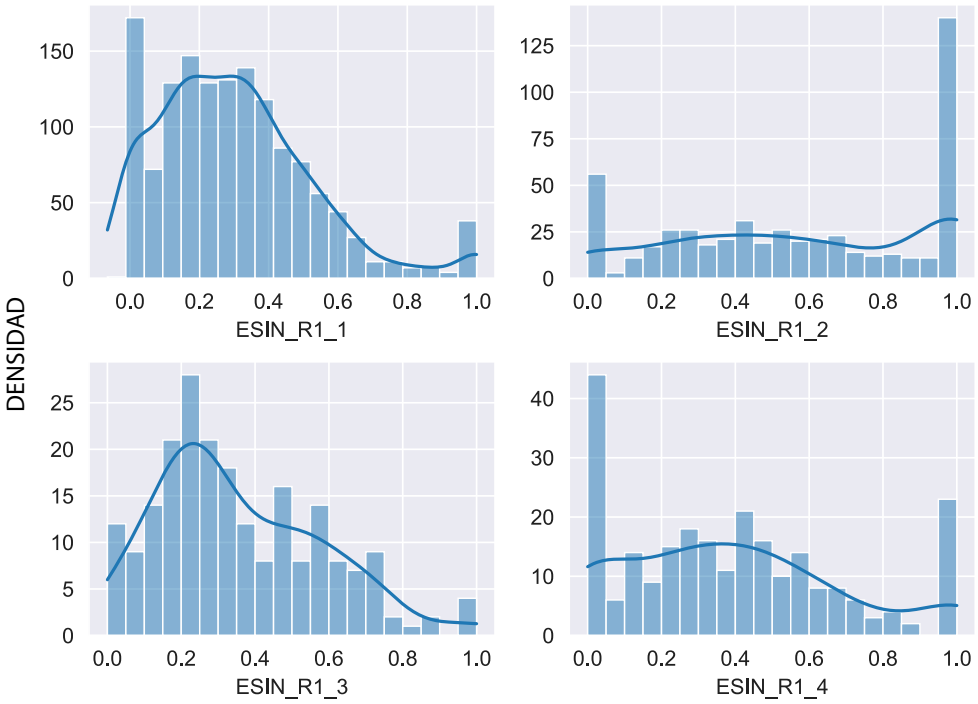
Histograma 5. Agrupación ESIN y Medida SPAN con bolsa Ventana-*n*



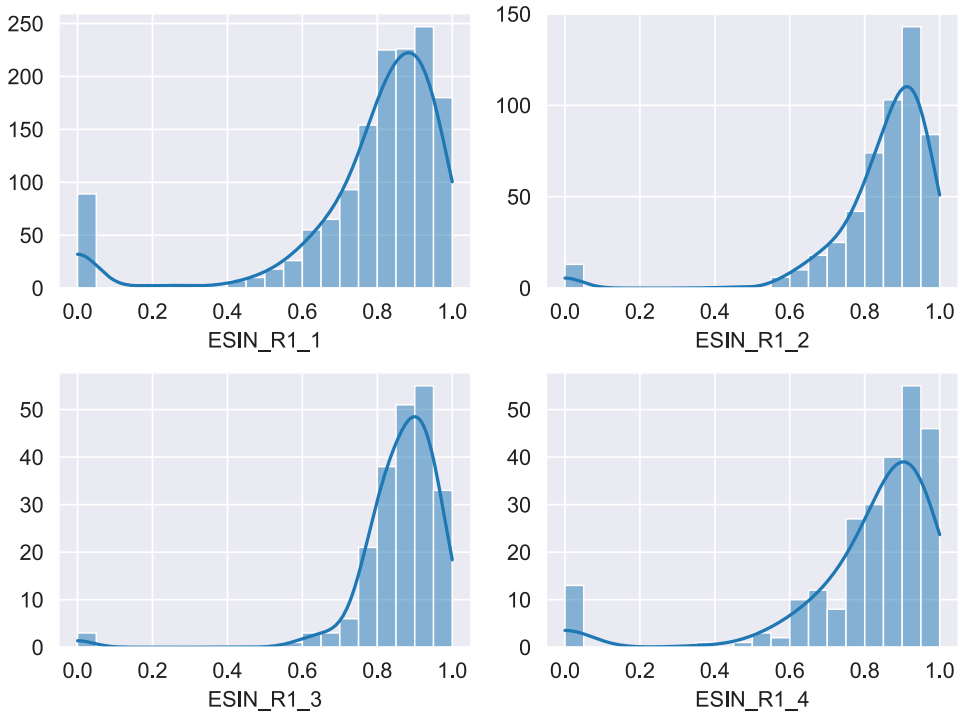
Histograma 6. Agrupación ESIN y Medida SPAN con bolsa Interior-*w* DEM



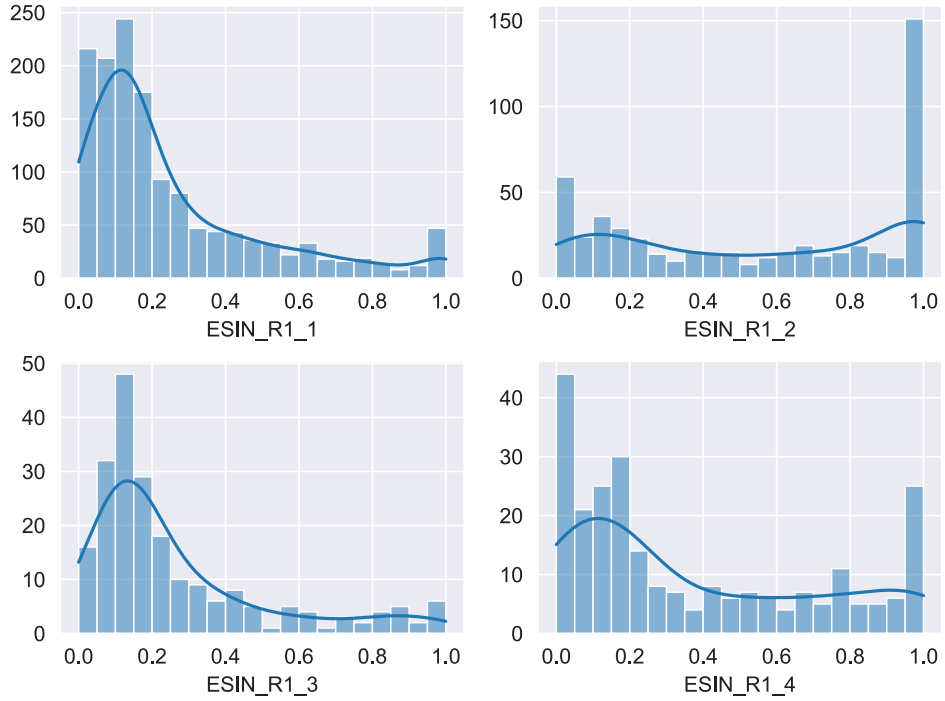
Histograma 7. Agrupación ESIN_R1 y Medida LSA con bolsa Interior-*w*



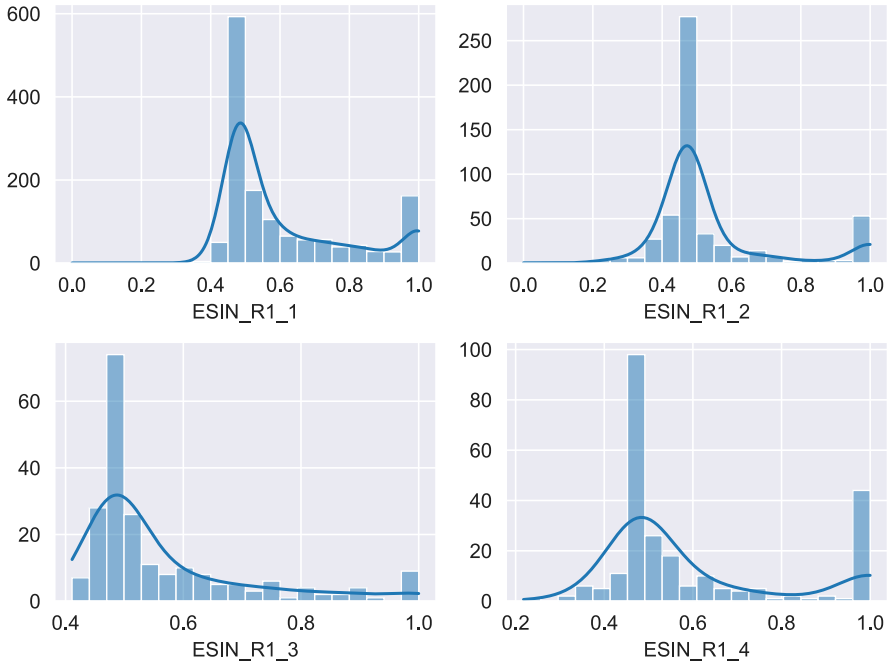
Histograma 8. Agrupación ESIN_R1 y Medida LSA con bolsa Ventana-*n*



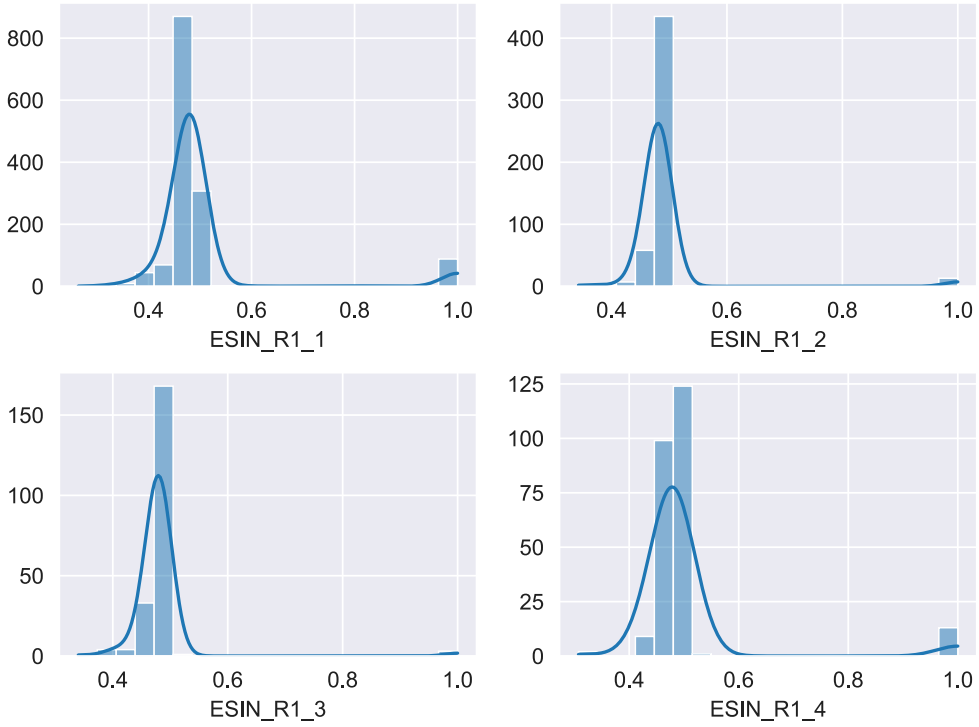
Histograma 9. Agrupación ESIN_R1 y Medida LSA con bolsa Interior-*w* DEM



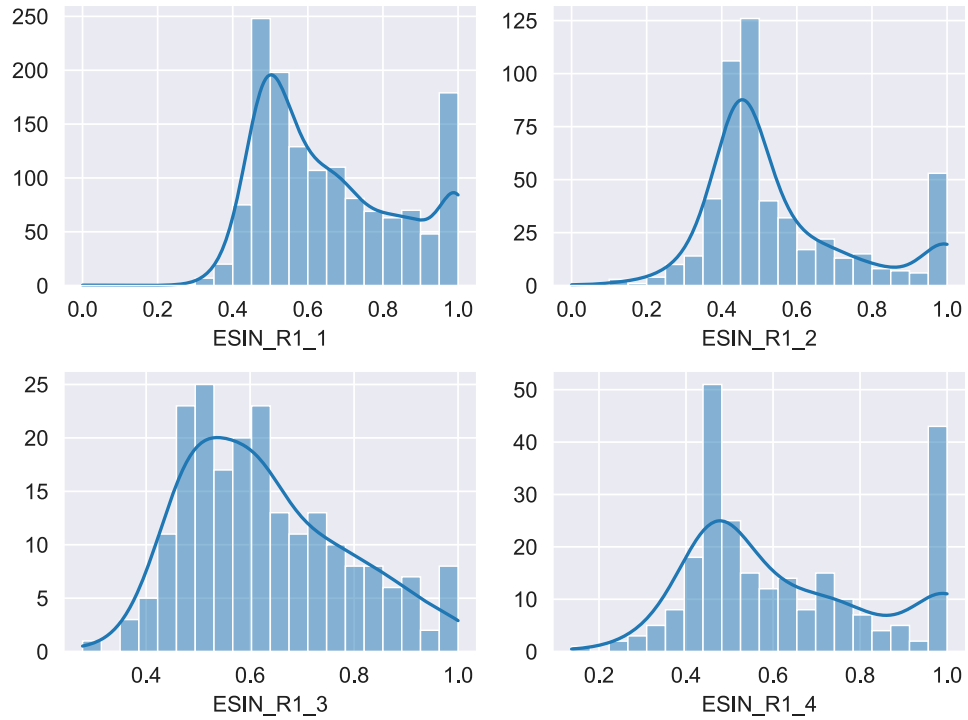
Histograma 10. Agrupación ESIN_R1 y Medida SPAN con bolsa Interior-*w*



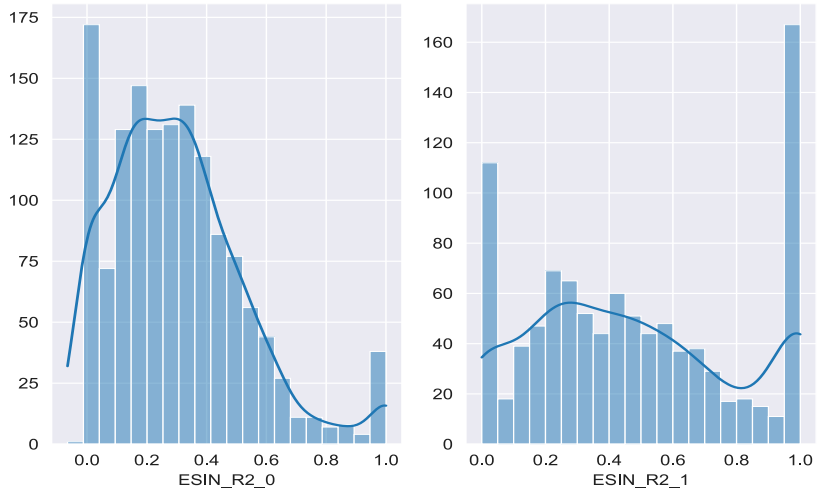
Histograma 11. Agrupación ESIN_R1 y Medida SPAN con bolsa Ventana-*n*



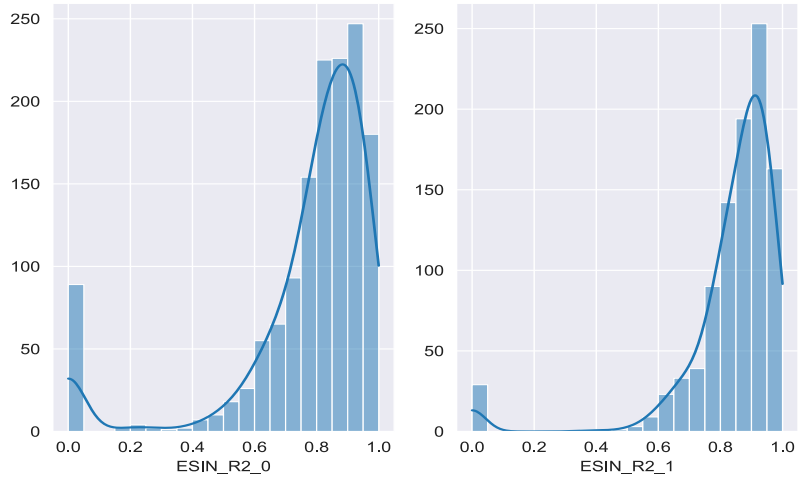
Histograma 12. Agrupación ESIN_R1 y Medida SPAN con bolsa Interior-*w* DEM



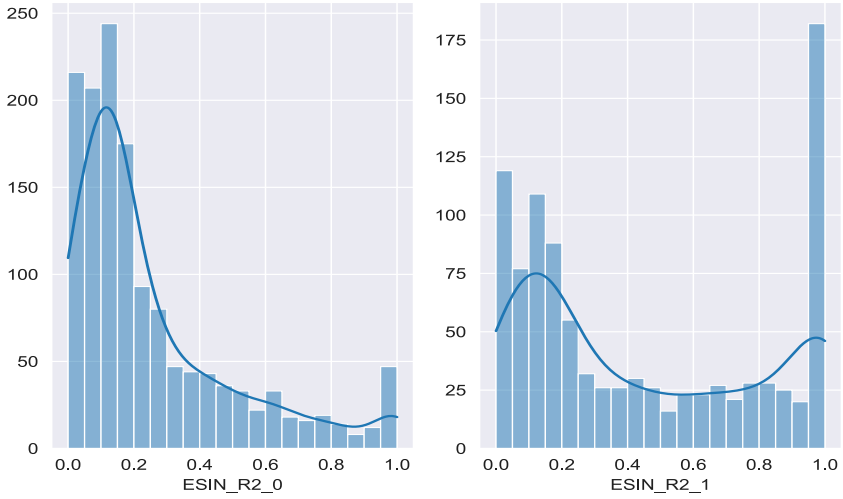
Histograma 13. Agrupación ESIN_R2 y Medida LSA con bolsa Interior-*w*



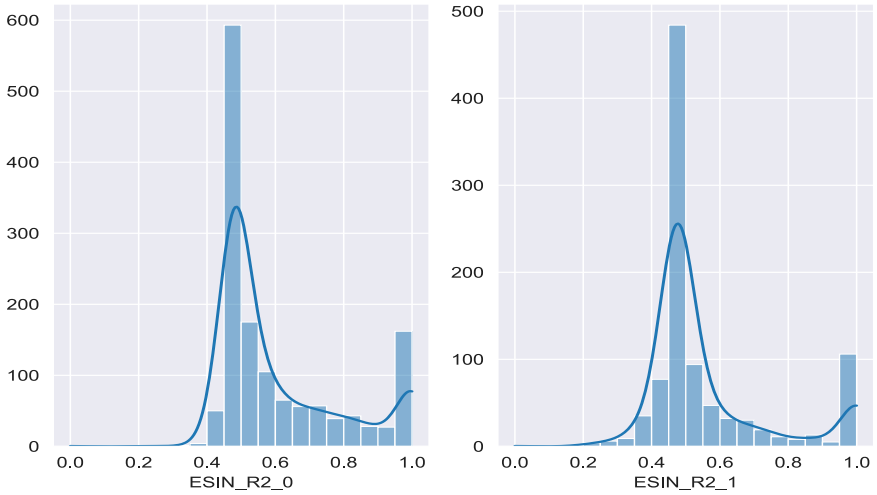
Histograma 14. Agrupación ESIN_R2 y Medida LSA con bolsa Ventana-*n*



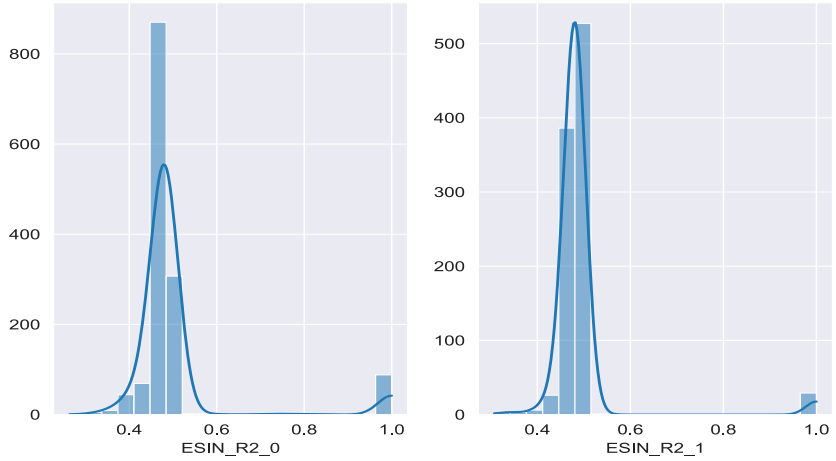
Histograma 15. Agrupación ESIN_R2 y Medida LSA con bolsa Interior-*w*



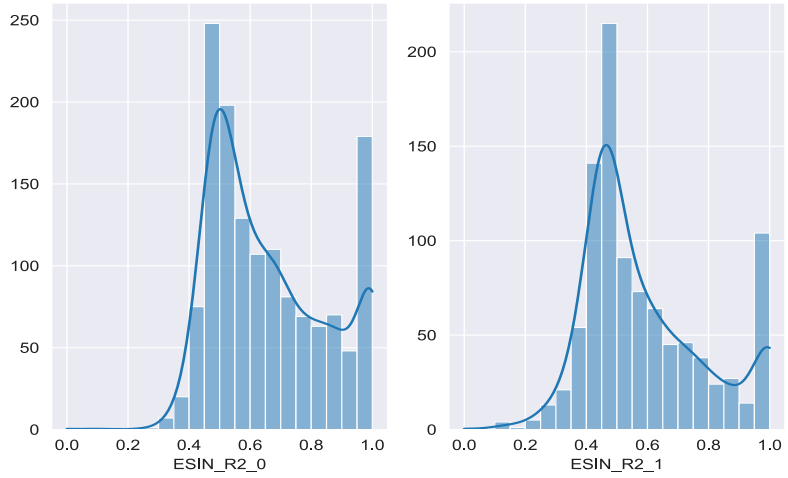
Histograma 16. Agrupación ESIN_R2 y Medida SPAN con bolsa Interior-*w*



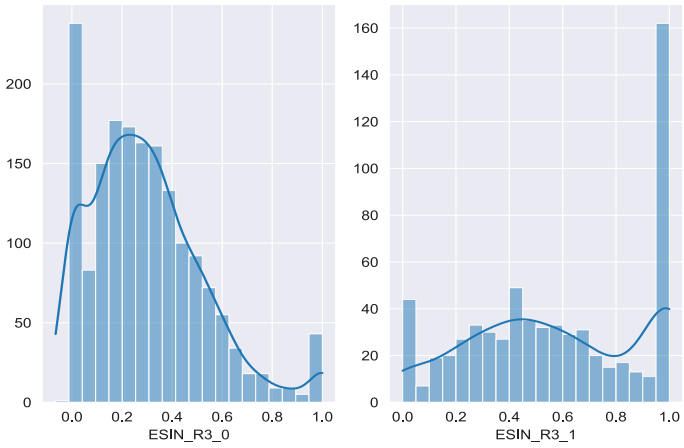
Histograma 17. Agrupación ESIN_R2 y Medida SPAN con bolsa Ventana-*n*



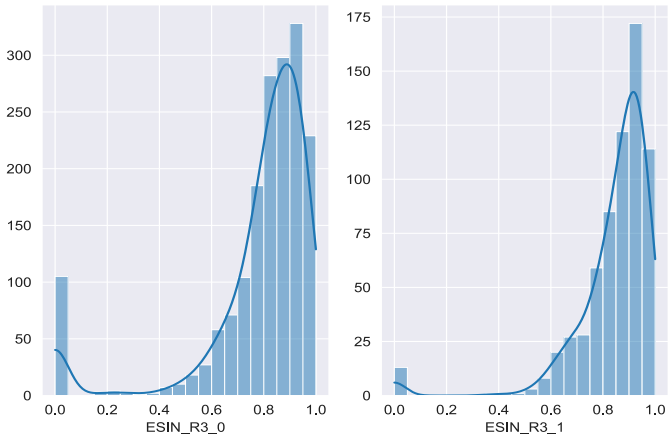
Histograma 18. Agrupación ESIN_R2 y Medida SPAN con bolsa Interior-*w*



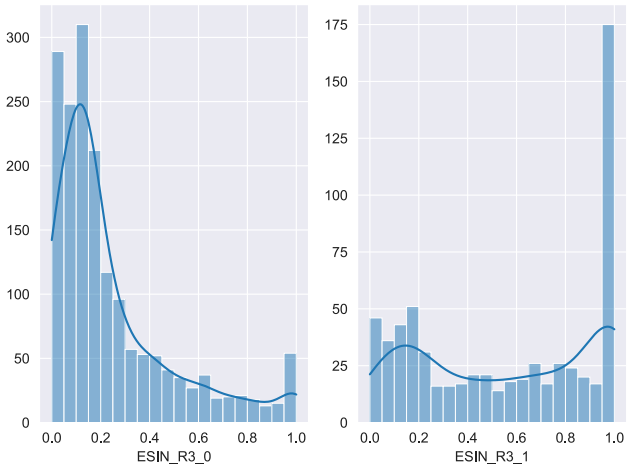
Histograma 19. Agrupación ESIN_R3 y Medida LSA con bolsa Interior-*w*



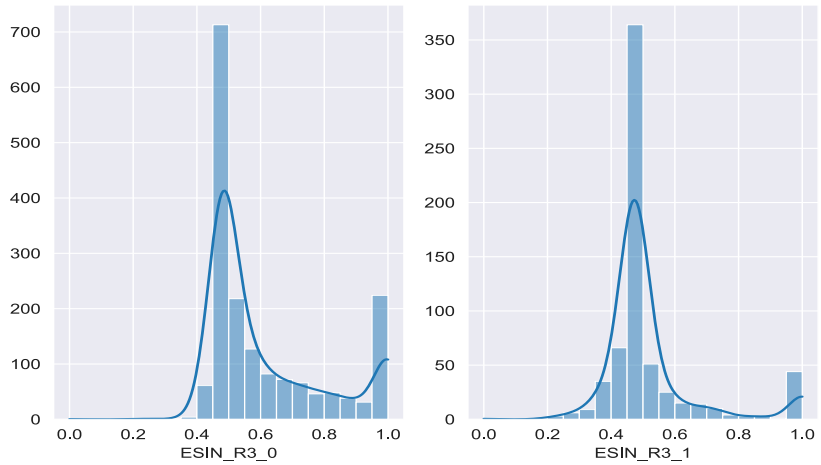
Histograma 20. Agrupación ESIN_R3 y Medida LSA con bolsa Ventana-*n*



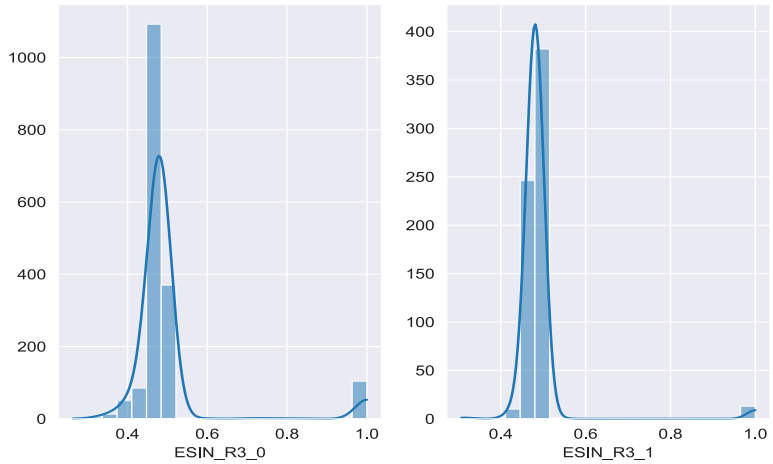
Histograma 21. Agrupación ESIN_R3 y Medida LSA con bolsa Interior-*w*



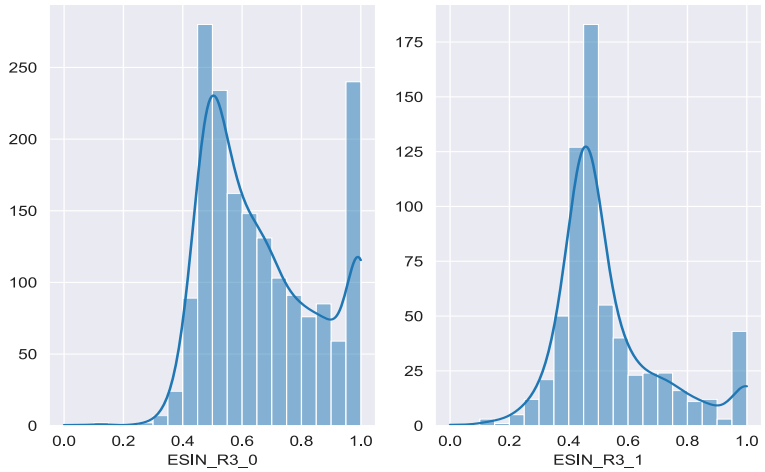
Histograma 22. Agrupación ESIN_R3 y Medida SPAN con bolsa Interior-*w*



Histograma 23. Agrupación ESIN_R3 y Medida SPAN con bolsa Ventana-*n*



Histograma 24. Agrupación ESIN_R3 y Medida SPAN con bolsa Interior-*w*



Anexo F. Matrices Conover-Iman

Matriz 1. Medidas LSA con bolsa Interior-w agrupada por ESIN

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	1	0.048164	$8.98e^{-08}$	$4.5e^{-66}$	$3.79e^{-06}$	0.536495	$1.96e^{-06}$	0.00097	$7.21e^{-08}$	0.021977
1	0.048164	1	$6.55e^{-15}$	$1.5e^{-110}$	$5.62e^{-05}$	0.151356	$2.41e^{-05}$	0.003117	$9.82e^{-07}$	0.288635
2	$8.98e^{-08}$	$6.55e^{-15}$	1	$2.77e^{-82}$	$1.34e^{-19}$	0.015909	$4.1e^{-14}$	$1.24e^{-06}$	$6.32e^{-16}$	$1.19e^{-12}$
3	$4.5e^{-66}$	$1.5e^{-110}$	$2.77e^{-82}$	1	$1.23e^{-30}$	$5.11e^{-17}$	0.001738	0.542353	0.024848	$5.32e^{-44}$
4	$3.79e^{-06}$	$5.62e^{-05}$	$1.34e^{-19}$	$1.23e^{-30}$	1	0.00324	0.020241	0.037178	0.002538	0.025177
5	0.536495	0.151356	0.015909	$5.11e^{-17}$	0.00324	1	$5.91e^{-05}$	0.000967	$5.92e^{-06}$	0.076566
6	$1.96e^{-06}$	$2.41e^{-05}$	$4.1e^{-14}$	0.001738	0.020241	$5.91e^{-05}$	1	0.40046	0.554663	0.000432
7	0.00097	0.003117	$1.24e^{-06}$	0.542353	0.037178	0.000967	0.40046	1	0.636129	0.00715
8	$7.21e^{-08}$	$9.82e^{-07}$	$6.32e^{-16}$	0.024848	0.002538	$5.92e^{-06}$	0.554663	0.636129	1	$2.91e^{-05}$
9	0.021977	0.288635	$1.19e^{-12}$	$5.32e^{-44}$	0.025177	0.076566	0.000432	0.00715	$2.91e^{-05}$	1

Matriz 2. Medidas LSA con bolsa Ventana-*n* agrupada por ESIN

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	1	$2.01e^{-12}$	0.038669	0.803975	0.794584	0.280562	0.170004	0.471329	0.37472	0.000161
1	$2.01e^{-12}$	1	0.003692	$2.41e^{-18}$	$2.76e^{-10}$	0.095625	$1.43e^{-05}$	0.034081	0.000189	0.07011
2	0.038669	0.003692	1	0.038765	0.072171	0.895754	0.012832	0.183902	0.045395	0.238023
3	0.803975	$2.41e^{-18}$	0.038765	1	0.953582	0.316297	0.126313	0.434777	0.305965	$5.42e^{-05}$
4	0.794584	$2.76e^{-10}$	0.072171	0.953582	1	0.347941	0.135048	0.431479	0.31197	0.000666
5	0.280562	0.095625	0.895754	0.316297	0.347941	1	0.064608	0.244284	0.13819	0.398852
6	0.170004	$1.43e^{-05}$	0.012832	0.126313	0.135048	0.064608	1	0.975563	0.732955	0.000896
7	0.471329	0.034081	0.183902	0.434777	0.431479	0.244284	0.975563	1	0.810521	0.087142
8	0.37472	0.000189	0.045395	0.305965	0.31197	0.13819	0.732955	0.810521	1	0.005297
9	0.000161	0.07011	0.238023	$5.42e^{-05}$	0.000666	0.398852	0.000896	0.087142	0.005297	1

Matriz 3. Medidas LSA con bolsa Interior-*w* DEM agrupada por ESIN

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	1	0.390955	$1.53e^{-08}$	$5.15e^{-56}$	0.026771	0.378303	$7.93e^{-06}$	0.006368	$3.64e^{-06}$	0.080552
1	0.390955	1	$2.35e^{-13}$	$5.5e^{-104}$	0.044245	0.204714	$1.23e^{-05}$	0.009419	$5.63e^{-06}$	0.156759
2	$1.53e^{-08}$	$2.35e^{-13}$	1	$6.72e^{-76}$	$1.79e^{-13}$	0.019417	$8.44e^{-14}$	$1.18e^{-05}$	$4.49e^{-14}$	$3.51e^{-12}$
3	$5.15e^{-56}$	$5.5e^{-104}$	$6.72e^{-76}$	1	$3.95e^{-37}$	$1.34e^{-15}$	0.006136	0.396949	0.018543	$2.27e^{-39}$
4	0.026771	0.044245	$1.79e^{-13}$	$3.95e^{-37}$	1	0.046707	0.001028	0.032036	0.0005	0.659648
5	0.378303	0.204714	0.019417	$1.34e^{-15}$	0.046707	1	$6.32e^{-05}$	0.003703	$3.24e^{-05}$	0.079082
6	$7.93e^{-06}$	$1.23e^{-05}$	$8.44e^{-14}$	0.006136	0.001028	$6.32e^{-05}$	1	0.653079	0.821778	0.000442
7	0.006368	0.009419	$1.18e^{-05}$	0.396949	0.032036	0.003703	0.653079	1	0.75772	0.02367
8	$3.64e^{-06}$	$5.63e^{-06}$	$4.49e^{-14}$	0.018543	0.0005	$3.24e^{-05}$	0.821778	0.75772	1	0.000212
9	0.080552	0.156759	$3.51e^{-12}$	$2.27e^{-39}$	0.659648	0.079082	0.000442	0.02367	0.000212	1

Matriz 4. Medidas SPAN con bolsa Interior-*w* agrupada por ESIN

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	1	0.136881	$8.21e^{-09}$	$1.85e^{-43}$	0.000318	0.333538	0.000596	0.001899	$8.14e^{-05}$	0.25612
1	0.136881	1	$2.38e^{-15}$	$3.5e^{-74}$	0.001453	0.107747	0.002516	0.004297	0.000366	0.959395
2	$8.21e^{-09}$	$2.38e^{-15}$	1	$4.28e^{-65}$	$7.11e^{-18}$	0.020593	$5.61e^{-11}$	$1.72e^{-06}$	$3.53e^{-12}$	$5.81e^{-11}$
3	$1.85e^{-43}$	$3.5e^{-74}$	$4.28e^{-65}$	1	$1.4e^{-20}$	$7.21e^{-13}$	0.00418	0.997935	0.031418	$1.08e^{-32}$
4	0.000318	0.001453	$7.11e^{-18}$	$1.4e^{-20}$	1	0.005562	0.124313	0.030891	0.036122	0.017861
5	0.333538	0.107747	0.020593	$7.21e^{-13}$	0.005562	1	0.00098	0.001013	0.000216	0.125629
6	0.000596	0.002516	$5.61e^{-11}$	0.00418	0.124313	0.00098	1	0.202696	0.641265	0.004997
7	0.001899	0.004297	$1.72e^{-06}$	0.997935	0.030891	0.001013	0.202696	1	0.327003	0.005117
8	$8.14e^{-05}$	0.000366	$3.53e^{-12}$	0.031418	0.036122	0.000216	0.641265	0.327003	1	0.000874
9	0.25612	0.959395	$5.81e^{-11}$	$1.08e^{-32}$	0.017861	0.125629	0.004997	0.005117	0.000874	1

Matriz 5. Medidas SPAN con bolsa Ventana-*n* agrupada por ESIN

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	1	0.207726	0.396906	0.00648	0.628768	0.218532	0.671396	0.960968	0.890327	0.132714
1	0.207726	1	0.050466	0.01718	0.071296	0.07257	0.33479	0.766226	0.723046	0.464934
2	0.396906	0.050466	1	0.001966	0.675146	0.486181	0.932771	0.837505	0.54506	0.036483
3	0.00648	0.01718	0.001966	1	0.001645	0.01362	0.093683	0.512528	0.289627	0.378746
4	0.628768	0.071296	0.675146	0.001645	1	0.328906	0.865243	0.938224	0.703766	0.054829
5	0.218532	0.07257	0.486181	0.01362	0.328906	1	0.530435	0.593392	0.302238	0.047904
6	0.671396	0.33479	0.932771	0.093683	0.865243	0.530435	1	0.880752	0.673434	0.229183
7	0.960968	0.766226	0.837505	0.512528	0.938224	0.593392	0.880752	1	0.912827	0.657122
8	0.890327	0.723046	0.54506	0.289627	0.703766	0.302238	0.673434	0.912827	1	0.532385
9	0.132714	0.464934	0.036483	0.378746	0.054829	0.047904	0.229183	0.657122	0.532385	1

Matriz 6. Medidas SPAN con bolsa Interior-*w* DEM agrupada por ESIN

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	1	0.586029	$3.62e^{-11}$	$3.2e^{-46}$	0.05438	0.20822	0.000658	0.002018	$2.42e^{-05}$	0.216881
1	0.586029	1	$1.99e^{-16}$	$1.71e^{-88}$	0.052833	0.123171	0.00073	0.00251	$2.31e^{-05}$	0.290011
2	$3.62e^{-11}$	$1.99e^{-16}$	1	$2.33e^{-74}$	$8.06e^{-16}$	0.011679	$2.25e^{-12}$	$5.08e^{-07}$	$1.41e^{-14}$	$8.75e^{-14}$
3	$3.2e^{-46}$	$1.71e^{-88}$	$2.33e^{-74}$	1	$4.97e^{-31}$	$1.61e^{-14}$	0.001875	0.898595	0.038309	$2.04e^{-34}$
4	0.05438	0.052833	$8.06e^{-16}$	$4.97e^{-31}$	1	0.026683	0.017294	0.009916	0.00128	0.511509
5	0.20822	0.123171	0.011679	$1.61e^{-14}$	0.026683	1	0.000467	0.000657	$3.62e^{-05}$	0.061467
6	0.000658	0.00073	$2.25e^{-12}$	0.001875	0.017294	0.000467	1	0.20421	0.482062	0.006352
7	0.002018	0.00251	$5.08e^{-07}$	0.898595	0.009916	0.000657	0.20421	1	0.406241	0.00586
8	$2.42e^{-05}$	$2.31e^{-05}$	$1.41e^{-14}$	0.038309	0.00128	$3.62e^{-05}$	0.482062	0.406241	1	0.000374
9	0.216881	0.290011	$8.75e^{-14}$	$2.04e^{-34}$	0.511509	0.061467	0.006352	0.00586	0.000374	1

Matriz 7. Medidas LSA con bolsa Interior-*w* agrupada por ESIN_R1

	1	2	3	4
1	1	$5.14e^{-57}$	0.001077	$2.05e^{-05}$
2	$5.14e^{-57}$	1	$2.21e^{-13}$	$2.05e^{-12}$
3	0.001077	$2.21e^{-13}$	1	0.564854
4	$2.05e^{-05}$	$2.05e^{-12}$	0.564854	1

Matriz 8. Medidas LSA con bolsa Ventana-*n* agrupada por ESIN_R1

	1	2	3	4
1	1	$4.81e^{-13}$	$1.12e^{-07}$	0.001014
2	$4.81e^{-13}$	1	0.833553	0.057331
3	$1.12e^{-07}$	0.833553	1	0.079148
4	0.001014	0.057331	0.079148	1

Matriz 9. Medidas LSA con bolsa Interior-*w* DEM agrupada por ESIN_R1

	1	2	3	4
1	1	$1.64e^{-53}$	0.14707	$1.77e^{-05}$
2	$1.64e^{-53}$	1	$7.6e^{-18}$	$3.17e^{-11}$
3	0.14707	$7.6e^{-18}$	1	0.042084
4	$1.77e^{-05}$	$3.17e^{-11}$	0.042084	1

Matriz 10. Medidas SPAN con bolsa Interior-*w* agrupada por ESIN_R1

	1	2	3	4
1	1	$1.4e^{-35}$	0.01363	0.007314
2	$1.4e^{-35}$	1	$8.88e^{-09}$	$1.95e^{-09}$
3	0.01363	$8.88e^{-09}$	1	0.96801
4	0.007314	$1.95e^{-09}$	0.96801	1

Matriz 11. Medidas SPAN con bolsa Ventana-*n* agrupada por ESIN_R1

	1	2	3	4
1	1	0.122991	0.035237	0.730253
2	0.122991	1	0.004035	0.472319
3	0.035237	0.004035	1	0.056079
4	0.730253	0.472319	0.056079	1

Matriz 12. Medidas SPAN con bolsa Interior-*w* DEM agrupada por ESIN_R1

	1	2	3	4
1	1	$8.44e^{-42}$	0.209565	0.000274
2	$8.44e^{-42}$	1	$4.3e^{-14}$	$3.31e^{-09}$
3	0.209565	$4.3e^{-14}$	1	0.088761
4	0.000274	$3.31e^{-09}$	0.088761	1

Anexo P. De Python

En un disco que se anexa a la tesis se encontrará una carpeta con el corpus tratado para la investigación. El tipo de contenido que se encuentra en cada carpeta del corpus es el explicado en la sección 2.6.3.

Además, se encuentra una carpeta con algunos *scripts* de Python mencionados en la tesis. A septiembre del 2021, se habilitó además el sitio https://gitlab.com/manuel.wortens/anexop para descargar el contenido de este anexo digital.