Análisis interdisciplinario de artículos científicos referentes a la pandemia COVID-19

Patricia Zapata, Sergio Peignier

Análisis interdisciplinario de artículos científicos referentes a la pandemia COVID-19

Patricia Zapata[1] y Sergio Peignier[2]

Resumen

En este trabajo se emplearon técnicas de procesamiento de lenguaje natural y de minería de datos, enmarcados por importantes conceptos de lingüística teórica y ciencias del lenguaje, como el principio de cooperación y las máximas de Grice, para estudiar un corpus de más de 52.000 artículos científicos de medicina y biología recientes, relacionados al problema de la lucha contra la pandemia COVID-19. El objetivo de este estudio es procesar un grupo de artículos científicos, con el fin de facilitar la extracción de conocimientos asociados a un grupo de términos de interés, permitiendo a los especialistas obtener una comprensión global y rápida de dichos conceptos.

Palabras Claves

Principio de cooperación, Máximas de Grice, Minería de Datos, Grafo de Conocimientos, COVID-19

Introducción

En la actualidad, los recientes avances tecnológicos ponen a disposición de la sociedad, una gran cantidad de información, de índole diversa. En este contexto, es imprescindible, para un receptor, disponer de herramientas que permitan filtrar y encontrar de manera rápida y resumida, la información esencial sobre un tema de interés específico. Esta necesidad es aún más imperiosa durante contextos históricos de gran trascendencia, como por ejemplo durante la pandemia COVID-19, que se atraviesa actualmente. Este fenómeno afecta no sólo al público en general, sino también a los especialistas. En efecto, cada día diferentes equipos de investigación, se dan a la tarea de publicar los resultados de sus trabajos. Sin embargo, el tiempo necesario para que un especialista lea toda esta información es prohibitivo. Por ende, los expertos necesitan tener acceso a una versión sintetizada de la información contenida en dichos artículos, para poder tomar decisiones, en beneficio de la sociedad, o para concentrar sus esfuerzos de investigación, en aspectos más prometedores. En este contexto, la lingüística teórica y las ciencias del lenguaje juegan un rol de gran importancia, ya que brindan un

[1] Master en Educación Superior y Desarrollo Boliviano, Licenciada en Lingüística e Idiomas.
Docente titular emérita en la Carrera de Lingüística e Idiomas en la Universidad Mayor de San Andrés (La Paz).
Correo electrónico: pzapata@umsa.bo

[2] Univ Lyon, INSA Lyon, INRAE, BF2I, UMR 203, Villeurbanne, 69100, France
Doctor en Ciencias de la Computación, Master en Informática Fundamental e Ingeniero en Bioinformática.
Docente investigador en el Departamento de Biociencias del Instituto Nacional de Ciencias Aplicadas (Lyon).
Correo electrónico: sergio.peignier@insa-lyon.fr

marco teórico e importantes líneas directrices, para diseñar sistemas informáticos eficientes, que permitan extraer, de dichos artículos, las estructuras lingüísticas esenciales del lenguaje natural, facilitando así el acceso a la información contenida en un gran volumen de datos. El presente trabajo se enmarca en la línea de investigación de la "lingüística teórica y ciencias del lenguaje", en su relación interdisciplinaria con las disciplinas conocidas como la Minería de Datos y el Procesamiento del Lenguaje Natural. En este trabajo, se analizó un corpus compuesto por 52.000 artículos académico, por medio de diferentes sistemas informáticos de procesamiento de lenguaje natural y de minería de datos, basados en conceptos lingüísticos, con el fin de identificar los principales medicamentos estudiados para tratar el coronavirus, y caracterizar su mecanismo de acción y sus beneficios. En este trabajo, se procedió a procesar e indexar el corpus de artículos científicos, para poder extraer todas las oraciones que contengan un conjunto de términos de interés particular, para luego elaborar grafos de conocimientos y extraer oraciones claves. La utilización de estos procedimientos inherentes al procesamiento de lenguajes naturales y a la minería de datos, se enmarca en el *principio de cooperación* definido por Grice (1975). El presente artículo se organiza de la manera siguiente: la sección 1 describe el corpus utilizado y la metodología de investigación desarrollada, la sección 2 presenta los resultados obtenidos con nuestra metodología, y finalmente en la sección 3 se concluye el artículo, con un resumen y perspectivas a futuro.

Corpus y Metodología

Corpus

Corpus CORD-19

El presente trabajo se basa en CORD-19, el conjunto de datos recopilados por Wang et al. (2020) y puesto a disposición por la Casa Blanca e importantes grupos de investigación tales como AI2, CZI, MSR y NIH. Este corpus está compuesto por más de 52.000 artículos académicos, en formato .json, sobre la pandemia COVID-19, y otras infecciones virales relacionadas. El objetivo de dicho corpus consiste en permitir a la comunidad de analistas de datos, aplicar técnicas de procesamiento de lenguaje natural e Inteligencia Artificial, para facilitar el acceso al masivo volumen de información relacionado al tema; permitiendo así a la comunidad científica y médica mundial coadyuvar en la lucha contra la pandemia del coronavirus, al generar nuevos conocimientos.

Corpus FDA

Con el fin de analizar la manera en que los artículos científicos presentan la acción terapéutica de ciertos compuestos, para tratar patologías virales, procedimos a descargar la lista de medicamentos y

compuestos químicos aprobados por la FDA[3] (2020), la agencia gubernamental estadounidense encargada de regular los alimentos y medicamentos. Este conjunto de datos contiene un total de 2.213 nombres de compuestos químicos y medicamentos. Con el fin de evitar el análisis de substancias muy generales, y nombres comunes que acompañan los nombres de los medicamentos, se procedió a verificar manualmente la lista obtenida, conservando únicamente 2.145 nombres de compuestos químicos.

Corpus ViralZone

De igual manera, procedimos a descargar una lista de enfermedades virales humanas, publicada por Hulo et al. (2011), a través del Instituto Suizo de Bioinformática (SIB), en el sitio denominado ViralZone, la cual reporta un total de 92 enfermedades virales.

Metodología

Con el fin de realizar la presente investigación, se procedió a elaborar un programa informático que permite: 1) cargar y procesar los artículos científicos, 2) crear una estructura de datos, que facilite la búsqueda de oraciones, que contengan vocablos específicos, 3) generar un grafo de conocimientos a partir de las oraciones seleccionadas 4) extraer citaciones claves para sustentar el análisis. A continuación se explicará con mayor detalle cada una de las etapas.

Fundamento Teórico

Este trabajo se fundamenta en la teoría de la pragmática lingüística elaborada por Grice (1975), con el nombre de pragmática conversacional. Según esta teoría, la comunicación humana se fundamenta en gran medida en el *principio de cooperación*, según el cual los interlocutores cooperarán para entenderse mutuamente. De acuerdo a la pragmática conversacional, el *principio de cooperación* se fundamenta en cuatro máximas, que tienden a maximizar la precisión del intercambio y minimizar su ambigüedad. A continuación se detallan las cuatro máximas que caracterizan el *principio de cooperación.*

- *Máxima de cantidad:* Según este principio, el mensaje debe contener únicamente la cantidad necesaria de información.
- *Máxima de calidad:* Según esta máxima, el mensaje debe contener información veraz, sustentada por un respaldo suficiente.
- *Máxima de relación*: Según este principio, el mensaje debe contener solamente información relevante.

[3] Food and Drug Administration [Administración de Medicamentos y Alimentos]

- *Máximas de modo*: Según esta máxima, el mensaje debe ser transmitido de manera clara, concisa, ordenada, y se deben evitar ambigüedades.

Es importante notar que según la teoría de la pragmática conversacional, las cuatro máximas tienen un valor descriptivo y no normativo. En efecto, pueden presentarse situaciones de comunicación en las cuales las máximas son rotas de manera implícita, con el fin de introducir un nuevo mensaje (implicatura). Sin embargo, el lenguaje científico tiene por objetivo principal transmitir informaciones de manera clara, limitando la inclusión de mensajes implícitos. Por ende, esta teoría lingüística se adapta al campo de estudio del presente trabajo, y las cuatro máximas de Grice, constituyen los ejes fundamentales en torno a los cuales se articularon las herramientas de extracción de información desarrolladas y empleadas en el presente estudio.

Pre-procesamiento

En primer lugar, se procedió a programar un conjunto de funciones con el fin de procesar el corpus de artículos de la base de datos. Se extrajeron de cada artículo: los nombres de los autores, las instituciones de afiliación de los mismos, el título del artículo, su resumen y el texto principal. Posteriormente, se procedió al pre-procesamiento de los datos, mediante la separación (tokenización) del texto en oraciones, y en palabras. Luego se eliminaron todos los símbolos que no fuesen letras. Estas funciones se programaron en en el lenguaje Python 3.7 y se utilizaron las librerías json y re (Van Rossum et al. (1995)), pandas (McKinney et al. (2010)) y nltk (Bird et al. (2009)).

Indexación de datos

Con el fin de realizar rápidamente búsquedas de oraciones que contengan palabras específicas, en los documentos del corpus, se procedió a indexar la base de datos, en una estructura de tipo *diccionario*. Esta estructura permite acceder directamente a todas las oraciones de los artículos que contengan el conjunto de términos deseados. Esto permite, por un lado facilitar y acelerar las búsquedas en el corpus, y por el otro, permite responder a la *máxima de relación*, ya que se brinda al receptor del mensaje la información que le es relevante. Esta indexación se programó igualmente en el lenguaje Python 3.7, gracias a la librería pickle (Van Rossum (1995)).

Grafo de conocimientos

Con el fin de extraer la información esencial, inherente a un subconjunto de documentos, se procedió a diseñar un programa que genere el grafo de conocimientos asociado. En este trabajo, un grafo de conocimientos se define como un conjunto de vocablos conectados entre sí, formando una red que plasma las ideas esenciales contenidas en los artículos analizados. La primera etapa en el diseño de dicho grafo, consiste en determinar el rol gramatical de cada palabra, mediante un programa de POS.

Posteriormente se conservan únicamente los verbos, sustantivos y conectores lógicos, con el fin de obtener únicamente el mensaje esencial de las oraciones consideradas, respetando la *máxima de modo.* A continuación se detectan palabras compuestas, con el fin de considerarlas como una sola entidad, y no perder su sentido, respetando una vez más la *máxima de modo*. En seguida, se calcula la frecuencia de cada palabra en el corpus considerado. Finalmente, se forman las conexiones del grafo, articulando todas las palabras que están adyacentes en los documentos. El peso de la conexión entre dos palabras corresponde a la co-ocurrencia de ambas palabras, es decir la frecuencia con la cual dichos términos están juntos en los textos. Esta representación busca transmitir el mensaje, contenido en las oraciones seleccionadas, de manera gráfica, clara, concisa y ordenada. En efecto, mediante esta representación, las repeticiones de asociaciones de palabras, presentes en múltiples oraciones, reciben una concisa representación gráfica, velando así por la *máxima de cantidad*. Por otra parte, al incorporar la información sobre la frecuencia de los términos y de sus coocurrencias, se brinda al receptor valiosa información, que le permite ponderar las relaciones entre los vocablos y evitar interpretaciones ambiguas, precautelando así la *máxima de modo*. Con el fin de simplificar y facilitar la lectura del grafo, se decidió incorporar una función que permite extraer un sub-grafo de conocimientos de mayor relevancia, al eliminar las conexiones que tengan un peso inferior a un cierto umbral. Esta funcionalidad brinda al receptor, la posibilidad de elegir la cantidad de información contenida en el mensaje, con el fin de promover la *máxima de cantidad*. Este sistema informático se programó igualmente en el lenguaje Python 3.7, gracias a la librería NetworkX (Hagberg et al. (2008)).

Extracción de citaciones claves

Para poder precautelar la *máxima de calidad*, respaldando las conclusiones que pudieran obtenerse mediante el análisis de los grafos de conocimiento, procedimos a la extracción de citaciones relevantes de los artículos seleccionados. Esta etapa se realizó por medio de dos algoritmos. El primero de ellos, denominado TextRank, fue desarrollado por Mihalcea et al. (2004), y se basa en un cálculo de similitud entre todas las oraciones, y utiliza dichas comparaciones para determinar la importancia de cada oración, mediante un cálculo de rangos. El segundo método, desarrollado por nosotros, selecciona iterativamente oraciones que sean cortas y representativas del vocabulario del corpus considerado. La extracción de citaciones importantes fue programada en el lenguaje Python 3.7, gracias a la librería Gensim (Rehurek et al. (2010)).

Resultados

Co-ocurrencias entre medicamentos y enfermedades

En este trabajo, se procedió a utilizar el programa anteriormente descrito, para buscar, en el corpus, todas las oraciones que contengan un nombre de enfermedad y un nombre de medicamento. En este sentido, si consideramos un medicamento ***m*** y una enfermedad ***e***, la coocurrencia entre ***m*** y ***e*** se define como el número de oraciones que contienen ambos términos simultáneamente. Las coocurrencias entre enfermedades y medicamentos pueden representarse en una matriz que posee ***M*** columnas y ***E*** líneas, donde ***M*** contabiliza el número de medicamentos en nuestro corpus, y ***E*** el número de enfermedades.

Dado el volumen de la matriz, con el fin de representar dichos resultados en este artículo, procedimos a extraer una submatriz compuesta únicamente por el subconjunto de enfermedades citadas en al menos 100 oraciones, y el subconjunto de medicamentos citados en al menos 200 oraciones. Dicha matriz se ilustra, mediante un "mapa de calor" en la figura 1. En dicha representación, el color de la casilla que une la enfermedad ***e*** con el medicamento ***m***, representa la importancia de la co-ocurrencia de las palabras ***m*** y ***e*** en el corpus. Los colores extremos, es decir el azul y el rojo, representan respectivamente una baja y una elevada co-ocurrencia.

Dicha representación nos permite analizar rápidamente, la posible conexión entre un medicamento y una enfermedad: Mientras más veces, se encuentren en una misma oración un medicamento y una enfermedad, mayor es la posibilidad de que exista una relación estrecha entre ambos. Dada la naturaleza de los conceptos (medicamentos versus enfermedades), es posible que el medicamento en cuestión haya sido probado, testeado o utilizado para tratar la enfermedad.

La matriz revela algunos procedimientos bien establecidos, como lo indica la fuerte coocurrencia entre los virus del *herpes zoster*, la *varicela* o el *herpes simplex* con el tratamiento antiviral a base de *aciclovir*. De la misma manera, otro tratamiento clásico aparece claramente indicado a través de la coocurrencia entre la *influenza* o el *H1N1* y antivirales como el *oseltamivir*, el *zanamivir*, la *rimantadina* o la *amantadina*. Finalmente podemos apreciar otro tratamiento clásico, mediante la coocurrencia entre la *hepatitis C* y el antiviral *ribavirin*.

Por otra parte, en este trabajo resulta interesante analizar, mediante la matriz, diferentes enfermedades o virus ligados a la familia de los *coronavirus* (*MERS*, *SARS*, *COVID*), que aparecen relacionados con los compuestos químicos siguientes: *ribavirin*, *angiotensina* y *cloroquina*. Dada la importancia de dichas coocurrencias, resulta importante analizar los grafos de conocimientos obtenidos a partir de las

oraciones donde aparecen dichos tratamientos, junto con el término *coronavirus*, con el fin de comprender la naturaleza de su relación entre dicha patología y los medicamentos ya mencionados.

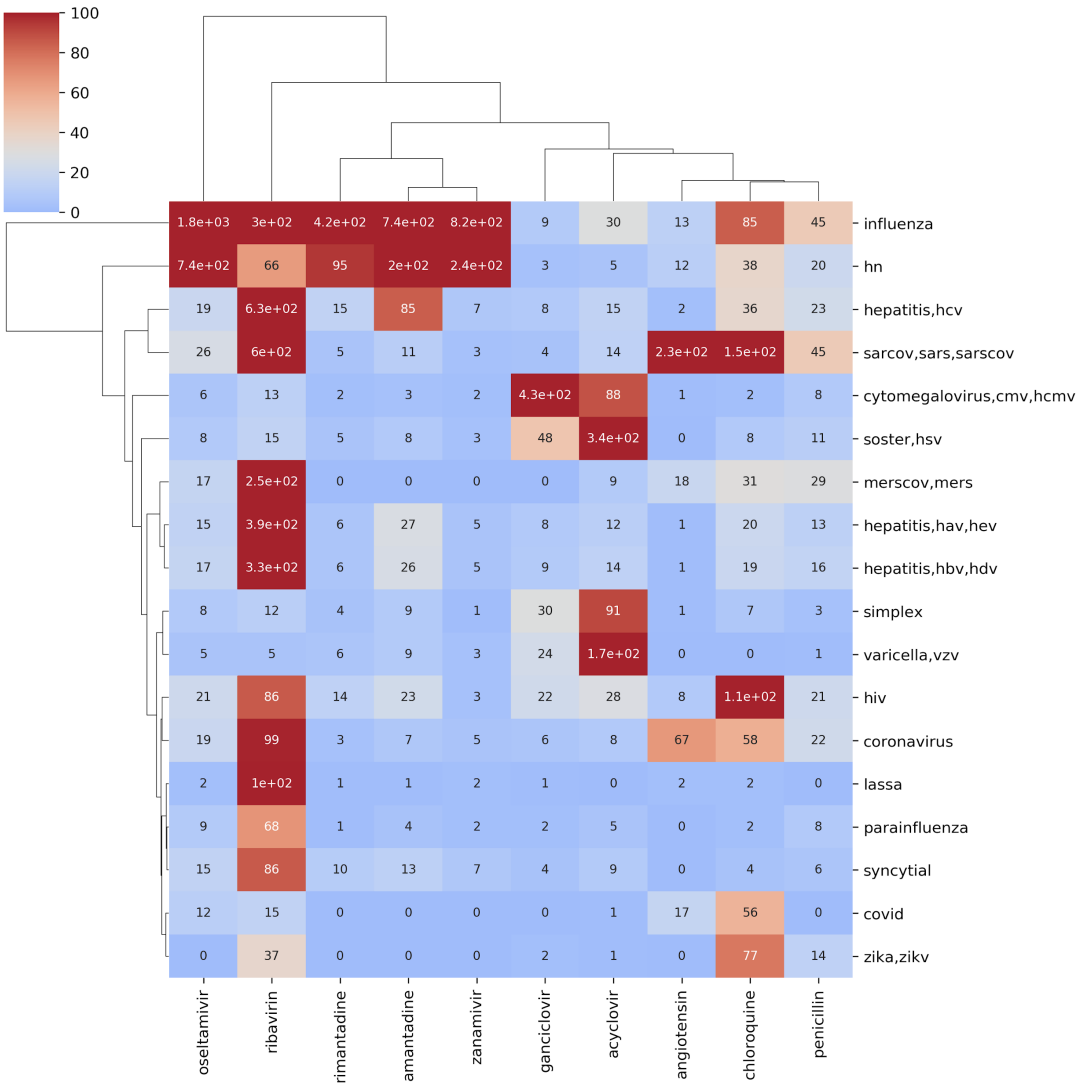

Figura 1 Matriz de coocurrencias entre las principales enfermedades virales y posibles medicamentos, representados en los artículos de la base de datos CORD-19

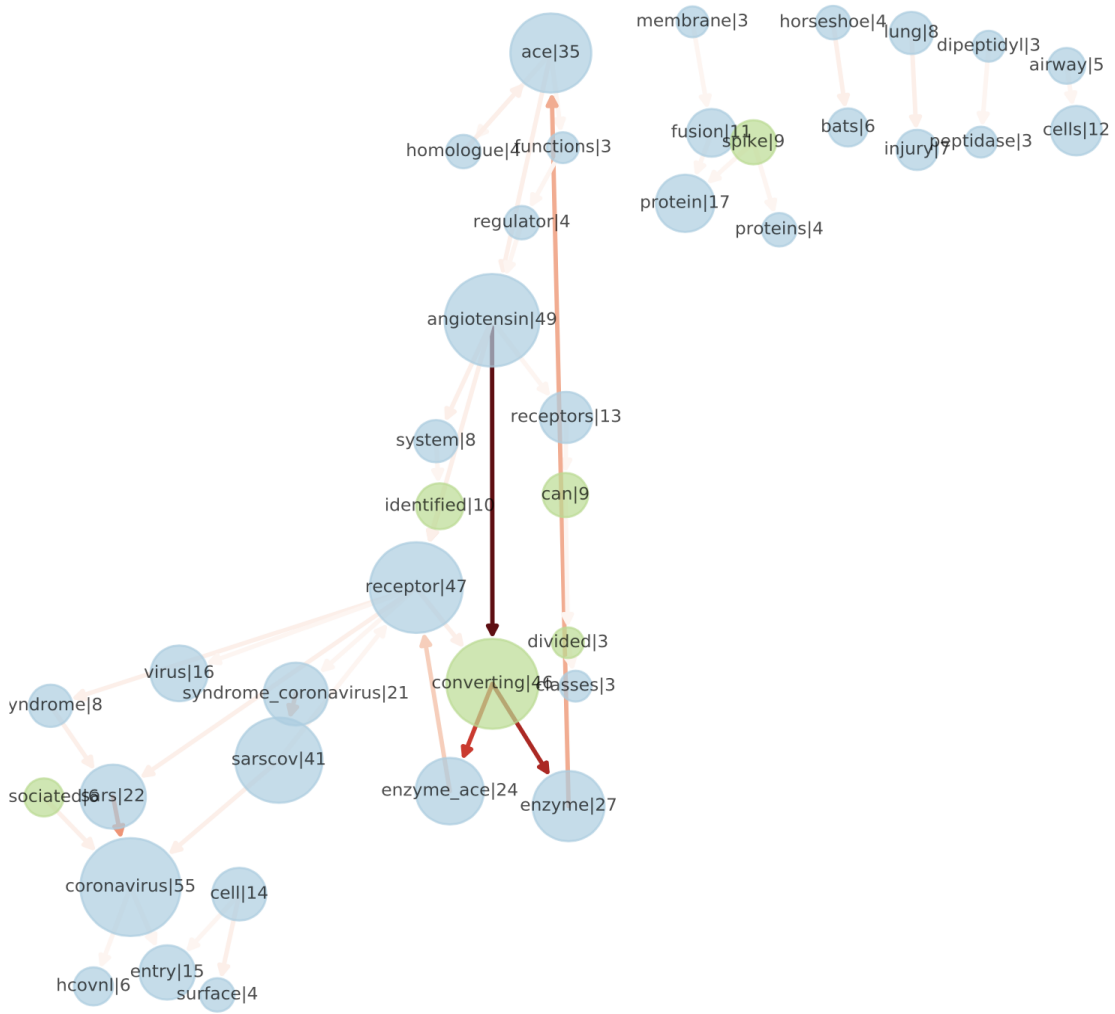

Figura 2 Grafo de conocimientos que representa las oraciones que contienen los vocablos *coronavirus* y *angiotensin.*

Grafos de conocimiento

Coronavirus - Angiotensina

En este grafo de conocimientos, la angiotensina se une estrechamente con diferentes sustantivos como *ACE*, *enzyme [enzima]*, *receptor, regulator [regulador]*, y el verbo *converting [convirtiendo],* lo que ilustra claramente la relación próxima entre la *angiotensina* y la *enzima ACE* y su acción. La Enzima Convertidora de Angiotensina (*ACE* por sus siglas en inglés), se encuentra en la superficie de la membrana de diferentes células del organismo, como ser en los pulmones, las arterias y el corazón; y se encarga de disminuir la presión arterial, regulando la cantidad de *angiotensina*, hormona responsable de la vasoconstricción y el posterior aumento de la presión arterial.

Por otra parte, la *angiotensina* se relaciona, mediante la palabra *receptor*, con diferentes enfermedades virales derivadas del coronavirus, tales como *virus, syndrome [síndrome], coronavirus, SARSCOV, SARS, HCOVN*. Esta relación de palabras no indica claramente una acción terapéutica de la *angiotensina* sobre el virus. Por otra parte existen algunas asociaciones de palabras que evocan la entrada del virus, a través de las membranas de las células como ser, *coronavirus, entry [entrada], cell [célula], surface [superficie], membrane [membrana], fusion [fusión], protein [proteina], proteins [proteinas], spike [pico - peplómero]*. Al estudiar citaciones importantes que se extrajeron de los artículos, se pudo verificar que la *enzima ACE* es el punto de entrada utilizado por los coronavirus para infectar las células. Finalmente, el estudio de dichas citaciones, confirma que la relación entre la *angiotensina* y el *coronavirus* no cumple una acción terapéutica, sino que se trata más bien de una relación indirecta, mediante la *enzima ACE*, que regula la cantidad de *angiotensina* por un lado y a su vez sirve de entrada al *coronavirus*.

Extractos textuales más representativos de los artículos

• "angiotensin converting enzyme has been shown to be a receptor for SARS coronavirus but the virus may also enter via lectins or by direct membrane fusion [La enzima convertidora de angiotensina ha demostrado ser un receptor para el coronavirus del SARS, pero el virus también puede entrar a través de lectinas o por fusión directa de membrana]"

• "the angiotensin converting enzyme ACE has been identified as a receptor for the severe acute respiratory syndrome associated coronavirus SARSCOV [la enzima convertidora de angiotensina ACE ha sido identificada como un receptor para el síndrome respiratorio agudo severo asociado al coronavirus SARSCOV]"

• "coronavirus entry is initiated by the binding of the spike protein S to cell receptors specifically dipeptidyl peptidase ddp and angiotensin converting enzyme ace for MERSCOV and SARSCOV respectively [La entrada del coronavirus se inicia por la unión de la proteína de pico S a los receptores celulares, específicamente la dipeptidil peptidasa DDP y la enzima convertidora de angiotensina ACE para MERSCOV y SARSCOV respectivamente]"

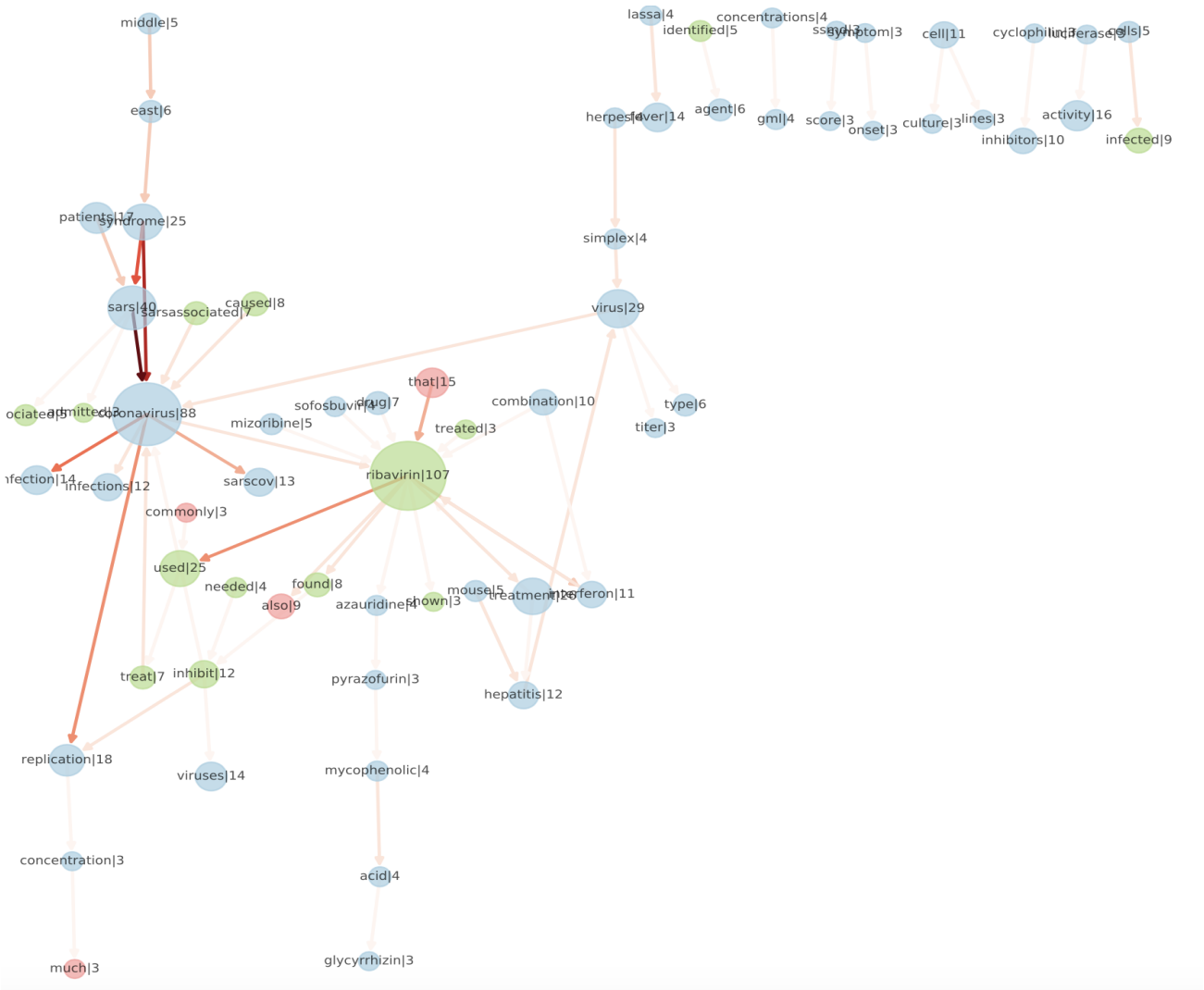

Figura 3 Grafo de conocimientos que representa las oraciones que contienen los vocablos *coronavirus* y *ribavirin.*

Coronavirus - Ribavirina

En este grafo de conocimientos, el *ribavirin [ribavirina]* se articula directamente con verbos en tiempo pasado y participio pasado (*found [encontrado], used [utilizado], shown [mostrado])*, enunciando los resultados de investigaciones, y poniendo en evidencia la posible acción terapéutica de la *ribavirina,* al articularla con el verbo en infinitivo *inhibit [inhibir]*, y el sustantivo *replication [replicación]*, término que denota el proceso biológico a través del cual el material genético de un organismo se duplica para formar más organismos. De igual manera, la *ribavirina* se relaciona con las enfermedades *hepatitis* y *herpes simplex*, mediante el sustantivo *treatment [tratamiento],* al igual que otros medicamentos que se utilizan de forma complementaria con la *ribavirina (mizoribine*

[mizoribina], sofosbuvir, ...). Sin embargo, la relación entre la *ribavirina* y estos otros medicamentos, y el *coronavirus* no se muestra de manera muy evidente. El sustantivo *coronavirus* se articula también al nombre de la afección viral *SARCOV,* al epicentro de esta enfermedad: *middle east [medio oriente]*, y a la patología misma, a través de los término *syndrome [síndrome]* y los sustantivos *infection [infección]*, *infections [infecciones]* y *patients [pacientes],* que caracterizan el accionar virus. Finalmente, el estudio de las oraciones más representativas extraídas de los artículos, mostró que si bien algunos trabajos valoraron como positiva la acción antiviral de la *ribavirina*, la mayor parte de las citaciones revelan su ineficiencia contra el coronavirus.

Extractos textuales más representativos de los artículos

• "Ribavirin has been extensively evaluated in severe acute respiratory syndrome coronavirus SARSCOV cell culture assays and used for actual treatment of SARS infections [La ribavirina ha sido ampliamente evaluada en ensayos de cultivo de células de coronavirus SARSCOV de síndrome respiratorio agudo severo y se utiliza para el tratamiento real de las infecciones por SARS]"

• "SARS associated coronavirus caused severe illnesses in most patients despite early treatment with ribavirin [El coronavirus asociado al SARS causó enfermedades graves en la mayoría de los pacientes a pesar del tratamiento temprano con ribavirina]"

• "Ribavirin might be ineffective against SARS associated coronavirus [la ribavirina podría ser ineficaz contra el coronavirus asociado al SARS]"

• "experimental study showed that ribavirin is ineffective against SARS associated coronavirus [un estudio experimental demostró que la ribavirina es ineficaz contra el coronavirus asociado al SARS]"

• "ribavirin does not inhibit severe acute respiratory virus SARS coronavirus in vitro [la ribavirina no inhibe el coronavirus respiratorio agudo severo SARS in vitro]"

• "however there was great debate as to whether ribavirin was effective against coronavirus [sin embargo, hubo un gran debate sobre la eficacia de la ribavirina contra el coronavirus]"

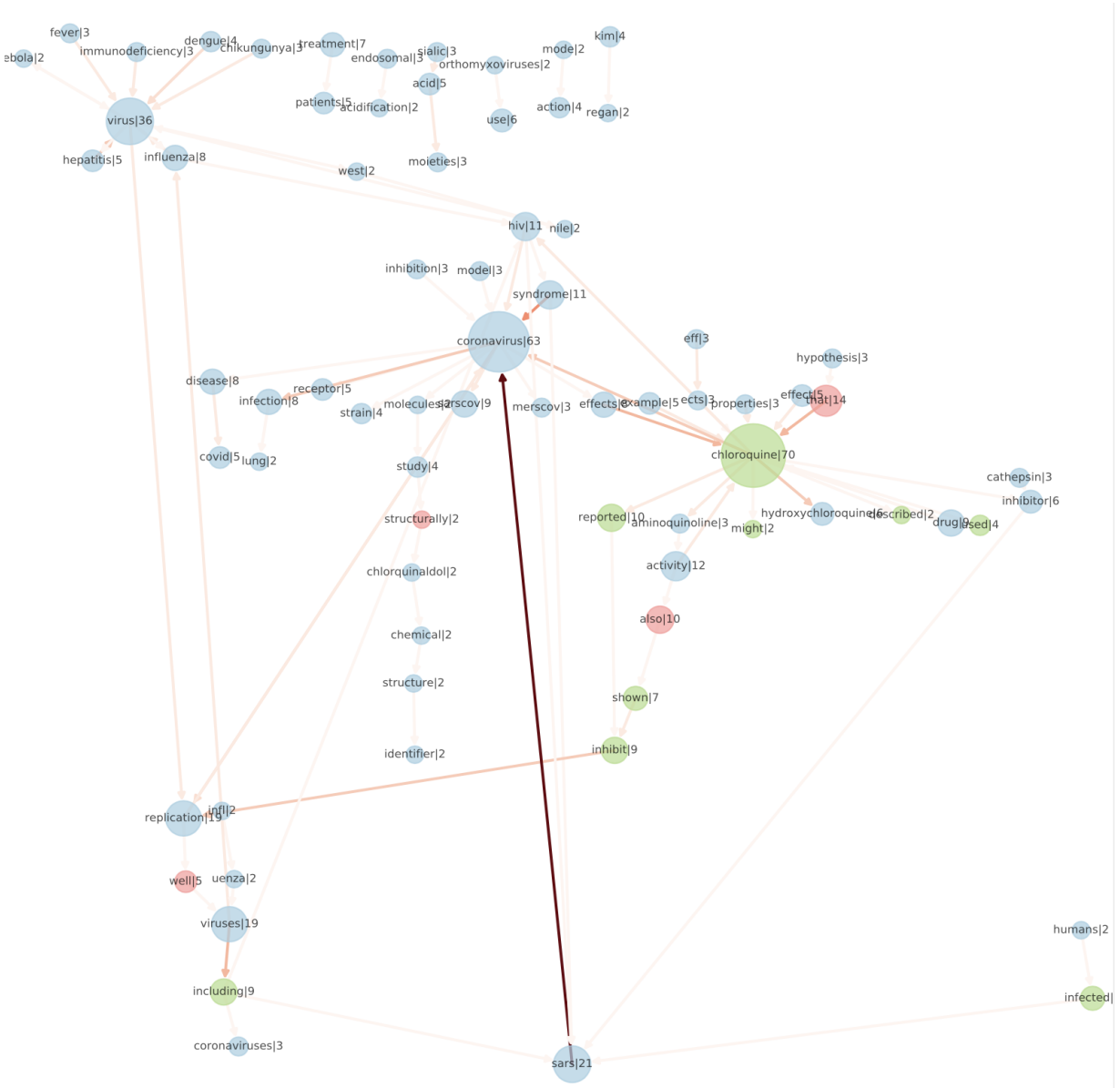

Figura 4 Grafo de conocimientos que representa las oraciones que contienen los vocablos *coronavirus* y *chloroquine.*

Coronavirus - Cloroquina

En este grafo de conocimientos, la *chloroquine [cloroquina]* se articula directamente con verbos en tiempo pasado y participio pasado (*reported [reportó]*, *described [descrito], used [utilizado], shown [mostrado])*, enunciando los resultados de investigaciones. La presencia del verbo en condicional *might [podría],* y del sustantivo *hypothesis [hipótesis]*, muestra que dichos resultados no se manifiestan categóricamente, lo cual es característico del discurso científico. Los verbos en pasado *reported [reportar]* y *shown [mostrar]*, ponen en evidencia la acción terapéutica de la *cloroquina,* al

articularla con el verbo en infinitivo *inhibit [inhibir]*, y el sustantivo *replication [replicación]*, término que denota el proceso biológico mediante el cual el material genético de un organismo se duplica para formar más organismos. De igual manera, la *cloroquina* se relaciona con las enfermedades *SARS* y *coronavirus*, mediante el sustantivo *inhibitor [inhibidor]*. Este conjunto de relaciones nos muestra, una vez más, que según los artículos considerados, la *cloroquina* actúa como un *inhibidor* de dichos virus (y de su *replicación*). El sustantivo *coronavirus* se articula con enfermedades causadas por virus pertenecientes a esta familia: *disease [enfermedad], COVID, SARCOV, MERSCOV, SARS*. El término *syndrome [síndrome]* y los sustantivos *infection [infección]* y *lung [pulmón]* caracterizan el cuadro clínico causado por este virus. Indirectamente, el sustantivo *coronavirus* se coordina con otras enfermedades virales como el *VIH,* la *influenza,* la *hepatitis,* el *ébola* y el *dengue*. Es interesante observar, en la matriz de la figura 1, la intensidad de la coocurrencia entre la *cloroquina* y estas afecciones virales, lo cual indica que posiblemente se probó este medicamento para luchar contra estas enfermedades. Finalmente, el estudio de las oraciones más representativas extraídas de los artículos, muestra que la acción terapéutica de la *cloroquina* ha sido estudiada in vitro, in vivo (en diferentes modelos animales) y en pacientes, subrayando su interés.

Citaciones importantes

• "chloroquine is a potent inhibitor of sars coronavirus infection and spread [La cloroquina es un potente inhibidor de la infección y de la propagación del coronavirus SARS]"

• "in vitro antiviral activity of chloroquine has been identified [Se ha identificado la actividad antiviral in vitro de la cloroquina]"

• "chloroquine is highly effective in the control of ncov infection in vitro [La cloroquina es muy eficaz en el control de la infección por ncov in vitro]"

• "different viruses can be inhibited in cell culture by both chloroquine and hydroxychloroquine including the sars coronavirus [diferentes virus pueden ser inhibidos en el cultivo celular tanto por la cloroquina como por la hidroxicloroquina, incluyendo el coronavirus SARS]"

• "in vitro experiments also showed a strong antiviral effect of chloroquine on a recombinant hcovo coronavirus [Los experimentos in vitro también mostraron un fuerte efecto antiviral de la cloroquina en un coronavirus HCOVO recombinante]"

• "our hypothesis that chloroquine might inhibit replication of the sars coronavirus has been confirmed in two independent in vitro studies [nuestra hipótesis según la cual, la cloroquina podría inhibir la replicación del coronavirus SARS, ha sido confirmada por dos estudios in vitro independientes]"

• "chloroquine potently inhibits the replication of a canine coronavirus" [la cloroquina inhibe potentemente la replicación de coronavirus canino]

• “effects of chloroquine in vivo have been shown only in a mouse model for coronavirus infection” [Los efectos de la cloroquina sobre la infección por coronavirus sólo se han demostrado en un modelo de ratón, in vivo]

• “possible benefit of chloroquine a broadly used antimalarial drug in the treatment of patients infected by the novel emerged coronavirus” [posible beneficio de la cloroquina, un fármaco antipalúdico de amplio uso, en el tratamiento de pacientes infectados por el nuevo coronavirus emergente]

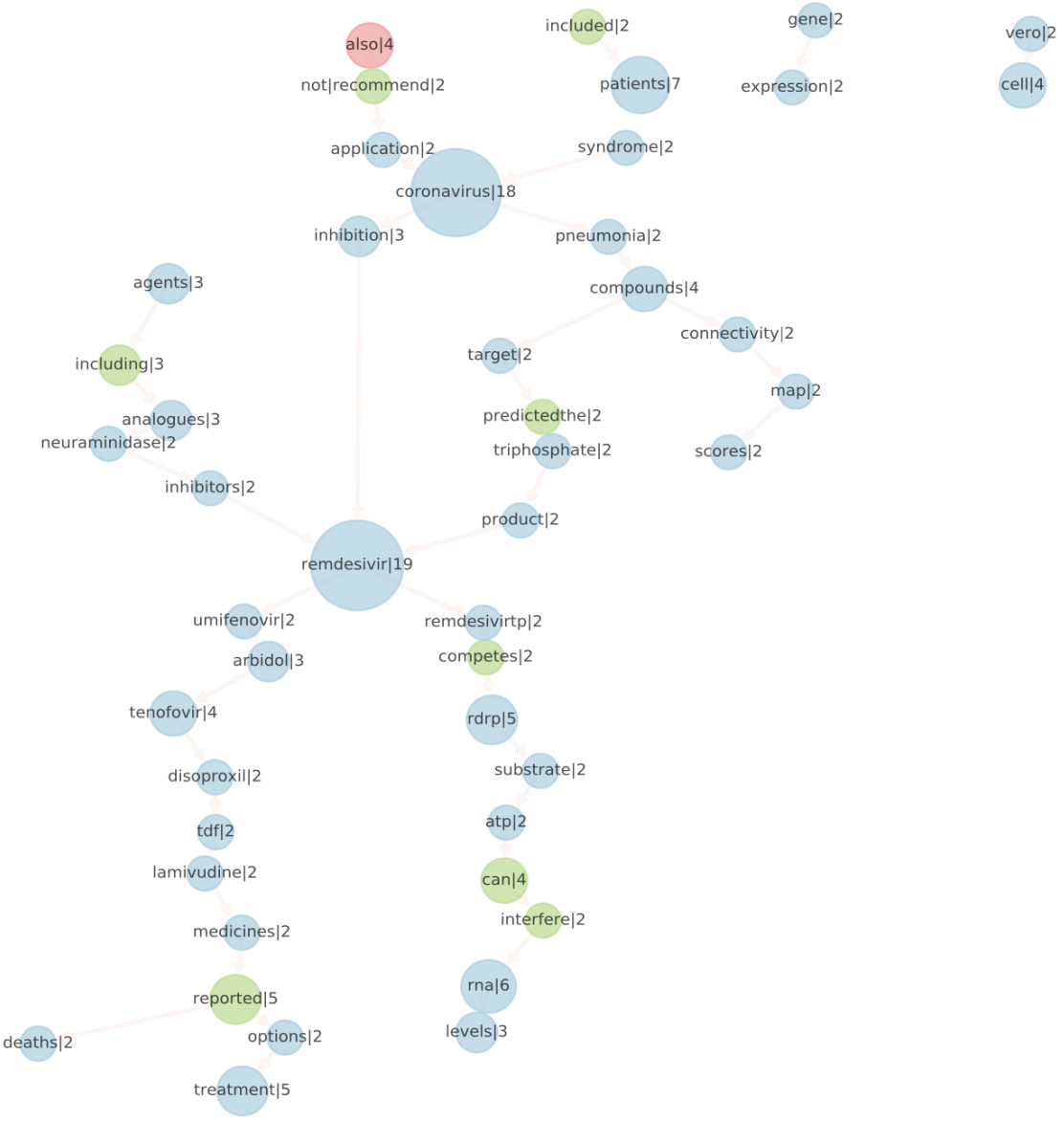

Figura 5 Grafo de conocimientos que representa las oraciones que contienen los vocablos *coronavirus* y *remdesivir.*

Coronavirus - Remdesivir

Dada la importancia que cobró el *remdesivir* en los últimos ensayos clínicos, se decidió incorporar en este estudio la relación entre el *coronavirus* y este medicamento. En este grafo de conocimientos, la relación entre el *remdesivir* y el *coronavirus* es muy tenue, al haber muy pocos artículos científicos que se refieren a ambos elementos en una misma oración. El *remdesivir* se une principalmente al *coronavirus* mediante el sustantivo *inhibition [inhibición],* lo cual puede hacer referencia a su acción terapéutica. Se observa una relación indirecta y débil, entre ambos conceptos, a través de una rama que contiene varios términos como *pneumonia [pneumonia], compounds [compuestos], target [diana], predicted [predicho], triphosphate [trifosfato], product [producto]*. Es interesante observar que el *remdesivir* está directamente relacionado, en las oraciones, a otros nombres de fármacos como ser la *neuraminidase [neuraminidasa],* el *umifenovir [umifenovir],* el *arbidol [arbidol]* y el *tenofovir [tenofovir]*. Finalmente, el estudio de las oraciones más representativas extraídas de los artículos, subrayan la calidad del *remdesivir* como antiviral de amplio espectro, y su posible rol terapéutico, pero no revelan muchas pruebas de la acción terapéutica (hasta la fecha de descarga del corpus).

Extractos textuales más representativos de los artículos

• “remdesivir is a promising broadspectrum antiviral that may prove useful in the treatment of highly pathogenic human coronavirus [remdesivir es un prometedor antiviral de amplio espectro que puede resultar útil en el tratamiento del coronavirus humano altamente patógeno]”

• “remdesivir potently inhibits human and zoonotic covs in vitro [remdesivir inhibe potentemente los coronavirus humanos y zoonóticos in vitro]”

• “coronavirus susceptibility to the antiviral remdesivir gs is mediated by the viral polymerase [La susceptibilidad del coronavirus al antivírico remdesivir gs está mediada por la polimerasa viral]”

• “Other agents including nucleoside analogues neuraminidase inhibitors remdesivi rumifenovir arbidol tenofovir disoproxil tdf and lamivudine tc along with several chinese traditional medicines are reported as viable options for antiviral treatment of human pathogenic coronavirus [Otros agentes, incluidos los análogos de los nucleósidos, los inhibidores de la neuraminidasa, el remdesivir, umifenovir, arbidol, tenofovir, disoproxil tdf y la lamivudina tc, junto con varios medicamentos tradicionales chinos, se consideran opciones viables para el tratamiento antiviral del coronavirus patógeno humano]”

• “remdesivir sofosbuvir galidesivir and tenofovir showed promising results for use against the newly emerged strain of coronavirus [remdesivir, sofosbuvir, galidesivir y tenofovir mostraron resultados @@prometedores para su uso contra la nueva cepa de coronavirus]”

• “certainly randomized controlled trials are needed to determine the safety and efficacy of remdesivir [ciertamente se necesitan ensayos controlados aleatorios para determinar la seguridad y la eficacia del remdesivir]”

Conclusiones

En este estudio se utilizó el *principio de cooperación* definido por Grice (1975) a través de sus cuatro *máximas*, para diseñar y emplear programas de minería de datos y de procesamiento de lenguaje natural, para facilitar la extracción de información relevante, contenida en un corpus de artículos científicos. Utilizando esta metodología interdisciplinaria, que combina la lingüística y la minería de datos, estudiamos un corpus de más de 52.000 artículos académicos recientes, relacionados a la pandemia COVID-19, con el fin de identificar los principales medicamentos estudiados para tratar el coronavirus, y caracterizar su mecanismo de acción y sus posibles beneficios. Tras cuantificar el grado de relación entre diferentes enfermedades virales, y medicamentos, mediante un análisis de coocurrencias de términos en los artículos, pudimos evidenciar que los medicamentos que estaban más fuertemente emparentados al término *coronavirus* eran: la *cloroquina*, la *ribavirina* y la *angiotensina*. Posteriormente, se procedió a efectuar un análisis de los grafos de conocimientos, generados a partir de todas las oraciones que contienen el vocablo *coronavirus* y cada uno de los tres medicamentos. Se añadió también al estudio el compuesto antiviral *remdesivir*, que cobró cierta notoriedad en los últimos ensayos clínicos. Este estudio reveló que: 1) La relación entre la *angiotensina* y el *coronavirus* no corresponde a una acción terapéutica, sino que más bien se trata de una relación indirecta, mediante la *enzima ACE*, que regula la cantidad de *angiotensina* por un lado y a su vez sirve de entrada al *coronavirus*. 2) La acción terapéutica de los otros compuestos parece haber sido estudiada en mayor o menor medida. Diversos estudios reportan la acción terapéutica de la *cloroquina* in vitro, in vivo (en diferentes modelos animales) y en pacientes. Si bien algunos trabajos valoraron como positiva la acción terapéutica de la *ribavirina*, la mayor parte de las citaciones revelan su ineficiencia contra el coronavirus. Las citaciones que conciernen al *remdesivir*, subrayan su calidad de antiviral de amplio espectro, y su posible rol terapéutico, pero no revelan demasiadas pruebas de la acción terapéutica.

Este estudio demuestra como teorías lingüísticas, articuladas de manera interdisciplinaria con otras disciplinas, pueden tener un importante rol social y científico. Esto se evidencia de manera tangible, en un contexto historico como el de la pandemia COVID-19. En efecto el desarrollo de esta metodología interdisciplinaria puede facilitar, a los especialistas, el acceso a una versión sintetizada, de la información contenida en diversos artículos científicos, para que estos puedan tomar decisiones, en beneficio de la sociedad, y orientar mejor sus esfuerzos de investigación. Por otra parte, es

interesante notar que un lingüista puede también complementar su formación, ampliando de esta manera su mercado laboral, ya que en esta época, el encargo social demanda la orientación de la lingüística hacia este tipo de especialización.

Bibliografía

Grice, H.P. (1975) Logic and conversation, Syntax and Semantics 3: Speech Acts, 3, 41-58.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., ... & Mooney, P. (2020). CORD-19: The Covid-19 Open Research Dataset. arXiv preprint arXiv:2004.10706.

FDA: Administración de Medicamentos y Alimentos, *FDA-Approved Drugs All Approvals and Tentative Approvals*, Versión 2020-04-03, https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=reportsSearch.process

Hulo, C., De Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., & Le Mercier, P. (2011). ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Research*, 39, D576-D582.

Van Rossum, G., & Drake, F. L. (1995). Python library reference.

McKinney, W. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51-56).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Hagberg, A., Swart, P., & S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404-411.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.