Locuciones Verbales del Español Mexicano

Un análisis desde la Lingüística Computacional

Angeles Belém Priego Sánchez
David Eduardo Pinto Avendaño

Libros BUAP

Primera edición: 2023
ISBN: 978-607-8957-02-6

DR Benemérita Universidad Autónoma de Puebla
4 Sur 104, Centro Histórico, Puebla, Puebla, C.P. 72000
Teléfono: 222 229 55.00 | www.buap.mx

Dirección General de Publicaciones
2 Norte 1404, Centro Histórico, Puebla, Puebla, C.P. 72000
Teléfono: 222 229 55 00, extensión 5768
libros.dgp@correo.buap.mx | www.publicaciones.buap.mx

Benemérita Universidad Autónoma de Puebla
Rectora: María Lilia Cedillo Ramírez | Secretario General: José Manuel Alonso Orozco | Vicerrector de Extensión y Difusión de la Cultura: José Carlos Bernal Suárez | Director General de Publicaciones: Luis Antonio Lucio Venegas

Soporte final PDF, 1.5 MB

Tabla de contenido

Índice de Tablas

Índice de Figuras

Capítulo 1. Introducción

El proceso de la comunicación se trata de un proceso complejo, dado que para llevarlo a cabo es necesario contar con un lenguaje (oral, escrito, señas, braille, etc.) que facilite expresar cualquier situación concreta y cada circunstancia determinada tomando como referente el horizonte cultural compartido. A lo largo de la vida, los seres humanos adquieren, captan y memorizan información que complementa la base de conocimiento léxica personal, permitiendo así, desarrollar la habilidad de comunicación.

En el lenguaje oral, esta información se representa en forma de oraciones, frases, expresiones, dichos, etc., los cuales permiten formular una idea o un concepto, visualizándolos como un todo capaz de expresar, de forma variada, una situación en contextos diferentes. Dicha información, en muchos casos se plasma de forma escrita en lenguaje natural en libros, periódicos, revistas u otros textos y en formato digital en idiomas como el español, francés, inglés o cualquier otro que sea de manejo de los seres humanos.

La información es un recurso importante para la humanidad, desde que ha sido factible almacenarla en una memoria secundaria (papiros, piedras, libros, dispositivos electrónicos, etc.), es posible analizarla y obtener resultados benéficos bajo la explotación de estas con la finalidad de obtener conocimiento de una lengua a través del acervo lingüístico y cultural que poseen.

Es a través de un campo de investigación denominado Lingüística Computacional (LC), también conocido como Procesamiento de Lenguaje Natural[1], que se desea plasmar en este libro los resultados obtenidos de diversos trabajos de investigación que tratan de un tópico específico relacionado con proceso de comunicación y que tiene que ver con expresiones verbales del español mexicano. Se tocan conceptos relacionados con la lingüística y la computación. En particular, ciertos temas vinculados con la computación y la inteligencia artificial, como la recuperación de información, el procesamiento masivo de información, la clasificación automática; pero también por el lado de la lingüística aplicada, se tocan temas como la lingüística de corpus y la fraseología.

En sí, el principal interés de investigación de este libro, se encuentra bajo el estudio de las expresiones del tipo *tomar el toro por los cuernos* (*coger el toro por los cuernos*, para el español de España; *prendre le taureau par les cornes*, para el francés de Francia), *colgar los tenis* (*acabarse la candela*, para el español de España; *casser sa pipe*, para el francés de Francia), entre otras, para la variante diatópica del español de México, en particular los mexicanismos. Es decir, se tiene interés por los sintagmas compuestos de dos o más palabras en las que al menos una de esas palabras es un verbo que juega el rol de predicado, denominadas locuciones verbales y en las que su significado global no se deduce necesariamente del significado de cada uno de sus componentes.

A partir de las locuciones verbales, se realiza un análisis de estas unidades fraseológicas y se construye una metodología capaz de validarlas de manera semi-automática en diferentes corpora y dominios. Se contribuye y se aporta así, una solución a una de las tareas de la lingüística computacional que tiene que ver con la validación automática de unidades fraseológicas.

[1] Un área formada por la intersección e interacción de la lingüística y la computación (Gelbukh & Sidorov, 2010)

1.1 Contextualización y motivación

La lengua permanece en constante cambio, normalmente está constituida por una gran cantidad de palabras, y en muchos casos uniones de éstas, que, aunque a veces no se considera una unión lógica, permite expresar un concepto determinado. Estas expresiones no son unidas libremente, sino que son utilizadas y repetidas durante años como bloque de combinaciones. Desde la antigüedad, la humanidad ha dedicado un gran interés por determinadas expresiones cotidianas; expresiones que quiso agrupar y estudiar. Este tipo de expresiones, que regularmente son fácilmente dominadas por los hablantes nativos de una lengua, plantean un gran desafío para las personas que desean aprender un segundo idioma, pero también es importante destacar la dificultad que implica estudiarlas e interpretarlas de manera automática por sistemas computacionales.

Estas expresiones no son tan frecuentes en los recursos léxicos como en los textos del mundo real y por lo tanto presentan un gran reto de estudio para diversas tareas de la lingüística computacional. Estos hechos motivan a que en este libro se tenga el interés particular de estudiar un tipo específico de uniones de palabras cotidianas en las cuales al menos una de esas palabras es un verbo: las locuciones verbales.

La fraseología, disciplina lingüística encargada del estudio de estas uniones de palabras, ha incrementado su importancia en los últimos años desde la perspectiva computacional y lingüística, dado que ha sido un espacio de interés para los investigadores atraídos en las combinaciones de palabras, por el hecho de ser un particular modo de expresión verbal y escrita de los pueblos. Como material de estudio científico implica tomar en cuenta diferentes dimensiones del lenguaje: lingüísticas, pragmáticas, culturales y muchas más. Este hecho refuerza más el interés de desarrollar este libro.

En la lingüística española se emplean diferentes denominaciones a la unión de estas palabras. A manera de ilustración se enumeran algunos nombres para estas secuencias de palabras, y sirva esta enumeración para dar una pequeña idea de la abundante

nomenclatura que existe: unidad fraseológica, fraseologismo, modismos, locuciones, expresiones fijas, dichos, frases hechas, expresiones idiomáticas, expresiones, frases, refranes, proverbios, modos de decir, aforismos, entre otras[2].

En este libro se utiliza el término "unidad fraseológica" para referirse a las combinaciones de palabras que tienen un significado como un todo. Esta denominación goza de una gran aceptación en la filología española, y es conocida en la fraseología internacional. Se hace hincapié de la existencia de diferentes tipos de unidades fraseológicas y se remarca que en este libro se analiza solamente un tipo particular de ellas, a decir, aquellas que contienen un verbo como el núcleo principal de sus componentes, llamadas también locuciones verbales.

Este libro está motivado en el estudio de las unidades fraseológicas dado que son ampliamente utilizadas por los humanos y representan un desafío para la lingüística computacional. Por un lado, por su comportamiento complejo, y por otro porque su tratamiento representa un objetivo importante para la amplia gama de fenómenos lingüísticos que son a menudo puntos débiles en muchas aplicaciones del procesamiento de lenguaje natural.

Existen estudios realizados por diferentes autores que dan una idea sobre la utilización de las unidades fraseológicas en una lengua. Así, Jackendoff (1997) estima que el número de unidades fraseológicas (esto expresado como *multiword expression*, en la terminología utilizada por el autor) en cualquier lengua humana es del mismo orden de magnitud que el orden de las palabras simples.

En el caso de las investigaciones de Danell et al. (1992:18), se estima que el número de *stock phrases* (traducido al español como frases hechas) en francés cubre entre un 20 y

[2] En el artículo de M. Martins-Baltar (1997) aparecen hasta 64 denominaciones diferentes.

30% de un texto. Según Mel'cuk (1993:83), el número de *phrasèmes* (traducción al español como frasemas) en la lengua francesa crece a decenas de millares. En un volumen consagrado a la locución (Martins-Baltar, 1997), G. Gross (1997:202) remarca que hay "aproximadamente 200,000 sustantivos compuestos o cuya variación no es libre, cerca de 15,000 adjetivos y al menos 30,000 verbos fijos". Estas cifras no son absolutas, dado que las expresiones referidas en estas citaciones no han sido calculadas a partir de la misma definición. Sin embargo, ofrecen indicios sobre la importancia de las unidades fraseológicas en una lengua.

Es importante destacar que este libro se centra principalmente en la investigación y desarrollo de una metodología para la validación semi-automática de locuciones verbales, la cual consiste en la utilización y análisis de los recursos léxicos recopilados (diccionario de locuciones verbales y corpora), a partir de los cuales se logra identificar y validar a estas unidades fraseológicas utilizando modelo computacionales, utilizando diferentes características extraídas a partir de los textos en donde aparecen las locuciones verbales.

1.2 Antecedentes

La técnica del discurso comprende las unidades del vocabulario de la lengua, más los elementos y reglas necesarios para su combinación en el hablar (Coseriu, 1966). En el dominio de la fraseología es evidente que se debe de tener en cuenta la existencia de palabras prefabricadas que constituyen texto repetido y que están formadas por una combinación fija de dos o más palabras. La frecuente repetición de combinaciones de palabras siempre idénticas es el elemento inherente del lenguaje natural que ocupa un papel central en la descripción de las unidades fraseológicas y que es denominada fijación, la cual señala el rasgo característico y definitorio de dichas unidades.

Las unidades fraseológicas son estables o fijas, ni el orden de sus componentes, ni las categorías gramaticales a las que pertenecen esos componentes, ni los componentes

mismos de una unidad fraseológica se pueden cambiar, pues pierde su significado. Por fijación o estabilidad formal se entiende "aquella propiedad que tienen ciertas expresiones de ser reproducidas en el hablar como combinaciones previamente hechas" (Zuluaga, 1975).

La problemática de la fijación es un fenómeno interesante para la lengua y ha tenido un auge remarcable, sobre todo en los estudios referentes al léxico, gracias a que los especialistas en la materia han entendido que se trata de un fenómeno transversal que abarca todas las partes de la oración (Mejri, 1997), lo que implica todas las dimensiones del lenguaje: lingüísticas, pragmáticas, culturales, entre otras (Mejri, 2010).
La mayoría de los estudiosos de la lengua concuerdan en decir que la fijación es una propiedad inherente a las lenguas naturales, lo que le permite ocupar un lugar central dentro de la descripción de éstas, y es por tanto tomada en cuenta en trabajos que tengan una perspectiva práctica como la traducción, la constitución de diccionarios, la enseñanza de idiomas, en tareas de procesamiento de lenguaje natural como la clasificación automática, etc.

Al hablar de procesamiento de lenguaje natural, es posible percatarse que la fijación impide contar con una modelización adaptada a las exigencias de los sistemas computacionales, y de ahí, surge la necesidad de contar con recursos lingüísticos suficientemente vastos y claramente estructurados para ser automatizados. Como se ha mencionado con anterioridad se tiene el interés en validar semi-automáticamente a locuciones verbales porque presentan un alto grado de fijación, en comparación con otras unidades fraseológicas (Mejri, 2008). Un proceso que implica analizar un texto plano y las características en el contexto de la unidad fraseológica para crear algoritmos computacionales capaces de completar la tarea mencionada en entornos escalables.

Al hablar de manera general del procesamiento automático de las unidades fraseológicas en un texto, se deben tomar en cuenta dos tareas de suma importancia: 1) la localización automática de las mismas y 2) su análisis sintáctico y semántico. La primera tarea debe

encarar la dificultad que se plantea en varios estudios realizados, la cual consiste en el hecho de que la mayoría de las unidades fraseológicas presentan una misma identidad formal como las secuencias libres (dar atole con el dedo vs dar un vaso de agua). La segunda tarea plantea la dificultad de la relación entre el sentido de la unidad fraseológica y su sintaxis.

La comunidad de lingüística computacional ha empezado a interesarse por la fijación de las estructuras sintácticas y a evaluar la importancia del aspecto semántico de este fenómeno. En este libro se toma en cuenta esta propiedad de las unidades fraseológicas porque implica la memorización de estas estructuras lingüísticas como si constituyeran un todo inseparable, tal y como se almacenan las unidades simples.

1.3 Descripción general del problema

El aumento en la popularidad de los repositorios de información, como bases de datos documentales o las bibliotecas digitales, Internet y la *World Wide Web* ha proporcionado tener un enorme repositorio de información para los estudiosos de la lingüística computacional, lo que les ha permitido tener más recursos para analizar una lengua.

La comunicación existe cada vez más es en formato escrito que oral, y con el crecimiento digital del que actualmente se forma parte, se ha provocado la necesidad de gestionar esta información de manera rápida y eficiente. Este ha sido uno de los factores clave en la consolidación de la computación, procesar grandes volúmenes de texto tratando de agruparlos según la información contenida. La meta es construir herramientas capaces de analizar la información y brindar un conocimiento.

A pesar de que la comunicación poco a poco se ha ido modificando, se posee el recurso más importante que es la información, la cual se guarda primordialmente en lenguaje

natural. La lengua y su conocimiento se mide generalmente por el grado de dominio con que un hablante utiliza las unidades fraseológicas propias y las características de esa lengua.

Las unidades fraseológicas forman, para los usuarios nativos, un saber lingüístico y cultural común depositado en su memoria por vía de la experiencia; sin embargo, para los usuarios no nativos, que aprenden un segundo idioma, son un reto por la dificultad, complejidad y la gran cantidad de información lingüística, social y cultural (referentes culturales) que deben de aprender a utilizar y a diferenciar de las características propias de su idioma materno.

Tómese, por ejemplo, el caso de la tarea de traducción, uno en el que el conocimiento de la lengua es importante. Si se desea traducir automáticamente del español al francés (o al inglés) una parte de la información, por ejemplo, la frase *Juan ha tomado el toro por los cuernos;* entonces surge la pregunta: ¿cómo se distinguiría el sentido literal del sentido composicional? Es decir, como se sabría si Juan realmente ha tomado el toro por los cuernos (significado literal o composicional) o que Juan ha enfrentado directamente un problema (significado figurativo o no composicional). Posiblemente, la respuesta sería que se podría deducir a través del análisis del contexto de dicha locución verbal.

La frase *Juan ha tomado el toro por los cuernos* será traducida en francés por *Juan a pris le taureau par les cornes*, o en inglés *Juan takes the bull by the horns*. En este ejemplo, la traducción no presenta dificultades porque el sentido literal y composicional se traduce de la misma manera. Sin embargo, no se puede generalizar estos casos. Si se toma otra frase en su sentido no composicional (figurativo), por ejemplo: *darse a conocer*; en francés está frase puede equivaler a *montrer patte blanche* que significa literalmente darse a conocer, pero ¿qué pasa al traducir de francés a español? Se tendría una traducción errónea porque *montrer patte blanche* sería traducido como *mostrar pata blanca* y esa traducción nada tiene que ver con el significado esperado de: *darse a conocer*.

Si el problema descrito anteriormente se lleva al ámbito del reconocimiento automático de información, ¿cómo la computación distinguiría el sentido global de una unidad fraseológica? Es a través de la incorporación de análisis léxicos, morfológicos, sintácticos y semánticos apropiados que se podría hacer distinción entre los sentidos o significados de una frase.

La información en un texto se conforma de párrafos, cada párrafo por un conjunto de oraciones, y éstas por unidades de texto más pequeñas que la oración llamadas frases que se obtienen a través de la descomposición de la oración. Cada frase tiene información independiente que puede ser usada como una unidad independiente (Hovy, Zhou, & Kwon, 2007). La inmensa mayoría de frases está formada por un verbo y una o varias variables. El verbo exige y realiza una rigurosa selección de los sujetos y de los componentes que pueden acompañarle. Estas frases se encuentran fusionadas en la oración para enunciar algo de manera más amplia, pero al separarse de la oración tienen sentido completo, es decir, tienen información semántica por ellas mismas.

La segmentación de una oración en palabras es tal vez la primera operación efectuada en un tratamiento automático de la lengua. Pero el término "palabra" es lingüísticamente inapropiado porque corresponde en el ámbito computacional a una entidad, llamada "token", delimitada por separadores gráficos (blancos, retorno de línea, entre otros delimitadores). La noción de palabra es mucho más compleja, y cuando se dice complejo se hace referencia a la dificultad posible que determina su polilexicalidad. En efecto, mientras que los estudiosos de la computación se concentran sobre la palabra simple, los lingüistas se concentran en las palabras complejas que son también importantes en el tratamiento de las lenguas.

Este libro propone la unión de los conocimientos computacional y lingüístico para el tratamiento automático de las unidades fraseológicas, que debe de ser tratado correctamente porque la clasificación correcta de estas unidades es útil para numerosas

aplicaciones como la traducción, la extracción de información, la clasificación, la constitución de diccionarios, la enseñanza de idiomas, entre otras.

A partir de lo descrito anteriormente, el interés de este libro es el de profundizar sobre las unidades fraseológicas, debido a que la mayoría de los trabajos ya elaborados (o en curso), están considerablemente centrados en las unidades simples de la lengua. Además de que estas unidades fraseológicas, al igual que las unidades simples, son numerosas.

Para contribuir al estudio de las unidades fraseológicas, en este libro se construye una metodología computacional capaz de validar semi-automáticamente a las locuciones verbales de la variante diatópica del español mexicano en corpora de diferentes dominios. Es decir, se analiza lingüísticamente a un conjunto de unidades fraseológicas de particular interés, y a partir de ese análisis son localizadas en texto plano. La extracción en sí no es una tarea fácil; sin embargo, la validación de locuciones verbales es en sí una tarea mucho más compleja.

1.4 Objetivos de investigación

Comencemos por definir un objetivo general de investigación, el cual puede ser expresado como sigue:

Desarrollar una metodología para la validación semi-automática de locuciones verbales candidatas del español de México con la interrelación de la lingüística y de la computación como campos interdisciplinarios de la investigación.

A partir del objetivo anterior, se plantean dos objetivos más: un objetivo computacional y un objetivo lingüístico.

1.4.1 Objetivo computacional

El objetivo de este libro por la parte computacional es modelar y desarrollar una metodología que sea capaz de validar semi-automáticamente a las locuciones verbales para la variante del español de México en diferentes corpora. De este objetivo se determinan los siguientes objetivos particulares:

1. *Construir y preprocesar recursos léxicos que servirán como base para el desarrollo de la metodología planteada.* Haciendo uso de técnicas computacionales se crea una base de conocimiento de locuciones verbales y un diccionario de contextos asociados a estas unidades fraseológicas para el español de México.

Se sabe por experiencia propia que uno de los principales obstáculos para abordar una tarea de procesamiento de lenguaje natural es la escasez de recursos (estos recursos pueden ser corpora, lexicones, ontologías, diccionarios, etc.), y a partir de este hecho se toma de apoyo los métodos computacionales para la creación de dichos recursos.

2. Producir y compartir recursos computacionales que posteriormente podrán ser utilizados por la comunidad científica de lingüística computacional. Se diseñan y ponen a disposición un conjunto de herramientas capaces de apoyar en el manejo y el entendimiento de locuciones verbales para el español de México.

El objetivo es crear recursos informáticos modulares y reutilizables para ciertas tareas que incluyen: la creación, el manejo y la consulta de bases de conocimiento y/o la compilación automática de corpora.

3. *Establecer la metodología para la validación semi-automática.* Se analiza, desde una perspectiva computacional, una metodología para la validación de locuciones verbales candidatas del español de México.

Para este punto se utilizan técnicas de procesamiento de lenguaje natural, basadas en Inteligencia Artificial, en particular de aprendizaje automático, las cuales están sustentadas en modelos computacionales que incluyen conceptos matemáticos y la estadística.

1.4.2 Objetivo lingüístico

El objeto de estudio de este libro pertenece al área de la lingüística aplicada y por tal motivo se plantean los siguientes objetivos particulares:

1. *Dotar al lector de una noción general del concepto "unidades fraseológicas".* Se estudia el estado del arte de las unidades fraseológicas haciendo principal énfasis en las locuciones verbales, con el fin de tener diferentes enfoques y determinar cuáles de ellos pueden ser aplicados computacionalmente.

2. *Analizar los recursos léxicos construidos a partir de la lingüística de corpus.* Se estudian los conceptos de lingüística de corpus con la finalidad de recolectar adecuadamente conjuntos representativos de información asociada a las unidades fraseológicas verbales, los cuales son objeto de investigación lingüística.

3. *Describir lingüísticamente a las locuciones verbales del español de México.* El objetivo lingüístico esencial es analizar las locuciones verbales del español de México y dar una descripción de su comportamiento.

1.5 Hipótesis propuestas para la validación de locuciones verbales

En este libro se introducen, un conjunto de hipótesis útiles para la validación de locuciones verbales, mismas que han sido propuestas, diseñadas y desarrolladas por los autores. Estas hipótesis asociadas con las unidades fraseológicas verbales dan soporte, guían y determinan el desarrollo del presente libro. Las hipótesis por verificar o refutar son las siguientes:

1. *Hipótesis de fijación.* Entre más fijación tenga una unidad fraseológica, es más alta la probabilidad de ser locución verbal.
   En esta hipótesis se sustituye cada componente de la locución verbal de entrada (LV candidata) por sus sinónimos cercanos y se verifica la nueva unidad verbal. En este sentido, se considera que el fenómeno de fijación existe, verificando que la LV candidata es realmente una locución verbal sin que pierda su significado. Con el fin de verificar el significado de la unidad verbal, se considera la utilización de un corpus de referencia en el que se busca evidencia de dicha frase.

2. *Hipótesis de traducción.* Entre más literal la traducción de una unidad fraseológica, menor su probabilidad de ser una locución verbal.

   En esta hipótesis se traduce la locución verbal a alguna otra lengua (en este caso, francés) utilizando diccionarios estadísticos de traducción y se verifica si la nueva locución verbal (locución verbal en francés) tiene evidencia de ser realmente una locución verbal en el corpus de referencia (corpus en francés).

3. *Hipótesis de atracción interna y correlación contextual.* Entre mayor sea la atracción interna y menor la correlación contextual en una frase verbal, es más alta la probabilidad de que la frase verbal sea una locución verbal.

   En esta hipótesis empleamos métodos estadísticos para determinar el nivel de atracción y correlación contextual entre los términos de la unidad fraseológica verbal y los de su contexto.

4. *Hipótesis del dominio de los documentos.* Entre mayor sea la cantidad de palabras de la locución verbal que se encuentren fuera del dominio al que pertenece el contexto, mayor su probabilidad de ser locución verbal.

En esta hipótesis se identifican los términos del vocabulario en el dominio actual, con el propósito de determinar si existe realmente o no una unidad fraseológica verbal.

1.6 Alcance e interés del libro

La lengua como sistema de comunicación verbal, gestual o escrita propia de una comunidad humana es una de las mayores riquezas que posee el ser humano, de la cual J. Sanmartín (1998) dice "la lengua es un continuo, un espacio sin límites ni fronteras, donde los lingüistas y estudiosos intentan la quimera de apresar con diversas etiquetas realidades heterogéneas y en permanente cambio".
Si la lengua está en permanente cambio, entonces se podría pensar ¿cuántas existen en el mundo? Y además las variantes que cada una tiene. Actualmente no se puede saber con exactitud el número de lenguas existentes, esto debido a la continua aparición y desaparición de lenguas que se va produciendo a lo largo de los siglos. Se calcula que en el mundo se hablan en la actualidad entre 3,000 y 5,000 lenguas, de las cuales solamente 600 cuentan con más de 100,000 hablantes, cifra que se considera mínima para garantizar su supervivencia a mediano plazo (Sanz, 2015).

Entre los idiomas más extendidos están el chino mandarín, usado por 900 millones de personas; el inglés, con 470 millones de hablantes; el hindi, hablado por más de 420 millones de personas; el español, utilizado por 360 millones; y el ruso, con casi 300 millones de hablantes (Cervantes, 2014; Sanz, 2015). Una vez que se saben estas cifras aproximadas, se delimita aún más el objeto de estudio de este libro, centrándose en el estudio del idioma español, con el objetivo de que la metodología propuesta pueda validarse y de manera esperanzadora, pueda funcionar también para otras lenguas.

El idioma español o castellano es una lengua romance procedente del latín vulgar, es una continuación moderna del latín hablado. Debido a su propagación por América, el

español es la lengua romance que ha logrado mayor difusión. El español ha tenido una gran expansión geográfica y por tal motivo presenta una gran variabilidad diatópica, entonces no existe un español, pero si diversas variedades del español.
En el español se distingue el español peninsular y el español de América. Paralelamente a esas variantes, se distingue en la primera: el andaluz, el canario y el castellano; la segunda comprende entre otras el mexicano, el peruviano, el cubano, el argentino, el colombiano, etc. Se encuentran diferencias significativas entre las dos grandes variedades diatópicas de español, tanto sobre en el plano lingüístico como en el plano extralingüístico.

Al saber que la lengua es un enorme repositorio que está en permanente cambio y crecimiento, se desearía estudiarla de una manera más amplia. Sin embargo, dadas las limitaciones físicas y temporales en este libro se tiene el interés particular de estudiar una parte específica del lenguaje que muestra una fracción de la cultura de una lengua; se estudia particularmente la fraseología. Se está consciente que la fraseología envuelve diferentes niveles de la lengua y que su principal objeto de estudio son las unidades fraseológicas. Dada la diversidad de unidades fraseológicas que existen en la lengua, esta investigación se limita en el estudio de las locuciones verbales. Con la ayuda de la lingüística computacional se desarrolla una metodología capaz de validar locuciones verbales semi-automáticamente en diferentes corpora y dominio.

En particular, el libro está enmarcado en el área de la lingüística computacional, comprendiendo el estudio de una parte de la fraseología que involucra a las locuciones verbales y centrándose para la construcción de la metodología para la validación semi-automática de estas unidades fraseológicas en el español hablado en México, el cual tiene cambios a nivel de pronunciación, de léxico y de morfosintaxis con respecto a las otras variantes del español.

La metodología podría ser capaz de identificar estas secuencias lingüísticas en diferentes idiomas, dado que se han realizado los experimentos con el español de México, el español de España y con el francés de Francia. Quizá se puede decir que es debido a que el español

y el francés pertenecen a las lenguas romance; sin embargo, con el idioma inglés que es una lengua germánica occidental se aplicó la metodología planteada y los resultados fueron satisfactorios.

Se abordan temas de interés del área que están relacionados con la fraseología, la lingüística computacional, la validación semi-automática de las locuciones verbales mexicanas. Se consideran características especiales inherentes a la naturaleza propia que tienen estas estructuras lingüísticas, a partir de las cuales se obtiene una base de datos de locuciones verbales de México, corpora del español mexicano y, para llevar a cabo una de las hipótesis planteadas, se construye una base de datos de locuciones verbales y corpora en francés de Francia.

1.7 Organización del libro

Para un claro entendimiento, el presente libro se compone de cinco capítulos. La descripción de estos sigue a continuación:

En el Capítulo 2 se presenta una recopilación y evaluación de los recursos léxicos utilizados para la realización de este libro, que incluye a las unidades fraseológicas verbales y los diferentes corpora utilizados.

El Capítulo 3 está dedicado a la identificación de contextos asociados a las locuciones verbales, es decir, se describe como se llevó a cabo el filtrado automático y manual de los contextos asociados a estas estructuras lingüísticas con la finalidad de construir corpora supervisado para las tareas de validación semi-automática de locuciones verbales. Adicionalmente, se presenta un lexicón de locuciones verbales, cada una con un valor asociado a su probabilidad de tener un significado figurativo o literal.

En el Capítulo 4 se presenta la metodología realizada para la validación semi-automática de locuciones verbales, la cual consiste en la utilización y análisis de los recursos léxicos recopilados (diccionario de locuciones verbales y corpora), a partir de los cuales se logra identificar y validar a las locuciones verbales utilizando un modelo de clasificación y añadiendo diferentes características a éste. Se considera a esta metodología una de las principales aportaciones científicas de esta obra literaria.

En el Capítulo 5 se presenta un estudio de polaridad de las locuciones verbales, que es otra de las tareas que derivó del desarrollo del presente libro. Para su análisis se consideran dos perspectivas, en un caso se toma en cuenta el contexto de la locución, y en el otro caso, no se considera el contexto de dicha unidad fraseológica.

Capítulo 2. Recuperación y evaluación de recursos léxicos

El interés por el estudio de las unidades fraseológicas ha demandado la construcción de recursos léxicos capaces de mostrar la riqueza que una lengua tiene y su descripción lexicográfica. Al hablar de recursos léxicos[3], se hace referencia a los diccionarios, lexicones, thesauri y corpora que han sido desarrollados con el objetivo de ofrecer un panorama general y, en algunos casos, particular de algunas de las unidades fraseológicas. A partir del estudio realizado en este libro, este capítulo tiene como objetivo mostrar los recursos léxicos desarrollados para poder dar solución a la problemática de la validación semi-automática de locuciones verbales del español mexicano.

Se encuentran diferencias significativas entre las dos grandes variantes diatópicas del español, tanto en el plano lingüístico como en el plano extralingüístico. De esta manera, la lengua española hablada en México ha sufrido cambios a nivel de la pronunciación, a nivel del léxico y a nivel de la morfosintaxis; por esta razón, se centra la atención de esta investigación en esta variante diatópica del español.

Se ha podido observar que existe una tendencia mundial en la generación de recursos léxicos para el idioma inglés, aunque existen también recursos para otras lenguas, por lo cual se puede decir que dicha lengua es una de las mayormente estudiadas y que el estudio de otras lenguas demanda precisamente la construcción de sus propios recursos léxicos. De manera particular, se ha notado la escasez de recursos léxicos para el español, posiblemente porque no han sido publicados o porque están en desarrollo; y qué decir, si

[3] Los recursos léxicos que han servido de apoyo para los objetivos propios de este libro se describen más adelante en este mismo capítulo.

se delimita más el estudio, en el caso de la variante diatópica del español hablado en México, se han encontrado comparaciones del mexicano con otras lenguas, pero no recursos léxicos.

A lo largo de la construcción de este libro, se ha podido observar que la necesidad principal y la parte más difícil, para realizar el análisis de una lengua, es la construcción de recursos léxicos. Los recursos léxicos son la materia prima principal que permite estudiar los fenómenos que existen en una lengua. En este libro se llevó a cabo la construcción de diferentes recursos léxicos con el objetivo de identificar y validar un tipo de unidades fraseológicas, las locuciones verbales, para el español con la variante mexicana. Los recursos léxicos, para el español de México, construidos en este libro comprenden:

- Una base de datos de locuciones verbales obtenida a partir del Diccionario de Mexicanismos (2010) con la traducción de equivalencia.

- Un corpus periodístico obtenido a partir del sitio de internet de la Organización Editorial Mexicana (OEM) que agrupa periódicos mexicanos.

- Un corpus de comentarios obtenido a partir de las bases de Twitter.

Los recursos léxicos arriba mencionados son descritos desde su recuperación (manual o automática) hasta su evaluación en las siguientes secciones de este Capítulo. La construcción de estos recursos ha permitido poner en evidencia el comportamiento del español de México. Estos recursos desarrollados han sido construidos para una tarea específica, sin embargo, podrían ser utilizados y servir para otros investigadores interesados por la variante del español mexicano, así como para otros tipos de tareas en el dominio de la lingüística computacional o en los dominios de investigación cubiertos de una manera independiente.

2.1 Corpora: el caso del español mexicano

La lengua presenta diferentes variedades (cf. Figura 2.1) producidas por distintas causas que se derivan del proceso de la comunicación; variedades como cambios a través del tiempo (variedades diacrónicas), variedades de acuerdo al lugar geográfico en el que se hable la lengua (variedades diatópicas), variedades de acuerdo a la situación (formal, informal o coloquial) en la que se encuentra el hablante (variantes diafásicas), variedades dependiendo de la cultura del hablante (variedades diastráticas). Estas variedades, en algunos casos, ocurren simultáneamente y en ocasiones no de forma independiente.

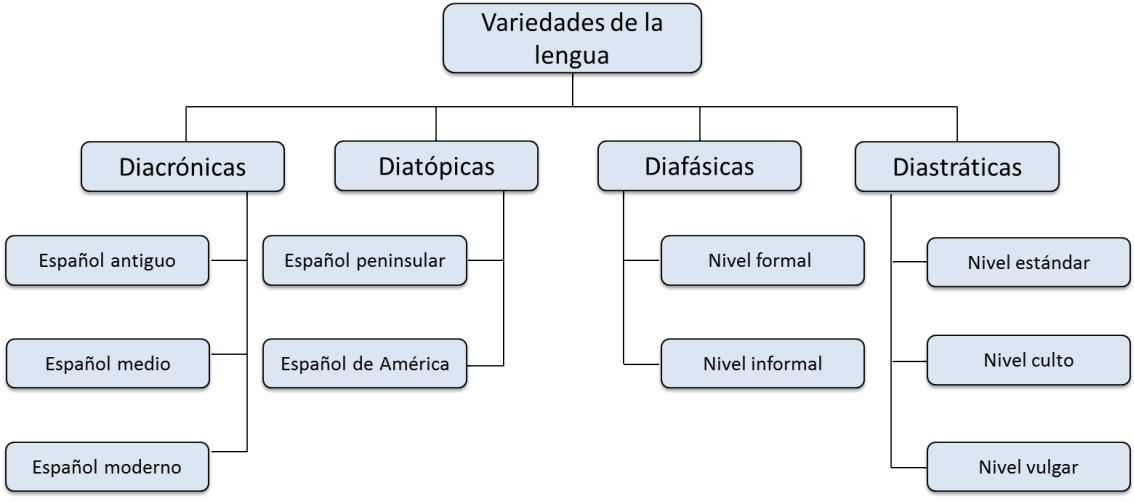

Figura 2.1. Variedades de la lengua: el caso del español.

Coseriu (1981) distingue tres tipos de variedades de la lengua: las diatópicas (en el espacio, dialectos), las diastráticas (niveles de lengua o sociolectos) y las diafásicas (estilos de lengua). Las primeras son las variedades geográficas; según Coseriu (1981), el dialecto se caracteriza por una forma de hablar, por subordinarse a una lengua histórica y por estar delimitado espacialmente. Las segundas son las variantes sociales o diastráticas, también llamadas "niveles de lengua o sociolectos" (Borrero & Cala, 2000:218), los cuales

identifican a los hablantes como miembros de un determinado grupo social. Estas variedades están motivadas por una pluralidad de factores: sexo, edad, clase social, nivel de instrucción, profesión, raza y etnia. Aquí se incluyen los grupos cuya forma de hablar se identifican con la lengua vulgar, o las diferentes jergas juveniles, o la forma de hablar de grupos sociales marginales (Romanelli, 2013). Las terceras son las variedades funcionales o diafásicas; son los llamados registros lingüísticos (estilos). Estos registros aparecen en función de las características de la situación, o del contexto comunicativo en el que se encuentra el hablante. Por ejemplo, no se le habla igual a un niño que a un anciano, a un conocido que a un desconocido; esto implica contextos comunicativos distintos, y para cada contexto se busca el registro más adecuado.

Las variedades pueden ser distinguidas, además de su vocabulario, por diferencias en su gramática, fonología y prosodia. Existen diversos factores de variación asociados a la geografía, la evolución lingüística, los factores sociolingüísticos o el registro lingüístico. Álvarez (2006) hace distinción entre las variedades de la lengua a través del tiempo, las variedades geográficas, las sociales y situacionales; estas tres últimas distinguidas también por Coseriu (1981). En la figura 2.1 pueden observarse estas variedades para el caso del español, se muestra de manera general los tipos que corresponden a cada variedad de la lengua.

En este libro, se ha seleccionado como caso de estudio la variedad diatópica del español, debido principalmente a la amplia gama que esta lengua contiene. En las variantes diatópicas del español, se distingue el español peninsular y el español de América[4]. Se

[4] El español es la lengua oficial de dieciocho repúblicas hispanoamericanas: Argentina, Uruguay, Paraguay, Chile, Bolivia, Perú, Ecuador, Colombia, Venezuela, Panamá, Costa Rica, Nicaragua, Honduras, El Salvador, Guatemala, México, Cuba y República Dominicana.

encuentran diferencias significativas entre las dos grandes variantes diatópicas del español, tanto en el plano lingüístico como en el plano extralingüístico.

Si se toma como ejemplo el verbo transitivo *coger*, sin importar el contexto en el que se aplique, en México puede ser malsonante; mientras que, en España, es simplemente una acción. Para alguien que aprende español como lengua extranjera, este hecho podría no ser relevante; sin embargo, existiría la necesidad de conocer este tipo de variantes que permitan diferenciar cada situación. De esta manera, la lengua española hablada en México ha sufrido cambios a nivel fonético, a nivel léxico y a nivel morfosintáctico; por esta razón, se centra la atención de esta investigación en esta variedad diatópica del español.

2.1.1 Español peninsular versus español mexicano

El español ha tenido una gran expansión geográfica, motivo por el cual presenta una gran variedad diatópica; distinguiendo, de manera general, el español peninsular y el español de América. De manera más particular se distingue el andaluz, el canario y el castellano, para el caso de la primera variedad diatópica y para el caso de la segunda variedad, el argentino, el cubano, el mexicano, el peruano, entre otros. Como se ha mencionado con anterioridad, en este libro se tiene particular interés en el español mexicano, particularmente el estudio de las locuciones verbales.

Las dos grandes variedades diatópicas presentan diferencias significativas, tanto lingüísticas como extralingüísticas; la lengua española hablada en México ha sufrido

---

El español (conviviendo con el inglés) también se habla en numerosas islas de las Antillas, en EE.UU. (en los estados de Nuevo México, Arizona, Texas, California y Florida) y en Puerto Rico, donde ha sido lengua oficial en diversos momentos.

cambios con respecto a la pronunciación, el léxico empleado y la morfosintaxis. Algunas de las divergencias entre el español peninsular (de aquí en adelante, ES) y el español mexicano (de aquí en adelante, MEX), se presentan a continuación.

- A nivel fonético.
  En el caso del MEX, se remarca que no existe diferencia de pronunciación entre la *s, z* y la *c* que son pronunciadas como /s/.

- A nivel sintáctico.
  El empleo del tratamiento "usted" es generalizado en el MEX incluido en niños y expresa respeto por alguien más; mientras que, en el ES, el tuteo es generalizado.

  El MEX utiliza muy frecuentemente el diminutivo *ahorita,* también emplea diminutivos como *ándele, pásele, órale* para dar un carácter enfático.

- A nivel léxico.
  Se remarcan diferencias significativas, como el caso de *hablar* para el ES mientras que para el MEX se dice *platicar*; *autobús* en el caso de ES y *camión* para MEX, entre otros.

  La significación de ciertos léxicos diferentes en ambas variantes, entre los cuales se tiene *coger*, que en ES significa *agarrar o tomar* y en el MEX significa hacer el amor (tomar a una mujer).

  El empleo de anglicismos, por ejemplo, *móvil* para el ES y *celular* para el MEX.

Con respecto a las unidades fraseológicas igual existen diferencias, las cuales pueden notarse a nivel de significado (cf. 1a, 1b y 1c) y por tanto a nivel de utilización, a nivel léxico (cf. 2a y 2b). La tabla 2.1 contiene una muestra de un breve estudio diatópico contrastivo de las locuciones verbales del ES y del MEX.

(1) a. No tener abuela.
    b. Pelar los ojos.

c. Estar hecho un cuero.

(2) a. Agarrar/tomar el toro por los cuernos.

b. Coger el toro por los cuernos.

Tabla 2.1 Locuciones verbales del español peninsular y del español mexicano.

Locución verbal en español peninsular	**Locución verbal en español mexicano**	**Significado**	**Traducción de equivalencia (francés)**
acabarse la candela	colgar los tenis	Morir	casser sa pipe
agarrar [coger/pillar] una chicuelina	ponerse hasta el gorro	Emborracharse	lever le coude
dar la murga	agarrar de su puerquito	Molestar continuamente	casser les oreilles à qqn
andar a la greña	agarrarse de la greña	Pelearse	s'arracher le blanc des yeux
estar sin una perra	andar roto	No tener dinero	n'avoir pas un rond
cascar/machacar las liendres a alguien	partir su mandarina	Golpear	casser la gueule
darla con queso	pasar gato por liebre	Burlar/Engañar	se payer la tête de qqn
pegar un parchazo	ver la cara de gringo	Burlar/Engañar	mener qqn en bateau
estar con el dogal al cuello	llegar el agua hasta el cuello	Situación difícil	avoir la corde au cou

El significado de (1a) es diferente tanto para ES como para MEX; en el primer caso, el diccionario DRAE la define como: "se dice irónicamente de la persona que se alaba a sí misma"; mientras que, en el caso del MEX, el diccionario DEM el significado corresponde

a: "no tener o tener pocos escrúpulos, honradez; tener poca consideración de las personas". En el caso de (2a) el significado para ES, de acuerdo al DRAE es, abrirlos desmesuradamente y para MEX, según DUE, abrir el ojo, estar advertido para que no lo engañen. Y finalmente en (1c), el significado en DRAE para ES se define como estar borracho, ebrio; mientras que, para MEX, es definida como hombre o mujer hermosos y atractivos sexualmente, por el DRAE.

Como se ha mencionado, con anterioridad, existen léxicos que son empleados de diferente manera en ambas variantes diatópicas y por lo cual las locuciones verbales sufren algunos cambios y en algunos casos conservan el mismo significado, tal es el caso de (2a y 2b), la primera unidad fraseológica es utilizada en MEX y la segunda empleada en ES.
También se podrían remarcar diferencias entre las variantes habladas en Hispanoamérica, entre las cuales se tiene, por ejemplo, el caso de Argentina, que emplea términos diferentes a los referenciados en MEX. Entre los cuales se tiene, la prenda de vestir *falda*, en MEX se emplea este término mientras que en Argentina emplean *pollera;* o que decir de *vereda* (Argentina) con respecto a *banqueta* (MEX). Estas y muchas diferencias más existen en las variantes del español de América, sin embargo, con estar conscientes de las variedades en el desarrollo de esta tesis es suficiente ya que se ha seleccionado de manera muy particular una parte de la lengua y para el MEX, en otras palabras, no se toman en cuenta este tipo de variedades, únicamente se analiza la variante diatópica del español mexicano.

2.2 Base de datos de locuciones verbales

Las locuciones, son definidas por Casares (1950) como "una combinación estable de dos o más términos, que funciona como elemento oracional y cuyo sentido unitario consabido no se justifica, sin más, como una suma del significado normal de los componentes". Las diferentes definiciones de locución en español han seguido esta

caracterización. Las locuciones se han divido según la función oracional que desempeñen, independientemente de que sean conmutables por palabras simples o por sintagmas.

Las locuciones verbales constituyen el núcleo de sintagmas verbales, es decir, están formadas por un núcleo verbal, acompañado por sus complementos. Desde el punto de vista sintáctico expresan procesos y actúan como los predicados, con o sin complementos. Estas unidades fraseológicas, igual que los verbos, se combinan con el sujeto y los complementos para formar una oración (Priego et al. 2015).

La base de datos de locuciones verbales que se describe a continuación consta de 1,219 unidades fraseológicas e incluye sólo una parte de la gran cantidad de estas unidades que responden a las características que poseen. Sin embargo, esta base de datos proporciona una idea del enorme número de unidades fraseológicas que corresponden al español y sobre todo proporciona una visión general de estas estructuras lingüísticas.

Para la elaboración de la base de datos, desde la construcción hasta la traducción, se han seguido las siguientes etapas:

1. Extracción de las locuciones verbales a partir del *Diccionario de Mexicanismos*.

2. Elaboración de las estructuras sintácticas, de manera automática y manual, que sirven para trabajar con estas unidades fraseológicas.

3. Construcción de patrones morfosintácticos de las locuciones verbales.

4. Traducción al francés de las locuciones verbales que conforman la base de datos.

Cada una de las etapas, mencionadas anteriormente, es descrita con mayor detalle en las siguientes secciones de este Capítulo.

2.2.1 Extracción

La información que forma parte de la base de datos de locuciones verbales ha sido obtenida a partir de la consulta de un diccionario; la construcción de este recurso léxico ha sido realizada de una manera manual. El diccionario de donde las locuciones han sido extraídas es el *Diccionario de Mexicanismos* (2010), diccionario que es el resultado de una investigación hecha por la Academia Mexicana de la Lengua. El diccionario tiene tres características esenciales: es sincrónico, es contrastivo y es descriptivo.

a) Sincrónico. Representa la lengua en su estado actual, es decir, que contiene los elementos léxicos de uso de la segunda mitad del siglo XX y principios del siglo XXI.

b) Contrastivo. Este diccionario es diferencial y a partir de esto, ha sido construido tratando de comparar lo que se dice en México con lo que se dice en otros países de habla española y sobre todo con el español de la Península Ibérica.

c) Descriptivo. Indica la realidad de uso; ya que no establece criterios normativos, no se excluyen prestamos de otras lenguas (que son principalmente del inglés), ni neologismos.

En este caso, se han obtenido 1,409 locuciones verbales del diccionario anteriormente mencionado. Sin embargo, existen algunas locuciones verbales con dos acepciones o más; por esta razón, la base de datos finalmente conformada contiene 1,219 locuciones que corresponden a las locuciones verbales sin repetición de significado; esto no significa que las locuciones con más de dos modos de empleo no sean tomadas en cuenta, lo que significa es que la base de datos no contiene duplicidad de locuciones verbales (PRIEGO, 2020). Al conjunto de estas unidades fraseológicas se les ha dado el nombre de *base de datos de locuciones verbales* y dan soporte a este libro. En la figura 2.2 se presenta una

muestra de la base de datos desarrollada, la cual contiene la información obtenida del *Diccionario de Mexicanismos.*

verbo	sustantivo	eXpresiÓN	classe de expresión	Español: hacer creer que una cosa es mucho mejor de lo que es en realidad	dic	Información	otros diccio	Hispanoamérica
ladrar	hambre	ladrar de hambre	loc verbal	Estar alguien muy hambriento	DM	pop.	México	Hisp
lanzar	calzón	lanzar el calzón	loc verbal	Insinuarse sexualmente una mujer a un hombre	DM	coloq/obsc/euf.	México	Hisp
lanzar	can	lanzar los canes	loc verbal		DM	coloq.	México	Hisp
lanzar	perro	lanzar los perros	loc verbal	Insinuarse alguien a otra persona, con la intención de enamorarla o seducirla	DM	coloq.	México	Hisp
lanzar	borrego	lanzar un borrego	loc verbal	Esparcir una noticia falsa	DM	coloq.	México	Hisp
largar	chingada	largarse a la chingada 1	loc verbal	Ser echado de un lugar por alguien, de manera violenta y grosera	DM	pop/coloq/vulg.	México	Hisp
largar	chingada	largarse a la chingada 2	loc verbal	Ser rechazado por alguien, de manera tajante y grosera	DM		México	Hisp
largar	goma	largarse a la goma	loc verbal	Ser echado de un lugar por alguien, de manera violenta y grosera	DM	pop/coloq.	México	Hisp
largar	verga	largarse a la verga	loc verbal	Ser echado de un lugar por alguien, de manera violenta y grosera	DM	pop/vulg.	México	Hisp
largar	carajo	largarse al carajo	loc verbal	Ser echado de un lugar por alguien, de manera violenta y grosera	DM	pop/coloq/vulg.	México	Hisp
levantar	ámpula	levantar ámpula	loc verbal	Irritar o provocar la indisposición de alguien con la atrevida de cierta idea o actor	DM	coloq.	México	Hisp
levantar	polvo	levantar el polvo	loc verbal	Huir	DM	coloq.	México	Hisp
levantar	canasta	levantar la canasta	loc verbal	Arrepentirse	DM	pop.	México	Hisp
levantar	catapulta	levantar la catapulta	loc verbal	Tener un hombre una erección	DM	pop/coloq/obsc/	México	Hisp
levantar	mano	levantar la mano	loc verbal	Ofrecerse alguien a prestar algún tipo de ayuda	DM		México	Hisp
levantar	muertos	levantar muertos	loc verbal	Dar vigencia o animar a alguien o algo que estaba decaído, apagado o inactivo	DM		México	Hisp
levantar	ronchas	levantar ronchas 1	loc verbal	Causar molestia o enfado a alguien	DM		México	Hisp
levantar	ronchas	levantar ronchas 2	loc verbal	Incitar a una multitud	DM		México	Hisp
levantar	cuello	levantarse el cuello	loc verbal	Presumir de algo, por lo general, sin fundamento	DM	coloq.	México	Hisp
limar	candado	limar el candado	loc verbal	Practicar el coito	DM	pop/coloq/obsc/	México	Hisp

Figura 2.2. Base de datos de locuciones verbales.

2.2.2 Análisis y estudio

El estudio de una lengua es un campo bastante amplio, en este caso, se ha seleccionado una parte de la lengua para estudiar y analizar lo que es una locución verbal. Como se ha mencionado con anterioridad, las locuciones verbales son un tipo de unidades fraseológicas que reflejan parte de la cultura y modo de hablar de una lengua.

En esta sección se presenta el estudio realizado a las locuciones verbales, el cual se llevó a cabo mediante el análisis de los elementos que conforman estas unidades fraseológicas y de las acepciones que estas unidades tienen. Cabe mencionar que este estudio se puede

generalizar también para las locuciones del ES e incluso casi para cualquier idioma, debido a que la mayoría de las locuciones verbales de una lengua poseen las características que se describirán a continuación.

2.2.2.1 Locuciones verbales ambiguas

Se comienza el análisis considerando a las locuciones ambiguas, es decir, locuciones susceptibles a dos o más significados o interpretaciones. Dicho en otras palabras, locuciones verbales varían su significado en función de su contexto. Los ejemplos de este tipo de locuciones son ilustrados en la tabla 2.2. En algunos casos, estas locuciones, podrían ser denominadas locuciones verbales polisémicas. Sin embargo, en este libro se ha adoptado el nombre de locuciones verbales ambiguas.

Tabla 2.2 Locuciones verbales ambiguas.

Locución verbal	**Significados**	**Ejemplos**
llevarse de corbata	Atropellar a alguien	El carro se llevó de corbata a María.
	Pasar por encima de los derechos de alguien	El gobierno se llevó de corbata a los ciudadanos al aumentar los impuestos.
	Tirar o arrastrar un objeto por ir alguien a prisa	María se llevó de corbata la jarra y la rompió.
mandar a volar	Denegar algo a alguien	María mandó a volar a su hermano al no prestarle la pelota.
	Terminar una relación con alguien	Juan mandó a volar a María porque él quiere a otra.
	Despedir a alguien de su trabajo	A María la mandaron a volar por no hacer bien su trabajo.
pegar con tubo	Decir o hacer algo contundente	María pego con tubo al dar la respuesta del examen.
	Tener éxito alguien o algo, impactar a alguien positivamente	María pego con tubo en la conferencia.

Las acepciones de las locuciones verbales fueron obtenidas del Diccionario de Mexicanismos (*DM)* y por esta razón se puede tener conocimiento de las locuciones con más de dos significados. Se determinaron alrededor de 262 locuciones verbales (ambiguas) que el *DM* proporcionó con sus acepciones para cada locución. Existe la posibilidad de que en algunos casos las locuciones verbales contenidas en la base de datos

desarrollada posean más modos de empleo. Sin embargo, la información contenida en la base de datos es proporcionada por el repositorio utilizado, que en este caso es el mismo diccionario de mexicanismos.

2.2.2.2 Locuciones verbales sinónimas

Al construir la base de datos, se observaron locuciones verbales que mantenía una relación de sinonimia entre ellas. Es decir, locuciones en las que el significado es similar al de otras locuciones, pero que sus elementos tienen variaciones o son completamente diferentes. Teniendo diferentes tipos de locuciones, las que por ejemplo sólo cambian o añaden elementos a ciertos verbos (***valer queso, valer sombrilla***); las que los verbos son hipónimos (***comer/tragar el orgullo***) y/o los elementos de las locuciones también son hipónimos (***parar el oído/las orejas***); las que todos los elementos de la locución son diferentes (***colgar los tenis, pegar botones***).

Es bien sabido que en la lingüística moderna no existen los sinónimos perfectos, por tal razón en este estudio no se consideran a las locuciones verbales como sinónimas perfectas. Sin embargo, son consideradas como sinónimas porque cumplen las características de tener el mismo significado, pero diferente locución empleada, es decir, alguno de sus elementos sufre algún cambio. En este caso se han obtenido 339 locuciones verbales que se han denominado sinónimas. En la tabla 2.3, se muestran algunos de los ejemplos que corresponden a este tipo de locuciones.

Se podría considerar a este tipo de locuciones verbales sinónimas como el caso inverso a las locuciones verbales ambiguas. Dado que, en las primeras, varias locuciones poseen un mismo significado; mientras que, en las segundas, una locución posee varios significados. Todo esto permite observar que, en general, las locuciones verbales y de acuerdo a lo percatado en el DM, pueden tener diferentes usos y diferentes acepciones.

Tabla 2.3 Locuciones verbales sinónimas.

Locución verbal	Significados	Ejemplos
comer el orgullo	Soportar una humillación	Me comí el orgullo y tuve que escuchar todo lo que me dijo.
tragar el orgullo		Me tragué el orgullo y tuve que escuchar todo lo que me dijo.
colgar los tenis	Morir	El pobre de Juan colgó los tenis ayer y hoy es el entierro.
pegar botones		El pobre de Juan pegó botones ayer y hoy es el entierro.
parar el oído	Poner atención para oír algo, especialmente las conversaciones ajenas	María paró el oído y se enteró que su hermana tenía problemas.
parar las orejas		María paró las orejas y se enteró que su hermana tenía problemas.
valer queso	Estropearse algo	Mis zapatos valieron queso por la lluvia
valer sombrilla		El vino valió sombrilla, ¿Cuántos días lleva abierto?

2.2.2.3 Locuciones verbales comparativas

Al decir locuciones verbales comparativas, se hace referencia precisamente a las locuciones que en ciertos casos realizan comparaciones, siendo estas comparaciones entre adjetivos. Utilizando un adjetivo seguido del adverbio *como*, el adverbio *más* unido con la preposición *que*, el adverbio *menos* unido con la preposición *que*, el adverbio como, en fin, todo el tipo de estructuras que realizan comparaciones. Para este caso, los

ejemplos son ilustrados en la tabla 2.4. En la base de datos se encuentran alrededor de 29 de este tipo de locuciones, que se han denominado como comparativas.

Tabla 2.4. Locuciones verbales comparativas.

Locución verbal	**Significados**	**Ejemplos**
andar como perro sin mecate	Sentirse alguien triste y desconsolado	Ando como perro sin mecate
bailar con la más fea	Llevarse alguien la peor parte	Me tocó bailar con la más fea
poner como lazo de cochino	Tratar mal de palabra a alguien, insultar	Me pusieron como lazo de cochino
quedar como araña fumigada	Terminar alguien en un estado alcoholizado	Nunca tomo, por eso ayer quede como araña fumigada con dos tequilas
reventar como chicharra	Morir	El delincuente reventó como chicharra por los balazos de la policía

2.2.2.4 Locuciones verbales formadas por una conjunción de coordinación

En la base de datos también existen locuciones formadas por una conjunción de coordinación (*y*), que sirve principalmente para realizar la unión de dos partes que forman la locución verbal. La principal característica de estas locuciones es que se recalca o se hace hincapié en algún aspecto que se quiere enfatizar en la unidad lingüística.

Para este tipo de locuciones verbales se han contabilizado 16 que forman el tipo de locuciones unidad por la conjunción *y* de las cuales algunos ejemplos se presentan en la tabla 2.5. En este tipo de unidades fraseológicas, en general, los elementos coordinados son elementos mínimos sin modificadores; en otras palabras, la conjunción de coordinación introduce un elemento o grupo de elementos estructurados en relación con la primera parte de la locución verbal.

Se tiene el caso de que, si en el primer elemento no hay un determinante, en el segundo tampoco habrá determinante; por ejemplo, el caso de la locución verbal *ser uña y mugre*. Al igual se observa que en el binomio formado por la conjunción de coordinación, en algunos casos, existe simetría; por ejemplo, la locución *arrastrar la cobija y ensuciar el apellido*. Y qué decir de las locuciones en las que se añade un valor de suma a los elementos que conforman la conjunción, por ejemplo, *quedar bien con dios y con el diablo*.

Tabla 2.5. Locuciones verbales formadas por una conjunción de coordinación.

Locución verbal	**Significados**	**Ejemplos**
arrastrar la cobija y ensuciar el apellido	Se usa para indicar que el comportamiento de alguien lo exhibe y pone en vergüenza	No tiene vergüenza, arrastró la cobija y ensució el apellido

poner de (a) vuelta y media	Dar su merecido a alguien	Se lo merecía, lo puso de vuelta y media
quedar bien con dios y con el diablo	Adherirse a un tiempo a dos posturas contrarias sin comprometerse en firme al final por ninguna de ellas	Es un hipócrita siempre quiere quedar bien con dios y con el diablo
ser uña y mugre	Tener gran amistad dos o más personas	En el jardín de niños éramos uña y mugre, pero después no nos volvimos a ver
traer entre ceja y ceja	Tener antipatía a alguien y tratar de hacerle daño	Mi suegro me trae entre ceja y ceja, nunca le caí bien

2.2.2.5 Locuciones verbales con negación

La base de datos también contiene locuciones verbales en las que existe una negación y como tal son negativas. A lo largo de este estudio se ha podido observar que algunas locuciones verbales con negación pueden optar la modalidad de positiva. Adicionalmente, en este grupo de locuciones se tiene las que incluyen la conjunción de coordinación "ni", lo que equivale a aumentar el significado negativo de la locución y generalmente a aumentar la negación. Este grupo de locuciones verbales con negación está formado por 35 locuciones y algunos ejemplos se muestran en la tabla 2.6.

Tabla 2.6. Locuciones verbales con negación.

Locución verbal	**Significados**	**Ejemplos**

no haber vuelta de hoja	Carecer una situación de la posibilidad de cambio o retorno a un estado previo	Ya lo hiciste, no hay vuelta de hoja
no ganar (ni) para los frijoles	Percibir alguien una cantidad de dinero insuficiente para satisfacer todas sus necesidades	Para que le hacemos al cuento, no gano ni para los frijoles
no querer estar en el pellejo	Compadecerse alguien de una persona que ha padecido algún hecho desafortunado	La verdad, no quisiera estar en su pellejo después de este problema
no quitar el dedo del renglón	Insistir alguien en algún propósito	Te lo aseguro, no quitaré el dedo del renglón
no servir (ni) para el arranque	Ser alguien o algo insuficiente o que dejar de ser útil demasiado pronto	Para eso me querías, no me sirves ni para el arranque

2.2.2.6 Idiomaticidad en las locuciones verbales

Una de las características semánticas de las locuciones verbales es la opacidad semántica o idiomaticidad (característica general de las unidades fraseológicas). Cabe recordar que es la característica por la cual el significado global de dicha unidad no es deducible del significado aislado de cada uno de sus elementos constitutivos.

Al estudiar a las locuciones verbales, se ha podido notar que existen locuciones que pueden ser deducidas con una lectura, es decir, el significado global se obtiene a partir de cada uno de los elementos que la componen. Sin embargo, existen otras locuciones que no pueden deducirse a partir de los elementos que la componen y el significado será deducido a partir de la lectura no literal, dicho en otras palabras, el significado global se

conocerá. En la tabla 2.7 se muestran algunos ejemplos que cumplen con este este tipo de características, y se notará, como se ha mencionado, que algunas locuciones son deducibles fácilmente y otras no. Cabe mencionar, que "no todas las unidades fraseológicas son idiomáticas, pues se trata de una característica potencial, no esencial de este tipo de unidades" (Corpas, 1996:27).

Tabla 2.7. Locuciones verbales idiomáticas.

Locución verbal	Significados	Ejemplos
agarrarse del chongo	Reñir, pelear	Las niñas de la secundaria se agarraron del chongo
asustar con el petate del muerto	Intimidar a alguien con una noticia falsa o exagerando un hecho	No me asustes con el petate del muerto
caer en la movida	Descubrir a alguien en un acto incorrecto o reprobable	Te caí en la movida
llevar hasta el límite	Llevar una situación hasta el extremo	Este asunto lo has llevado al límite
morirse en la raya	Hacer alguien hasta el último esfuerzo por cumplir algo	Y si hay que morirse en la raya, yo lo haré
parecer retrato	Vestir alguien de manera continuada la misma ropa	Con esa ropa, pareces retrato
pedir para las aguas	Pedir alguien una propina	Y después de estacionarme, me pidió para las aguas
quebrarse la cabeza	Reflexionar alguien sobre un problema para tratar de resolverlo	Me estuve quebrando la cabeza toda la noche para saber cómo ayudarte

Anteriormente, la gramática tradicional había establecido que no se podía deducir el significado de las unidades fraseológicas a partir del significado que las componen. Actualmente, esto no es una verdad absoluta debido a que el significado de unidades puede deducirse fácilmente a partir del significado de las palabras que la componen. Sin embargo, es cierto que existen, en su mayoría, unidades fraseológicas en las cuales resulta

difícil comprender el significado. Para el caso de estudio presentado en este libro: las locuciones verbales, se ha notado que la base de datos contiene ambos tipos de unidades (cf. 1 y 2).

(1) Entendí lo que dijiste, así que sale sobrando tu sarcasmo.
LV: salir sobrando
Traducción: *ne pas importer*

(2) Se quedó muy fresco a pesar de la terrible noticia.
LV: quedar muy fresco
Traducción: *être tranquille*

En (1) las palabras que forman la locución verbal y las palabras de la locución verbal permiten deducir que algo está de más, es innecesario. Mientras que en (2), las palabras que forman la locución verbal no permiten deducir que el conjunto de palabras significa que alguien está tranquilo. A partir de los ejemplos citados, se tienen dos significados diferentes en las locuciones verbales, el primero un significado transparente, es decir, que se pueden deducir del significado de las palabras que la componen y el segundo un significado opaco, dicho en otras palabras, el significado no se puede deducir de la suma de los componentes. En las tablas 2.8 y 2.9, respectivamente, se muestran algunos ejemplos correspondientes a estos significados.

Tabla 2.8. Locuciones verbales con significado transparente.

Locución verbal	Significados	Ejemplos
ser un pan de Dios	Poseer alguien mucha bondad y nobleza	Ese profesor es un pan de dios, inscríbete con el
tirarse de risa	Reírse mucho, sin control	Me tiré de risa cuando vi la foto de la credencial de Saúl
vender como pan caliente	Vender algo alguien con gran rapidez	Hoy las flores se vendieron como pan caliente
ver la cara	Engañar a alguien	¿Me quieres ver la cara con esos precios?

Tabla 2.9. Locuciones verbales con significado opaco.

Locución verbal	Significados	Ejemplos
tirar los perros	Cortejar a alguien	Llevo dos años tirándole los perros y nada que me hace caso
ver si como ronca duerme	Comprobar si alguien es coherente con lo que dice y con lo que hace	Vamos a ver si como roncan duermen los diputados, ojalá que cumplan con frenar la delincuencia
volverse ojo de hormiga	Desaparecer alguien mañosa e intencionalmente	Cuando vi llegar a mi suegra, me volví ojo de hormiga y regrese a la sala
zafarse un tornillo	Perder alguien el juicio, volverse loco	Desde que se murió su mamá, a Pepe se le zafo un tornillo

2.2.3 Estructuras sintácticas

De manera general, la sintaxis es, la rama de la lingüística que estudia el modo en el que las palabras se combinan para formar frases o enunciados en una lengua. En este apartado, se hace la descripción de las reglas por las cuales las locuciones verbales se combinan como unidades lingüísticas. El objetivo es conocer cuáles son las características sintácticas de este tipo de unidades fraseológicas, dicho en otras palabras, se hace el análisis sintáctico de las locuciones verbales que servirá para trabajar posteriormente con estas unidades.

El proceso de análisis sintáctico consiste en asignar a cada frase de un texto su estructura sintáctica; este proceso se lleva a cabo con las locuciones verbales. En este estudio se aborda de manera general el análisis sintáctico automático y el análisis sintáctico manual. Para la parte automática se utilizan dos herramientas capaces de realizar esta tarea, *TreeTagger* y *FreeLing*. En el caso de la parte manual, el estudio está basado en las nueve partes de la oración: el sustantivo, el pronombre, el verbo, el adjetivo, el artículo, el adverbio, la preposición, la conjunción y la interjección. Posteriormente, se realiza una comparación entre los tipos de análisis sintácticos realizados, con el objetivo de observar las irregularidades existentes y las diferencias a nivel de etiquetado entre las herramientas utilizadas.

2.2.3.1 Análisis sintáctico automático

El análisis sintáctico automático es obtenido por una herramienta que toma una frase como entrada y la herramienta le asigna la estructura sintáctica. Existen una gran cantidad de herramientas que permiten realizar esta tarea; sin embargo, para esta investigación se han seleccionado dos: *TreeTagger*[5] y *FreeLing*[6] debido a que la utilización y el funcionamiento, entre otros idiomas, cubre la lengua española.

2.2.3.1.1 Análisis sintáctico con *TreeTagger*

TreeTagger es una herramienta automática para la anotación de textos con la parte del discurso (PoS, por sus siglas en inglés) y la información del lema. Esta herramienta ha sido desarrollada por Helmut Schmid en el proyecto de TC en el Instituto para la Lingüística Computacional de la Universidad de Stuttgart (Schmid, 1994; 1995). *TreeTagger* ha sido utilizado con éxito para etiquetar las lenguas como alemán, inglés, francés, italiano, holandés, español, búlgaro, ruso, portugués, gallego, chino, eslovaco, latino, estonio, polaco y antiguos textos franceses; es posible adaptar esta herramienta a otros lenguajes, si el léxico y el corpus de aprendizaje etiquetado manualmente están disponibles.

El funcionamiento de *TreeTagger* es el siguiente: a partir de texto de entrada (frase, oración, etc.) del cual se requiere conocer la estructura sintáctica, se obtiene como resultado la palabra, la estructura sintáctica y la raíz de la palabra. El funcionamiento de esta herramienta será más visible a partir de un ejemplo: se toma la frase "*TreeTagger es*

[5] Para más información concerniente a la herramienta TreeTagger, consultar http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[6] Para más información concerniente a la herramienta FreeLing, consultar http://nlp.lsi.upc.edu/freeling/

fácil de usar" (*Le TreeTagger est facile pour user.*), y el resultado obtenido se muestra en la tabla 2.10.

Tabla 2.10. Resultado de etiquetado con TreeTagger de la frase "TreeTager es fácil de usar.".

Palabra	**Estructura sintáctica**	**Lema**
TreeTagger	NP	TreeTagger
Es	VSfin	Ser
fácil	ADJ	Fácil
de	CSUBI	De
Usar	VLinf	Usar
.	FS	.

Las estructuras sintácticas[7] que se pueden observar en la tabla 2.10, corresponden a las partes de la oración y son: *NP* corresponde a un nombre propio, VSfin al verbo finito *ser*, *ADJ* a un adjetivo, *CSUBI* a una preposición, VLinf a un verbo lexical infinitivo y finalmente, FS a un signo de puntuación (punto final).

[7] Una descripción completa de las estructuras sintácticas en español, puede consultarse directamente en: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/spanish-tagset.txt

Para obtener la estructura sintáctica de cada locución verbal con *TreeTagger* se utilizó el parámetro de archivos para el español UTF-8[8] con sus estructuras sintácticas correspondientes. En la tabla 2.11 se presenta una muestra del número de locuciones verbales que corresponde a cada tipo de estructura sintáctica obtenida por la herramienta y algunos ejemplos.

Tabla 2.11. Muestra de la clasificación de las estructuras sintácticas a partir de TreeTagger.

Estructura sintáctica		**Locuciones verbales**	
Abreviación	**Significado**	**Número**	**Ejemplos**
VLinf ART NC	Verbo infinitivo + Artículo + Sustantivo común	193	abrir la boca
VLinf NC	Verbo infinitivo + Sustantivo común	167	comer ansias
VCLIinf ART NC	Verbo cíclico + Artículo + Sustantivo común	66	dar el azotón
VLinf PREP NC	Verbo infinitivo + Preposición + sustantivo común	44	ladrar de hambre
VLfin ART NC	Verbo finito + Artículo + Sustantivo común	37	sobarse el lomo
VCLIinf NC	Verbo cíclico + Sustantivo común	29	ponerse águila

[8] El archivo de parámetros está disponible en:
http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/spanish-par-linux-3.2-utf8.bin.gz

VLinf PREP ART NC	Verbo infinitivo + Preposición + Artículo + Sustantivo común	28	pedir para las aguas
VCLIinf PREP NC	Verbo cíclico + Preposición + Sustantivo común	18	seguirse de frente
VLinf Vladj	Verbo infinitivo + Verbo en participio pasado	18	llegar barriéndose
VCLIinf PREP ART NC	Verbo cíclico + Preposición + Artículo + Sustantivo común	16	ponerse a las vivas

2.2.3.1.2 Análisis sintáctico con *FreeLing*

FreeLing es una herramienta informática que ofrece servicio de análisis del lenguaje y está compuesta por una biblioteca de funciones preprogramadas. El proyecto *FreeLing* ha sido creado y mantenido por Lluís Padró (Pradó, 2011). Esta herramienta está desarrollada por TALP Research Center de la Universitat Politècnica de Catalunya. El principal objetivo y ventaja de utilización de esta herramienta es que ha sido desarrollada para el español. Los principales servicios ofrecidos por la herramienta son:

- Tokenización de texto.
- Separación de sentencias.
- Análisis morfológico.
- Reconocimiento de unidades poliléxicas.
- Reconocimiento de fechas y/o horas.
- Reconocimiento de expresiones de moneda.
- Reconocimiento de expresiones numéricas (números, cantidades, porcentajes, etc).
- Reconocimiento de expresiones de medidas físicas (velocidad, longitud, presión, frecuencia, temperatura, densidad, etc).
- Reconocimiento de nombres propios.
- Etiquetado Part-of-Speech.
- Análisis de dependencias basado en reglas.

Las lenguas actualmente soportadas por *FreeLing* son: el español, el catalán, el francés, el gallego, el italiano, el inglés, el ruso, el portugués, el galés y el asturiano; el checo y el esloveno son lenguas parcialmente soportadas. El funcionamiento *de FreeLing*, será más visible a partir de un ejemplo: la frase a etiquetar será "*El gato come pescado y bebe agua*" (*Le chat mange du poisson et boit de l'eau*). El análisis morfológico obtenido por la herramienta se muestra a través de la tabla 2.12 y en esa misma tabla se presentan el etiquetado de las partes de la oración que corresponden a la columna 2 y 3, respectivamente.

El analizador morfológico para el español de *FreeLing* utiliza un conjunto de etiquetas (EagV2.0) para representar la información morfológica de las palabras. Este conjunto de etiquetas se basa en las etiquetas propuestas por el grupo EAGLES (por sus siglas en inglés, *Expert Advisory Group on Language Engineering Standards*) para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas.

Las categorías que el analizador morfológico utiliza para el español son las siguientes:

- Adjetivos
- Adverbios
- Determinantes
- Sustantivos
- Verbos
- Pronombres
- Conjunciones
- Interjecciones
- Preposiciones
- Signos de puntuación
- Cifras
- Fechas/Horas

Tabla 2.12. Análisis morfológico de la oración "El gato come pescado y bebe agua." con FreeLing.

	Análisis morfológico		
Palabra	**Lema**	**Etiqueta Morfológica**	**Probabilidad de transición**
El	el	DA0MS0	1
gato	gato	NCMS000	1
come	comer	VMIP3S0	0.997101
pescado	pescado	NCMS000	0.777534
y	y	CC	0.999963
bebe	beber	VMIP3S0	0.997101
agua	agua	NCCS000	0.992095
.	.	Fp	1

Para cada categoría morfológica, el valor de la etiqueta es diferente y los valores muestran los atributos, los valores y los códigos que pueden tomar. Por ejemplo, si se toma el verbo de la frase *El gato come pescado*, la etiqueta morfológica correspondiente es *VMIP3S0*, que significa Verbo Principal Indicativo Presente de la 3 persona del Singular. Los valores del conjunto de la etiqueta son definidos en las categorías EagV2.0, es decir, dependiendo de la categoría de la palabra en el nodo. Los valores que corresponden a las etiquetas de la categoría del verbo son presentados en la tabla 2.13 y si un valor es indeterminado la asignación es cero.

Tabla 2.13. Valores de FreeLing para la categoría verbo.

Verbos			
Posición	**Atributo**	**Valor**	**Código**
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
		Semi- auxiliar	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
		Condicional	C
		-	0
5	Persona	Primera	1
		Segunda	2
		Tercera	3

6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

Una vez que se conocen las categorías que *FreeLing* contiene, para identificar cada etiqueta morfológica de cada palabra se debe revisar las construcciones formadas para cada valor del discurso[9]. *FreeLing*, al igual que *TreeTagger*, han ayudado a conocer la estructura sintáctica de cada locución verbal que forma la base de datos construida y esto ha sido con el objetivo de conocer y tener más información de estas unidades fraseológicas.

En la tabla 2.14, se puede observar el número de locuciones verbales para cada tipo de estructura sintáctica asignada por *FreeLing* y algunos ejemplos. El resultado final, es que existen 469 tipos de estructuras sintácticas diferentes en la base de datos de locuciones verbales. Sin embargo, estos resultados no son absolutos ya que al realizar el etiquetado manual de los resultados, los tipos de estructuras sintácticas pueden ser diferentes.

Tabla 2.14. Muestra de la clasificación de las estructuras sintácticas a partir de FreeLing.

Estructura sintáctica	**Locuciones verbales**

[9] Para el caso de los valores de cada categoría de FreeLing para el español, la lista completa puede consultarse directamente en http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

Abreviación	Significado	Número	Ejemplos
VMN0000 DA0MS0 NCMS000	Verbo principal infinitivo + Artículo masculino singular + Sustantivo común masculino singular	82	agarrar el modo
VMN0000 NCMS000	Verbo principal infinitivo + Sustantivo común masculino singular	72	blanquear dinero
VMN0000 DA0FS0 NCFS000	Verbo principal infinitivo + Artículo femenino singular + Sustantivo común femenino singular	70	cachetear la banqueta
VMN0000 NCFS000	Verbo principal infinitivo + Sustantivo común femenino singular	48	comer gente
VMN0000 PP3CN000 DA0FS0 NCFS000	Verbo principal infinitivo + Pronombre personal + Articulo femenino singular + sustantivo común femenino singular	27	llevar la tostada
VMN0000 SPS00 NCFS000	Verbo principal infinitivo + Preposición simple + Sustantivo común femenino singular	23	pasar de noche
VMN0000 DA0MP0 NCMP000	Verbo principal infinitivo + Artículo masculino plural + Sustantivo común masculino plural	19	parar los tacos

VMN0000 NCFP000	Verbo principal infinitivo + Sustantivo común femenino plural	18	quedar tablas

2.2.3.2 Análisis sintáctico manual

Una vez que se ha realizado el etiquetado automático de las locuciones verbales, se procede al etiquetado manual. Se podría decir que esta es una etapa de evaluación de los etiquetados con las diferentes herramientas. El estudio del etiquetado manual se basa en las nueve partes generales del discurso: el sustantivo, el pronombre, el verbo, el adjetivo, el artículo, el adverbio, la preposición, la conjunción y la interjección.

Este etiquetado se lleva a cabo a partir de las definiciones de cada parte del discurso; los resultados automáticos sirven como base para el desarrollo del etiquetado manual. Se ha podido percatar que, como se sabe, los etiquetadores automáticos llegan a tener algunos errores; así, con el etiquetado manual se garantiza la calidad de la información.

Para cada componente de la locución verbal se le asigna la estructura sintáctica correspondiente. En la tabla 2.15, se presenta una muestra del tipo de locuciones verbales identificadas a partir de la estructura sintáctica, así como algunos ejemplos. En total se identificaron 160 estructuras sintácticas diferentes. Estas unidades fraseológicas son tomadas en cuenta como punto de partida para la identificación de las locuciones verbales candidatas.

Como se puede observar, los resultados entre el etiquetado manual y el etiquetado automático son completamente diferentes, alrededor del 50% menos con respecto a *TreeTagger* y 34.11% menos con respecto a *FreeLing*. Con este resultado se observa la

variabilidad de etiquetado entre las tres formas llevadas a cabo y la diferencia de etiquetado que cada herramienta tiene.

El etiquetado de la estructura sintáctica de las locuciones verbales, que conforman la base de datos desarrollada, ha permitido conocer las diferencias existentes entre este tipo de unidades fraseológicas y ha proporcionado características importantes a considerar en la metodología para la validación de estas locuciones. En algunos casos, se observó que las estructuras no tienen un sentido totalmente lógico, sin embargo, forman parte de la lengua.

Tabla 2.15. Muestra de estructuras sintácticas de las locuciones verbales mexicanas.

Estructuras sintácticas	**Número de locuciones verbales correspondiente**	**Ejemplo**
V Dét N	331	colgar los tenis (étendre les baskets)
V N	228	dar cisca (donner cisca)
V Prép N	121	agarrar de puerquito (attraper un petit porc)
V Prép Dét N	59	agarrarse de la greña (s'attraper la tignasse)
V Adj	32	andar roto (marcher cassé)
V Pron N	23	partir su mandarina (diviser sa mandarine)
V N Prép N	20	pasar gato por liebre (prendre un chat pour un lièvre)
V Dét N Prép N	15	ver la cara de gringo (voir le visage d'un étranger)
V Dét N Prép Dét N	15	llegar el agua hasta el cuello (avoir de l'eau jusqu'au cou)

2.2.4 Patrones morfosintácticos

A partir del análisis realizado a las locuciones verbales, en esta sección, se presentan los patrones morfosintácticos de las locuciones verbales, cuyo elemento principal es un sintagma verbal y presentan una gran diversidad morfosintáctica.

Entre los autores interesados en el estudio de las unidades fraseológicas, podemos encontrar por un lado grupos vinculados a corrientes de lingüística teórica (Sager, 1990; Cabre, 1999; Cabre & Estopá, 2005) y, por otro lado, corrientes vinculadas a la práctica terminográfica y la estandarización de unidades fraseológicas (Wüster, 1979; Arntz & Picht, 1988). En las últimas décadas ambas corrientes comparten el interés por las tecnologías de extracción automática de unidades fraseológicas.

A partir del interés por la extracción de estas unidades fraseológicas, algunos autores se han centrado en identificar patrones sintácticos, morfológicos o la mezcla de ambos que ayuden a determinar la estructura interna de esta combinación de palabras (Dagan, 1994; Jacquemin, 1997; Michiels, 1998).

Para el caso del español, algunos trabajos dedicados al análisis morfosintáctico de las locuciones y que determinan diferentes tipos de ellas, se presentan en (Casares, 1950; Coseriu, 1966; Zuluaga, 1980; Haensch, 1982; Carneado & Tristá, 1983; Corpas, 1996). Adicionalmente, en (McCarthy et al., 2003; Piao et al., 2003; Baldwin, 2005; Zhang et al., 2006; Van de Cruys & Moirón, 2007; Davis & Barrett, 2013; Nissim & Zaninello, 2013) se analizan otro tipo de patrones (semántico, composicional, léxico), con el fin de extraer a estas unidades y determinar las características que podrían generalizarse en estas unidades lingüísticas.

Como se ha observado, el principal objetivo de extraer patrones morfosintácticos de las unidades fraseológicas es por un lado conocer las características que estas estructuras lingüísticas tienen, y por otro, el de identificarlas en corpora. Al igual que los investigadores, anteriormente descritos, en esta sección del libro se han extraído los patrones morfosintácticos de las locuciones verbales para conocer más acerca de las características que poseen, y efectivamente buscar estos patrones en corpora que permitan identificar estas unidades. En la siguiente sección se describe la metodología

llevada a cabo para el análisis de la diversidad morfosintáctica de las locuciones verbales y como estos patrones han sido de utilidad para recuperar locuciones verbales candidatas.

2.2.4.1 Metodología utilizada para el análisis de la diversidad morfosintáctica

Con el fin de identificar los patrones morfosintácticos en las locuciones verbales, se parte de la taxonomía de las locuciones realizada por Corpas (1996) en su clasificación de las unidades fraseológicas en español. De acuerdo a la taxonomía, como se ha dicho en múltiples ocasiones, se decide centrarse en las locuciones verbales debido a que la mayoría de las frases está formada por un verbo y una o varias variables. El verbo exige y realiza una rigurosa selección de los sujetos y de los componentes que pueden acompañarle. Estas frases se encuentran fusionadas en la oración para enunciar algo de manera más amplia, pero al separarse de la oración tienen sentido completo, es decir, tienen información semántica por ellas mismas y constituyen el núcleo de sintagmas verbales.

En dicha taxonomía se clasifican los tipos de locuciones verbales de acuerdo a su variedad morfosintáctica, los cuales comprenden: a) Locuciones formadas por dos núcleos verbales unidos por conjunción, b) Locuciones compuestas de verbo y pronombre, c) Locuciones compuestas de verbo, pronombre y partícula, d) Locuciones de verbo más partícula asociada a este, con complemento opcional, e) Locuciones formadas por verbo copulativo más atributo, f) Locuciones formadas por verbo más complemento circunstancial, g) Locuciones formadas por verbo más suplemento h) Locuciones formadas por verbo más objeto directo y i) Locuciones negativas. En este libro son denominadas como Tipo 1, Tipo 2, ..., Tipo 9; respectivamente. Con base en esta taxonomía, se prosigue a adquirir ejemplos de locuciones verbales que cumplan con la variedad morfosintáctica y que ayuden a determinar los patrones morfosintácticos.

En cuanto a los ejemplos utilizados, se emplean las locuciones verbales presentadas en Priego et al. (2014), debido a que estas fueron recuperadas manualmente y son la base de

todo lo realizado y presentado en este libro. Posteriormente, estas locuciones verbales se clasifican de acuerdo a los tipos de la taxonomía empleada según sus componentes. Una vez clasificadas es necesario conocer su estructura morfosintáctica para de esta manera obtener los patrones, así que las locuciones fueron etiquetadas con *FreeLing*[10].

En la tabla 2.16 se presenta una muestra de las locuciones verbales identificadas de acuerdo a su tipo y sus respectivas etiquetas morfosintácticas[11]. Para la búsqueda de los patrones morfosintácticos identificados, se seleccionó un fragmento del corpus periodístico presentado en Priego et al. (2014), que será descrito con mayor detalle en la sección 4.3, el cual contiene aproximadamente 1,960,373 palabras.

[10] *FreeLing* se ha descrito con mayor detalle en la sección 2.2.3.1 de este capítulo.

[11] Para una referencia del significado del etiquetado morfológico de *Freeling* referirse a http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

Tabla 2.16. Ejemplo de locuciones verbales identificadas de acuerdo a sus etiquetas morfosintácticas.

Tipo de locución verbal	**Ejemplos**	**Etiquetas morfosintácticas (resultados de *FreeLing*)**
Tipo 1	dar y tomar	VMN0000 CC VMN0000
	ir y venir	VMN0000 CC VMN0000
	llevar y traer	VMN0000 CC VMN0000
Tipo 2	apañársela	VMN0000 PP3CN000 PP3FSA00
	arreglársela	VMN0000 PP3CN000 PP3FSA00
	Cargársela	VMN0000 PP3CN000 PP3FSA00
Tipo 3	brincarse la barda	VMN0000 PP3CN000 DA0FS0 NCFS000
	darse su taco	VMN0000 PP3CN000 DP3CS0 NCMS000
	tomarla con (alguien/algo)	VMN0000 PP3FSA00 SPS00 (PI0CS000/PI0CS000)
Tipo 4	dar de sí	VMN0000 SPS00 CS
	ir con (uno)	VMN0000 SPS00 (PI0MS000)
	tomar (algo/a alguien) por	VMN0000 (PI0CS000 / SPS00 PI0CS000) SPS00
Tipo 5	ser ajonjolí de todos los moles	VSN0000 AQ0CS0 SPS00 DI0MP0 NCMP000
	ser el vivo retrato de alguien	VSN0000 DA0MS0 AQ0MS0 NCMS000 SPS00 PI0CS000
	ser gacho	VSN0000 AQ0CS0
Tipo 6	decir hasta la despedida	VMN0000 SPS00 DA0FS0 NCFS000

	dormir como un tronco	VMN0000 CS DI0MS0 NCMS000
	meter a alguien en cintura	VMN0000 SPS00 PI0CS000 SPS00 NCFS000
Tipo 7	meter las cuatro	VMN0000 DA0FP0 Z
	oler a cuero quemado	VMN0000 SPS00 NCMS000 VMP00SM
	pagar el pato	VMN0000 DA0MS0 NCMS000
Tipo 8	chuparse el dedo	VMN0000 PP3CN000 DA0MS0 NCMS000
	mover cielo y tierra	VMN0000 NCMS000 CC NCFS000
	saber de qué pie cojea alguien	VMN0000 SPS00 DT0CN0 NCMS000 VMIP3S0 PI0CS000
Tipo 9	no haber vuelta de hoja	RN VMN0000 NCFS000 SPS00 NCFS000
	no poder ver ni en pintura a alguien	RN VMN0000 VMN0000 CC SPS00 NCFS000 SPS00 PI0CS000
	no tener un pelo de tonto	RN VMN0000 DI0MS0 NCMS000 SPS00 NCMS000

La identificación de los patrones morfosintácticos en el corpus se ha realizado de dos diferentes maneras, por un lado, tomando en cuenta el contexto y la otra sin tomarlo en cuenta. En la primera aproximación, se ha utiliza una ventana de cinco palabras a la izquierda de la locución verbal y cinco palabras a la derecha, denominándolas *contexto izquierdo* y *contexto derecho*, respectivamente.

Básicamente, en la metodología propuesta en este libro se considera tener dos elementos esenciales: 1) Una lista de locuciones verbales, y 2) Un conjunto de textos, ambos etiquetados morfosintácticamente. Del primer recurso léxico se obtienen los patrones morfosintácticos, y éstos son buscados en el corpus de textos con la finalidad de obtener una lista de posibles locuciones verbales, las cuales concuerdan con los patrones morfosintácticos obtenidos de las locuciones semilla (ver figura 2.3).

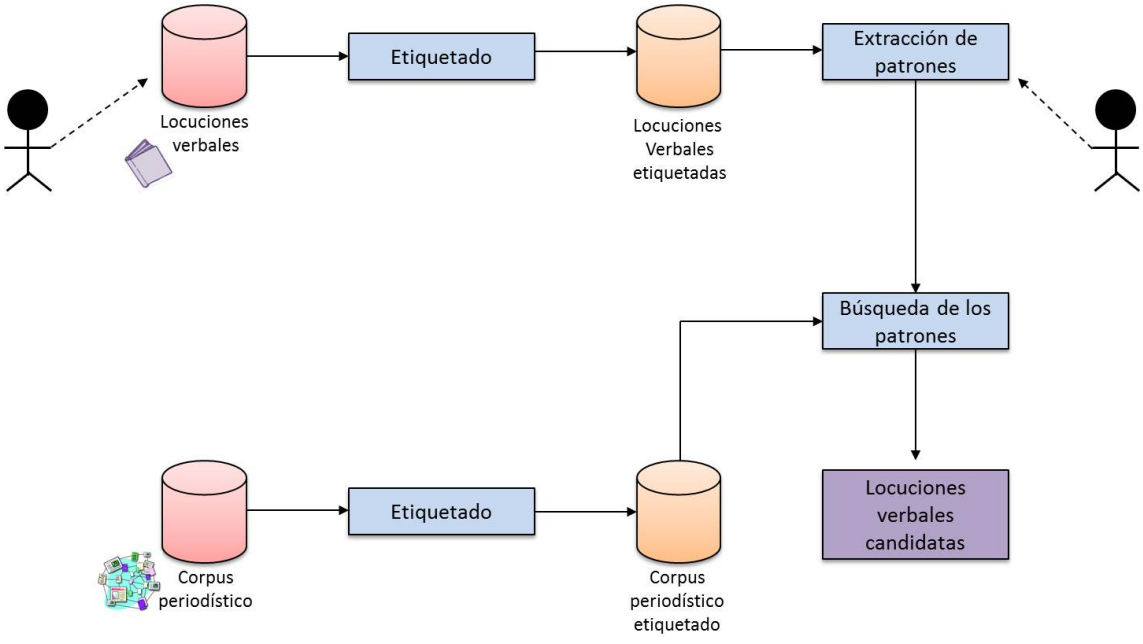

Figura 2.3. Metodología empleada para la identificación de patrones morfosintácticos en las locuciones verbales.

2.2.4.2 Resultados

En este libro se han identificado 34 patrones morfosintácticos que sirven como semilla para encontrar posibles locuciones verbales dentro de un corpus de textos. Dichos patrones han sido obtenidos mediante el etiquetado morfosintáctico de una lista semilla de 43 locuciones verbales. En la tabla 2.17 se presenta una muestra de los patrones morfosintácticos identificados como más frecuentes en el corpus de textos periodísticos.

Tabla 2.17. Muestra de patrones morfosintácticos de las locuciones verbales.

Estructura sintáctica	Patrón morfosintáctico	Ejemplos
V Prep	VMN0000 SPS00	ir con
V Det Nom Adj	VMN0000 DA0FS0 NCFS000 AQ0CS0	llevar la voz cantante
V Pron Prep	VMN0000 PP3FSA00 SPS00	tomarlo con
V Prep Conj	VMN0000 SPS00 CS	dar de sí
V Pron Det Nom	VMN0000 PP3CN000 DA0MS0 NCMS000	chuparse el dedo
V Det N Prep Det Nom	VMN0000 DI0MS0 NCMS000 SPS00 DA0FS0 NCFS000	costar un ojo de la cara
V Det Nom	VSN0000 DA0FS0 NCFS000	ser la moda
V Prep Pron	VMN0000 SPS00 PI0CS000	dar sobre alguien
V Pron Prep	VMN0000 PI0CS000 SPS00	tomar algo por

En la tabla 2.18 se presenta un ejemplo de las 10 locuciones verbales encontradas como más frecuentes en el corpus y que empatan con el patrón morfosintáctico indicado en la misma tabla. En total, se extrajeron 3,083 resultados coincidentes con los patrones registrados.

Tabla 2.18. Ejemplo de las locuciones verbales encontradas en el corpus periodístico.

Frecuencia de aparición	Locución verbal candidata
357	llegar/llegar/VMN0000 a/a/SPS00
201	contar/contar/VMN0000 con/con/SPS00
152	participar/participar/VMN0000 en/en/SPS00
117	tratar/tratar/VMN0000 de/de/SPS00
113	apoyar/apoyar/VMN0000 a/a/SPS00
110	cumplir/cumplir/VMN0000 con/con/SPS00
106	salir/salir/VMN0000 de/de/SPS00
99	ir/ir/VMN0000 a/a/SPS00
93	ver/ver/VMN0000 con/con/SPS00
90	acudir/acudir/VMN0000 a/a/SPS00

En la figura 2.4 se puede observar que de los 10 patrones morfosintácticos más frecuentes (ver tabla 2.18), el primero obtiene un 80% de cobertura con respecto a los demás. Este hecho se encuentra derivado de que es un patrón demasiado general que parte de locuciones verbales semilla tales como: *ir con*.

Cabe mencionar que de los 34 patrones morfosintácticos detectados a partir de las locuciones semilla, solamente se encontraron coincidencias sobre 18. Esto significa, que

16 patrones no han arrojado posibles locuciones verbales. En la tabla 2.19 se muestran tales patrones; una discusión sobre los mismos sigue a continuación.

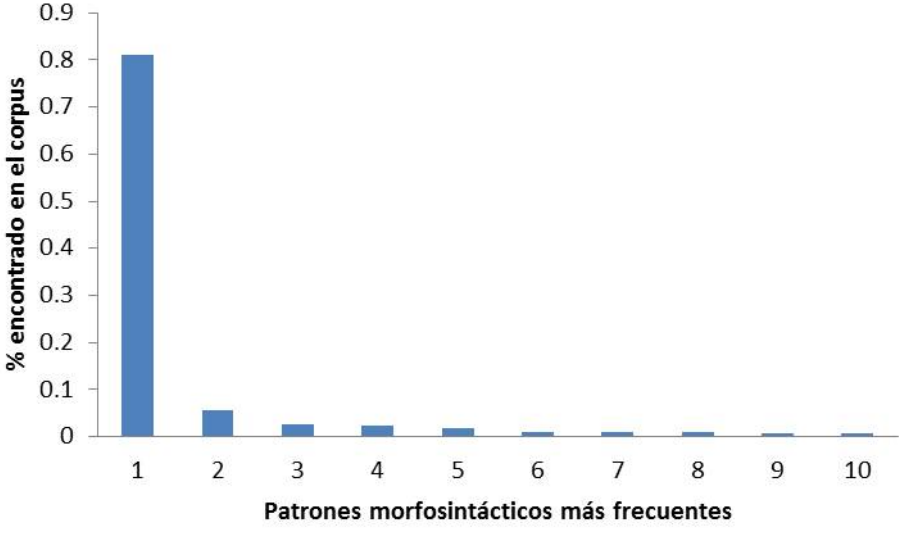

Figura 2.4. Porcentaje de las locuciones verbales más frecuentes encontradas en el corpus.

Tabla 2.19. Conjunto de patrones morfosintácticos de los cuales no se encontraron coincidencias en el corpus de textos.

Patrón morfosintáctico	**Ejemplos**
VMN0000 PP3FSA00 SPS00 PI0CS000	tomarla con alguien
VSN0000 DA0MS0 AQ0MS0 NCMS000 SPS00 PI0CS000	ser el vivo retrato de alguien
VMN0000 Z NCFP000 SPS00 PI0CS000	dar cien vueltas a alguien
VMN0000 SPS00 PI0CS000 SPS00 NCFS000	meter a alguien en cintura
VMN0000 SPS00 PI0CS000 CS VMIP3S0 SPS00 NCMS000	meter a alguien como chupa de dómine
VMN0000 SPS00 DT0CN0 NCMS000 VMIP3S0 PI0CS000	saber de qué pie cojea alguien
RN VMN0000 Z NCMP000 SPS00 RG	no tener dos dedos de frente
RN VMN0000 NCFS000 SPS00 NCFS000	no haber vuelta de hoja

Observando los patrones que no encontraron coincidencias en el corpus de textos podemos ver que en general se trata de una secuencia no habitual de etiquetas morfológicas, cuya frecuencia es muy baja en los corpus textuales. El corpus utilizado tiene únicamente 5,000 noticias (361 palabras en promedio por noticia) y, por tanto, la probabilidad de encontrar una de estas secuencias es muy baja.

En esta parte del libro se presentan experimentos tendientes a la identificación de locuciones verbales a partir de textos planos. La metodología propuesta indica tomar como entrada un conjunto de locuciones verbales semilla que sirvan para encontrar un conjunto de patrones morfosintácticos, los cuales son posteriormente utilizados para encontrar coincidencias de estos sobre un corpus de textos (en este caso, fue del género periodístico).

El experimento realizado sobre un conjunto inicial de 43 locuciones verbales permitió encontrar 34 patrones morfosintácticos. De éstos, únicamente 18 encontraron coincidencias en el corpus de textos. Como trabajo a futuro se desea ampliar el corpus de textos para determinar si es posible encontrar coincidencias de todos y cada uno de los patrones morfosintácticos detectados a partir de las locuciones verbales semilla.

2.2.5 Traducción de las locuciones verbales

La base de datos creada es de gran utilidad para el estudio de las locuciones verbales mexicanas, sin embargo, se ha decidido incluir más información que permita realizar estudios con otras lenguas y que sea capaz de proporcionar un recurso léxico utilizable para otros investigadores interesados en la lengua. Por tal motivo, se decidió realizar la traducción en francés de estas unidades fraseológicas, por lo que se tiene una base de datos de locuciones verbales con la correspondiente traducción en francés. La traducción se realiza en francés debido a que ambas lenguas pertenecen a las lenguas romances.

La traducción consiste en buscar la mejor forma posible de pasar de una lengua a otra un texto, examinando la equivalencia más próxima en la otra lengua. La Real Academia Española (2014) define traducir en una de las acepciones como: "Expresar en una lengua lo que está escrito o se ha expresado antes en otra". Y precisamente es lo que se pretende realizar con las locuciones verbales, expresarlas en francés y transmitir el sentido de la locución. Con el fin de realizar la traducción de las locuciones verbales mexicanas en francés, se buscó la equivalencia más próxima de estas unidades fraseológicas de la base de datos con el fin de encontrar la mejor traducción en francés. La traducción se realizó mediante tres etapas (Tricás, 1995: 148 – 151):

1. Reproducir el segmento (es decir, la locución verbal) de manera literal mediante un segmento simétrico desde el punto de vista morfológico y semántico.

Por ejemplo, para la frase Ahora es cuando comenzamos a hacer la traducción, la traducción literal en francés es *C'est maintenant que nous commençons à faire la traduction.*

2. Transferir la semanticidad del conjunto (es decir, la locución verbal) recogiendo la misma intencionalidad por medio de una construcción morfológica y semántica diferente.

   En el caso de la locución verbal *meter en cintura* entonces la traducción es *mettre au pas*.

3. Traducir por una paráfrasis, si es que no se ha logrado obtener una traducción de equivalencia, es decir, transmitir únicamente la idea que contiene la locución verbal.

   Para la frase *París bien vale una misa* se tiene la traducción *Paris vaut bien une messe*.

Para llevar a cabo las etapas anteriores, primero se realizó la utilización de traductores automáticos, para posteriormente proceder a la revisión manual de estas traducciones literales y finalmente utilizar el *Dictionnaire des expressions et locutions* (Rey & Chantreau, 1993) para obtener la traducción de equivalencia. Entonces, la traducción se llevó a cabo inicialmente de manera automática y completada de manera manual, para garantizar la calidad de la traducción de los datos. En la tabla 2.20 se presenta una muestra de algunas de las locuciones verbales mexicanas con la correspondiente traducción.

A lo largo de la realización de la traducción, se observó que entre ambos idiomas existen diferencias significativas de acuerdo a las preposiciones, artículos que son utilizados; efectivamente, por el empleo y uso diferente de estas partes del discurso. En algunos casos tienen preposiciones de más, en otros casos artículos, en fin, esto remarca la diversidad morfológica que existe entre ambas lenguas.

Otro hecho importante a remarcar es la variedad existente entre ambas lenguas, con variedad se hace referencia a que las locuciones verbales cubren una parte de la lengua y al realizar la traducción de estas estructuras del lenguaje han permitido ampliar el

panorama que se tenía de ambas lenguas. La diversidad entre las locuciones es amplia, al realizar la traducción al francés se observaron locuciones en las que la traducción y acepción son las mismas, es decir, los elementos de las locuciones no cambian en ningún idioma, otras en las existía bastante variedad en los componentes de la locución, unas más en las que al menos un elemento cambiaba, y otras en las que los elementos eran totalmente diferentes. Este estudio ha permitido, de cierto modo, comparar ambas lenguas y conocer las diferentes variedades que las locuciones verbales pueden llegar a tener.

Tabla 2.20. Muestra de la traducción de las locuciones verbales mexicanas.

Locuciones verbales	**Traducción Literal**	**Traducción de Equivalencia**
agarrar de puerquito	attraper un petit porc	casser les oreilles à qqn
agarrarse de la greña	s'attraper la tignasse	s'arracher le blanc des yeux
andar roto	marcher cassé	n'avoir pas un rond
colgar los tenis	étendre les baskets	casser sa pipe
dar cisca	donner cisca	avoir la tremblote
llegar el agua hasta el cuello	avoir de l'eau jusqu'au cou	avoir la corde au cou
partir su mandarina	diviser sa mandarine	casser la gueule
pasar gato por liebre	prendre un chat pour un lièvre	se payer la tête de qqn
ver la cara de gringo	voir le visage d'un étranger	mener qqn en bateau

2.3 Corpus periodístico

En esta sección se presenta la descripción del corpus de noticias de periódicos de la República Mexicana, el cual sirve como base para la identificación de locuciones verbales candidatas en diferentes géneros textuales.

Un corpus es un conjunto de textos compilados, ya sea de un mismo tema o varios. El objetivo particular de este corpus es convertirse en un conjunto de datos que proporcione ejemplos de uso (con sus respectivos contextos) de algunas locuciones verbales con el fin de analizar el uso y frecuencia en diversos géneros textuales. Adicionalmente, este corpus podría ser útil en algoritmos de aprendizaje automático para generar modelos que identifiquen automáticamente este tipo de estructuras lingüísticas. Dependiendo de la naturaleza de los algoritmos y de la tarea, las expresiones existentes en el corpus podrían estar previamente desambiguadas o no.

El material de trabajo es un corpus escrito –contiene solo el idioma español–, abierto –en constante crecimiento–, especializado –corresponde al género de noticias– y finalmente periódico, es decir, la colección de documentos (noticias) utilizada consta de relatos periodísticos ocurridos a partir del año 2007 y hasta el año 2013, estos documentos han sido recopilados por una agencia mexicana especializada en noticias. Si bien, los documentos obtenidos presentan diferentes metadatos, para este caso han resultado útiles los siguientes:

- Título de la noticia
- Sección del periódico (Sociedad, Espectáculos, Política, Futbol, entre otros)
- Fecha de la noticia
- El texto correspondiente

Es cierto que a partir de los documentos de noticias se pueden obtener más datos; sin embargo, para la identificación de locuciones verbales, con la información mencionada anteriormente es suficiente.

La obtención del corpus ha sido realizada de una manera automática con la construcción de un software capaz de efectuar dicha tarea. La herramienta informática ha sido

programada con el lenguaje de programación Java[12] y con las bibliotecas que este lenguaje de programación posee.

En total 1,895,983 noticias han sido agrupadas, sin embargo, se ha estudiado aproximadamente el 20% del corpus total. El corpus analizado se compone de 378,890 noticias, un total de 4,579,284 oraciones y alrededor de 1,159,571 palabras (ver tabla 2.21). A este corpus se le ha denominado *corpus periodístico*.

[12] Para más información concerniente al lenguaje de programación Java, ver https://www.java.com/fr/

Tabla 2.21. Características generales del corpus periodístico del español mexicano.

Características	Descripción
Repositorio	Organización Editorial Mexicana
Periódicos agrupados	70 periódicos mexicanos
Lengua	Español
Año	A partir de 2007
Número de noticias	1,895,983
Número de noticias	1,895,983
Número de noticias utilizadas	378,890
Número de frases de las noticias utilizadas	4,579,284
Número de palabras en las noticias utilizadas	140,386,809
Vocabulario de las noticias utilizadas	1,483,524

La primera parte corresponde a la construcción de corpora que permite conocer de una manera estándar la lengua mexicana, por esta razón, se construyó el corpus de noticias. Posteriormente, se procede a la construcción de un corpus que proporcione la oportunidad de analizar la lengua utilizada por los hablantes mexicanos. Por lo tanto, a partir de comentarios publicados en Twitter, se aborda esta parte encontrada a lo largo de la investigación. En la sección siguiente, se describe el corpus de comentarios que se ha construido.

2.4 Corpus de comentarios

El corpus de comentarios se construyó con el objetivo de conocer y de analizar el lenguaje utilizado por hablantes mexicanos y a partir de esto evidenciar el uso de las locuciones verbales.

La construcción del corpus ha sido realizada de manera automática con la ayuda de un módulo Web para el lenguaje de programación Python. Se garantiza que estos comentarios son escritos en la República Mexicana dado que fueron extraídos de acuerdo a la ubicación geográfica y el corpus contiene alrededor de 21,559 comentarios. En la tabla 2.22 se presentan las características del corpus de comentarios.

Una vez recuperados los comentarios de las personas, denominados tweets, se procede a verificar la legitimidad del corpus para lo cual se utilizó sólo una parte del corpus. El objetivo principal de utilizar sólo una parte del corpus es debido a que la verificación de la existencia de locuciones verbales ha sido realizada de manera manual. Se tomaron alrededor de 2,000 tweets y se encontró que aproximadamente el 50% de estos contienen realmente al menos una locución verbal.

Los resultados obtenidos revelan que la utilización de las locuciones verbales es más común en textos escritos por los hablantes de la lengua que los que se escriben en las noticias. Sin embargo, no podemos generalizar este hecho dado que se sabe que la utilización de las unidades fraseológicas es más común en lenguaje oral.

Tabla 2.22. Características generales del corpus de comentarios del español mexicano.

Características	**Descripción**
Repositorio	Twitter

Lengua	Español
Número de comentarios (tweets)	21,559
Número de tweets utilizados	1,000
Número de palabras en los tweets utilizados	17,160
Vocabulario de los tweets utilizados	5,474

Los tweets obtenidos ponen en evidencia a las locuciones verbales, pero no se sabe qué pasa con el resto de los tweets que no fueron evaluados o que no han sido extraídos. Así que, se necesita ya sea obtener más tweets o continuar evaluándolos. Sin embargo, cabe mencionar que los 1,000 tweets que han sido ya evaluados han servido para evidenciar la utilización de las locuciones verbales y para tener un corpus de referencia que servirá para futuros experimentos aplicados a la metodología para la identificación de locuciones verbales candidatas.

Capítulo 3. Construcción de corpora supervisado

Para un estudio adecuado del proceso de comunicación y la lengua, es necesario contar con recursos léxicos adecuados. El tema que atañe a este libro no es la excepción. Con la intención de investigar el estudio de la validación de locuciones verbales, es necesario contar, de manera particular, con recursos léxicos para el español mexicano y por tal motivo, se han construido recursos léxicos propios. En este capítulo, se presenta la metodología para la construcción de corpora supervisado, es decir, un conjunto de información que ha sido etiquetada automáticamente para propósitos de uso en métodos de aprendizaje automático.

3.1 Recuperación de contextos asociados a las locuciones verbales

En el proceso de construcción de recursos léxicos se han adquirido una base de locuciones verbales, un corpus periodístico y un corpus de comentarios. Estos recursos han sido desarrollados con el objetivo de extraer contextos asociados a las locuciones verbales, de donde a su vez sea posible extraer características para la identificación y validación automática de estas estructuras lingüísticas en un texto. En esta sección se explica cómo se ha llevado a cabo la recuperación de los contextos asociados a las locuciones verbales.

La recuperación de contextos asociados a las locuciones verbales se ha llevado a cabo de manera automática. En el proceso de recopilación han intervenido únicamente procesos computacionales, es decir, que no existe intervención humana. El proceso general de recuperación de contextos puede observarse de manera gráfica en la figura 3.1.

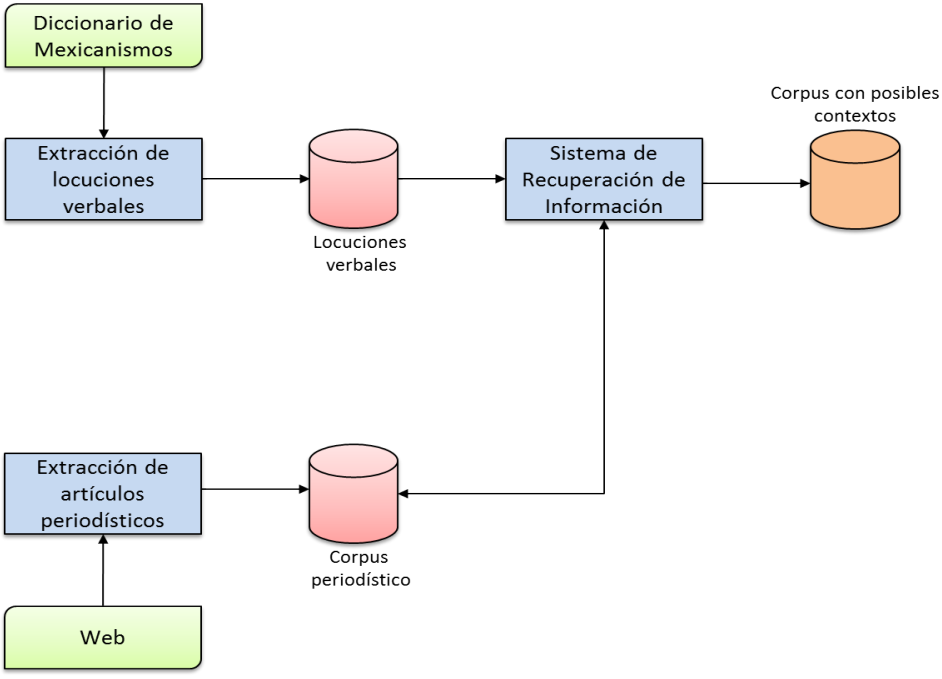

Figura 3.1. Proceso general de recuperación de contextos asociados a las locuciones verbales.

El objetivo de esta etapa fue el de obtener una serie de contextos asociados a cada locución verbal. Este proceso se puede llevar a cabo automáticamente, a partir de procesos asociados con los sistemas de recuperación de información. Así, usando métodos computacionales es posible extraer, automáticamente, ejemplos de contextos en donde ocurran las variantes morfológicas de cierta locución verbal. El proceso propuesto y utilizado se describe a continuación.

Inicialmente se debe de contar con un conjunto de locuciones verbales válidas. Tal y como se ha comentado con anterioridad, en este caso se ha recopilado un conjunto de este tipo de locuciones a partir del Diccionario de Mexicanismos. Por otro lado, se ha

extraído de Internet, un conjunto de textos de donde posteriormente se seleccionan contextos en donde ocurra, al menos, una locución verbal.

Las etapas de recuperación de contextos asociados a las locuciones verbales se desenvuelven en, básicamente, dos etapas principales: 1) Pre-procesamiento de los recursos léxicos y 2) Recuperación de contextos; estas etapas se presentan gráficamente en la figura 3.2. Ahí se observa que el pre-procesamiento fue realizado para el corpus periodístico, esto es con fines ilustrativos. Sin embargo, la misma etapa se realizó para el corpus de comentarios.

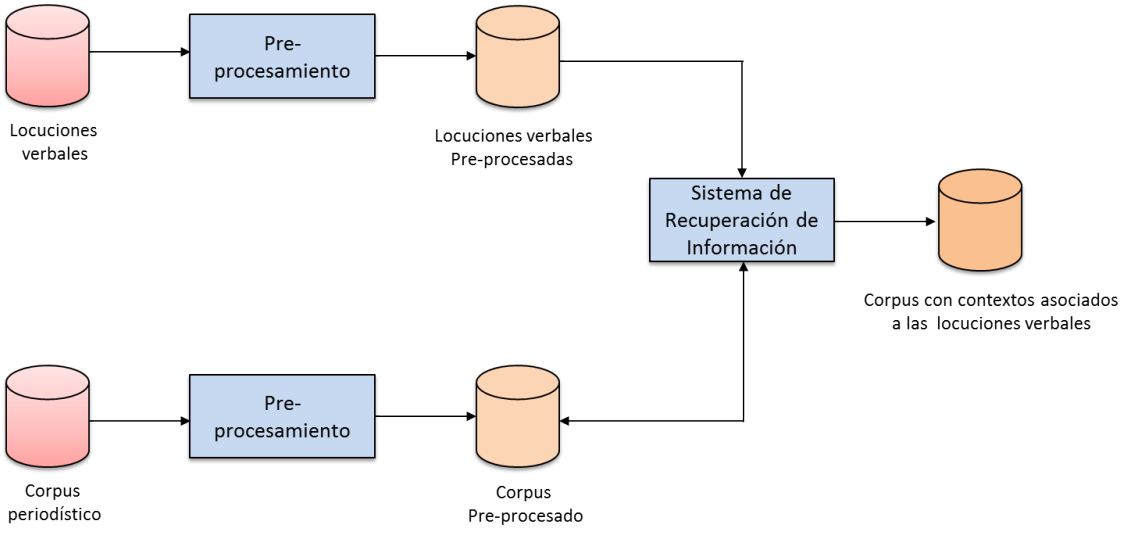

Figura 3.2. Etapas de recuperación de contextos asociados a las locuciones verbales.

La primera etapa que se ha llevado a cabo es la de pre-procesamiento, la cual tiene como propósito preparar los datos para poder ser manipulados, transformados y de fácil acceso; tal como menciona Doriam Pyle (1999:88):

"The fundamental purpose of data preparation is to manipulate and transform raw data so that the information content enfolded in the data set can be exposed, or made more easily accessible".

Efectivamente, en esta etapa se han preparado los datos (recursos léxicos), con el objetivo de generalizarlos para poder ser manipulados, transformados y utilizados posteriormente en la metodología propuesta para la identificación de locuciones verbales. La normalización de los datos consistió principalmente en la eliminación de noticias, comentarios o locuciones verbales repetidas; posteriormente se realizó la lematización (obtención del lema de la palabra) de cada uno de los recursos léxicos con las herramientas computacionales de *FreeLing* y *TreeTager*.

La primera normalización realizada al corpus periodístico fue la inserción de una etiqueta artificial (a la que denominamos *TDIVT*) para separar cada metadato de la noticia, obteniendo la estructura presentada en la figura 3.3. Esto con el fin, precisamente como su nombre lo indica, de generalizar y uniformizar la información que se posee en el corpus. Este proceso fue realizado sobre todo el corpus periodístico, sin embargo, para la identificación de contextos fue solamente utilizado alrededor de un 20%, con el objetivo de mostrar la presencia y evidencia de estas unidades fraseológicas.

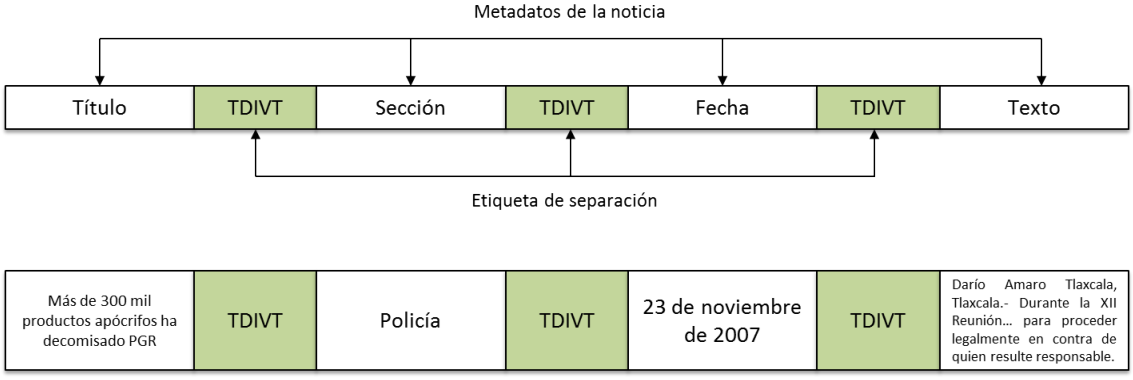

Figura 3.3. Estructura de los metadatos de un artículo periodístico.

Para poder manipular la información de las noticias se hizo necesaria la división de estos textos periodísticos en oraciones, obteniendo al final aproximadamente 4,579,284 oraciones. Además de que este proceso permitía, de cierto modo, tener textos de longitud[13] equilibrados. Es decir, que los textos contengan aproximadamente la misma cantidad de palabras.

Se usa un sistema de recuperación de información, el cual posee la capacidad de regresar oraciones en donde ocurre una determinada locución verbal. El proceso global requiere que se introduzca la versión no lematizada de dicha locución verbal como consulta del sistema de recuperación de información; esta locución verbal es entonces lematizada y buscada en el corpus base (en este caso un corpus periodístico). Un corpus que requiere ser también lematizado para procurar un empate adecuado entre la locución verbal fuente y los contextos buscados en dicho corpus. Por lo tanto, se realizó la lematización de cada texto de manera automática; en esta etapa se dirá que el corpus periodístico ha sido transformado. En la tabla 3.1 se presentan ejemplos de oraciones lematizadas

[13] En este libro se denomina longitud textual a la cantidad de palabras contenidas en él.

pertenecientes a una noticia, mostrando de esta manera las oraciones y como fueron transformadas.

Tabla 3.1. Muestra de oraciones lematizadas del corpus periodístico.

Identificador	**Oraciones sin lematización**
	Oraciones lematizadas
ID001	La comisión de obreros de la Cervecería Cuauhtémoc Moctezuma inicia las pláticas conciliatorias con los representantes de la empresa el 15 de enero.
	El comisión de obrero de el cervecería Cuauhtémoc Moctezuma iniciar el plática conciliar con el representante de el empresa el 15 de enero.
ID002	El emplazamiento vence el día 20, aunque los delegados de los trabajadores tienen confianza de lograr un incremento salarial.
	El emplazamiento vencer el día 20, aunque el delegar de el trabajador tener confianza de lograr un incremento salarial.
ID003	Integrantes de la delegación que viajarán la próxima semana a la Ciudad de México para iniciar las negociaciones.
	Integrante de el delegación que viajar el próximo semana a el ciudad de México para iniciar el negociación.
ID004	Comentaron que solicitan un aumento salarial igual o superior a la inflación que es por arriba del 6.5 por ciento.
	Comentar que solicitar un aumento salarial igual o superior a el inflación que ser por arriba del 6.5 por ciento.
ID005	El pliego de demandas salariales, incluye que la compañía cervecera ratifique su compromiso de que no habrá recorte de personal.
	El pliego de demanda salarial, incluir que el compañía cervecero ratificar suyo compromiso de que no haber recorte de personal.

ID006	Los delegados de la comisión revisora de la cervecería, en tanto, actualizan sus conocimientos mediante un curso de capacitación sobre revisión del contrato salarial y la Ley Federal del Trabajo.
	El delegar de el comisión revisar de el cervecería, en tanto, actualizar suyo conocimiento mediante un curso de capacitación sobre revisión del contrato salarial y el ley federal del trabajo.

En el corpus de comentarios se realizan las mismas etapas, a excepción de que los comentarios no son divididos en oraciones debido a que existen comentarios con una longitud muy pequeña. Con respecto a las locuciones verbales, como se ha mencionado, estas han sido lematizadas con el objetivo de incluir la mayoría de contextos con las variantes morfológicas de estas unidades fraseológicas. Así, por ejemplo, si se tiene la locución verbal (1a), un conjunto de posibles variantes morfológicas sería, por ejemplo, (1b, 1c, 1d, 1e, 1f).

(1) a. darse por vencido
b. darse por vencida
c. darse por vencidos
d. darse por vencidas
e. darnos por vencidos
f. darnos por vencidas

El proceso de lematización, tanto de cada uno de los componentes de dichas locuciones, como de los componentes del corpus es necesario, pues tiene por objetivo buscar cualquier tipo de variante morfológica de las locuciones verbales. En la tabla 3.2 se muestra una locución verbal con cuatro posibles contextos de ocurrencia. Tal y como puede observarse, dichos contextos no podrían haber sido encontrados si la búsqueda se hubiese realizado sin llevar a cabo el proceso de lematización.

Tabla 3.2. Locución verbal con contextos.

Locución verbal	Contextos
Poner el dedo en el renglón	Estos tiempos no concuerdan con los derechos humanos enarbolados en algunas dependencias y habrán de poner el dedo en el renglón para que se agilicen.
	Destacó sus proyectos en empleo, seguridad, educación y cultura enfatizando que se pondrá el dedo en el renglón en la transparencia y honestidad.
	Digo que en diversas ocasiones se ha conocido que son los empleados municipales los que ponen el dedo en el renglón cuando el alcalde o la primera dama empiezan con su actitud de perdonavidas que les cambia el carácter.
	La Coordinadora Nacional de Trabajadores de la Educación CNTE que ya ha puesto el dedo en el renglón de los contratos individuales lo cual advierten podría ir en contra de los derechos laborales de los docentes.

En resumen, se han construido los recursos léxicos, éstos a su vez han sido pre-procesados y la etapa final es la búsqueda de contextos. En la figura 3.4 se observa gráficamente el proceso.

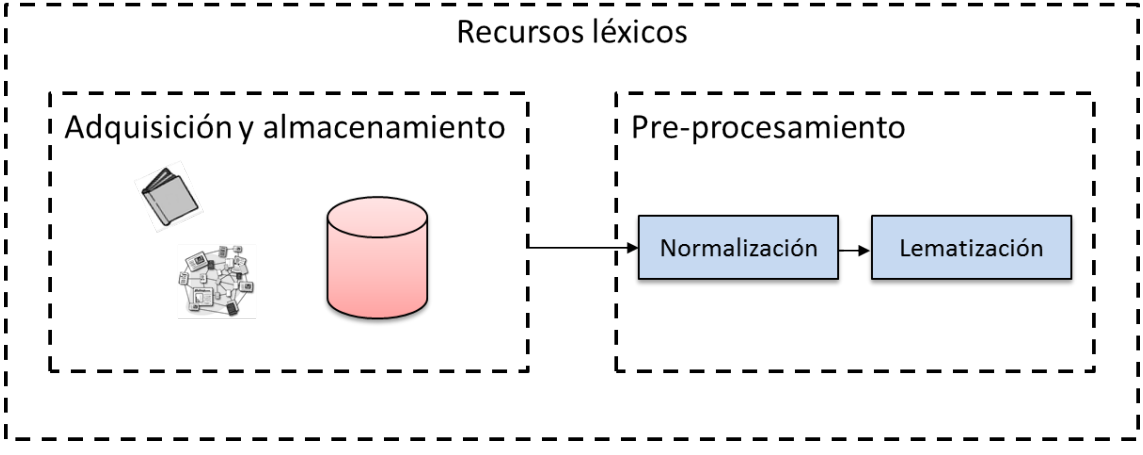

Figura 3.4. Etapas de pre-procesamiento de los datos.

El resultado de la etapa de pre-procesamiento es utilizado como insumo para recuperar los contextos asociados a las locuciones verbales. Para llevar a cabo este objetivo, se utilizan técnicas de recuperación de información en las que se identifican textos con al menos una ocurrencia de alguna de las locuciones verbales. Este proceso considera la ocurrencia original de la locución verbal y/o alguna de sus variantes morfológicas; por tal motivo antes de realizar la recuperación de contextos se lematizaron los recursos léxicos. Se considera además la separación entre las palabras que conforman la locución, es decir, se recuperan contextos en los que los componentes de las locuciones no son contiguos.

El proceso de recuperación de información consiste en identificar los elementos que coinciden con las características dadas en la consulta realizada al sistema de recuperación de información; en otras palabras, las características de la consulta son las locuciones verbales y el resultado son los contextos que contienen a estas locuciones verbales (ver figura 3.5).

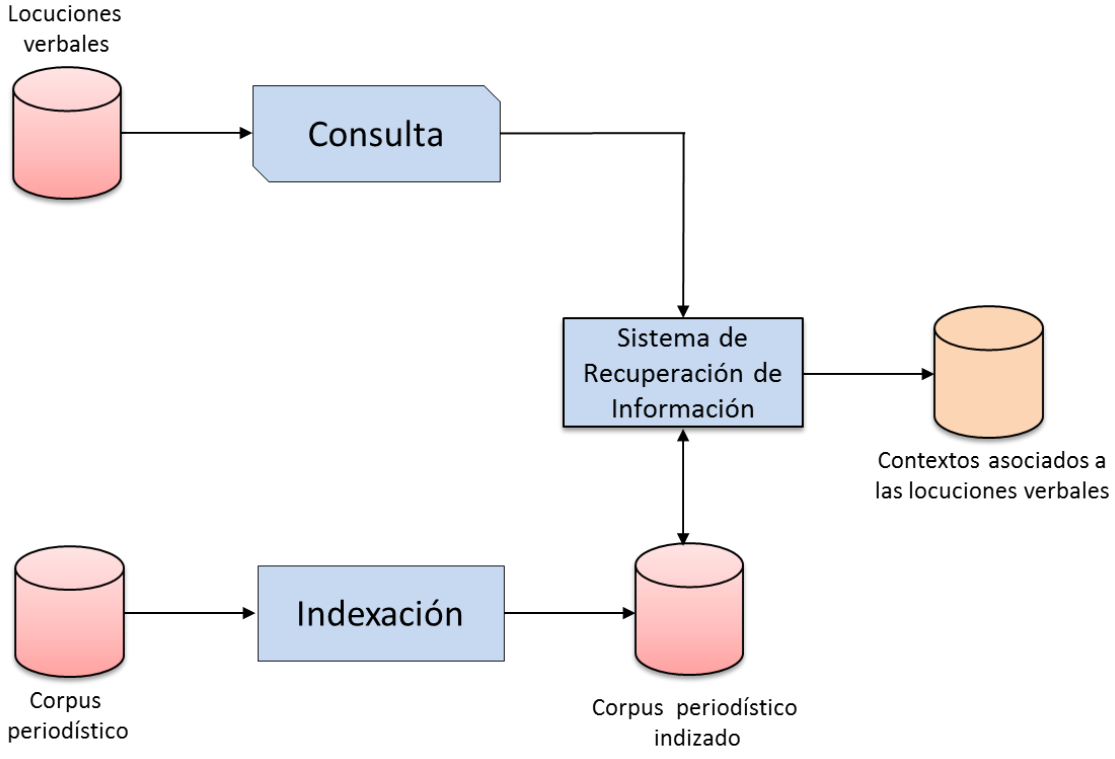

Figura 3.5. Sistema de recuperación de información de contextos asociados a las locuciones verbales.

Después de aplicar las técnicas de recuperación de información se han encontrado 3,164 textos informativos que contienen al menos una ocurrencia de alguna de las locuciones verbales. Como se ha mencionado con anterioridad, este proceso considera la ocurrencia original de la locución y/o alguna de sus variantes morfológicas. Como consecuencia del conteo de la ocurrencia de las locuciones verbales mexicanas en el corpus recopilado, en este proceso, se ha construido un corpus etiquetado de locuciones verbales. Este corpus puede utilizarse como conjunto de datos de entrenamiento para los métodos de

aprendizaje automático supervisado, con el objetivo de identificar si una noticia contiene o no una locución verbal.

Se debe aclarar que dichos contextos solamente indican que la secuencia de palabras asociadas a una cierta locución verbal tiene dicho contexto, sin embargo, no existe una garantía de que esos contextos se refieran a una verdadera locución verbal (sentido figurado), pues podrían ser contextos asociados a una expresión en su sentido literal. En los dos ejemplos siguientes se puede observar que en un caso se refiere a un contexto real alrededor de una locución verbal (ver 3), pero en el otro caso, se refiere a una expresión literal (ver 2).

(2) Juan y María se vieron la cara para observar si tenían arrugas en el rostro.

(3) Juan le vio la cara a María al burlarse de ella en el patio de su casa.

En este sentido, este proceso de evaluación automática requerirá una etapa posterior de filtrado que permita determinar los casos en los cuales los contextos están asociados a una verdadera locución verbal, y los casos en los que dichos contextos solamente hacen referencia a una expresión literal.

3.2 Evaluación de contextos

Una vez que han extraído los contextos asociados a las locuciones verbales, la siguiente etapa es la evaluación de estos contextos, es decir, mediante métodos manuales revisar lo que el sistema de recuperación de información ha proporcionado como respuesta a la consulta realizada. En la figura 3.6 se presenta gráficamente el proceso. Este proceso se realiza con el objetivo de construir corpora supervisado que se utilizará posteriormente en la metodología propuesta en este libro.

Como resultado del sistema de recuperación de información, se ha proporcionado evidencia del uso de las locuciones verbales, sin embargo, no en todos los casos se trata realmente de una locución verbal. Por tal motivo, es necesaria la revisión manual de los resultados proporcionados, además de que este proceso dará como resultado final un corpus supervisado, que hasta donde se sabe, no existe actualmente en la literatura un recurso léxico de este tipo. Así, se considera este recurso como una aportación más de este libro.

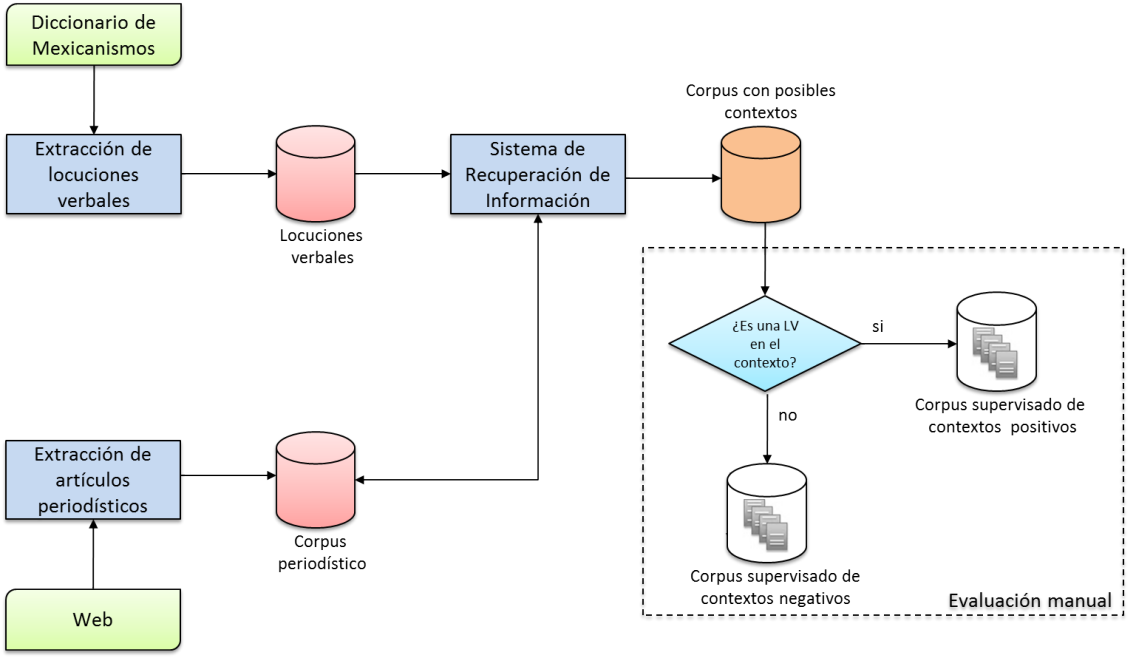

Figura 3.6. Evaluación de contextos asociados a las locuciones verbales.

Como puede observarse en la figura 3.6, los recursos que se han creado sirven de entrada para alimentar al sistema de recuperación de información, obteniendo de esta manera un corpus con posibles contextos asociados a las locuciones verbales. Una vez que se tienen

estos posibles contextos, se prosigue con el proceso de evaluación manual, el cual consiste principalmente en evaluar la pregunta: ¿Existe realmente una locución verbal en este contexto? La respuesta se da basándose en la definición de locución verbal, proporcionada anteriormente, y sólo tiene dos respuestas posibles, *sí* o *no*. Ambas respuestas, son de utilidad ya que, si la respuesta es un *sí*, lo que se tendrá son contextos realmente asociados a las locuciones verbales. Sin embargo, si la respuesta es *no* también se tendrá información que es útil para las máquinas de aprendizaje supervisado. En otras palabras, se tendrán contextos en los cuales no existe una locución verbal, llamados contextos negativos (en realidad se puede referir a ellos como muestras negativas), en los cuales se obtendrán también características de textos en los que no aparece una locución verbal.

En resumen, se constata que los resultados proporcionados por el sistema de recuperación de información aportan, de cierta manera, características que en conjunto pueden ayudar a discriminar si hay o no realmente una locución verbal, ya sean características para contextos positivos o características para contextos negativos.

Los resultados obtenidos hasta el momento son almacenados de manera que puedan ser manipulados de una manera eficiente. Para este propósito, el contexto se dividió en tres partes que son el *contexto izquierdo*, la *locución verbal* y el *contexto derecho*. La primera parte corresponde a lo que está del lado izquierdo de la locución verbal, que es lo que se ha denominado contexto izquierdo. Posteriormente se tiene la locución verbal y finalmente la parte derecha que compone al contexto, a lo que se le ha denominado contexto derecho. Existen algunas locuciones que no contienen ya sea el contexto izquierdo o el derecho, debido a que la locución puede aparecer al inicio o al final del texto; además de que pueden existir más de dos contextos en ambos lados debido a que en el texto posiblemente se encuentra la existencia de al menos dos locuciones verbales.

El dividir los contextos en tres partes ha permitido corroborar que se tiene elementos oracionales antes y después de la locución verbal. Por lo tanto, es posible agregar el mayor

número de características para cumplir con el objetivo final de este libro, proporcionando elementos de estudio que se adjuntan como características de análisis. En la tabla 3.3 se observan algunos ejemplos de la división realizada de los contextos.

Los contextos que se han obtenido contienen una serie de palabras que presumiblemente son una locución verbal. Sin embargo, es en el proceso de evaluación o validación de estos textos cuando se decidirá si realmente contienen o no una locución verbal. En la siguiente sección, se describe el proceso de validación manual de los contextos asociados a las locuciones verbales que se han considerado para este libro.

Tabla 3.3. División de los contextos asociados a las locuciones verbales.

		División del contexto		
Identificador	**Locución verbal**	**Contexto izquierdo**	**Locución verbal**	**Contexto derecho**
3355	poner la cara	No	puso la cara	pero del mismo movimiento de las manos le brincó y le hizo infección en toda la cara
3357	poner la cara	Con esa diferencia de 22 unidades los Indios hicieron un tímido intento por	poner la cara	en el último período sin apenas les alcanzó para acortar la distancia final a quince puntos
3368	poner el cuerno	Finalmente Azela Robinson encarnará a una esposa abnegada casada con un cronista deportivo que cree que su matrimonio y la vida son perfectas sin darse cuenta que le	ponen el cuerno	
3379	poner el dedo en el renglón	Leonel Araiza León Guanajuato Juan Ignacio Torres Landa ex candidato a la Gubernatura del Estado	poner el dedo en el renglón	para llegar a lo más profundo sin tapar a nadie

		afirmó que las finanzas estatales y municipales de la Entidad requieren una revisión profunda		
3405	ponerse el saco	Nuestra integración es contra el autoritarismo si el PRI quiere	ponerse el saco	acotó. El senador Manuel Camacho Solís resaltó que se construye desde ese momento un frente favor de la democracia en este sexenio
3408	ponerse guapo	Les comunico a todos los amigos amigas y familiares que la guapísima Griselda Ramos Marín que el día 21 de este mes festejará su cumpleaños. A	ponerse guapos	con el regalito

3.2.1 Validación de contextos

Los métodos de aprendizaje automático asumen que se tienen datos supervisados con los cuales se puede adquirir (aprender) conocimiento. En este caso, se necesita corpora etiquetado manualmente por expertos indicando si cierto texto tiene o no una locución verbal. Por lo tanto, se ha construido un corpus supervisado para los experimentos propuestos en este capítulo, seleccionando un número de noticias (del corpus periodístico) que contengan y otros que no contengan a estas unidades fraseológicas. Así que, para lograr el objetivo planteado, lo primero es utilizar la base de datos de locuciones

verbales creada a partir del Diccionario de Mexicanismos. La evaluación manual se llevó a cabo revisando cada contexto y en cada uno de ellos evaluando si realmente correspondía a una locución verbal o a una frase libre.

Los resultados obtenidos por el sistema de recuperación de información han sido evaluados manualmente. Para este fin, los contextos recopilados han sido manualmente anotados por 5 anotadores humanos con un grado de acuerdo entre anotadores superior al 80%. Cada anotador humano fue encargado de clasificar manualmente cuando un texto plano contenía una locución verbal (Clase 1) y cuando el texto no contenía una locución verbal (Clase 2). La evaluación, como se ha mencionado, se realizó tomando en cuenta principalmente la definición de locución verbal. La descripción del corpus supervisado obtenido se muestra en la tabla 3.4.

Tabla 3.4. Descripción del corpus anotado manualmente.

Característica	Clase 1 (Locución Verbal)	Clase 2 (No es Locución Verbal)	Total
Instancias	1,959	1,205	3,164
Palabras	117,715	63,600	181,315
Vocabulario	16,359	10,817	20,953
Longitud mínima	3	3	3
Longitud máxima	2,291	302	2,291
Longitud promedio	60.09	52.78	57.31

3.3 Lexicón supervisado de locuciones verbales

La comunicación humana depende del conocimiento que se va adquiriendo e incorporando al repertorio de expresiones, frases y dichos que se poseen a lo largo de la vida, las cuales son matizadas por cada cultura de manera diferente y específica. Este repertorio es utilizado en diferentes momentos, de acuerdo a diversos temas y circunstancias, distinguiéndose así las distintas comunidades sociales. A partir de este hecho, en esta sección se analizan las probabilidades de uso de las locuciones verbales en el género textual periodístico, teniendo como repertorio a estas unidades fraseológicas y como comunidad a estudiar, las noticias recolectadas de México.

El estudio realizado proporciona como conclusión un lexicón supervisado de locuciones verbales. Así, se considera este recurso léxico como un resultado y aportación sobresaliente del libro. Suministrando, de esta manera, acceso a un nuevo lexicón con frases como entradas (en lugar de palabras individuales), en la que cada entrada es

asociada a un valor real (normalizado entre cero y uno) que indica la probabilidad de ser una locución verbal.

Este recurso léxico puede ser altamente benéfico para la comunidad de lingüística computacional, puesto que hasta donde se sabe no se ha construido para un corpus de dominio restringido o por lo menos no se ha considerado con esta cantidad de datos. En particular, se tienen recolectadas 1,219 locuciones verbales del Diccionario de Mexicanismos, sin embargo, para la construcción del lexicón se han seleccionado las locuciones verbales más significativas.

Hasta ahora, este lexicón contiene 69 entradas, debido a que se han seleccionado sólo las locuciones verbales más frecuentes en el corpus periodístico, es decir, se tomó en cuenta la frecuencia de ocurrencia de las locuciones verbales en el corpus, seleccionando al final las más frecuentes. Un ejemplo de entradas de este lexicón se muestra en la tabla 3.5.

Tabla 3.5. Lexicón de UFV con probabilidades de ser y no ser UFV en contextos del dominio noticiero.

Locución verbal	**Probabilidad de ser una LV real P(LV)**	**Probabilidad de no ser una LV real P(¬LV)**
comer el mandado	0.94	0.06
dar el ancho / no dar el ancho	0.95	0.05
darse por vencido	0.51	0.49
dejar colgado	0.59	0.41
echar porras	0.52	0.48

no quitar el dedo del renglón	0.99	0.01
pegar su chicle	0.95	0.05
ponerse la camiseta	0.57	0.43
salir a flote	0.83	0.17
valer madre	0.98	0.02

3.3.1 Identificación de probabilidades de uso de las locuciones verbales

Los relatos periodísticos fueron recolectados de la Web, posteriormente, se extrajeron contextos por medio de un sistema de recuperación de información empleando a las locuciones verbales como consulta de entrada. Por lo tanto, se obtuvieron relatos periodísticos de internet en los cuales hay contextos que pueden contener o no en su interior a estas unidades fraseológicas. En otras palabras, la distribución de ocurrencia de cierta locución verbal se puede aproximar contando el número de veces que la expresión candidata es realmente una locución verbal. Al hacerlo, es posible estimar la probabilidad de una secuencia de palabras (locución verbal candidata) de ser realmente una locución verbal en textos reales. Este proceso se realizó a partir de la evaluación manual descrita anteriormente, y como resultado se obtuvo un lexicón de probabilidades de uso de las locuciones verbales.

3.3.2 Perspectivas de uso del lexicón supervisado

Las utilidades del lexicón pueden ser variadas (traducción automática, enseñanza del idioma, etc.), pero indudablemente es un recurso léxico valioso puesto que contempla

unidades poliléxicas en un contexto específico (dominio noticiero) y su probabilidad de ser una locución verbal en ese dominio.

Capítulo 4. Validación semiautomática de locuciones verbales

El humano interesado en la lengua ha querido investigar, desde diferentes perspectivas, el comportamiento que esta tiene. El enfoque común consiste analizar problemas discretos, que en su conjunto conforman la inmensidad que es la lengua. En este libro, el objeto de estudio de la lengua son las locuciones verbales para la lengua mexicana, una variante diatópica del español que se caracteriza por su propios aspectos fonéticos, léxicos y morfosintácticos.

Se tiene como meta proponer una metodología para validar lingüística-computacionalmente a las locuciones verbales en diferentes géneros textuales. Se sabe que, en general, este tipo de unidades fraseológicas no son tan frecuentes en recursos léxicos como lo son en textos del mundo real. Se considera que este problema de cobertura puede tener un impacto negativo en la realización de muchas tareas que requieren un procesamiento del lenguaje natural. Así, una de las etapas de la metodología a proponer necesariamente tendrá que ver con la construcción de recursos léxicos en donde se encuentren presentes estos tipos de unidades lingüísticas. De manera particular, para llevar a cabo esta tarea se han seleccionado documentos provenientes del género de noticias, debido a que cubren la lengua de una manera estándar, es decir, los artículos periodísticos son entendidos por todas las personas que hablan la lengua en la que son escritos; en este caso el español mexicano. Y porque existen evidencias y estudios de que las noticias son un buen referente para el análisis de estas unidades fraseológicas (Priego et al. 2014).

El reconocimiento automático de las locuciones verbales es una tarea importante para el área de lingüística computacional, debido a que expresan de forma estándar un concepto o una idea. Y precisamente se ha decidido centrarse en la identificación de las locuciones

verbales mexicanas dado que se desea colaborar significativamente con la comunidad de lingüística computacional en el desarrollo de una metodología capaz de solucionar esta tarea.

La metodología propuesta, ha sido desarrollada para que se pueda generalizar y se utilice en otros idiomas. Los resultados que se presentan cubren al idioma español; sin embargo, es posible aplicar las mismas etapas para otros idiomas, como es el caso del francés y del inglés. En este capítulo de este libro, se incorpora todo lo analizado y desarrollado en cuanto a los recursos léxicos, tratando de proyectar una de las principales aportaciones: la metodología propuesta. Adicionalmente, se presenta a partir de cuatro hipótesis la metodología para la validación de locuciones verbales.

4.1 Hacia la identificación de locuciones verbales

En este libro se está particularmente interesado en estudiar unidades fraseológicas escritas en español, un tipo de sintagmas que están formados por un núcleo verbal acompañado por sus complementos. En otras palabras, se tiene el interés en las locuciones verbales, que desde el punto de vista sintáctico expresan procesos y actúan como los predicados, con o sin complementos. Estas unidades fraseológicas, igual que los verbos, se combinan con el sujeto y los complementos para formar una oración. Cabe mencionar que las locuciones verbales presentan un alto grado de fijación, en comparación con otras unidades fraseológicas (Mejri 2008), por ejemplo: *leer entre líneas, costar un ojo de la cara, hacer de chivo los tamales,* entre otras.

En cuanto a los avances actuales en la tarea de identificación automática de locuciones verbales para el español, se considera el dominio periodístico mexicano y las locuciones verbales mexicanas. Esta primera sección está dedicada a los primeros experimentos realizados para la validación de locuciones verbales. Se presenta un enfoque para la validación de Locuciones Verbales que está basado en técnicas supervisadas de

aprendizaje automático, un área de la inteligencia artificial que concierne al estudio de sistemas computacionales que pueden llegar a aprender a partir de datos supervisados (manualmente etiquetados) y, por lo tanto, también se incluye a los clasificadores empleados en los experimentos. Finalmente, se presentan y discuten los resultados obtenidos en los experimentos llevados a cabo.

Así, el objetivo de esta sección es el de presentar los resultados obtenidos cuando diferentes métodos de aprendizaje automático supervisado se utilizan para determinar si una locución verbal existe o no en un artículo periodístico. Por lo tanto, se muestran los resultados obtenidos al llevar a cabo los primeros experimentos realizados contemplando: los recursos léxicos que incluyen la base de datos de locuciones verbales y el corpus periodístico; los métodos de aprendizaje supervisado empleados para el aprendizaje automático, y como tal los resultados. Toda esta información es descrita en las siguientes secciones.

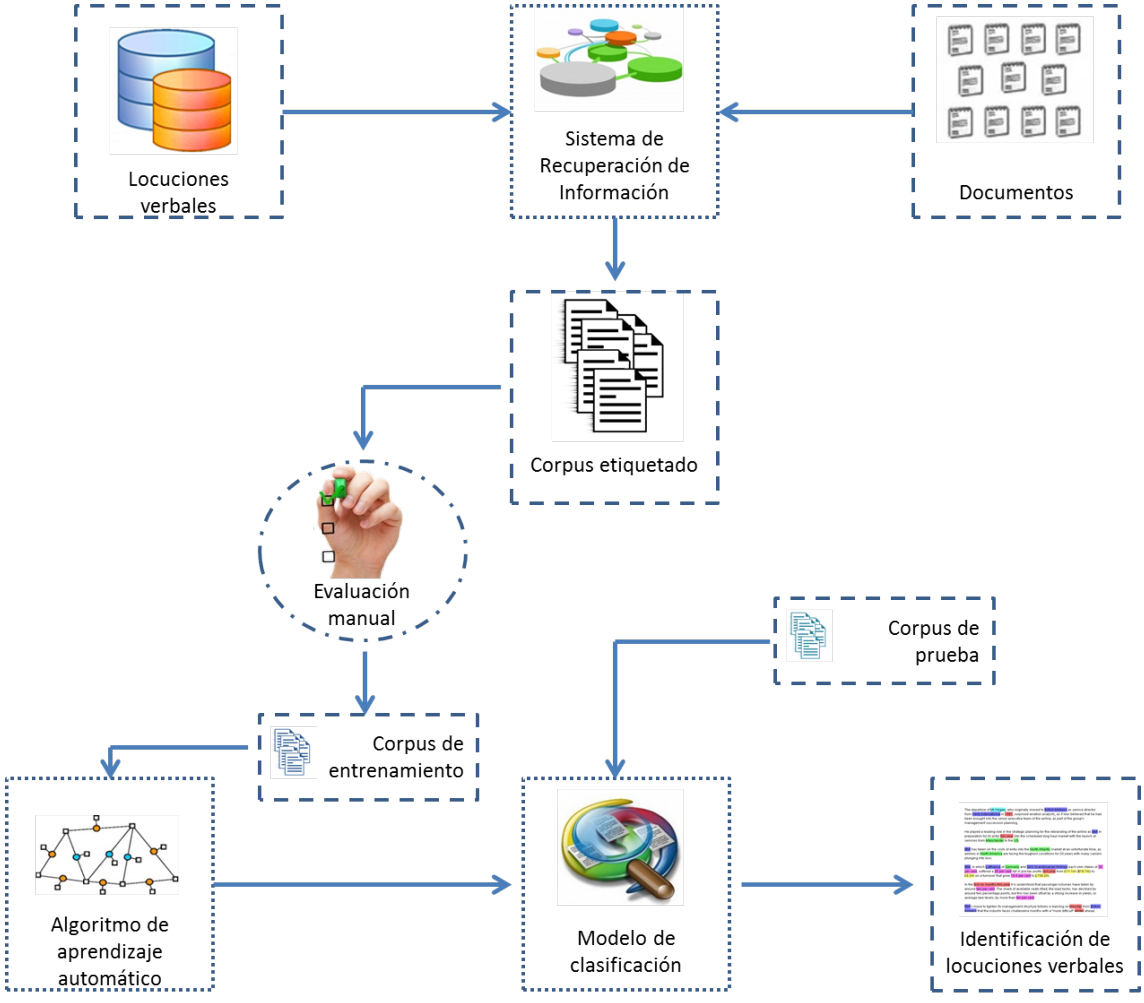

Figura 4.1. Metodología para la identificación de locuciones verbales.

4.1.1 Conjunto de datos

Los métodos de aprendizaje automático supervisado asumen que se tienen datos supervisados con los cuales se puede adquirir (aprender) conocimiento. En este caso, se necesita corpora etiquetado manualmente por expertos indicando si cierto texto tiene o no una unidad fraseológica verbal. Por lo tanto, para lograr el objetivo se ha utilizado el corpora descrito anteriormente, seleccionando un número de noticias (de un periódico mexicano) que contengan y no contengan unidades fraseológicas verbales. Dichas noticias son posteriormente evaluadas manualmente con la finalidad de obtener un corpus supervisado. Con respecto a la construcción del corpus supervisado, primero se han extraído las locuciones verbales del Diccionario de Mexicanismos.

Por lo tanto, los principales conjuntos de datos a utilizar para los primeros experimentos son los recursos léxicos que se han desarrollado y presentado en este libro, los cuales han sido descritos en capítulos previos.

En los experimentos realizados, todos los textos fueron representados por un vector de frecuencia de *n-gramas*, con n=1, 2 y 3. Sólo frecuencias de ocurrencia superiores a dos para los *n-gramas* fueron considerados para el vector de características. El corpus fue utilizado tanto para el entrenamiento como para las pruebas por medio de un mecanismo de validación cruzada a 10 pliegues (*10-fold cross-validation*).

4.1.2 Clasificadores empleados

Las técnicas de aprendizaje automático supervisado son capaces de aprender el proceso humano para identificar las unidades fraseológicas verbales con base en las características alimentadas por el corpus manualmente anotado. Con el fin de tener una perspectiva del tipo de clasificador que puede tratar mejor el problema de la identificación automática de unidades fraseológicas verbales, se han seleccionado algoritmos de aprendizaje para uno

de cuatro tipos de clasificadores diferentes: Bayes, Lazy, Functions y Trees. Los siguientes cuatro algoritmos de aprendizaje fueron los elegidos:

- **Naïve Bayes**: es un clasificador probabilístico basado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales.
- **K-Star:** Este es el clasificador de los k vecinos más cercanos con una función de distancia generalizada.
- **SMO:** Se trata de un algoritmo de optimización secuencial mínima para la clasificación de vectores soporte.
- **J48:** Es un algoritmo usado para generar un árbol de decisión, el árbol generado de C4.5 aprende con la implementación de la revisión 8 de C4.5.

4.1.3 Resultados experimentales

En esta sección se presenta la exactitud obtenida por cada clasificador cuando se trata de clasificar si existe o no una locución verbal en un texto plano. Se han empleado las siguientes medidas de evaluación estándares: verdaderos positivos, falsos positivos, precisión, exhaustividad, y Medida-F (en inglés *TP Rate - True Positive Rate, FP Rate - False Positive Rate, Precision, Recall y F-Measure,* respectivamente) (Manning & Schütze, 1999).

Las tablas 4.1, 4.2, 4.3 y 4.4 muestran los detalles de exactitud por clase utilizando los clasificadores supervisados *Naïve Bayes, K-Star, SMO* y *J48*, respectivamente. Como se puede observar, en todos los casos de clasificación de la Clase 1 (cuando el conjunto palabras es realmente una locución verbal) obtiene mejores resultados de clasificación con respecto a la Clase 2 (cuando el conjunto de palabras no es una locución verbal). Incluso si la diferencia no es tan significativa, este tema es importante.

Por un lado, el clasificador *K-Star* obtuvo los peores resultados de clasificación, con respecto a los otros clasificadores, resultando tener un valor de *F-Measure* igual a 0.712. Y por el otro lado, el clasificador *J48* obtiene los mejores resultados con un *F-Measure* promedio ponderado igual a *0.766*. En este caso, *J48* obtiene una exactitud de verdaderos positivos, precisión, exhaustividad y Medida-F mayor a 0.8 para la Clase 1, la cual consideramos que es bastante aceptable.

Tabla 4.1. Exactitud detallada por clase utilizando el clasificador de Naïve Bayes.

Clase	**Verdaderos positivos**	**Falsos positivos**	**Precisión**	**Exhaustividad**	**Medida-F**
Clase 1 (LV)	0.795	0.349	0.788	0.795	0.791
Clase 2 (¬LV)	0.651	0.205	0.662	0.651	0.657
Promedio ponderado	0.741	0.294	0.740	0.741	0.740

Tabla 4.2. Exactitud detallada por clase utilizando el clasificador K-Star.

Clase	Verdaderos positivos	Falsos positivos	Precisión	Exhaustividad	Medida-F
Clase 1 (LV)	0.760	0.367	0.771	0.760	0.765
Clase 2 (¬LV)	0.633	0.240	0.618	0.633	0.626
Promedio ponderado	0.711	0.319	0.713	0.711	0.712

Tabla 4.3. Exactitud detallada por clase utilizando el clasificador SMO.

Clase	Verdaderos positivos	Falsos positivos	Precisión	Exhaustividad	Medida-F
Clase 1 (LV)	0.831	0.336	0.801	0.831	0.816
Clase 2 (¬LV)	0.664	0.169	0.707	0.664	0.685
Promedio ponderado	0.767	0.272	0.765	0.767	0.766

Tabla 4.4. Exactitud detallada por clase utilizando el clasificador J48.

Clase	Verdaderos positivos	Falsos positivos	Precisión	Exhaustividad	Medida-F
Clase 1 (LV)	0.795	0.349	0.788	0.795	0.791
Clase 2 (¬LV)	0.651	0.205	0.662	0.651	0.657
Promedio ponderado	0.741	0.294	0.740	0.741	0.740

En la tabla 4.5 se muestran los porcentajes de las instancias que fueron clasificadas correcta e incorrectamente. Esta tabla básicamente resume el promedio ponderado de las tablas de resultados previamente mostradas. De hecho, se incluye un gráfico (ver figura 4.2) con el objetivo de mostrar de una manera más clara el resumen de los resultados promedio.

Tabla 4.5. Porcentaje de instancias clasificadas correcta versus incorrectamente.

Clasificador	Tipo	Instancias correctas (%)	Instancias incorrectas (%)
Naïve Bayes	Bayes	74.05	25.95
K-Star	Lazy	71.14	28.86
SMO	Functions	75.32	24.68
J48	Trees	76.74	23.26

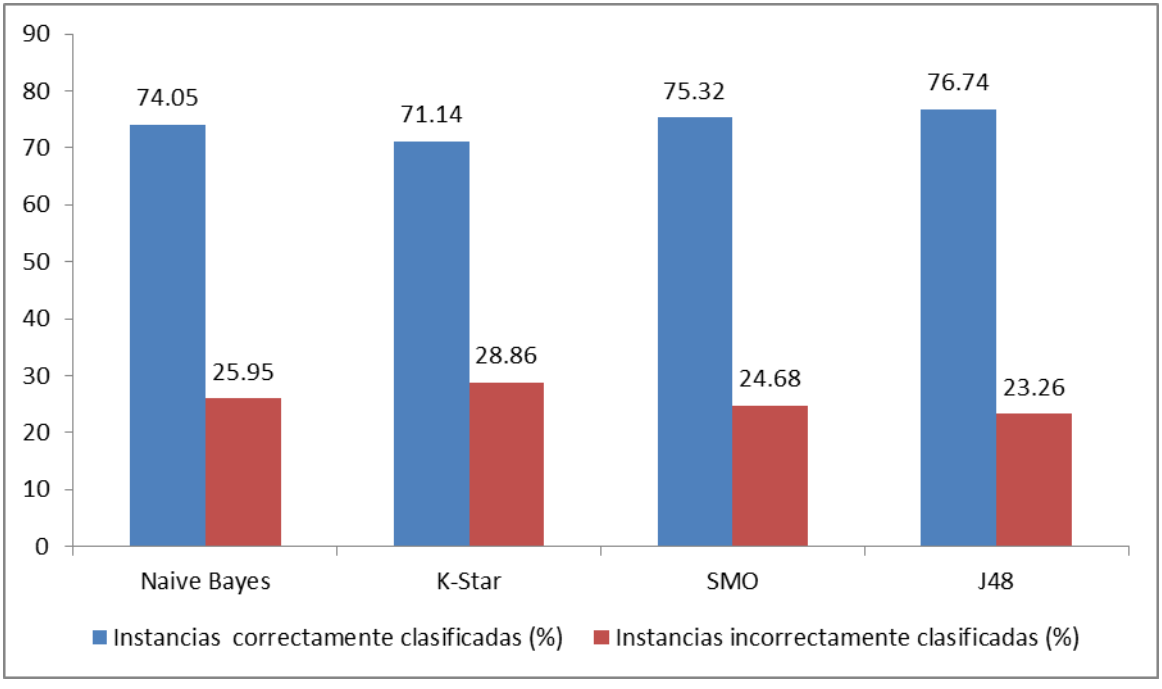

Figura 4.2. Comparación entre los diferentes clasificadores empleados en los experimentos.

Los resultados obtenidos a través del uso de técnicas de aprendizaje automático parecen ser bastante efectivos. Es seguro que el número de muestras en el corpus de entrenamiento es un factor fundamental para obtener aún mejores resultados. Sin embargo, tal y como sucede con estos métodos computacionales, el esfuerzo o costo humano necesario para llevar a cabo la anotación manual en el conjunto supervisado es muy alto. En función de la existencia de material humano que lleve a cabo el proceso de etiquetado manual de corpora para este problema, será posible decidir usar o no la metodología propuesta.

De cualquier manera, los experimentos realizados muestran que esta metodología proporciona resultados aceptables para la tarea en cuestión. Los modelos de aprendizaje automáticos son capaces de discriminar las características que ayudan a determinar, con un alto grado de probabilidad, si un texto dado contiene o no una locución verbal. Se debe aclarar en este punto, de acuerdo a los resultados obtenidos, el hecho de que los métodos supervisados son capaces de discriminar cuando una unidad fraseológica verbal tiene un significado literal o si el significado es idiomático.

Una de las aportaciones más importantes de este libro es la conformación de cuatro hipótesis que fundamentan experimentos adicionales para validar si una cierta unidad fraseológica verbal es o no una locución verbal. Estas hipótesis son de gran importancia pues emanan de un análisis detallado del comportamiento de las estructuras lingüísticas que son objetivo de estudio en este libro. Es importante remarcar que además de describir la metodología para la validación de las hipótesis propuestas, se presenta un ejercicio de evaluación. Seguramente, el lector podrá imaginar otras formas de uso y evaluación en corpora de diversos géneros o idiomas.

4.2 Hipótesis de fijación

"Entre más fijación tenga una unidad fraseológica verbal, es más alta la probabilidad de ser locución verbal".

4.2.1 Metodología para la validación o rechazo de la hipótesis planteada

Se considera que la manera de averiguar que una combinación de palabras presenta fijación fraseológica es mediante la aplicación de operaciones lingüísticas, como permutaciones, inserciones, sustituciones pronominales, conmutaciones, modificaciones gramaticales y transformaciones sintácticas. Por lo tanto, en esta hipótesis

se sustituyen algunos de los componentes de la unidad fraseológica verbal de entrada (locución verbal candidata) por sus sinónimos cercanos y se verifica la nueva unidad verbal.

En este sentido, se considera que, si el fenómeno de fijación existe, entonces la locución verbal candidata tiene una alta probabilidad de ser una locución verbal. La secuencia de palabras que contiene a la unidad fraseológica verbal junto con su transformación sinonímica es buscada en un corpus de referencia con la finalidad de encontrar evidencia de dicha unidad fraseológica.

La unidad fraseológica transformada debería ser gramaticalmente correcta, y por esta razón, se realiza su búsqueda en el corpus de referencia. Si hay evidencia de esta, significa que no está principalmente sujeta a las reglas de fijación y, por tanto, no es posible determinar si esta unidad es una locución verbal. Sin embargo, si no existe evidencia en el corpus de referencia de la unidad fraseológica transformada, entonces, es bastante claro que la locución verbal candidata tiene un alto grado de ser una locución verbal (LV) debido a que no está sujeta a la regla de modificabilidad. Este proceso se esquematiza a continuación:

Paso 1: Sustituir uno de los componentes de la LV candidata por un sinónimo cercano para producir una Unidad Fraseológica Transformada (UFT).
Paso 2: Calcular la frecuencia de ocurrencia de la LV candidata en el corpus de referencia (*Frec(LVC)*).
Paso 3: Calcular la frecuencia de ocurrencia de la UFT en el corpus de referencia (*Frec(UFT)*).
Paso 4: Si el cociente (*Frec(UFT)/Frec(LVC)*) tiende a cero, entonces se asume que la LV candidata tiene un alto grado de ser una locución verbal, en caso contrario, se repite el proceso hasta agotar todas las sustituciones de los elementos de la LV candidata.

4.2.2 Evaluación

A continuación, se presentan experimentos sobre un conjunto breve de unidades fraseológicas verbales con la finalidad de validar o rechazar la hipótesis planteada. La tabla 4.6 presenta una serie de ejemplos que muestran que la hipótesis planteada puede aceptarse. La implementación computacional del proceso de validación es bastante simple pues requiere la existencia de un diccionario de sinónimos y un corpus de referencia indizado por un sistema de recuperación de información.

Tabla 4.6. Ejemplos para la validación de la hipótesis de fijación.

Locución Verbal Candidata		**Unidad Fraseológica Transformada**		**Resultado**
Texto	**Frecuencia de ocurrencia**	**Texto**	**Frecuencia de ocurrencia**	**¿Es una Locución Verbal?**
Colgar los tenis	15,300	Tender los tenis	3	SI
Agarrar de puerquito	50	Agarrar de marranito	0	SI
Agarrarse de la greña	43	Tomarse de la greña	2	SI
Partir su mandarina	3,580	Cortar su mandarina	1	SI
Pasar gato por liebre	15,700	Pasar gato por conejo	12	SI
Comprar balones	11,500	Comprar pelotas	16,500	NO

4.2.3 Resultados

La metodología propuesta para la validación de la hipótesis de fijación ha permitido constatar que una de las características de las locuciones verbales es precisamente la fijación. Es normal que las locuciones verbales tengan un cierto grado de fijación y, por ende, se deben aún estimar los umbrales adecuados para caracterizar el nivel de fijación que una cierta locución verbal tenga. Este problema reviste una gran importancia y puede derivar un tema de investigación a futuro.

4.3 Hipótesis de traducción

"Entre más literal la traducción de una unidad fraseológica verbal, menor su probabilidad de ser una locución verbal".

4.3.1 Metodología para la validación o rechazo de la hipótesis planteada

En esta hipótesis se traduce la locución verbal a alguna otra lengua (en este caso, francés) utilizando diccionarios estadísticos de traducción y se verifica si la nueva locución verbal (locución verbal en francés) tiene evidencia de ser realmente una locución verbal en el corpus de referencia (corpus en francés). Así, se sustituyen todos los componentes de la unidad fraseológica verbal de entrada (locución verbal candidata) por sus traducciones literales y se verifica la nueva unidad verbal.

En este sentido, se considera que, si el fenómeno de traducción no existe, entonces la LV candidata tiene una alta probabilidad de ser una locución verbal. La secuencia de palabras que contiene a la unidad fraseológica verbal junto con su transformación llevada a cabo por una traducción literal es buscada en un corpus de referencia (en otro idioma) con la finalidad de encontrar evidencia de dicha unidad fraseológica.

La unidad fraseológica transformada debería ser gramaticalmente correcta, y por esta razón, se realiza su búsqueda en el corpus de referencia. Si no hay evidencia de la misma, significa que no está principalmente sujeta a las reglas de traducción literal y, por tanto, es posible determinar con alto grado de probabilidad que esta unidad es una locución verbal. Sin embargo, si existe evidencia en el corpus de referencia de la unidad fraseológica transformada, entonces, puede suceder que la LV candidata no es una locución verbal, o que la traducción de la locución verbal en el otro idioma corresponde literalmente, es decir, el significado idiomático es el mismo en ambos idiomas. Este proceso se esquematiza a continuación:

Paso 1: Sustituir cada uno de los componentes de la LV candidata por su traducción literal para producir una Unidad Fraseológica Transformada (UFT).
Paso 2: Calcular la frecuencia de ocurrencia de la LV candidata en el corpus de referencia en el idioma fuente (*Frec(LVC)*).
Paso 3: Calcular la frecuencia de ocurrencia de la UFT en un corpus de referencia en el idioma destino (*Frec(UFT)*).
Paso 4: Si el cociente (*Frec(UFT)/Frec(LVC)*) tiende a cero, entonces se asume que la LV candidata tiene un alto grado de ser una locución verbal, en caso contrario, no es posible afirmar nada acerca de la LV candidata.

4.3.2 Evaluación

A continuación, se presentan experimentos sobre un conjunto breve de unidades fraseológicas verbales, con la finalidad de validar o rechazar la hipótesis planteada. La tabla 4.7 presenta una serie de ejemplos que muestran que la hipótesis planteada puede aceptarse. La implementación computacional del proceso de validación no es complicada pues requiere la existencia de dos corpora de referencia, uno en el idioma fuente y otro en el idioma destino, además de un diccionario estadístico de traducción que puede ser construido con facilidad.

Tabla 4.7. Ejemplos para la validación de la hipótesis de traducción.

Locución Verbal Candidata		**Unidad Fraseológica Transformada**		**Resultado**
Texto	**Frecuencia de ocurrencia**	**Texto**	**Frecuencia de ocurrencia**	**¿Es una Locución Verbal?**
Colgar los tenis	15,300	étendre les baskets	0	SI
Agarrar de puerquito	50	attraper de petit porc	0	SI
Agarrarse de la greña	43	s'attraper la tignasse	1	SI
Partir su mandarina	3,580	diviser sa mandarine	0	SI
Pasar gato por liebre	15,700	prendre un chat pour un lièvre	1	SI
Comprar balones	11,500	acheter des balles	49,500	NO

4.3.3 Resultados

La metodología propuesta para la validación de la hipótesis de traducción ha permitido constatar que una de las características de las locuciones verbales es precisamente la idiomaticidad. Es normal que las locuciones verbales estén asociadas a un entorno geográfico y en particular a una lengua y, por ende, cuando se realiza una traducción literal de la locución verbal, normalmente no se obtendrá evidencia de la existencia de la transformación en el lenguaje destino.

4.4 Hipótesis de atracción interna y correlación contextual

"Entre mayor sea la atracción interna y menor la correlación contextual en una unidad fraseológica verbal, es más alta la probabilidad de que la unidad sea una locución verbal".

4.4.1 Metodología para la validación o rechazo de la hipótesis planteada

En esta hipótesis se emplean métodos estadísticos para determinar el nivel de atracción y correlación contextual entre los términos de la unidad fraseológica verbal y los de su contexto. El procedimiento planteado para validar la hipótesis se basa entonces en medir la correlación entre las palabras que constituyen a la Unidad Fraseológica Verbal (UFV).

Formalmente, sean dos palabras p_1 y p_2, entonces, en este libro se estima la probabilidad de que dos palabras estén juntas a partir de la ecuación siguiente:

$$P(p_i, p_j) = \frac{frec(p_i, p_j)}{frec(p_i)}$$

(6.1)

Siendo $frec(p_i, p_j)$ la frecuencia de ocurrencia de las dos palabras p_i y p_j, en un contexto determinado por un cierto tamaño de ventana (normalmente del tamaño de una oración). Entonces, para toda palabra p_i que pertenece a la unidad fraseológica verbal, podemos estimar su *Nivel de Atracción Interna* (NAI) como:

$$NAI(UFV) = \frac{1}{|UFV|} \sum_{\forall p_i \neq p_j \,;\, p_i, p_j \in UFV} P(p_i, p_j)$$

(6.2)

Por otro lado, el *Nivel de Correlación Contextual* (NCC) de un texto *T* compuesto por un Contexto Izquierdo (CI), una UFV, y un Contexto Derecho (CD) se puede estimar utilizando la siguiente ecuación:

$$NCC(T) = \frac{1}{|T|} \sum_{p_i \in UFV \ \wedge p_j \in \{CI \ \cup CD\}} P(p_i, p_j)$$

(6.3)

Podemos decir que, si *NAI(UFV)* es mayor que *NCC(T)*, entonces es altamente probable que la unidad fraseológica verbal sea una locución verbal (Priego & Pinto, 2019).

A continuación, se presentan ejemplos de evaluación sobre la metodología propuesta para la validación de la hipótesis de atracción interna y correlación contextual.

4.4.2 Evaluación

Considere el siguiente texto:

(1) *T* = "Si me mientes, **te parto tu mandarina en gajos**, me dijo mi amigo"

A partir de (1), se identifican los componentes del texto, a decir, la unidad fraseológica verbal, el contexto izquierdo y el contexto derecho:

UFV = "te parto tu mandarina en gajos"
CI = "Si me mientes"
CD = "me dijo mi amigo"

Tomando en consideración los componentes detectados, se realiza la evaluación de la hipótesis, usando en este ejemplo, únicamente aquellas palabras que se encuentren en alguna de las siguientes categorías morfológicas: verbos, adjetivos, sustantivos. Así, los componentes anteriores quedarían como sigue:

UFV = {parto, mandarina, gajos}
CI = {mientes}
CD = {dijo, amigo}

Para obtener el nivel de atracción interna, *NAI(UFV)*, es necesario calcular las probabilidades siguientes:

P(parto, mandarina) = frec(parto, mandarina) / frec(parto) = 0.0011
P(parto, gajos) = frec(parto, gajos) / frec(parto) = 0.2907
P(mandarina, gajos) = frec(mandarina, gajos) / frec(mandarina) = 0.2059

Para estimar la probabilidad de que dos palabras aparezcan juntas en un contexto se puede calcular, tal y como se ha mencionado anteriormente, usando la frecuencia de que ambas palabras aparezcan en el contexto y dividir este valor entre las veces que aparece la primera palabra en el corpus. Si se cuenta con un sistema propio de recuperación de información, se puede realizar este cálculo de una manera muy fácil utilizando un operador tipo NEAR[14]. Sin embargo, si no se cuenta con un sistema de recuperación información personalizable, entonces se puede hacer uso de los sistemas comerciales (*google, yahoo, msn*, etc), considerando que dicha estimación pueda realizarse usando *n*-gramas de longitud mayor. Por ejemplo, la probabilidad (parto, mandarina), puede estimarse considerando la secuencia completa de palabras "parto su mandarina". Así, *P("parto su mandarina") = frec("parto su mandarina") / frec("parto su")*.

Así, NAI(UFV)= 1/3 * (0.0011 + 0.2907 + 0.2059) = 0.1659

[14] El operador NEAR es conocido como un operador de cercanía y permite determinar cuándo dos palabras se encuentran a una distancia de *n* palabras una de la otra. En google, por ejemplo, se puede usar el operador AROUND para llevar a cabo este cálculo.

Por otro lado, para obtener el nivel de correlación contextual, *NCC(T)*, es necesario calcular las probabilidades siguientes:

P(parto, mientes) = frec(parto, mientes) / frec(parto) = $1.136x10^{-5}$
P(mandarina, mientes) = frec(mandarina, mientes) / frec(mandarina) = $7.57x10^{-6}$
P(gajos, mientes) = frec(gajos, mientes) / frec(gajos) = $1.51x10^{-6}$
P(parto, dijo) = frec(parto, dijo) / frec(parto) = $2.62x10^{-5}$
P(mandarina, dijo) = frec(mandarina, dijo) / frec(mandarina) = 0.0004
P(gajos, dijo) = frec(gajos, dijo) / frec(gajos) = $3.50x10^{-5}$
P(parto, amigo) = frec(parto, amigo) / frec(parto) = 0
P(mandarina, amigo) = frec(mandarina, amigo) / frec(amigo) = 0
P(gajos, amigo) = frec(gajos, amigo) / frec(gajos) = $6.58x10^{-5}$

Asi, NCC(T) = 1/9 * ($1.136x10^{-5}$ + $7.57x10^{-6}$ + $1.51x10^{-6}$ + $2.62x10^{-5}$ + 0.0004 + $3.50x10^{-5}$ + 0 + 0 + $6.58x10^{-5}$) = $6.28x10^{-5}$

En resumen, NAI(UFV) = 0.1659 y NCC(T) = $6.28x10^{-5}$, por lo tanto, se puede afirmar que existe una mayor atracción interna que una correlación contextual. Así, la unidad fraseológica verbal "te parto tu mandarina en gajos" tiene una muy alta probabilidad de ser una locución verbal.

4.4.3 Resultados

Se ha presentado únicamente un ejemplo para demostrar que la hipótesis planteada está bien fundamentada. Si bien, los cálculos asociados a la probabilidad de ocurrencia de los pares de palabras considerados pueden ser estimados de diversas maneras, no es el objetivo hacer un estudio exhaustivo de las métricas de co-ocurrencia o de estimación de probabilidad, sino más bien, plantear una hipótesis válida para la identificación semiautomática de locuciones verbales a partir de textos planos.

4.5 Hipótesis del dominio terminológico

"Entre mayor sea la cantidad de palabras de la locución verbal que se encuentran fuera del dominio al que pertenece el contexto, mayor su probabilidad de ser locución verbal".

4.5.1 Metodología para la validación o rechazo de la hipótesis planteada

En esta hipótesis se identifican los términos del vocabulario en el dominio actual, con el propósito de determinar si existe realmente o una Unidad Fraseológica Verbal (UFV) dentro del texto a analizar (Priego et al., 2018b; Pinto & Priego, 2020).

Así, la hipótesis se pretende validar calculando que tanto el dominio semántico de las palabras de la UFV está ligado o desligado del dominio terminológico de las palabras que se encuentran en su contexto. Este proceso podría ser llevado a cabo siguiendo la siguiente metodología:

Paso 1: Identificar el dominio o género de la UFV, *Dom(UFV)*.
Paso 2: Identificar el dominio o género del contexto en el que se encuentra inmerso la UFV, *Dom(Contexto)*.
Paso 3: Calcular la distancia entre Dom(UFV) y Dom(Contexto), Distancia(Dom(UFV), Dom(Contexto)).
Paso 4: Si *Distancia(Dom(UFV), Dom(Contexto))* es un valor pequeño, entonces, el dominio es similar y por tanto, la probabilidad de UFV de ser una locución verbal será baja; en caso contrario, la probabilidad de ser una locución verbal será alta.

Esta metodología es bastante genérica, pues no establece de manera particular cómo calcular *Dom(UFV)*, *Dom(Contexto)*, así como tampoco la distancia entre estos dos valores. Sin embargo, a continuación, se proponen algunas alternativas.

Una de las alternativas lógicas es el uso de ontologías de dominio. En caso de contar con éstas, es posible localizar cada una de las palabras que ocurren dentro de la UFV y calcular su densidad conceptual (Agirre & Rigau, 1996). Se puede hacer lo mismo para las palabras que ocurren en el contexto y de esta manera obtener los valores deseados. La métrica de distancia podría ser incluso la euclidiana, cuidando solamente el umbral necesario para determinar cuándo estos dos valores de dominio están lo suficientemente alejados.

Si bien, esta propuesta es adecuada, tiene el inconveniente de tener que contar con las ontologías de dominio. Otra propuesta sería utilizar métricas para el cálculo de la similitud semántica entre palabras (Vilariño et al., 2014), tales como las siguientes medidas de similitud de palabras ofrecidas por la herramienta NLTK[15]: Lesk (1986), Wu & Palmer (1994), Resnik (1995), Jiang & Conrath (1997), Leacock & Chodorow (1998), y Lin (1998).

Una última propuesta considera el aprendizaje del dominio basado en métricas de coocurrencia estadística de palabras (Pinto, 2008). En ese caso, de calcula un conjunto de palabras que se encuentran semánticamente relacionadas con una palabra fuente, obteniendo así un thesauri. La unión de las palabras relacionadas con cada palabra de la UFV daría como resultado un conjunto de palabras que definen el dominio de dicha UFV. Se puede realizar el mismo procedimiento para el contexto de la UFV, para finalmente comparar, posiblemente usando métricas basadas en el coeficiente de Jaccard[16], la cantidad de palabras que ambos dominios comparten.

[15] Herramienta de Python para el procesamiento de Lenguaje Natural; http://www.nltk.org/

[16] El coeficiente de Jaccard se calcula mediante la intersección entre dos conjuntos dividido entre la unión de los mismos conjuntos

4.5.2 Evaluación

Considere la siguiente oración que contiene una unidad fraseológica verbal:

(2) Cuando los dos coches se estrellaron, se **llevaron de corbata** al ciclista que iba al lado.

Los términos que aparecen en el contexto (por ejemplo: coches, estrellaron, ciclista), son evidentemente de un dominio diferente al de "corbata". Así, es fácilmente deducible que la UFV tiene un sentido idiomático y, por tanto, tiene una alta probabilidad de ser una locución verbal.

4.5.3 Resultados

Se ha presentado una metodología para la validación de esta hipótesis, y un ejemplo que demuestra la manera en que se puede utilizar dicha hipótesis para determinar cuándo una cierta unidad fraseológica verbal en realmente una locución verbal.

Como en las hipótesis anteriores, no en todos los casos funcionará completamente, sin embargo, es un elemento más que puede ser considerado para llevar a cabo la meta planteada en este tema de investigación, el cual es: la validación semiautomática de locuciones verbales para el español de México.

Capítulo 5. Estudio de la polaridad en las locuciones verbales

Uno de los fenómenos gramaticales que recientemente ha despertado el interés de los investigadores dedicados al estudio de una lengua es el de la polaridad. El fenómeno de la polaridad consiste en la sensibilidad que presentan algunas unidades léxicas a ciertos contextos, esto es, que algunos elementos no pueden aparecer libremente en cualquier entorno. Tales piezas léxicas se caracterizan por exigir la presencia de otro elemento en la oración en la que aparecen o por ser incompatibles con ciertos elementos (González Rodríguez, 2008).

Dependiendo del tipo de contexto al que es sensible un término, la polaridad se distingue por ser polaridad negativa, polaridad positiva o polaridad modal (Bosque, 1980; 1996). En este trabajo se abordan los tres tipos de polaridad y se hace notar que a la polaridad modal muy regularmente se le denomina también como polaridad neutral. En la polaridad negativa, los términos exigen la presencia de una negación en su oración, y en concreto, estar bajo el dominio sintáctico de dicho operador (ver 1a). Para la polaridad positiva los términos presentan la situación opuesta, teniendo como principal característica no estar en el ámbito de la negación (ver 1b). Por último, la polaridad neutral alude a la dependencia que se establece entre ciertas unidades léxicas y los contextos neutrales (ver 1c).

(1) a. No me ha gustado nada y ni creo que valga la pena.
b. Me ha gustado mucho y ha valido la pena.
c. No sé si me ha gustado y si realmente vale la pena.

Como puede observarse en (1a, 1b, y 1c), el término subrayado corresponde a una locución verbal, sin embargo, de acuerdo al contexto la polaridad es diferente en cada

frase. Este contexto es el que regularmente permite determinar si dicha polaridad es positiva, negativa o neutral.

La bibliografía que ha abordado el estudio de la polaridad se ha centrado principalmente en la polaridad negativa; citando algunos trabajos como Klima, 1964; Fauconnier, 1975; Ladusaw, 1979; Kadmon & Landman, 1993; Zwarts, 1995 y 1998; Israel, 1996; Van der Wouden, 1997; Giannakidou, 1998; Lahiri, 1998; Tovena, 1998; Chierchia, 2004 y 2006; entre otros. El amplio estudio de la polaridad negativa ha permitido conocer mejor cómo funciona este fenómeno gramatical. La polaridad positiva, en cambio, ha pasado prácticamente desapercibida en la bibliografía. Aunque algunos trabajos sobre la polaridad negativa aluden a la variante afirmativa, son pocos los trabajos dedicados a su estudio, entre los que destacan Progovac (1994), Van der Wouden (1997), Nilsen (2004) y Szabolcsi (2004).

La detección de polaridad es una tarea básica de lo que se conoce como análisis de sentimientos, también conocida como minería de opinión[17]. Esta tarea tiene por objetivo identificar automáticamente si una opinión expresada en un documento, una frase, una entidad característica o aspecto es positiva, negativa o neutra. Hasta donde se sabe, no hay trabajos relacionados para la identificación automática de la polaridad de locuciones verbales, pero hay algunas otras obras que se ocupan de estudiar la polaridad de frases u oraciones en general.

Los primeros trabajos en abordar la detección de polaridad son los de Turney (2002) y Pang et al. (2002), los cuales aplican diferentes métodos (a nivel de documento) para detectar la polaridad de críticas de productos y de películas, respectivamente. Es posible clasificar la polaridad de un documento en una escala de varios valores, lo cual fue

[17] El análisis de sentimientos o minería de opinión, es una disciplina que se centra en detectar la información subjetiva de un texto y clasificarla.

intentado por Pang & Lee (2005) y Snyder & Barzilay (2007). En el primer trabajo (Pang & Lee, 2005) se expande la tarea básica de clasificar una crítica de película como positiva o negativa a predecir evaluaciones en una escala de 3 o 4 estrellas; mientras que en el segundo (Snyder & Barzilay, 2007), se realizó un análisis en profundidad de críticas a restaurantes, prediciendo evaluaciones para varios aspectos del restaurante dado, tales como la comida y atmósfera (en una escala de 5 estrellas).

A pesar de que en la mayoría de los métodos de clasificación estadísticos la clase neutral es ignorada bajo la suposición de que los textos neutrales se encuentran cerca de la frontera del clasificador binario, varios investigadores sugieren que, al igual que en todo problema de polaridad, tres categorías deben ser identificadas.

Otros autores como Hu & Liu (2004) han tratado de identificar la polaridad de adjetivos usando algunos recursos lingüísticos como WordNet[18]. Este enfoque es muy rápido y no necesita de la conformación de datos de entrenamiento para obtener una buena exactitud predictiva (alrededor del 69%), pero las principales desventajas es que no se ocupa del sentido de múltiples de las palabras, las cuestiones de contexto, y no funciona para frases de varias palabras (unidades fraseológicas) o palabras que no sean adjetivos.
El auge de los medios sociales como blogs y redes sociales ha impulsado el interés en el análisis de sentimientos. En particular, existen una serie de obras de la literatura asociadas a la identificación automática de las emociones en *Twitter*, debido principalmente a la masificación de esta red social en todo el mundo y la manera fácil en que se puede acceder a los *tweets* a través la API (del inglés: *Application Programming Interface*) proporcionada por el propio *Twitter*. Algunos de estos trabajos se han centrado en la contribución de algunas características particulares, tales como etiquetas de las partes de la oración (PoS), emoticones, entre otros, en esta tarea.

[18] https://wordnet.princeton.edu/

En el trabajo de Agarwal et al. (2011), por ejemplo, se calcula la probabilidad a priori de cada etiqueta PoS, se utilizan hasta 100 características adicionales que incluyen emoticones, un diccionario de palabras positivas y negativas. Los autores en esta investigación reportaron un 60% de precisión. Por otro lado, en el trabajo de Mukherjee, & Bhattacharyya (2012), se propone una estrategia basada en las relaciones discursivas, como conectividad y condicionalidad, con un bajo número de recursos léxicos. Estas relaciones están integradas en los modelos clásicos de representación, como la bolsa de palabras, con el objetivo de mejorar los valores de precisión obtenidos en el proceso de clasificación. Se analiza la influencia de ciertos operadores semánticos como modales y negaciones, en particular, el grado en que afectan estos operadores a la emoción presente dado un párrafo o una frase.

Uno de los principales avances obtenidos en la tarea de análisis de sentimientos ha sido hecho en el marco de la competencia SemEval[19]. En 2013, por ejemplo, varios equipos participaron aplicando diferentes enfoques (Balahur &Turchi, 2013; Balage & Pardo, 2013; Chawla et al., 2013; Clark & Wicentwoski, 2013; Hamdan et al., 2013; Martínez-Cámara et al., 2013; Levallois, 2013; Marchand et al., 2013; Moreira et al., 2013; Reckman et al., 2013; Tiantian et al., 2013). La mayoría de estas obras han contribuido en la tarea de identificación de la polaridad proponiendo métodos, técnicas de representación y clasificación de documentos hacia la clasificación automática de sentimiento en los tweets.

Los algoritmos de análisis de sentimientos reportados en la literatura, en su mayoría, utilizan términos simples para expresar el sentimiento acerca de un producto o servicio. Sin embargo, factores culturales, matices lingüísticos y diferentes contextos hacen que sea extremadamente difícil convertir una cadena de texto escrito en un simple sentimiento a favor o en contra.

[19] Para más información de la competencia, visitar el sitio https://en.wikipedia.org/wiki/SemEval.

El hecho de que los seres humanos a menudo estén en desacuerdo sobre el sentimiento de un texto dado ilustra lo difícil que es automatizar esta tarea en una computadora, y para que se lleve a cabo realmente bien el análisis de sentimientos se requieren algoritmos mucho más eficientes y complejos. Entre más corta sea la cadena de texto a analizar, más difícil se vuelve el poder resolver esta tarea. En particular, la detección de la polaridad de las locuciones verbales supone ser una tarea con una complejidad mayor.

5.1 Detección de la polaridad en las locuciones verbales

En este capítulo se presentan un conjunto de experimentos realizados para la identificación de la polaridad de las locuciones verbales. Se ha seleccionado un enfoque basado en aprendizaje automático, por tanto, se construyeron un conjunto de muestras etiquetadas manualmente. En muchas ocasiones el campo de actuación del aprendizaje automático se traslapa con el de la estadística, ya que las dos disciplinas se basan en el análisis de datos. Sin embargo, el aprendizaje automático se centra más en el estudio de la complejidad computacional de los problemas. El objetivo de llevar a cabo estos experimentos es el de conocer más acerca de lo que estas unidades fraseológicas en el español envuelven. A partir del conocimiento que se ha logrado adquirir en estos experimentos, se ha conseguido contribuir de una manera positiva a las investigaciones realizadas en el marco de la detección de sentimientos.

La metodología desarrollada para lograr detectar la polaridad de las locuciones verbales consiste de dos etapas principales: La primera fase es el **entrenamiento** y se basa en tener un conjunto de muestras que permitan a la computadora *aprender* mediante el uso de algoritmos de aprendizaje automático, con lo cual se generaliza el comportamiento a partir de una información no estructurada suministrada en forma de ejemplos; para posteriormente inducir conocimiento. Es este caso, se han etiquetado manualmente ejemplos de artículos periodísticos que contienen locuciones verbales y que de acuerdo a

éstas el sentimiento incluido es positivo, negativo o neutro. La segunda consiste en las **pruebas** y reside en tener nuevos datos que permitirán inducir la polaridad de la locución verbal, tomando o no en cuenta su contexto.

En la figura 5.1, se puede observar gráficamente el procedimiento que se llevó a cabo para la identificación automática de la polaridad de locuciones verbales (Priego & Pinto, 2018a), el cual consiste en una etapa de entrenamiento y una de pruebas. En ambas etapas se necesitan documentos (corpora), en un caso que permitan alimentar a los algoritmos de aprendizaje, siendo estos documentos los relatos periodísticos y las locuciones verbales etiquetadas. Y en otro los documentos que son el conjunto de prueba, dicho en otras palabras, los documentos en los que el modelo de clasificación predecirá la polaridad para las locuciones verbales.

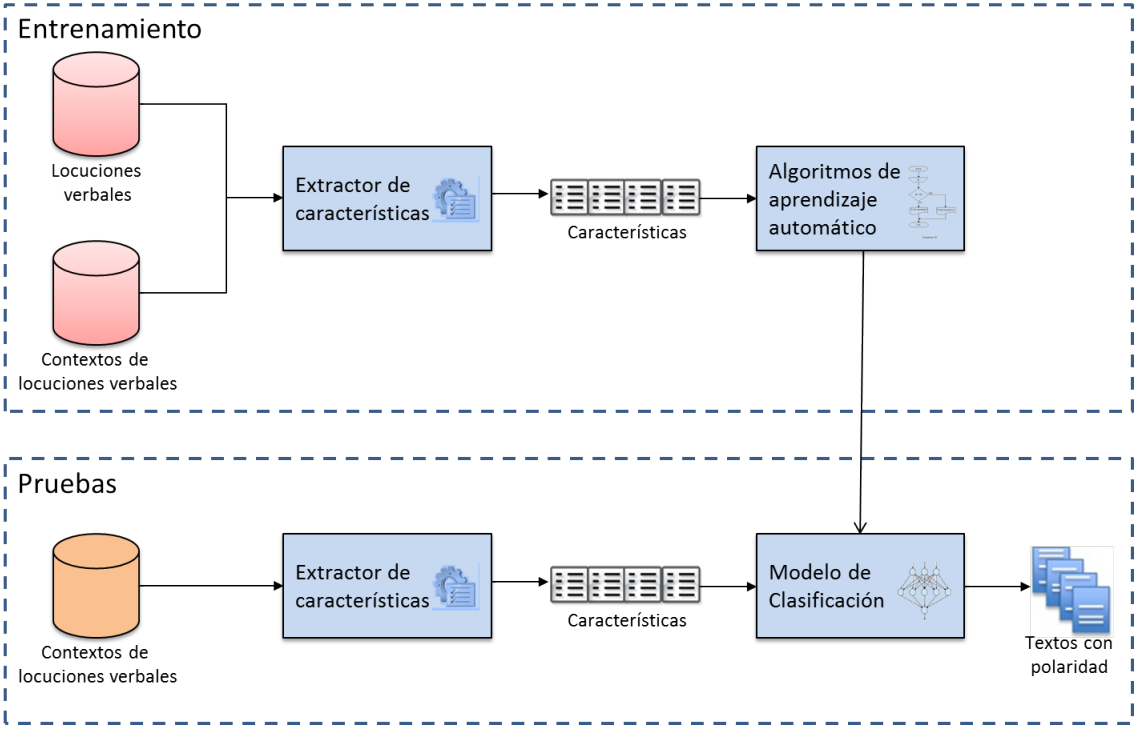

Figura 5.1. Metodología empleada para la identificación de la polaridad de las locuciones verbales.

En las siguientes secciones (Sección 5.2.1 y Sección 5.2.2) se describe tanto el conjunto de datos utilizado para el entrenamiento como los clasificadores utilizados en los experimentos. Finalmente se presentan los resultados alcanzados al llevar a cabo los experimentos.

5.2 Conjunto de datos

El conjunto de datos empleado en los experimentos está conformado por dos corpora: 1) Una lista de locuciones verbales etiquetadas manualmente con una clase particular de polaridad (positiva, negativa, neutra), y 2) Un conjunto de oraciones en las que aparece al menos una unidad fraseológica, en particular, una locución verbal con su contexto correspondiente. Se ha querido aprovechar al máximo los recursos léxicos creados en este libro y por tal motivo en el desarrollo de tareas particulares son utilizados, como se verá a continuación.

5.2.1 Corpus etiquetado de locuciones verbales

La construcción del corpus etiquetado de locuciones verbales con una clase particular se llevó a cabo mediante la extracción de estas unidades fraseológicas del Diccionario de Mexicanismos. En particular, se reunieron 1,219 locuciones verbales diferentes del diccionario, y cada una de ellas fue etiquetada con una de las siguientes clases de polaridad: positiva, negativa, neutra. El proceso de desarrollo de este corpus es completamente manual y consiste, como se ha mencionado anteriormente, en la

recopilación de las locuciones verbales del Diccionario de Mexicanismos, el etiquetado de cada locución verbal con una clase de polaridad y el resultado es un corpus etiquetado manualmente de locuciones verbales (cf. Figura 5.2).

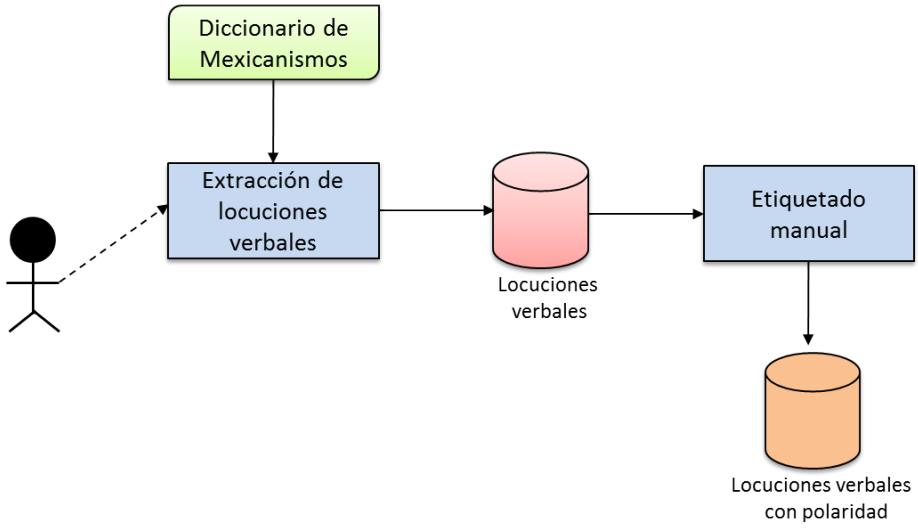

Figura 5.2. Proceso de etiquetado manual del corpus de locuciones verbales.

La selección de la clase de polaridad con que se etiquetó a cada locución verbal fue de acuerdo a la información que la definición de cada una estas unidades fraseológicas, la cual se proporciona en el diccionario de Mexicanismos; en otras palabras, la polaridad es asignada a las locuciones verbales de acuerdo a la definición y/o significado de la unidad fraseológica. En la tabla 5.1, se puede observar una muestra de la polaridad etiquetada a algunas locuciones verbales que son empleadas frecuentemente. Para determinar la

frecuencia de ocurrencia[20] de las locuciones verbales, estas fueron buscadas en el corpus periodístico desarrollado en ese libro. La locución verbal *quedar bien* (*rester bien*) es muy frecuente en México, regularmente puede tener una connotación negativa y sugiere un grado de la falta de sinceridad o algo más superficial y "deshonesto" como, por ejemplo, "para dar una impresión".

El motivo de haber etiquetado las locuciones verbales con la clase de polaridad es debido a que los algoritmos de aprendizaje automático necesitan *enseñarle* a la computadora a que, de cierta manera, aprenda a distinguir entre las diferentes clases (etiquetas) que se tienen en el corpus y que se pretenden identificar en una etapa posterior. Una vez etiquetada manualmente cada locución verbal con la polaridad correspondiente, los resultados para este etiquetado son: 581 locuciones para la clase negativa, 339 locuciones para la clase positiva y 299 para la clase neutra. En la figura 5.3, se presenta una gráfica con estos resultados.

Tabla 5.1. Muestra de locuciones verbales etiquetadas con su correspondiente polaridad.

Locuciones verbales	**Polaridad**
agarrar de puerquito (*attraper un petit porc*)	negativa
agarrarse de la greña (*s'attraper la tignasse*)	negativa
andar roto (*marcher cassé*)	negativa
colgar los tenis (*étendre les baskets*)	negativa
dar cisca (*donner cisca*)	negativa

[20] Recordar que la frecuencia de ocurrencia de las locuciones verbales es una característica que estas unidades fraseológicas poseen.

dar de alta (donner de haute)	neutra
dar el grito (*donner le cri*)	negativa
dejar así (laisser comme ça)	neutra
llegar el agua hasta en cuello (*avoir de l'eau jusqu'au cou*)	negativo
partir su mandarina (*diviser sa mandarine*)	negativa
pasar gato por liebre (prendre un chat pour un lièvre)	negativa
poner en su lugar (*mettre en lieu*)	neutra
ponerse las pilas (*mettre les piles*)	positiva
quedar bien (*rester bien*)	negativa
remar contra corriente (*pagayer contre le courant*)	negativa
ver la cara de gringo (voir le visage d'un étranger)	negativa

Es cierto que la clase más abundante es la negativa, sin embargo, no se puede generalizar que la mayoría de locuciones verbales sean negativas dado que sólo se ha utilizado un diccionario como referencia de extracción de estas unidades fraseológicas.

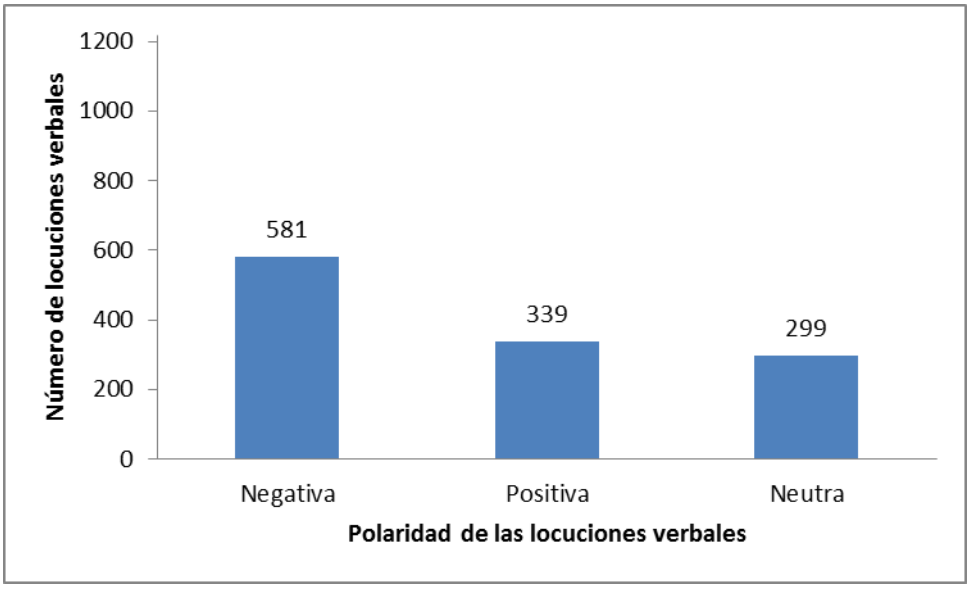

Figura 5.3. Distribución de la polaridad de las locuciones verbales.

5.2.2 Corpus de contextos de locuciones verbales

El segundo corpus por utilizar es un conjunto de textos cortos en los existe al menos una locución verbal. Como ha sido mencionado con anterioridad, el objetivo es el de evaluar si las palabras en el contexto de la locución verbal pueden mejorar o empeorar la predicción de la polaridad de estas unidades fraseológicas.

Para obtener contextos del gran repositorio que es la Web, se realizó la búsqueda de las locuciones verbales en dicho repositorio y se revisó cada uno de estos textos extraídos con el fin de garantizar que realmente se emplea una locución verbal en dicho texto. Este corpus desarrollado contiene contextos que son etiquetados a partir del modelo de clasificación que ha sido entrenado a partir de las unidades fraseológicas y los cuales

sirven para entrenar a los algoritmos de aprendizaje automático. En la figura 5.4 se presenta el proceso de etiquetado de los contextos de las locuciones verbales.

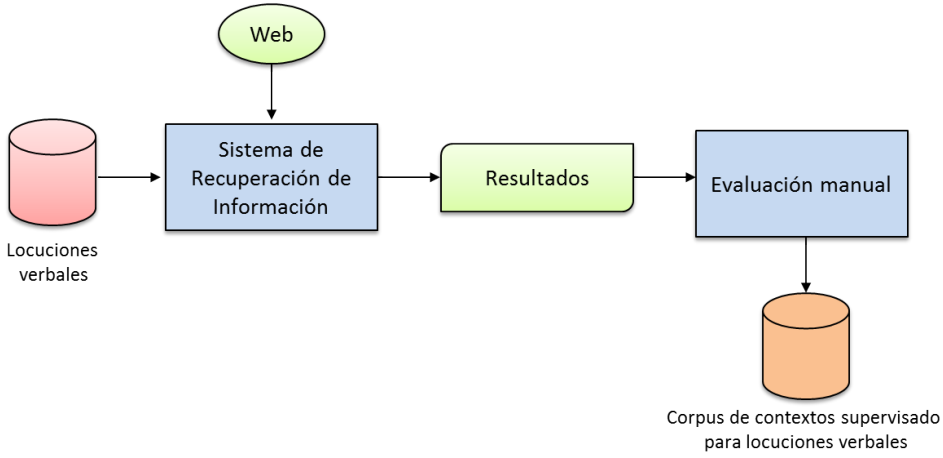

Figura 5.4. Proceso de etiquetado de los contextos de las locuciones verbales.

En estos experimentos, el conjunto de datos está compuesto de 265 textos, y las características del corpus se describen en la tabla 5.2. Es posible expandir el número de contextos utilizados para la tarea de la identificación de la polaridad en las locuciones verbales, ya sea con los contextos del corpus periodístico o con más textos de la Web. Sin embargo, el corpus que hasta el momento se ha trabajado ha sido suficiente para, proporcionar evidencia de la polaridad de las locuciones verbales, y ha servido para tratar de resolver esta tarea de análisis de sentimientos, logrando de esta manera contribuir a la identificación de polaridad en las unidades fraseológicas. En la tabla 5.3 se presentan algunos ejemplos de los contextos de las locuciones verbales, la traducción palabra a palabra en francés de estos textos y su correspondiente polaridad. La locución verbal en esta tabla puede identificarse en los textos debido a que está resaltada en negritas.

Tabla 5.2. Características del corpus en el que aparece al menos una locución verbal.

Característica	**Descripción**
Repositorio	Web
Tipo	Textos libres
Idioma	Español
Número de textos	265
Número de palabras	6,590
Tamaño del vocabulario	2,743

Tabla 5.3. Muestra de contextos con locuciones verbales con su correspondiente polaridad.

Textos que contienen una locución verbal	**Traducción en francés (palabra a palabra)**	**Polaridad**
Se **vendieron como pan caliente** los 8 Ferrar 458 Speciale.	Elles **sont vendus comme pain chaud** les 8 Ferrar 458 Speciale.	Positiva
El Papa **tomo el toro por los cuernos.**	Le pape **a pris le taureau pour les cornes.**	Positiva
Mi cell **valió queso**.!!	Mon portable **a valu fromage**.!!	Negativa
En la escuela ya **me agarraron de su puerquito**. Todos me dicen apodos.	Dans l'école ils **m'ont** déjà **attrapé de son petit porc**. Tous me disent des surnoms.	Negativa
Europa tiene dificultades para **abrir paso** a nuevas industrias de alta tecnología.	L'Europe a les difficultés d'**ouvrir voie** aux nouvelles industries de haute technologie.	Neutra

5.3 Clasificadores empleados

Las técnicas de aprendizaje automático supervisado son capaces de aprender el proceso humano para identificar las unidades fraseológicas verbales con base en las características alimentadas por un corpus manualmente anotado. Con el fin de tener una perspectiva del tipo de clasificador que puede tratar mejor el problema de la validación automática de locuciones verbales, se ha seleccionado un algoritmo de aprendizaje representativo de cada uno de los siguientes cuatro tipos de clasificadores: Bayes, Lazy, Functions y Trees. Los siguientes cuatro algoritmos de aprendizaje fueron los elegidos:

- **Naïve Bayes**: es un clasificador probabilístico basado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales.
- **K-Star:** Este es el clasificador de los k vecinos más cercanos con una función de distancia generalizada.
- **SMO:** Se trata de un algoritmo de optimización secuencial mínima para la clasificación de vectores soporte.
- **J48:** Es un algoritmo usado para generar un árbol de decisión, el árbol generado de C4.5 aprende con la implementación de la revisión 8 de C4.5.

5.4 Resultados experimentales

En parte del libro, se presentan resultados obtenidos a partir de la ejecución de dos experimentos diferentes: en primer lugar, se ha calculado la polaridad de las locuciones verbales por sí mismas, es decir, sin tomar en cuenta las palabras que rodean a la locución verbal. El objetivo de este experimento es determinar el máximo valor de precisión que puede obtenerse cuando ninguna otra característica de clasificación es utilizada pero sí son utilizados los componentes de la locución verbal. Y, en segundo lugar, se ha ejecutado un experimento con un conjunto de textos que contienen, cada una, al menos una locución verbal. En este caso, se está interesado en la evaluar el impacto de las palabras en el contexto de la locución verbal y así determinar si el uso del contexto puede ayudar o afectar la precisión al predecir la polaridad de una locución verbal.

En la tabla 5.4, pueden observarse los resultados obtenidos cuando la clasificación de la polaridad de la locución verbal utiliza solamente los componentes internos de estas unidades fraseológicas. Como se nota, el porcentaje máximo de precisión obtenido es alrededor del 62% con ambos algoritmos de clasificación, el de lazy K-Star y el de máquinas de soporte vectorial.

Tabla 5.4. Resultados obtenidos al clasificar la polaridad de las locuciones verbales fijas (sin contexto).

Clasificador	Tipo	Clasificación correcta (%)	Clasificación incorrecta (%)
J48	Árboles	52.33	47.66
Naïve Bayes	Bayes	61.44	38.55
SMO	Funciones	62.75	37.24
K-Star	Lazy	62.67	37.32

En la tabla 5.5, se pueden observar los resultados obtenidos cuando la clasificación de la polaridad de locución verbal está usando como características, sus componentes internos y aquellos que se encuentran en su contexto. Como se nota, el porcentaje máximo de precisión obtenida es de alrededor de 80% para el caso de las máquinas de soporte vectorial. El valor más bajo obtenido fue con el algoritmo clasificador de lazy K-Star pero, aun así, se ha obtenido un rendimiento mucho mejor que el mejor resultado obtenido cuando solo se ha considerado la locución verbal. Por lo tanto, se puede decir que el contexto parece ayudar a mejorar el rendimiento al clasificar la polaridad de la locución verbal. Se considera que este resultado se deriva del hecho de que el contexto está bien definido (un número pequeño de palabras relacionadas).

Tabla 5.5. Resultados obtenidos al clasificar la polaridad de las locuciones verbales fijas (con contexto).

Clasificador	Tipo	Clasificación correcta (%)	Clasificación incorrecta (%)
J48	Árboles	78.94	21.05

Naïve Bayes	Bayes	78.57	21.42
SMO	Funciones	80.45	19.54
K-Star	Lazy	78.94	21.05

En resumen, se han presentado experimentos realizados hacia la identificación automática de la polaridad de las locuciones verbales. Dos experimentos diferentes fueron realizados: en el primero se ha evaluado el rendimiento de diferentes clasificadores cuando los datos de entrenamiento sólo consideran a la locución verbal. En el segundo experimento, se ha considerado incluir el contexto de la locución verbal con el fin de determinar si las palabras del contexto pueden ayudar o perjudicar al rendimiento del proceso de clasificación. Los experimentos muestran que las palabras incluidas en el contexto mejoran la exactitud de la clasificación de polaridad de la locución verbal. Incluso el peor resultado obtenido cuando se utiliza la locución verbal y su contexto, supera el mejor resultado cuando el contexto no es considerado.

Bibliografía

AGARWAL, A., XIE, B., VOVSHA, I., RAMBOW, O., & PASSONNEAU, R. (2001). Sentiment analysis of twitter data. *Proceedings of the Workshop on Language in Social Media (LSM 2011),* Portland, Oregon, pp. 30–38.

AGIRRE, E., & RIGAU, G. (1996). Word Sense Disambiguation Using Conceptual Density. *Proceedings of the 16th conference on Computational Linguistics-Volume 1*, pp. 16 - 22.

ÁLVAREZ GONZÁLEZ, A., (2006). La variación lingüística y el léxico: conceptos fundamentales y problemas metodológicos. Hermosillo: Universidad de Sonora. ISBN 970-689-287-7

ARNTZ, R., & PICHT, H., (1988). *Introducción a la terminología*. Fundación Germán Sánchez Ruipérez. Barcelona.

BALAGE FILHO, P., & PARDO, T. (2013). NILC USP: A Hybrid System for Sentiment Analysis in Twitter Messages. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA, pp. 568–572.

BALAHUR, A., & TURCHI, M. (2013). Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013,* Hissar, Bulgaria, pp. 49–55. INCOMA Ltd, Shoumen.

BALDWIN, T. (2005). Deep Lexical Acquisition of Verb-Particle Constructions. *Comput. Speech Lang. 19(4)*, pp. 398 - 414 (Oct 2005), http://dx.doi.org/10.1016/j.csl.2005.02.004.

BORRERO BARRERA, M. J. & CALA CARVAJAL, R., (2000). Norma y diccionario. Las variedades diatópicas del español en la enseñanza de ELE. *ASELE actas XI.* Disponible en:

http://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/11/11_0217.pdf

BOSQUE, I. (1980). *Sobre la negación*, Madrid, Cátedra.

BOSQUE, I. (1996). La polaridad modal. Actas del IV Congreso de Hispanistas de Asia, Seúl, Asociación Asiática de Hispanistas, pp. 7-14.

CABRÉ, T., (1999). La terminología: representación y comunicación.

CABRÉ, T., & ESTOPÁ, R., (2005). Unidades de conocimiento especializado, caracterización y tipología. *Cabré, M. T.; Bach, C. (eds.) Coneixement, llenguatge i discurs especialitzat*. Barcelona.

CARNEADO MORÉ, Z., & TRISTÁ PÉREZ, A. M., (1983). *Estudios de la fraseología*. La Habana: Academia de Ciencias de Cuba. Instituto de literatura y lingüística.

CASARES, J. (1992 [1950]). *Introducción a la lexicología moderna.* 3ª edición, C.S.I.C. Madrid.

CERVANTES, I. (2014). El español: Una lengua viva, Informe 2014. Edición Instituto Cervantes.

CHAWLA, K., RAMTEKE, A., & BHATTACHARYYA, P. (2013). Iitb-sentiment-analysts: Participation in Sentiment Analysis in Twitter Semeval 2013 Task. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013),* Atlanta, Georgia, USA, pp. 495–500.

CHIERCHIA, G., (2004). Scalar Implicatures, Polarity Phenomena, and the Syntax/Pragmatics Interface. *A. Belleti (ed.), Structures and Beyond. The Cartography of Syntactic Structures,* Oxford, Oxford University Press, pp. 39-103.

CHIERCHIA, G., (2006). Broaden Your Views: Implicatures of Domain Widening and the Logicality of Language. *Linguistic Inquiry 37,* pp. 535-590.

CLARK, S., & WICENTWOSKI, R. (2013). Swatcs: Combining Simple Classifiers with Estimated Accuracy. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA, pp. 425–429.

CORPAS PASTOR, G. (1996). *Manual de fraseología española.* Madrid: Gredos. 337 páginas.

COSERIU, E. (1966). Structure lexical et enseignement du vocabulaire. *Actes du premier colloque international de linguistique appliquée,* pp. 175-217.

COSERIU, E., (1981a). *Lecciones de lingüística general.* Madrid, Gredos.

COSERIU, E., (1981b). Los conceptos de "dialecto", "nivel" y "estilo de lengua" y el sentido propio de la dialectología. *LEA, III, 1*, pp. 1–32.

DAGAN, I., & CHURCH, K. (1994). Termight: Identifying and Translating technical Terminology. *Proceedings of the 4th Conference on Applied Natural Language Processing*, 34-40. Stuttgard, German.

DANELL, KARL JOHAN, & OLSSON, HUGO (1992). *La grammaire française*. Almqvist & Wiksell/Liber AB. Uppsala et Stockholm.

DAVIS, A.R., & BARRETT, L. (2013). Lexical Semantic Factors in the Acceptability of English Support-Verb-Nominalization Constructions. *ACM Trans. Speech Lang. Process. 10(2),* 5:1- 5:15 (Jun 2013), http://doi.acm.org/10.1145/2483691.2483694.

DICCIONARIO DE MEXICANISMOS (2010). Academia Mexicana de la lengua.

FAUCONNIER, G., (1975). Polarity and the Scale Principle. *Chicago Linguistic Society 13*, pp. 188-199.

GELBUKH, A. & SIDOROV, G. (2010). Procesamiento automático del español, con enfoque en recursos léxicos grandes (Segunda ed.). México, D.F., México: Instituto Politécnico Nacional.

GIANNAKIDOU, A., (1998). Polarity sensitivity as (Non)veridical Dependency. Amsterdam, John Benjamins.

GONZÁLEZ RODRÍGUEZ, R. (2008). La polaridad positiva en español. (Tesis doctoral). Universidad Complutense de Madrid. Madrid.

GROSS, GASTON (1997). Du bon usage de la notion de locution. *La locution entre langue et usages, Martins-Baltar (éd.)*, ENS Éditions, Fontenay Saint-Cloud.

HAENSCH, WOLG, G., ETTINGER, L., & WERNER, S. (1982). La lexicografía. De la lingüística teórica a la lexicografía práctica. Gredos, Madrid.

HAMDAN, H., BÉCHET, F., & BELLOT, P. (2013). Experiments with Dbpedia, Wordnet and Sentiwordnet as Resources for Sentiment Analysis in Micro-Blogging. *Second Joint*

*Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013),* Atlanta, Georgia, USA, pp. 455–459.

HOVY, E., ZHOU, L., & KWON, N. (2007). A Semi-Automatic Evaluation Scheme: Automated Nuggetization for Manual Annotation. *Proceedings of NAACL HLT 2007*. Companion Volume, Rochester, NY: Association for Computational Linguistics, 217–220.

HU, M., LIU, B. (2004). Mining and Summarizing Customer Reviews. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2004, pp. 168–177. ACM, New York.

ISRAEL, M., (1996). Polarity Sensitivity as Lexical Semantics. *Linguistics and Philosophy 19,* pp. 619-666.

JACKENDOFF, R. (1997). The Architecture of the Language Faculty. *Number 28 in Linguistic Inquiry Monographs.* MIT Press, Cambridge, MA.

JACQUEMIN, C., KLAVANS, J., & TZOUKERMANN, E. (1997). Expansion of Multi-word Terms for Indexing and Retrieval Using Morphology and Syntax. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics,* 24–31. Association for Computational Linguistics.

JIANG, J. & CONRATH, D. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of 10th International Conference on Research in Computational Linguistics,* ROCLING'97, pp. 19 – 33.

KADMON, N., & LANDMAN, F., (1993). Any. *Linguistics and Philosophy 1,* pp. 3-44.

KLIMA, E., (1964). Negation in English. J. Fodor y J. Katz (eds.), The Structure of Language, Englewood Cliffs, Prentice Hall, pp. 246-323.

LADUSAW, W., (1979). *Polarity sensitivity as inherent scope relations.* (Tesis doctoral), University of Texas at Austin.

LAHIRI, U., (1998). Focus and Negative Polarity in Hindi. *Natural Language Semantics 6*, pp. 57-125.

LEACOCK, C. & CHODOROW, M. (1998). Combining Local Context and Wordnet Similarity for Word Sense Identification. *Christiane Fellbaum, editor, MIT Press*, pp. 265 – 283.

LESK, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How To Tell A Pine Cone From An Ice Cream Cone. *Proceedings of the 5th Annual International Conference on Systems Documentation, ACM*, pp. 24 – 26.

LEVALLOIS, C. (2013). Umigon: Sentiment Analysis for Tweets Based on Terms Lists and Heuristics. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA, pp. 414–417.

LIN, D. (1998). An Information-Theoretic Definition of Similarity. *Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98,* pp. 296 – 304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

MANNING, C., & SCHÜTZE, H. (1999a). Collocation. *Draft*, 141–177.

MANNING, C., & SCHÜTZE, H. (1999b). *Foundations of statistical natural language processing*. MIT Press, Cambridge, USA.

MARCHAND, M., GINSCA, A., BESANCON, R., & MESNARD, O. (2013). [lvic-limsi]: Using Syntactic Features and Multi-Polarity Words for Sentiment Analysis in twitter. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013),* Atlanta, Georgia, USA, pp. 418–424.

MARTÍNEZ-CÁMARA, E., MONTEJO-RÁEZ, A., MARTÍN-VALDIVIA, M.T., & UREÑA LÓPEZ, L.A. (2013). Sinai: Machine learning and emotion of the crowd for sentiment analysis in microblogs. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013),* Atlanta, Georgia, USA, pp. 402–407.

MARTINS BALTAR, M., (1997). La locution entre langue et usages. *ENS Editions, Fontenay- St. Cloud.*

MCCARTHY, D., KELLER, B., & CARROLL, J., (2003). Detecting a Continuum of Compositionality in Phrasal Verbs. *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18.* pp. 73 - 80. MWE '03, Association for Computational Linguistics, Stroudsburg, PA, US, http://dx.doi.org/10.3115/1119282.1119292

MEJRI, S. (1997). Le figement lexical. Descriptions linguistiques et structuration sémantique. Publications de la faculté des lettres de Manouba, Tunis.

MEJRI, S. (2008a). Constructions à verbes supports, collocations et locutions verbales. MOGORRON HUERTA Pedro, MEJRI Salah, (eds), *Las construcciones verbo-nominales libres y fijas. Aproximación contrastiva y traductológica*, pp. 191-202. Universidad de Alicante.

MEJRI, S. (2008b). La traduction des séquences figées, *META 53(2)* Les Presses de l'Université de Montréal,

MEJRI, S. (2010). Traduction et fixité idiomatique. *Meta 55 (1)*, pp. 31-41.

MEL'CUK, I. (1993). La phraséologie et son rôle dans l'enseignement / apprentissage d'unelangue étrangère. *Étude de Linguistique Appliquée, 92*, pp. 82-113.

MICHIELS, A. & DUFOUR, N. (1998). Defi, a Tool for Automatic Multi-Word Unit Recognition, Meaning Assignment and Translation Selection. *Proceedings of the first international conference on language resources & evaluation*, pp. 1179–1186.

MOIRON, M.V. (2005): Data-driven Identification of Fixed Expressions and their Modifiability. (Tesis doctoral), University of Groningen, Pays-Bas.

MOREIRA, S., FILGUEIRAS, J.A., MARTINS, B., COUTO, F., & SILVA, M.J. (2013). Reaction: A Naïve Machine Learning Approach for Sentiment Classification. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013),* Atlanta, Georgia, USA, pp. 490–494.

MUKHERJEE, S., & BHATTACHARYYA, P. (2012). Sentiment Analysis in Twitter with Lightweight Discourse Analysis. *Proceedings of COLING 2012,* Mumbai, India, pp. 1847–1864. The COLING 2012 Organizing Committee.

NILSEN, O., (2004). Domains for adverbs. *Lingua 114,* pp. 809-847.

NISSIM, M., & ZANINELLO, A. (2013). Modeling the Internal Variability of Multiword Expressions Through a Pattern-Based Method. *ACM Trans. Speech Lang. Process. 10(2)-7*, pp. 7:1 - 7:26, http://doi.acm.org/10.1145/2483691.2483696

PADRÓ, L. (2011). *Analizadores multilingües en FreeLing.* Centro de Investigación TALP / Universitat Politècnica de Catalunya, Lenguajes y Sistemas Informáticos.

PANG, B., & LEE, L. (2005). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. *Proceedings of the Association for Computational Linguistics (ACL).* pp. 115–124.

PANG, B., LEE, L., & VAITHYANATHAN, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of EMNLP*, pp. 79–86.

PIAO, S., RAYSON, P., ARCHER, D., WILSON, A., & MCENERY, T. (2003). Extracting Multiword Expressions with a semantic Tagger. *In Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*. Volume 18, pp. 49–56. Association for Computational Linguistics. PA, USA, http://dx.doi.org/10.3115/1119282.1119289

PINTO, D. (2008). On Clustering and Evaluation of Narrow Domain Short-Text Corpora. (Tesis doctoral). Universidad Politécnica de Valencia. España.

PINTO, D. & PRIEGO SÁNCHEZ, B. (2020). Using automatic constructed thesauri instead of dictionaries in the verbal phraseological units validation task. *J. Intell. Fuz)zy Syst*. 39(2), pp. 2061-2070.

PRIEGO SÁNCHEZ, B. (2015). Análisis de la diversidad morfosintáctica en las locuciones verbales. *Research in Computing Science, 97(1)*, pp. 113-124.

PRIEGO SÁNCHEZ, B. & PINTO, D. (2018a). Idiom Polarity Identification using Contextual Information. *Computación y Sistemas,* 22(1).

PRIEGO SÁNCHEZ, B., PINTO, D. & VIVEK KUMAR SINGH (2018b). Compositionality versus non-compositionality verification based on lexical domain for verbal phraseological units. *J. Intell. Fuzzy Syst*. 34(5), pp. 3059-3067.

PRIEGO SÁNCHEZ, B. & PINTO, D. (2019). An unsupervised method for automatic validation of verbal phraseological units. *J. Intell. Fuzzy Syst*. 36(5), pp. 4579-4585.

PRIEGO SÁNCHEZ, B. (2020). A New Lexical Resource for Evaluating Polarity in Spanish Verbal Phrases. *Computación y Sistemas*, 24(2).

PRIEGO SÁNCHEZ, B., PINTO, D., & MEJRI, S. (2014). Metodología para la identificación de secuencias verbales fijas. *Research in Computing Science 85*, pp. 45-56, http://rcs.cic.ipn.mx/2014_85/Metodologia%20para%20la%20identificacion%20de%20secuencias%20verbales%20fijas.pdf

PRIEGO SÁNCHEZ, B., PINTO, D., & MEJRI, S. (2015). Towards the automatic identification of Spanish verbal phraseological units. *Research in Computing Science 96,* pp. 65-73, http://rcs.cic.ipn.mx/2015_96/Towards%20the%20Automatic%20Identi_cation%20of%20Spanish%20Verbal%20Phraseological%20Units.pdf

PROGOVAC, L., (1994). *Positive and negative polarity: a binding approach.* Cambridge, Cambridge University Press.

PYLE, D., (1999). *Data Preparation for Data Mining.* Morgan Kaufmann Publishers, 466 páginas.

REAL ACADEMIA ESPAÑOLA, (2014). Diccionario de la lengua española (23ª edición), Madrid: Espasa, consultado el 16 de noviembre de 2015.

RECKMAN, H., BAIRD, C., CRAWFORD, J., CROWELL, R., MICCIULLA, L., SETHI, S., & VERESS, F. (2013). Teragram: Rule-Based Detection of Sentiment Phrases Using SAS Sentiment Analysis. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013),* Atlanta, Georgia, USA, pp. 513–519.

RESNIK, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95*, pp. 448 – 453, San Francisco, CA, USA.

REY, A. & CHANTREAU, S. (1993). *Dictionnaire des expressions et locutions.* La deuxième édition.

ROMANELLI, S. (2013). El español y sus variedades. *European Scientific Journal. Edition Vol. 9, No. 32.* ISSN: 1857 – 7881, e-ISSN 1857 – 7431, pp. 45 – 58.

SAGER, J., (1990). *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

SANMARTÍN, J. (1998). Lenguaje y cultura marginal: El argot de la delincuencia. *Valencia, Cuadernos de filología, XXV*.

SANZ, H. (2015). *¿Cuántos idiomas se hablan en el mundo?* Consultado el 24 de octubre de 2015.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

SCHMID, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

SNYDER, B., & BARZILAY, R. (2007). Multiple Aspect Ranking using the Good Grief Algorithm. Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL). pp. 300–307.

SZABOLCSI, A., (2004). Positive Polarity-Negative Polarity. *Natural Language and Linguistic Theory 22,* pp. 409-452.

TIANTIAN, Z., FANGXI, Z., LAN, M. (2013). Ecnucs: A Surface Information Based System Description of Sentiment Analysis in Twitter in the Semeval-2013 (task 2). *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013),* Atlanta, Georgia, USA, pp. 408–413.

TOVENA, L., (1998). *The fine structure of polarity ítems.* Nueva York, Garland.

TRICÁS PRECKLER, M. 1995. *Manual de traducción francés-castellano*. Barcelona, Gedisa.

TURNEY, P.D. (2002). Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002*, pp. 417–424. Association for Computational Linguistics, Stroudsburg.

VAN DE CRUYS, T., & MOIRÓN, B. N. V., (2007). Semantics-Based Multiword Expression Extraction. *Proceedings of the Workshop on a Broader Perspective on Multiword*

Expressions, pp. 25 - 32. MWE '07, Association for Computational Linguistics, Stroudsburg, PA, USA, http://dl.acm.org/citation.cfm?id=1613704.1613708

VAN DER WOUDEN, T., (1997). Negative Contexts. Collocation, polarity and multiple negation. London, Routledge.

VILARIÑO, D., TOVAR, M., BELTRÁN, B., & LEÓN, S. (2014). Un modelo para detectar la similitud semántica entre textos de diferentes longitudes. Research in Computing Science 85, pp. 57 – 64.

WU, Z. & STONE PALMER, M. (1994). Verb Semantics and Lexical Selection. *James Pustejovsky, editor, ACL,* pp. 133 – 138. Morgan Kaufmann Publishers / ACL.

WÜSTER, E., (1979). *Introducción a la teoría general de la terminología y a la lexicografía*. Institut Universitari de Lingüística Aplicada. Barcelona.

ZHANG, Y., KORDONI, V., VILLAVICENCIO, A., & IDIART, M. (2006). Automated Multiword Expression Prediction for Grammar Engineering. *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties,* pp. 36 - 44. MWE '06, Association for Computational Linguistics, Stroudsburg, PA, USA, http://dl.acm.org/citation.cfm?id=1613692.1613700

ZULUAGA, A. (1975). Estudios generativo-transformacionistas de las expresiones idiomáticas. *Thesaurus, XXX, 1*, 1-48.

ZULUAGA, A. (1980). La función del diminutivo en español. *Thesaurus XXV*. pp. 23 – 48.

ZWARTS, F. (1998). Three Types of Polarity. F. Hamm y E. Hinrichs (eds.), Plurality and Quantification, Dordrecht, Kluwer, pp. 177-238.

ZWARTS, F., (1995). Nonveridical Contexts. *Linguistic Analysis 25,* pp. 286-312.

Ángeles Belém Priego Sánchez

Adscripción:
Departamento de Sistemas
División de Ciencias Básicas e Ingeniería
Universidad Autónoma Metropolitana Unidad Azcapotzalco

Doctora en Ciencias del Lenguaje por la Universidad París 13, Sorbonne Paris Cité y Maestra en Ciencias de la Computación por la Benemérita Universidad Autónoma de Puebla (BUAP). Actualmente, es profesora-investigadora y coordinadora divisional de docencia en la Universidad Autónoma Metropolitana unidad Azcapotzalco; es investigadora nacional nivel I formando parte del Sistema Nacional de Investigadoras e Investigadores y cuenta con el Reconocimiento a Profesores de Tiempo Completo con Perfil Deseable. Su principal área de investigación es la interacción entre las computadoras y el lenguaje humano, centrándose en lingüística computacional, procesamiento de lenguaje natural e inteligencia artificial.

David Eduardo Pinto Avendaño
Adscripción:
Dirección de Innovación y Transferencia del Conocimiento
Vicerrectoría de Investigación y Estudios de Posgrado
Benemérita Universidad Autónoma de Puebla

Doctor en Inteligencia Artificial y Reconocimiento de Formas por la Universidad Politécnica de Valencia, España y Maestro en Ciencias de la Computación por la Benemérita Universidad Autónoma de Puebla (BUAP).

Actualmente, es profesor investigador y director de innovación y transferencia de conocimiento en la Benemérita Universidad Autónoma de Puebla; sus actividades de investigación le han permitido tener un reconocimiento nacional e internacional en el ámbito del tratamiento automático de la información, en particular en sus áreas de expertise que están relacionadas con la inteligencia artificial y el tratamiento de información en general.

Agradecimientos:

Los autores de este libro deseamos expresar nuestro más sincero y profundo agradecimiento por el invaluable apoyo y la generosidad que han brindado nuestras instituciones académicas de adscripción durante el proceso de escritura de este libro. Su contribución ha sido fundamental para hacer posible la realización de esta obra, que seguramente enriquecerá el ámbito académico y científico.

La dedicación de nuestros departamentos académicos al facilitar el tiempo y los recursos necesarios para que los autores pudieran concentrarse en la creación de su libro es un testimonio del compromiso institucional con la excelencia académica y el desarrollo de la investigación en nuestras universidades. La disposición para respaldar proyectos académicos y científicos demuestra el verdadero espíritu de colaboración y apoyo que caracteriza a ambas instituciones.

Sabemos que la escritura de un libro requiere no solo esfuerzo intelectual, sino también tiempo y recursos que a menudo pueden ser escasos. El hecho de que tanto la Benemérita Universidad Autónoma de Puebla (BUAP) como la Universidad Autónoma Metropolitana (UAM) hayan brindado a los autores de este libro el respaldo necesario demuestra su reconocimiento a la importancia de su trabajo y su contribución al conocimiento en su campo.
Este acto de generosidad y apoyo no solo beneficia a los autores, sino que también impacta positivamente en toda la comunidad académica al enriquecer el acervo de conocimiento disponible. Su labor en el ámbito educativo y científico deja huella y marca la diferencia en la formación de profesionales comprometidos y en la promoción del avance del saber.

Una vez más, en nombre de la Dra. Belém Priego y del Dr. David Pinto extendemos nuestro más sincero agradecimiento por el valioso respaldo de nuestras instituciones de adscripción. Su gesto no pasará desapercibido y sin duda será un pilar en el camino hacia un mayor entendimiento y progreso en el campo de las ciencias.

Atentamente,

Dra. Ángeles Belém Priego Sánchez
Departamento de Sistemas
División de Ciencias Básicas e Ingeniería
Universidad Autónoma Metropolitana Unidad Azcapotzalco

Dr. David Eduardo Pinto Avendaño
Dirección de Innovación y Transferencia del Conocimiento
Vicerrectoría de Investigación y Estudios de Posgrado
Benemérita Universidad Autónoma de Puebla