Metodología para la identificación de secuencias verbales fijas

Belém Priego Sánchez[1], David Pinto[2], Salah Mejri[1]

[1] Laboratoire Lexiques, Dictionnaires, Informatique, CNRS (UMR 7187)
Université Paris 13, Sorbonne Paris Cité, Francia

[2] Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla, México
belemps@gmail.com, dpinto@cs.buap.mx, smejri@ldi.univ-paris13.fr

Resumen Las locuciones verbales fijas designan un tipo particular de construcciones fijadas. En nuestro enfoque concebimos una secuencia verbal fija como un grupo de palabras en las que al menos una es un verbo que funciona como núcleo del predicado. En este artículo se presenta una aproximación computacional que permite reconocer automáticamente este tipo de estructuras lingüísticas en corpus de diferentes dominios. En el contexto de esta investigación, cuando hablamos de "reconocer" nos referimos al hecho de identificar los limites inferior y superior que enmarcan una secuencia de palabras que tienen un alto grado de probabilidad de ser una expresion verbal fija.

Palabras clave: Secuencia verbal fija, aprendizaje automático, léxico

1. Introducción

La fijación ha tenido un auge remarcable en los últimos años, sobre todo en los estudios referentes al léxico, esto 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 [20], lo que implica todas las dimensiones del lenguaje: lingüísticas, pragmáticas, culturales, entre otras [21]. 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 todos los trabajos que tengan una perspectiva práctica (la traducción, la constitución de diccionarios, la enseñanza de idiomas y el tratamiento automático del lenguaje en tareas como la clasificación automática).

Al hablar de tratamiento automático del lenguaje, nos percatamos que la fijación impide contar con una modelización adaptada a las exigencias de los sistemas informáticos, y de ahí, surge la necesidad de contar con recursos lingüísticos suficientemente vastos y claramente estructurados para ser automatizados.

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 informática a una entidad,

llamada token, delimitada por separadores gráficos (blancos, retorno de línea, etc.). La noción de palabra es mucho más compleja, y cuando decimos complejo nos referimos a la dificultad posible que determina su polilexicalidad. En efecto, mientras que los informáticos nos concentramos sobre la palabra simple, los lingüistas se concentran en las palabras complejas que son también importantes en el tratamiento de las lenguas. Este artículo propone la unión de ambos conocimientos (informático, lingüístico) para el tratamiento automático de este tipo de secuencias, que debe de ser tratado correctamente porque la clasificación correcta de éstas secuencias dichas fijas es útil para numerosas aplicaciones como la traducción, la extracción de información, la clasificación, la constitución de diccionarios, la enseñanza de idiomas, entre otras.

El tratamiento automático de las secuencias fijas en un texto implica dos tareas de suma importancia:

1. La localización automática de las secuencias polilexicales.
2. El 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 secuencias fijas presentan una misma identidad formal tal como lo hacen las secuencias libres. La segunda tarea plantea la dificultad de la relación entre el sentido de la Secuencia Fija (SF) y su sintaxis [4]. Muy a menudo se ha dado preferencia, en cuanto a los estudios sobre la fijación, a las secuencias que rechazan cualquier tipo de variación sintáctica y cuya significación es opaca. Incluso los estudios que han unido la sintaxis con la semántica han privilegiado un aspecto semántico, la opacidad, asociándolo con el bloqueo sintáctico. Todo esto nos indica que se debe contar con una teoría que tome en cuenta, al momento de describir la SF, su entorno sintáctico.

Nuestra aproximación consiste en un recurso informático que localiza automáticamente las secuencias verbales fijas del Español Mexicano en un corpus de diferentes dominios; es decir, toma en cuenta la primer tarea implicada en el tratamiento automático de las secuencias fijas en un texto. Con respecto a la segunda tarea por el momento solo se han etiquetado las expresiones verbales fijas y analizado patrones morfosintácticos que sirven como base para identificar éstas construcciones en un texto.

El objetivo de este proyecto es el de reconocer secuencias verbales fijas, extraerlas y validarlas. La extracción no ha sido una tarea fácil pero hemos llegado a identificar expresiones verbales fijas en un texto escrito en lenguaje natural. La validación es una tarea más compleja, por tanto requiere de más tiempo y de un estudio profundo tanto sintáctico como semántico, por tal ese análisis lo tratamos como trabajo futuro pero partiendo de diferentes hipótesis. En el contexto de esta investigacion, cuando hablamos de “reconocer” nos referimos al hecho de identificar los limites inferior y superior que enmarcan una secuencia de palabras que tienen un alto grado de probabilidad de ser una expresión verbal fija.

2. Expresiones fijas

La comunicación humana depende, en gran medida, del conocimiento enciclopédico del mundo que tienen los hablantes. Todos tenemos en común universales humanos (escenarios, frames), que, sin embargo, vienen matizados por cada cultura, de manera que poseen valores específicos y diferenciados. Así, las sociedades poseen sus propias visiones del mundo, actitudes y conductas sobre diversos temas y circunstancias, distinguiéndose así las distintas comunidades culturales. Una de las disciplinas lingüísticas que mejor recoge éstas especificidades lingüísticas es la fraseología, pues pone de relieve numerosos aspectos socio-culturales [28].

Las Unidades Fraseológicas (UF), tambien llamadas fraseologismos, pertenecen a lo que Coseriu [6] denomina "discurso repetido", caracterizándose, sobre todo, por los tres rasgos siguientes:

1. Su caracter poliléxico, que las distingue de las palabras aisladas, simples o compuestas, de la lengua.
2. Su fijación, que implica su memorización como si constituyeran un todo inseparable, tal y como se almacenan las unidades simples.
3. Su idiomaticidad u opacidad léxica, rasgo que, sin embargo, puede faltar, como ocurre en las llamadas colocaciones, clasificación que aclararemos en seguida.

En efecto, las UF suelen clasificarse atendiendo a dos parámetros :

1. Que se trate de oraciones o de sintagmas.
2. Que sean idiomáticas o no.

Nuestro objeto de estudio concierne a las secuencias fijas, conocidas también con el nombre de de expresiones fijas, locuciones, expresiones idiomáticas, etc. Una secuencia fija es un grupo de palabras, no necesariamente contiguas, poseedoras de una unidad semántica (sentido global), una fijación a la vez morfológica (bloqueo del número), lexical (bloqueo del paradigma conmutacional) y sintáctica (bloqueo de la pasivación, de la relativización para las secuencias verbales) [16]

En el estudio del 2011 de Mejri [23], se hace distinción entre las secuencias fijas y otros dos tipos de conceptos:

1. Las *secuencias totalmente fijas*, no aceptan ninguna modificación. El conjunto es un bloque inmutable y cuyo tratamiento necesita referente simple en un diccionario.
2. Las *colocaciones*: secuencias repetidas que aparecen frecuentemente en conjuntos. Ellas pueden ser propias de un dominio (colocación terminológica según [27]) o típicas de una lengua (como los verbos soporte o los verbos apropiados).

Como se menciona en [26], las locuciones verbales, colocaciones, construcciones con verbo soporte apropiados, entre otros, son denominaciones variadas

para designar un tipo particular de construcciones fijadas. Nosotros estudiamos más precisamente las *Secuencias Verbales Fijas (SVF)*, que en este trabajo se conciben como un grupo de palabras en las que al menos una es un verbo que funciona como núcleo del predicado, es decir, expresiones idiomáticas de significado no composicional. Son sintagmas fijos e idiomáticos cuya interpretación no se obtiene de la suma de sus partes, tomadas por separado [19].

La problemática de las secuencias fijas, y de manera más particular de las SVF, viene del hecho que no son totalmente fijas [1,13,12,15]. De hecho, las SVF permiten ciertas modificaciones de orden sintagmático y/o paradigmático creando también los grados de fijación [12]. Sin embargo, no es aparentemente posible definir a priori las transformaciones realizables de una secuencia. En el trabajo de Villada [24] remarca que "no hay presencia uniforme o ausencia de restricciones sintácticas en todas las expresiones fijas dado que no todas las expresiones fijas exhiben la misma versatilidad sintáctica". En [2] se postula que las secuencias de la misma estructura sintáctica no aceptan las mismas libertades tranformacionales.

Estos problemas desencadenan un reto importante en términos de la identificación y clasificación semántica. Por esta razón consideramos relevante la construcción de recursos, técnicas y herramientas para el tratamiento y análisis de expresiones verbales fijas.

2.1. Trabajo relacionado

Existen diversos trabajos que se centran principalmente en el estudio del grupo estable de dos o más palabras que funcionan como una unidad léxica con significado propio, no derivado de la sumas de sus componentes, es decir, su estudio se basa en las locuciones; una locución es una secuencia fija de palabras con un sentido unitario que en muchos casos no se puede deducir el significado de cada una de las palabras por separado. Si nos centramos en uno de los muchos trabajos que existen, tal como se menciona en [22] , las secuencias fijas se asemejan debido a su funcionamiento sintáctico-semántico, a predicados monolexicales, así, su descripción lingüística se realiza con la ayuda de las mismas herramientas utilizadas para describir las unidades léxicas simples. Las locuciones verbales ilustran perfectamente la saturación total, tal como se ha indicado en [22], y partiendo del hecho que remarca que las locuciones verbales tienen una ruptura paradigmática, concentramos nuestra atención en las locuciones verbales.

En el trabajo de Mogorron [14] se menciona que el significado de las locuciones verbales no puede deducirse de la suma de los significados individuales de cada uno de sus componentes. En el mismo artículo se menciona que las locuciones verbales tienen como principal propiedad a la idiomaticidad, centrándose en la elaboración de un diccionario bajo la forma de una base de datos que incorpor el mayor número de sentencias posibles de este tipo de expresiones. En general, la recolección de locuciones verbales que se ha hecho en este trabajo demuestra el interés de estudio de este campo lingüístico.

La pregunta ahora es: ¿cómo pueden extraerse este tipo de secuencias?, existen diferentes aproximaciones y en este artículo, lo abordamos desde el punto de vista sintáctico, estadístico, híbrido y basado en diccionarios.

La primera aproximación es puramente sintáctica. En el trabajo de Laporte et al. [10] se utilizan patrones sintácticos producidos de sustantivos compuestos y se los propone a un transductor (con la herramienta Unitex). Permitiendo ciertas transformaciones (inserción, coordinación y otras), recuperan así secuencias nominales que corresponden sintácticamente a sustantivos compuestos.

La segunda aproximación es puramente estadística. Estos métodos utilizan medidas estadísticas para determinar la relación entre los elementos de las secuencias. En este tipo de aproximación podemos citar a Caseli et al. [5] que en su trabajo se enfoca al uso estadístico basado en la alineación de la identificación de expresiones multipalabras en corpora. Utilizan varias fuentes de datos: incluido un corpus paralelo (inglés-portugués), corpus basado en dominio (pediátrico) y examinan con un segundo lenguaje que puede proporcionar pistas para resolver este tipo de tareas.

La aproximación más utilizada es la híbrida que involucra la sintaxis y la estadística. Algunos trabajos como [18,7,29] comienzan por un filtro lingüístico (selección de lexemas, patrones sintácticos) para enseguida tomar una decisión basada sobre el cálculo probabilístico (información mutua, logaritmo de máxima verosímilitud, entre otros). Otros a la inversa, generan el primer filtrado por criterios estadísticos para efectuar luego su elección sobre criterios lingüísticos [27]. Otro trabajo basado en este tipo de aproximación es el de Dias [9] en el cual se describe un sistema hibrido que extrae multipabras candidatas, la solución propuesta en este trabajo identifica automáticamente patrones sintácticos relevantes en el corpus, evalua el sistema con el *Brown Corpus* y propone una medida de asociación que llama GenLocalMax [8] la cual permite calcular el grado de fijación de una secuencia de más de dos palabras no necesariamente contiguas. Sin embargo el enfoque depende del tamaño del corpus.

Este tipo de métodos híbridos son más precisos, no obstante, permiten la extracción de datos terminológicos (a menudo nominales) más que la extracción de secuencias fijas; es decir, que pueden encontrarse en cualquier texto sea cual sea el dominio. Las posibles modificaciones integradas son del orden de la expansión de la secuencia.

Otra aproximación permite extraer unidades fraseológicas, basándose en el uso de diccionarios electrónicos, como son los trabajos de [17,11,3], por citar algunos. Existe también la aproximación basada en las gramáticas de adjunción de árboles, un método propuesto por [1], gracias al cual se pueden extraer secuencias fijas a pesar de su discontinuidad (inserción, modificación) y sus cambios sintácticos. Esto implica que la descripción transformacional sea completa.

El término expresiones multipalabra se ha hecho más polular a partir del 2000, probablemente por las diferentes iniciativas [25]. Sin embargo, el estudio de este tipo de estructuras es viejo para el campo de la lingüística. Al tratar de clasificar lingüísticamente este fenómeno involucramos conocimientos de léxicos, sintaxis, semántica, y muchas más áreas lingüísticas pero también observamos

que se necesita unir con la computación, es decir, el estudio de las secuencias fijas se encuentra entre los dos niveles. La consecuencia de ello, es que cualquier aproximación lingüística computacional incluye ambos niveles en los modelos con el fin de lograr más robustez.

3. Creación del recurso: Diccionario de Mexicanismos

El diccionario del cual se han extraído las secuencias verbales fijas con las que se han trabajado en este artículo es el Diccionario de Mexicanismos, el cual podríamos denominar base del conocimiento. Este diccionario es un resultado de una investigación realizada por la Academia Mexicana y tiene tres características esenciales : es sincrónico, contrastivo y descriptivo.

- Sincrónico. Representa lo actual, los elementos léxicos de uso en la segunda mitad del siglo XX, y principios del XXI.
- Contrastivo. Es diferencial ; se preparó el diccionario 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.
- Descriptivo. Indica la realidad del uso ; ya que no establece criterios normativos, no se excluyen prestamos de otras lenguas (que son principalmente del inglés), ni neologismos.

En nuestro caso, se obtuvieron 1,157 expresiones verbales fijas del Diccionario de Mexicanismos, que nosotros llamamos base del conocimiento de expresiones verbales fijas y las cuales forman la base del presente trabajo. Éstas expresiones son tomadas como punto de partida para la identificación de expresiones verbales fijas candidatas en el dominio noticiero. En la siguiente sección se describe la herramienta utilizada para la realización de dicha tarea.

4. Herramienta de idenficación de las expresiones verbales fijas

En este artículo se propone un método que considera el uso de dos recursos importantes: la base del conocimiento de expresiones verbales fijas y el corpus del dominio de noticias. Ambos recursos han sido etiquetados morfosintácticamente y lematizados por TreeTagger.

Con la lematización realizada se toman en cuenta diferentes transformaciones de las secuencias verbales fijas. Una de las primeras transformaciones realizada es la *conjugación* del verbo; es decir, dado que se ha lematizado se toma en cuenta las diferentes flexiones verbales que tiene. Continuando con las transformaciones tenemos la *flexión* que consiste en que la secuencia es modificada cambiando el número de sustantivos y su actualizador asociado; por ejemplo: *vender como pan caliente, vender como panes calientes*. Si bien sabemos que necesitan ser tomadas en cuenta otras transformaciones como la sustitución, inserción, supresión de elementos, negación/afirmación, inversión, entre otras. El trabajo sigue desarrollándose y estan considerándose este tipo de características.

4.1. Descripción del corpus

En esta sección se describe el corpus en español de noticias de los periódicos de la República Mexicana, el cual sirve como recurso para la identificación de expresiones verbales fijas candidatas en los diferentes dominios analizados.

Un corpus es un conjunto de textos recopilados, ya sea de un mismo tema o varios. El propósito de este corpus en particular, es convertirse en un conjunto de datos que proporcione ejemplos de uso (con sus respectivos contextos) de varias expresiones verbales fijas a fin de analizar su uso y frecuencia en diversos dominios. 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 podrian estar previamente desambigüadas 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, y recopilados de una agencia mexicana de noticias.

Si bien, los documentos obtenidos presentan diferentes metadatos, para nuestro caso han resultado útiles los siguientes:

- Título de la noticia
- Dominio (Sociedad, Espectáculos, Política, Futbol, entre otros)
- Fecha de la noticia
- Noticia (el texto mismo de la noticia informativa)

Es cierto que de los documentos de noticias se pueden obtener más datos, sin embargo, dada la tarea presentada en este artículo, con la informacioń mencionada anteriormente es suficiente. El corpus utilizado para esta tarea consta de 378,890 noticias, un total de 4,579,284 oraciones y alrededor de 1,159,571 palabras. Cabe mencionar que solo se utilizó una parte del corpus total de 1,895,983 de noticias para poder balancear la base del conocimiento de expresiones verbales fijas con respecto al corpus.

4.2. Metodología propuesta

En esta sección se describe la metodología para la identificación automática de expresiones verbales fijas. La aproximación está basada en técnicas de aprendizaje automático (machine learning), una rama de la inteligencia artificial que se refiere a la construcción y estudio de los sistemas computacionales que pueden aprender a partir de datos. En este caso, necesitamos un corpus etiquetado por expertos que indique si un conjunto de palabras es o no una expresión verbal fija. El aprendizaje automático por lo general construye un modelo de clasificación que se utiliza para el etiquetado automático de nuevas muestras. Comenzamos este enfoque teniendo en cuenta que existe dependencia entre las palabras en

el contexto de las expresiones verbales fijas. Por lo tanto, para nuestros experimentos preliminares hemos utilizado las técnicas de aprendizaje automático que tienen en cuenta el orden de las palabras conocido como *Conditional Random Fields* que fue originalmente propuesto por la Universidad de Stanford.

Las técnicas de aprendizaje automático son capaces de aprender el proceso humano de la identificación de las expresiones verbales fijas basadas en características alimentados en el clasificador. Además, es capaz de detectar completamente nuevos tipos de expresiones verbales fijas que comparten propiedades con las almacenadas en el corpus de entrenamiento, la cual destaca que tan poderosa puede ser la Computación basada en métodos automáticos de aprendizaje.

Como mencionamos antes, los métodos de aprendizaje automático son alimentados por los datos etiquetados que necesitan ser construidos manualmente. Pero, tener un gran corpus de expresiones fijas marcadas con su respectivo contexto no es una tarea fácil. Por este motivo, se propone un mecanismo automático para la construcción de dicho recurso. Primero necesitamos la construcción de una base del conocimiento de expresiones verbales fijas que luego sea utilizada para identificar construcciones similares en un corpus de textos. Después utilizamos un sistema de recuperación de información para la búsqueda de contextos en el que se produce una estructura similar a una expresión verbal fija.

En resumen, la metodología propuesta para la identificación automática de expresiones verbales fijas candidatas del español mexicano es presentada de la siguiente manera:

1. Construir una base del conocimiento de Expresiones Verbales Fijas para el Español (SVFE).
2. Reunir un conjunto de documentos escritos en español en el que se espera encontrar SVFE.
3. Construir un gran corpus etiquetado de SVFE utilizando técnicas de recuperación de información.
4. Construir un modelo de clasificación para identificar SVFE candidatas usando técnicas de aprendizaje automático.
5. Identificar SVFE candidatas en textos no etiquetados.

En la figura 1 se puede observar la metodología que proponemos para la identificación automática de las expresiones verbales fijas; en la cual se puede observar que existe un proceso de aprendizaje automático y para el proceso de etiquetamiento automático utilizamos un sistema de recuperación de información para la búsqueda de contextos en los que se produce una estructura similar a una expresión verbal fija.

Partiendo de la metodología propuesta, a continuación presentamos una muestra de expresiones verbales fijas lo cual corresponde a la construcción de la base del conocimiento de SVFE.

poner a disposición, salir a relucir, dar a conocer, ganar terreno, llevar a cabo, estar detrás, hacer acto de presencia...

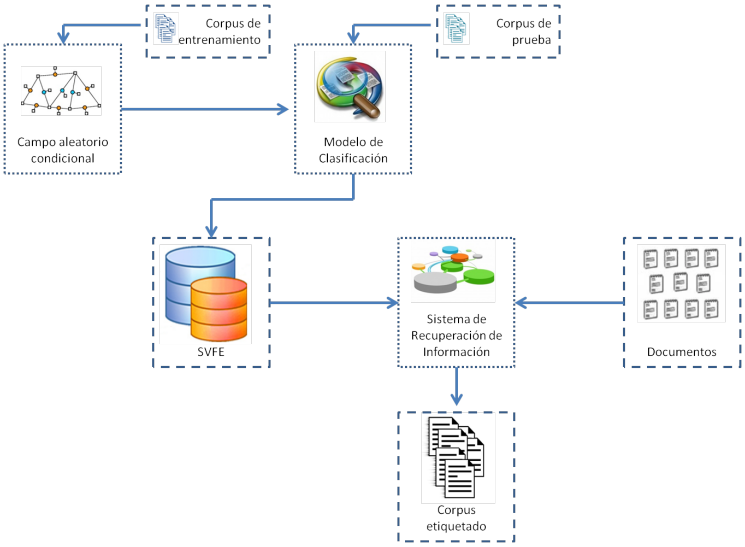

Fig. 1. Metodología propuesta para la identificación de expresiones verbales fijas.

Como se mencionó en la sección 4.1 para la realización de esta tarea hemos recolectado relatos periodísticos de una agencia mexicana de noticias y esa recopilación corresponde al punto 2 de la metodología.

Continuando con la metodología y aplicando los pasos 3, 4 y 5; damos un ejemplo del etiquetado que se realizó y lo que tenemos como resultado final en una noticia que contiene una expresión verbal fija, esto se visualiza en la imagen 2.

poner a disposición *-> Agotados los trámites legales en la dependencia policial el presunto delincuente junto el arma de fuego cartuchos y el vehículo asegurado fueron <EVF>**puestos a disposición**</EVF> del ministerio Público*

Fig. 2. Ejemplo de una noticia que contiene una expresión verbal fija.

Del total de documentos periodísticos que utilizamos para llevar a cabo nuestros experimentos, hemos encontrado que 222,857 noticias contienen expresiones verbales fijas, la etapa siguiente es validar éstas expresiones ya sea sintáctica o semánticamente e incluyendo más características.

5. Conclusiones y Perspectivas

En este trabajo se ha presentado una metodología para la identificación automática de expresiones verbales fijas candidatas para el español de México.

La metodología ha sido probada preliminarmente empleando recursos léxicos del mismo idioma. Como resultado hemos compilado 1,157 expresiones verbales fijas en solamente 378,890 de los 1,895,983 textos del dominio periodistico, de lo cual se obtuvo una muestra de textos anotados con las ocurrencias de las expresiones. Este último recurso debe ser ampliado y perfeccionado con el fin de tener un corpus de entrenamiento que pueda ser utilizado para la construcción de modelos de clasificación que permitirá identificar automáticamente la presencia de expresiones verbales fijas en texto plano y ampliar el experimento con todos los textos disponibles en el corpus.

La utilización de un método computacional que identifique ciertas expresiones de palabras en específico expresiones verbales fijas, es una tarea difícil y compleja, en este trabajo se presentó una aproximación computacional para tratar de resolver dicha tarea. Sin embargo, en el trabajo es necesario tomar en cuenta características semánticas y sintácticas, así como las diferentes transformaciones que tienen las expresiones. Las expresiones verbales fijas pueden ser categorizadas en numerosos tipos no solo por su grado de fijación si no por su literalidad, su duplicidad de sentido, su opacidad o según el dominio en el cual se encuentra, así que este tipo de características también podrían ser tomadas en cuenta para que la aproximación presentada tenga mejores resultados a la hora de identificar las expresiones. Consideramos que sería de gran utilidad tener un corpus de referencia, con el fin de realizar más pruebas y que los resultados puedan ser comparados con medidas estadísticas, así como con otros recursos.

Agradecimientos. Este trabajo ha sido parcialmente apoyado por el Consejo Nacional de Ciencia y Tecnología - CONACYT referencia 218862/314461.

Referencias

1. Abeillé, A., Schabes, Y.: Parsing idioms in lexicalized tags. In: Somers, H.L., Wood, M.M. (eds.) EACL. pp. 1–9. The Association for Computer Linguistics (1989), http://dblp.uni-trier.de/db/conf/eacl/eacl1989.html#AbeilleS89
2. Balibar-Mrabti, A.: Semi-figement et limites de la phrase figée. In: LINX. pp. 34–54 (2005)
3. Bungum, L., Gambäck, B., Lynum, A., Marsi, E.: Improving word translation disambiguation by capturing multiword expressions with dictionaries (2013)
4. Buvet, P.A.: Vers l'elaboration d'un dictionnaire unique des predicats du franais : Deesse. dictionnaire electronique syntactico-semantique. In: Description linguistique pour le traitement automatique du franais. pp. 23–42 (2008)
5. Caseli, H.d.M., Villavicencio, A., Machado, A., Finatto, M.J.: Statistically-driven alignment-based multiword expression identification for technical domains. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications. pp. 1–8. MWE '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1698239.1698241
6. Coseriu, E.: Principios de semántica estructural. In: Gredos. p. 113. Madrid (1977)
7. Daille, B.: Study and implementation of combined techniques for automatic extraction of terminology. In: Klavans et Resnik 1996. pp. 29–36 (1996)

8. Dias, G.: Extraction Automatique dAssociations Lexicales a partir de Corpora. Ph.D. thesis, New University of Lisbon (Portugal) and LIFO University of Orleans (France (2002)
9. Dias, G.: Multiword unit hybrid extraction. In: Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18. pp. 41–48. MWE '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003), http://dx.doi.org/10.3115/1119282.1119288
10. Eric Laporte, T.N., Voyatzi, S.: A french corpus annotated for multiword nouns. In: LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008). pp. 27–30 (2008)
11. Grégoire, N.: Design and implementation of a lexicon of dutch multiword expressions. In: Proceedings of the ACL 2007 Workshop on A Broader Perspective on Multiword Expressions, Prague, Czech Republic. pp. 17–24 (2007)
12. Gross, G.: Les expressions figées en francais noms composés et autres locutions. In: Ophrys. Paris, France (1996)
13. Gross, M.: Une classification des phrases figées du francais. In: Revenue québécoise de linguistique (1982)
14. Huerta, P.M.: Estudio contrastivo lingüístico y semántico de las construcciones verbales fijas diatópicas mexicanas/espaola. In: Las construcciones verbo-nominales libres y fijas. pp. 179–198 (2010)
15. Lamiroy, B.: Le problème central du figement est le semi figement. In: LINX (2005)
16. Lamiroy, B.: Les expressions figées: à la recherche d'une définition. In: Blumental et Mejri 2008. pp. 85–98 (2008)
17. Laporte, E., Voyatzi, S.: An electronic dictionary of french multiword adverbs. In Language Resources and Evaluation Conference. Workshop Towards a Shared Task on Multiword Expressions pp. 31–34 (2008)
18. Manning, C.D., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA (1999)
19. Martinez-Blasco, I.: Verbos soporte y fijacin lexica. In: Las construcciones verbo-nominales libres y fijas. pp. 47–59 (2008)
20. Mejri, S.: Le figement lexical. descriptions linguistiques et structuration semantique. In: Publications de la faculte des lettres de Manouba, Tunis (1997)
21. Mejri, S.: Categories linguistiques et etiquetage de corpus. In: Linformation grammaticale, Peeters, Paris (2007)
22. Mejri, S.: Constructions verbes supports, collocations et locutions verbales. In: La traduction des MEJRI Salah (2008)
23. Mejri, S.: Les dictionnaires électroniques sémantico-syntaxiques. In: Cardoro er al. 2011. pp. 159–188 (2011)
24. Moiron, M.V.: Data-driven identification pf fixed expressions and their modifiability. Ph.D. thesis, University of Groningen, Pays-Bas (2005)
25. Ramisch, C., Villavicencio, A., Kordoni, V.: Introduction to the special issue on multiword expressions: From theory to practice and use. TSLP 10(2), 3 (2013), http://doi.acm.org/10.1145/2483691.2483692
26. Sfar, I.: Polylexicalite et continuite predicative: le cas des locutions verbales figées. In: Las construcciones verbo-nominales libres y fijas. Aproximación contrastiva y traductológica. pp. 213–221 (2008)
27. Smadja, F.: Retrieving collocations from text: Xtract. Comput. Linguist. 19(1), 143–177 (Mar 1993), http://dl.acm.org/citation.cfm?id=972450.972458
28. Soler, N.P., Rodríguez, J.J.B.: Unidades fraseológicas y variación. In: Ogigia. Revista electrónica de estudios hispánicos. pp. 43–52 (2008)

29. Watrin, P.: Collocations et traitement automatique des langues. In: Lexis and Grammar, Bonifacio (2007)