

Análisis e identificación de locuciones verbales mexicanas

Belém Priego Sánchez

Universidad Autónoma Metropolitana unidad Azcapotzalco,
Departamento de Sistemas, CDMX, México
abps@azc.uam.mx

Resumen. Esta investigación presenta los resultados obtenidos tras analizar e identificar locuciones verbales en la variante diatópica general del español mexicano. Centrándose en el estudio morfosintáctico de estas locuciones, las cuales son un tipo específico de unidades fraseológicas, que se componen de un verbo y una o más partes de la oración (variables). El propósito de esta investigación es identificar automáticamente la existencia de locuciones verbales en textos planos (noticias), utilizando diferentes métodos supervisados de aprendizaje automático. Entre los hallazgos más relevantes es posible destacar que el clasificador J48 fue el que mejor desempeño demostró con un 76.74% de exactitud, superando a los métodos de clasificación probabilísticos y vectoriales. Los hallazgos demuestran que la existencia y uso de estas expresiones en textos escritos en español mexicano es frecuente; así también, comprueban que la identificación automática de locuciones verbales es una tarea lingüística-computacional compleja.

Palabras clave: unidades fraseológicas, locuciones verbales, identificación automática, métodos supervisados, aprendizaje automático.

Analysis and Identification of Mexican Verbal Phrases

Abstract. This research presents the results obtained after analyzing and identify verbal locutions in the general diatopic variant of Mexican Spanish. Focusing on the morphosyntactic study of these locutions, which are a specific type of phraseological units, which are composed of a verb and one or more parts of the sentence (variables). The purpose of this research is to automatically identify the existence of verbal phrases in plain texts (news), using different supervised methods of machine learning. Among the most relevant findings, it is possible to highlight that the J48 classifier showed the best performance with 76.74% accuracy, surpassing the probabilistic and vector classification methods. The findings show that the existence and use of these expressions in texts written in Mexican Spanish is frequent; likewise, they verify that the automatic

identification of verbal phrases is a complex linguistic-computational task.

Keywords: unidades fraseológicas, locuciones verbales, identificación automática, métodos supervisados, aprendizaje automático.

1. Introducción

La habilidad de la comunicación se desarrolla mediante información que se adquiere, capta y memoriza a lo largo de la vida. Dicha información es representada en forma de frases, expresiones, dichos, que permiten formular una idea o un concepto, y que son el campo de estudio de la fraseología, el cual es independiente de la lingüística que agrupa diversas disciplinas como la morfología, la sintaxis, la semántica, la pragmática, la sociolingüística y la psicolingüística [21]. La palabra “unidad” hace alusión a la propiedad, de estas expresiones, de actuar como un bloque y “fraseológica” se refiere a que están formadas por varias palabras [14]. éstas son unidades léxicas compuestas, o combinaciones de palabras, que poseen una amplia gama de construcciones lingüísticas como los modismos (tormenta en un vaso de agua, barrer bajo la alfombra), las expresiones fijas (*in vitro*, *por lo general*, *rock and roll*), sustantivos compuestos (*aceite de oliva*, *abrelatas*), y muchas más.

En la lingüística española se emplean diferentes variantes terminológicas que hacen referencia a las unidades fraseológicas tales como: “unidad fraseológica, fraseologismo, modismos, locuciones, expresiones fijas, dichos, frases hechas, expresiones idiomáticas, expresiones, frases, refranes, proverbios, modos de decir, aforismos”, entre otras. En [18] aparecen hasta 64 denominaciones diferentes. En este trabajo, se empleará el término “unidad fraseológica” debido a que goza de una gran aceptación en la filología española y es conocido en la fraseología internacional. Además, de que es el término más adoptado por la mayoría de los investigadores y aceptado por la lingüística soviética, pionera en el estudio y consolidación de la fraseología como disciplina científica [6].

Este trabajo de investigación está delimitado en analizar locuciones verbales, como objeto de estudio, para la variante diatópica general del español mexicano. Algunos ejemplos, que ilustran este tipo de expresiones, son los siguientes:

1. a) *colgar los tenis*
b) *acabarse la candela*
2. a) *zafarse un tornillo*
b) *perder el norte*

El ejemplo (1) significa morir y el segundo, perder la razón. Las expresiones (1a) y (2a) son locuciones verbales para el caso del español de México, mientras que (1b) y (2b) son locuciones verbales para el español de España; note que existe variabilidad léxica entre ambas variantes diatópicas del español. Como se nota en los ejemplos (1) y (2), éstos proporcionan un significado diferente al que los léxicos, separadamente, aportan. Este hecho nos remite a la no composicionalidad como rasgo propio de las unidades fraseológicas. Tal como menciona

González Rey [11] “el sentido idiomático de la expresión no reside, pues, en la no composicionalidad de sus componentes, sino en la no interpretación literal del enunciado”.

El estudio realizado, consiste en el análisis morfosintáctico de las locuciones verbales con el objetivo principal de identificar semi-automáticamente en un texto plano (noticias) si existe presencia o no de este tipo de estructuras lingüísticas, basándonos en la no composicionalidad. Tome en consideración el siguiente ejemplo:

3. *María colgó los tenis anoche.*

El ejemplo (3) tendría dos significados, uno literal y uno composicional, que dependen del contexto y utilización que se esté dando. Entonces la problemática reside en identificar de manera automática y computacional la interpretación de “colgar los tenis” en el contexto dado, es decir, determinar el sentido composicional de la expresión identificada.

2. Unidades fraseológicas

La fraseología, considerada como el reflejo del patrimonio cultural incluso folklórico de una comunidad lingüística, ha adquirido durante las últimas décadas un estado de investigación en la lingüística teórica [16]. Por mencionar algunos de los trabajos centrados en la fraseología española se puede encontrar un trabajo pionero presentado en [5], así como los trabajos relacionados [25,6,10]. Todos estos autores coinciden en que la unidad básica de análisis de la fraseología es la Unidad Fraseológica (de aquí en adelante UF), también llamada fraseologismo. Las Unidades Fraseológicas (denotadas por UFS) pertenecen a lo que en [7] se denomina “discurso repetido”, caracterizándose, sobre todo, por los tres rasgos siguientes:

- Su carácter poliléxico, es decir, que están formadas de al menos dos categorías gramaticales; característica que las distingue de las palabras simples.
- Su fijación, que implica su memorización como si constituyeran un todo inseparable, tal y como se almacenan las unidades simples.
- Su idiomática u opacidad léxica, al ser una agrupación de por lo menos dos palabras, su significado no es necesariamente deducido del significado de cada una de las palabras que la componen.

Para ejemplificar estas características considere la siguiente oración:

4. *Juan dejó con el ojo cuadro a María al regalarle flores.*

En el ejemplo (4), la expresión: “dejó con el ojo cuadrado” está formada de cinco categorías gramaticales, un verbo, una preposición, un artículo, un sustantivo y un adjetivo, ese es su carácter poliléxico. Si se cambiara la palabra “ojo” por otra parte de la cara, para los hablantes nativos del español mexicano,

sería una expresión totalmente diferente y posiblemente no entendida dado que la expresión es concebida y entendida como un todo y, en algunos casos, los elementos que la componen no se pueden modificar; a eso se refiere la fijación. El significado de la frase es “impresionar”, el cual no puede ser semánticamente inferido de los componentes léxicos “dejó”, “con”, “el”, “ojo” y “cuadrado”, justamente esa es la idiomatidad de la expresión.

Los autores [24,19], dedicados al estudio de las Variantes Fraseológicas (que denotaremos como VF), reiteran que las VF vienen prefijadas, es decir, que la variación está determinada y limitada, por lo que no se puede alterar y es conocida por los hablantes. Zuloaga [25] distingue dos tipos de variantes de las unidades fraseológicas (UF): las sinónimas y las que varían completamente su designación con la modificación de algunos de sus componentes. Ejemplos para la primera variación, son el (5) y el (6); mientras que, para el segundo grupo, el (7) y (8).

5. a) *llorar como una Magdalena*
b) *estar en un mar de lágrimas*
6. a) *estirar la pata*
b) *colgar los tenis*
c) *entregar el equipo*
7. a) *tener buena fama*
b) *tener mala fama*
8. a) *bajar el listón*
b) *subir el listón*

Para el caso del ejemplo (5), la unidad fraseológica (5a) es sinónima de la unidad fraseológica (5b), el significado es llorar demasiado. En la unidad fraseológica (6), el significado es morir y éstas (6a, 6b y 6c) son sinónimas. El ejemplo (7a), significa tener buena reputación, mientras que el (7b) lo contrario. En (8a) el significado es disminuir las exigencias y en (8b) aumentarlas.

Con este hecho, se nota que las unidades fraseológicas tienen cambios morfológicos, sintácticos y léxicos. Un cambio morfológico es en el cual uno de los componentes de la unidad fraseológica sufre alguna alteración. Estos cambios pueden ser de género, de número, de cuantificación, de determinación. Por ejemplo: (9) *más pobre que un perro* por (10) *más pobre que los perros*; (11) *romper en pedazos* por (12) *romper en mil pedazos*; (13) *ponerse gallo* por (14) *ponerse gallito*; (15) *andar de pata de perro* por (16) *andar de pata de perrito*; (17) *volver el agua al río* por (18) *volver las aguas a los ríos*. Las variantes sintácticas se dan cuando surgen cambios o alteraciones en el orden de los elementos de la UF, pero que no influyen en la lexicalización de la UF. Por ejemplo: (19) *mover cielo y tierra* por (20) *mover tierra y cielo*; (21) *del derecho y del revés* por (22) *del revés y del derecho*. Las variantes fraseológicas más frecuentes son aquellas que sustituyen un elemento léxico por otro. Por ejemplo: (23) *me importa un pepino* por (24) *me importa un comino* o por (25) *me importa rábano*. Por tanto, la problemática de las unidades fraseológicas viene del hecho que no son totalmente fijas [13,1,8,16]. Además de los cambios que las unidades fraseológicas tienen, existen diferentes clasificaciones de éstas.

Las clasificaciones de las unidades fraseológicas han surgido, la mayoría, como resultado de los problemas prácticos a los que se ha tenido que enfrentar el lexicógrafo al incluir la información fraseológica en la elaboración de diccionarios [6]. Las investigaciones sobre la fraseología española han surgido a comienzos de los años cincuenta con [5] y que dio la primera clasificación de las UFS. Posteriormente, en [7] se estableció la distinción entre la técnica libre del discurso y el discurso repetido. Trabajos basados en las investigaciones alemanas y soviéticas son los presentados en [22,25,15,3,4,23]. En el trabajo [22] se abordó la fraseología de las lenguas romances, mientras que en [25] se mejora y complementa la clasificación del trabajo presentado en [5]. Posteriormente, en el trabajo de [6] se propone una clasificación de las UFS a partir del trabajo [12], y además toma en cuenta características de los trabajos anteriormente mencionados; enseguida a esta clasificación, aparece la clasificación de [21].

2.1. Locuciones verbales

Las locuciones, son definidas en [5] 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 y se han dividido según la función oracional que desempeñen, independientemente de que sean conmutables por palabras simples o por sintagmas. En el trabajo [6] se distinguen los siguientes tipos de locuciones: las nominales, las adjetivas, las adverbiales, las verbales, las prepositivas, las conjuntivas y las clausales.

Una Locución Verbal (denotada, de aquí en adelante, por LV) es una UF que contiene un verbo en el centro de su gramática. Desde el punto de vista sintáctico, expresan procesos y actúan como los predicados, con o sin complementos. Estas UFS, al igual que los verbos, se combinan con el sujeto y los complementos para formar una oración. En el trabajo [20] se define a una locución verbal 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. Definición que se considera a lo largo del desarrollo de esta investigación.

Las LV permiten ciertas modificaciones de orden sintagmático y/o paradigmático creando también los grados de fijación [13]. Por ejemplo, la LV (26) *dormir como un tronco*, significa dormir profundamente, y acepta modificaciones paradigmáticas (dormir como un bendito, dormir como piedra), observando que su grado de fijación no es alto (no totalmente fija). Sin embargo, no es aparentemente posible definir a priori las variaciones realizables de una unidad fraseológica verbal. En el trabajo de [5] se hace énfasis que en las UF no hay presencia uniforme o ausencia de restricciones sintácticas, dado que no todas las unidades fraseológicas exhiben la misma versatilidad sintáctica. En [6] se postula que las secuencias de la misma estructura sintáctica no aceptan las mismas variaciones. Tomando en cuenta estas características, centramos nuestra

atención en la identificación de locuciones verbales, pues es una tarea que implica un reto importante en términos de la identificación y clasificación semántica. En [2] se realiza un estudio gramatical asociado a las locuciones verbales en español. Dicho estudio, “recoge y analiza una muestra de locuciones verbales del español distinguiéndolas claramente como tales de otras unidades fraseológicas que suelen confundirse con ellas”.

Las siguientes secciones presentan los experimentos computacionales realizados al intentar detectar automáticamente si existe una locución verbal en un texto plano, es decir, sin etiquetas o preprocesamiento alguno.

3. Marco metodológico

En cuanto a los avances actuales en la tarea de identificación automática de Locuciones Verbales (denotadas por LVS) para el español, se ha considerado el género periodístico mexicano y las LVS mexicanas (al denotar locuciones verbales mexicanas, se hace referencia a que son entendidas y utilizadas en México; esto no excluye que se empleen en otro país de habla española). Por lo tanto, en esta sección primero se describirán los recursos léxicos construidos para la tarea propuesta. El enfoque para la identificación de LVS 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 supervisado. Por lo tanto, también se incluye en esta sección una parte que describe los clasificadores empleados en los experimentos.

3.1. Recursos léxicos

Los métodos de aprendizaje automático supervisado asumen que se tienen datos anotados manualmente con los cuales se puede adquirir (aprender) conocimiento. En este caso, se necesitan corpora etiquetados 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 artículo, seleccionando relatos periodísticos que contengan y no contengan locuciones verbales. Los relatos periodísticos están escritos en español mexicano y han sido recolectados durante el periodo del 2007 al 2013 de los periódicos mexicanos pertenecientes a la Organización Editorial Mexicana¹. Para este fin, primero se han extraído todas las locuciones verbales de un diccionario, versión papel, denominado “Diccionario de Mexicanismos”. Se ha seleccionado este diccionario porque sus elementos léxicos datan de la mitad del siglo XX y principios del siglo XXI (sincrónico), y compara 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 (contrastivo), indicando la realidad de uso sin excluir neologismos ni palabras de otras lenguas (descriptivo).

¹ Organización Editorial Mexicana, conocida como OEM, es la compañía de medios impresos más grande de México y una de las editoras de periódicos más grande de Latinoamérica.

En particular, se han recolectado 1,406 locuciones verbales de dicho diccionario que han sido almacenadas en una base de datos, considerando que éstas podrían ser posteriormente empleadas para la identificación de su uso regular en el género periodístico mexicano.

Con el uso de técnicas de recuperación de información se han encontrado 3,164 textos informativos, recolectados de la Organización Editorial Mexicana, que contienen al menos una ocurrencia de alguna de las unidades fraseológicas seleccionadas. Este proceso considera la ocurrencia original de la LV y/o alguna de sus variantes morfológicas. Por ejemplo, la LV (27) *darse por vencido* con sus variantes morfológicas (*darse por vencida*, *darse por vencidos*, *darse por vencidas*, *darnos por vencidos*, *darnos por vencidas*). Para este propósito, se han lematizado ambos conjuntos de datos, la LV y el texto de la noticia, con la finalidad de ser capaces de encontrar las variaciones de la LV en el texto objetivo.

Como consecuencia del conteo de la ocurrencia de las LVS mexicanas en el corpus recopilado, se ha construido un corpus etiquetado que además puede utilizarse como corpus de entrenamiento para los métodos de aprendizaje automático supervisado, con el objetivo de identificar si una noticia contiene o no una LV. Los contextos recopilados han sido manualmente anotados por 5 anotadores humanos, con un grado de acuerdo entre anotadores superior al 80 %. El proceso de la anotación manual realizado por estudiantes de computación dedicados a la lingüística computacional ha consistido en que un experto en lingüística explique que es una locución verbal y proporcione detalladamente sus características. Esta explicación fue basada en ejemplos, con el fin de diferenciar estas unidades fraseológicas de otras. Cada anotador humano fue encargado de clasificar manualmente, de acuerdo a las características de las LVS (descritas en la sección 2), cuando un texto plano contenía una LV (Clase 1) y cuando el texto no contenía una LV (Clase 2). El resultado de clasificación de las 3,164 noticias inicialmente recuperadas correspondió a 1,959 textos informativos para la Clase 1 y 1,205 para la Clase 2. Cada clase con 117,715 y 63,600 palabras, respectivamente.

En los experimentos realizados, todos los textos fueron representados mediante un vector de frecuencia de “n-gramas” de palabras, con $n = 1, 2$ y 3 . Frecuencias superiores a dos, para los “n-gramas” fueron consideradas en 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*).

3.2. 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 locuciones verbales, se han seleccionado los siguientes cuatro algoritmos de aprendizaje (cada uno perteneciente a un tipo de clasificador diferente: Bayes, Lazy, Functions y Trees [17]):

1. Naïve Bayes: es un clasificador probabilístico basado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales.
2. K-Star: Este es el clasificador de los k vecinos más cercanos con una función de distancia generalizada.
3. SMO: Se trata de un algoritmo de optimización secuencial mínima para la clasificación de vectores de soporte.
4. J48: Es un algoritmo usado para generar un árbol de decisión.

4. Resultados experimentales y discusión

En esta sección se presenta la exactitud obtenida por cada clasificador cuando se trata de clasificar si existe o no una LV en un texto plano. Se ha usado la exactitud que una medida de evaluación estándar y una descripción completa de ésta puede ser consultada en [9].

En la Tabla 1, se muestran los porcentajes de las instancias que fueron clasificadas correcta e incorrectamente. Esta tabla, básicamente, muestra el promedio ponderado de los diferentes clasificadores empleados. La diferencia entre los valores de exactitud obtenidos por los clasificadores muestra que el rendimiento es bastante aceptable para ser una técnica computacional que se ejecuta de manera automática.

Tabla 1. Porcentaje de instancias clasificadas correcta versus incorrectamente.

Calsificador	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

Es importante remarcar que, en el estado del arte, las técnicas basadas en vectores de soporte obtienen regularmente los mejores resultados, lo que en este caso no resultó así. Es altamente probable que la técnica basada en árboles de decisión (J48) haya tomado ventaja de las estructuras de decisión para proponer un mejor rendimiento sobre el resto de las aproximaciones. En particular, el uso de múltiples tipos de “n-gramas” es un factor relevante en el proceso de clasificación que utiliza J48, puesto que asume la presencia de una gramática al considerar que una cierta palabra va antes o después de otra. Los valores obtenidos (alrededor del 76 %) indican que se han podido extraer algunas relaciones suficientemente generales como para generar hipótesis de comportamientos de los algoritmos de árboles J48 ante distintas situaciones, lo cual ha validado que se obtiene un porcentaje de efectividad similar para cada ejecución promediada mediante la aplicación del mecanismo de validación cruzada.

En cualquier caso, el experimento realizado demuestra la factibilidad de aplicar métodos computacionales para automatizar una tarea que puede incluso ser compleja para un ser humano, puesto que determinar sin ayuda de los seres humanos si una determinada frase es una locución verbal o no, definitivamente no es una tarea sencilla. Los resultados obtenidos son alentadores y se espera mejorarlos con la ayuda del enfoque lingüístico.

5. Conclusión

En los últimos años se ha incrementado el interés por la fraseología, una disciplina de la lingüística aplicada, debido a su importancia en diversas áreas del procesamiento de lenguaje natural como la traducción y el análisis de sentimientos, por citar algunos ejemplos. ésta pone en relieve diferentes aspectos tanto sociales como culturales de estas combinaciones de palabras. Las unidades básicas de análisis en la fraseología son las unidades fraseológicas o frasemas. Ciertas combinaciones de palabras no se pueden separar; pueden existir ciertas variantes, usos y modos de empleo de acuerdo a cada lengua, sin embargo, siguen estando unidas.

El análisis y estudio de las unidades fraseológicas, en específico locuciones verbales (que en este trabajo se han denominado locuciones verbales mexicanas), pone en relieve la complejidad y riqueza de la lengua española. Este artículo se ha enfocado principalmente a su estudio e identificación, expresándolo al presentar un conjunto de experimentos dedicados a su identificación automática en textos planos.

Para poder aplicar el enfoque basado en máquinas de aprendizaje supervisado, se ha recolectado una serie de unidades fraseológicas mexicanas usando el Diccionario de Mexicanismos. Estas unidades fraseológicas han servido como consultas de un sistema de recuperación de información que arrojó como respuesta un conjunto de textos (noticias) que contienen ocurrencias de dichas secuencias de palabras. Sin embargo, tal y como era esperado, no todas las secuencias hacen referencia a una locución verbal, y por tanto, fue necesario etiquetar manualmente todos y cada uno de los textos recuperados de Internet. El corpus supervisado que se obtuvo se utilizó como conjunto de entrada para cada uno de los clasificadores a evaluar.

Los siguientes, fueron los cuatro tipos de algoritmos de aprendizaje supervisado que se analizaron: Naïve Bayes, K-Star, SMO y J48. Se compararon estos clasificadores supervisados con el objetivo de determinar cuál de ellos presenta el mejor rendimiento. Se observa que la revisión 8 del árbol de decisión C4.5 fue la que obtuvo el mejor resultado para la tarea ejecutada en este artículo, obteniendo un valor de 76.74% de exactitud (instancias correctamente clasificadas). Este resultado es ligeramente mayor que el algoritmo SMO el cual obtuvo un valor de 75.32%, mientras que Naïve-Bayes y K-Star obtuvieron, 74.05% y 71.14%, respectivamente. Se está interesado en mejorar el rendimiento obtenido mediante el análisis de otras características que se pueden utilizar en el proceso de clasificación, por ejemplo, el grado de cohesión interna entre los compo-

de la locución verbal, o la posible baja correlación de dichos componentes con respecto a las palabras que se encuentran en el contexto de la locución verbal, sin embargo, este tema es considerado como trabajo futuro.

Referencias

1. Abeillé, A., Schabes, Y.: Parsing idioms in tree adjoining grammars. In: Proceedings of the Fourth Conference of the European Chapter of the Association for Computational Linguistics, 1–9 (1989)
2. Bobes Soler, E.: Información gramatical asociada a las locuciones verbales del español. Tesis de Doctorado. Universidad de Barcelona, España (2016)
3. Carneado Moré, Z.: Algunas consideraciones sobre el caudal fraseológico del español hablado en Cuba. *Estudios de fraseología*, 7–38 (1985)
4. Carneado Moré, Z.: Algunas clasificaciones de la composición fraseológica de la lengua. *Anuario L/L*, 18, 34–45 (1987)
5. Casares, J. (Eds.): Introducción a la lexicología moderna. C.S.I.C. Madrid (1992 [1950])
6. Corpas Pastor, G. (Eds.): Manual de fraseología española. Madrid: Gredos (1996)
7. Coseriu, E.: Structure lexical et enseignement du vocabulaire. Dans: Actes du premier colloque international de linguistique appliquée, 175–217 (1966)
8. Dias, G.: Extraction automatique d’associations lexicales à partir de corpora. Tesis doctoral, New University of Lisbon, Portugal and LIFO University of Orleans, France (2002)
9. Fawcett, T.: Roc graphs: Notes and practical considerations for researchers. Technical report, HP Labs (2004)
10. García-Page Sánchez, M.: Introducción a la fraseología española. Estudio de las locuciones. Rubí (Barcelona), Antrhopos (2008)
11. González Rey, M.: La idiomaticidad en las expresiones fraseológicas. En Wotjak, G. (ed.) *Estudios de fraseología y fraseografía del español actual*, Frankfurt-Madrid: Vervuert - Iberoamericana, pp. 57–73 (1998)
12. Gramley, S., Pätzold, K.M. (Eds.): *A Survey of Modern English*. Londres-Nueva York, Routledge (1992)
13. Gross, M.: Une classification des phrases “figée” du français. *Revue québécoise de linguistique* 11(2), 151–185 (1982)
14. Haddouch, B.: Las unidades fraseológicas en la enseñanza del español (caso del alumnado marroquí del instituto cervantes de tetuán). Tesis doctoral. Universidad Mohammed Vagdal. Facultad de letras y ciencias humanas, Departamento de estudios hispánicos-Rabat (1992)
15. Haensch, Wolg, G., Ettinger, L., Werner, S. (Eds.): *La lexicografía. De la lingüística teórica a la lexicografía práctica*. Gredos, Madrid (1982)
16. Lamiroy, B.: Le problème central du figement est le semi figement. *LINX*, pp. 135–153 (2005)
17. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA (2008)
18. Martins Baltar, M. (Eds.): *La locution entre langue et usages*. Fontenay- St. Cloud (1997)
19. Ortega, G., Gonzáles A.: En torno a la variación de las unidades fraseológicas, *Fraseología contrastiva: con ejemplos del alemán, español, francés e italiano*, pp. 91–109 (2005)

20. Priego Sánchez, B., Pinto, D., Mejri, S.: Metodología para la identificación de secuencias verbales fijas. *Research in Computing Science* 85, pp. 45–56 (2014)
21. Ruiz Gurillo, L. (Eds.): Un enfoque didáctico de la fraseología para extranjeros. Valencia: Quaderns de Valencia: Quaderns de Filología, Universitat de Valencia (2001)
22. Thun, H.: Probleme der phraseologie. Beihefte zur zeitschrift für romanische philologie 168 (1978)
23. Tristán Pérez, A.M.: Fuentes de las unidades fraseológicas. *Estudios de fraseología*, pp. 67–90 (1985)
24. Wotjak, G. (Eds.): *Estudios de fraseología y fraseografía actual*. Frankfurt: Vervuert/Iberoamericana (1985)
25. Zuloaga, A. (Eds.): *Introducción al estudio de las expresiones fijas*. Frankfurt: Verlag Peter Lang (1980)