Extracción semiautomática de locuciones especializadas de Economía en español[1]

Pedro Patiño
Universidad de Antioquia
(Medellín, Colombia)

La constitución de recursos léxicos, reutilizables e interoperables, es una necesidad para el área de las tecnologías lingüísticas. Este trabajo aborda en concreto la extracción semiautomática de las unidades fraseológicas del tipo locución, es decir, que presentan fijación léxica y morfosintáctica, e idiomaticidad en muchos casos. Los ejemplos se tomaron de locuciones que aparecen en textos del área de Economía en español.

Palabras clave: locución, unidad fraseológica, recursos lingüísticos, fraseología computacional, procesamiento del lenguaje natural, terminología.

Semiautomatic Extraction of Specialized Spanish Economics Idioms
Language technology is in need of reusable and interoperable lexical resources. This article deals with the semiautomatic extraction of phraseological units, like idioms, i.e. units which present lexical and morphosyntactic fixedness and idiomaticity in many cases. Examples of idioms are taken from Spanish economics texts.

Keywords: idiom, phraseological unit, language resources, computational phraseology, natural language processing, terminology.

Extraction semi-automatique de locutions spécialisées d'Économie en espagnol
La constitution de ressources lexicales réutilisables et interopérables est une nécessité dans le domaine des technologies linguistiques. En somme ce travail aborde l'extraction semi-automatique des unités phraséologiques de type locution, c'est à dire celles qui présentent une fixation lexicale et morphosyntaxique, ainsi que de nombreux cas d'idiomacité. Nous prendrons des exemples de locutions qui apparaissent dans des textes du domaine de l'économie en espagnol.

1 Este artículo resume un aspecto abordado en el trabajo de fin de Máster Oficial en Lingüística y Aplicaciones Tecnológicas, cursado entre 2006-2008 y financiado con una beca IULA-UPF del Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona (España), defendido el 10 de septiembre de 2008.

Mots clés : locution, unité phraséologique, ressources linguistiques, phraséologie computationnelle, traitement automatique des langues, terminologie.

> Il y a entre autres toute une série de phrases qui sont toutes faites pour la langue. (F. de Saussure)

INTRODUCCIÓN

Las disciplinas que forman parte de las llamadas industrias de la lengua, como la traducción, la terminología y la lexicografía, requieren recursos léxicos y diccionarios donde se almacene información de una lengua que sean lo más exhaustivos posible y que ofrezcan información lingüística detallada que pueda ser usada tanto por parte de humanos como en el procesamiento del lenguaje natural. Este artículo aporta una aproximación para la extracción semiautomática de las unidades fraseológicas del tipo locución, tomando como ejemplo unidades que aparecen en corpus y diccionarios del área de Economía en español.

El tratamiento de las locuciones en los recursos lingüísticos, aplicable también en herramientas para la traducción asistida o automática, es pertinente debido a que son un fenómeno de la lengua que no se puede tratar de la misma forma que las unidades simples, por las razones que se exponen en la sección "Unidades fraseológicas: locuciones" del presente artículo.

El principal interrogante que motiva este trabajo es: ¿cómo identificar y extraer las locuciones teniendo en cuenta las piezas léxicas con las que presentan una mayor coocurrencia o que forman una colocación (e.g. "{*ac tuar*|*destinar*|*distribuir*|*entregar*|*recibir*} a modo de subvención)" de forma tal que puedan utilizarse en traducción asistida, traducción automática, terminografía o lexicografía? En este trabajo se sugiere que la inclusión de estas unidades fraseológicas en los recursos lingüísticos, mediante un tratamiento que las incluya de forma sistemática y permita gestionar los datos (importarlos y exportarlos) entre recursos diferentes sin pérdida de información, sería sumamente útil en áreas como traducción asistida, traducción automática, terminografía, lexicografía y enseñanza del español con fines específicos.

REFERENTES TEÓRICOS

Unidades fraseológicas: locuciones

Ahora se procede a definir la noción de locución y luego se da razón de las diversas denominaciones para ese fenómeno de la lengua que han sido empleadas en la literatura. A renglón seguido, se comentan algunas diferencias entre los diccionarios en formato electrónico y los que están concebidos para el procesamiento computacional.

Los investigadores del área de la fraseología adoptan una u otra concepción respecto a su enfoque de investigación de la fraseología, bien sea desde una perspectiva "ancha" o desde una perspectiva "estrecha." La primera concepción incluye en el objeto de estudio los refranes y proverbios, mientras que en la segunda sólo interesan las unidades pluriverbales que se comportan como las palabras o los sintagmas (Sosinski, 2006, p. 23). Desde la concepción "estrecha" se definen tres tipos de unidades fraseológicas en función de su fijación: sintagmas libres, colocaciones y locuciones.

A continuación se presentan algunas definiciones de ´locución` propuestas por varios investigadores de la fraseología de la lengua española. Casares (1992[1969], p. 167) define las locuciones de este modo: "Combinación estable de dos o más términos, que funciona como elemento oracional y cuyo sentido unitario consabido no se justifica, sin más, como suma del significado normal de sus componentes." Corpas (1997, p. 88) define las locuciones como "unidades fraseológicas del sistema de la lengua con los siguientes rasgos distintivos: fijación interna, unidad de significado y fijación externa pasemática. Estas unidades no constituyen enunciados completos, y, generalmente, funcionan como elementos oracionales." Y Navarro (2002, p. 200) anota:

> las locuciones son combinaciones especializadas en expresar contenidos de gran complejidad a pesar de su brevedad y simplicidad para lo cual las unidades monolexemáticas están, en cierto modo, incapacitadas, razón por la que constituye un recurso léxico de uso frecuente, sobre todo, en los lenguajes sectoriales.

Como lo señala Ruiz (2001, p. 47), la caracterización de las locuciones, en tanto unidades lingüísticas se puede realizar de diversas

maneras, al igual que se estudian las palabras: atendiéndose a su forma, a su función o a su significación. Dicho de otro modo, se pueden describir las unidades fraseológicas desde una perspectiva morfológica, sintáctica o semántica. En la primera perspectiva, se atiende a rasgos como la fijación o la idiomaticidad, o los componentes que las integran, en la segunda, se estudia la función que desempeñan en el discurso, y de ahí proviene la categorización de locuciones verbales, nominales, adjetivales y preposicionales. En tercer lugar, en la perspectiva semántica se mira su significado para describir o agrupar las unidades fraseológicas y hacer distinciones que la morfología o la sintaxis no permiten.

Respecto a las locuciones adverbiales, Pavón (1999, p. 614) comenta que en su mayoría están formadas a partir de una preposición que tiene por término un nombre que a su vez puede estar modificado por determinantes o adjetivos u otros complementos. Además anota que el nombre o adjetivo que forma parte de ellas puede variar, por lo que puede dar lugar a diversas locuciones formadas sobre un mismo esquema sintáctico, aunque muchas de ellas son invariables, fuertemente cohesionadas y el nombre que las integra no admite las expansiones propias de los sintagmas nominales.

En este trabajo nos referimos a las locuciones como aquellas unidades de la lengua que constan de dos o más palabras, que presentan fijación morfosintáctica y que no permiten una lectura composicional.

Denominaciones de estas unidades

La fraseología se ha ido consolidando como una disciplina muy dinámica dentro de la lingüística, prueba de ello son las múltiples denominaciones que han recibido las unidades fraseológicas por parte de distintos investigadores. Montero (2003) menciona una lista de denominaciones para el fenómeno de las unidades fraseológicas que diversos autores han empleado en sus publicaciones en inglés y en español:

> Otras denominaciones comunes en inglés son (Corpas Pastor, 1997; Cowie 1998; Pawley, 2001): multiword units; multiword lexemes; multiword lexical units; multi-word lexical phenomena; phrasemes; conventional expressions, formulae, prefabs, composites, fixed expressions, set expressions, set phrase, word combinations, phrasal lexemes, etc. En el caso del español, se han propuesto casos como (Corpas Pastor, 1997, p. 17): expresión pluriverbal; unidad pluriverbal lexicalizada y habitualizada;

unidad léxica pluriverbal; expresión fija; unidad fraseológica; fraseologismo; frasema, etc.

Adicionalmente, en el caso del español tenemos estas otras denominaciones: expresiones idiomáticas, expresiones fijas, frases hechas y modismos. Ruiz (2001, p. 15) aclara que ya en 1950 Casares había rechazado la denominación "modismo", por considerar que no está claramente delimitada y por tanto no es apta para un trabajo sistemático. Aunque puede servir para referirse a lo más idiomático o característico de una lengua, no sirve para precisar y describir lingüísticamente este tipo de unidades. Igual sucede con "frases hechas", pues Ruiz las considera demasiado amplias y vagas. Respecto a "expresiones fijas", ha sido ampliamente usada en la literatura (Zuluaga, 1980), aunque atiende más al rasgo de invariabilidad de las locuciones que al de opacidad semántica. Sin embargo, el nombre locución es empleado por muchos autores en lengua castellana, como Corpas (1997), Penadés (1999), Ruiz (2001) y es el que se emplea en el presente trabajo, debido a su carácter incluyente de los rasgos de fijación y de idiomaticidad.

Recursos lingüísticos y normalización

La noción de recursos lingüísticos alude a los conjuntos de datos lingüísticos y descripciones en formato electrónico, que se usan para construir, mejorar o evaluar sistemas o algoritmos para el tratamiento del lenguaje natural, según lo definen Godfrey y Zampolli (1997, p. 381). Como ejemplos de estos recursos, se puede mencionar los corpus escritos u orales, bases de datos léxicas, gramáticas, terminologías, ontologías o incluso herramientas de software para la preparación, recolección, gestión o uso de otros recursos. Cunningham y Bontcheva (2006:734) los llaman "la materia prima de la ingeniería del lenguaje" y establecen la diferencia entre recursos lingüísticos y recursos para el procesamiento, como lematizadores, generadores, traductores automáticos, *parsers* o sistemas para reconocimiento de la voz.

Con el fin de suplir la necesidad de producir recursos en formato electrónico que sean reutilizables, ha cobrado importancia la normalización, que resulta imprescindible para la creación de un diccionario que se pueda procesar computacionalmente, y que luego se pueda intercambiar, actualizar o combinar con otros recursos existentes

de manera transparente (Hanks, 2003, p. 54). Si cada proyecto para la constitución de recursos lingüísticos emplea una sintaxis particular para codificar su información, como ha sucedido a lo largo de los años, a la hora de combinar el recurso existente con otros recursos o de exportar o importar datos, se dificulta su reutilización, pues el usuario ha de adaptarse a la nueva estructura de datos si desea reutilizarlos. Francopoulo, Declerck, Monachini y Romary (2006, p. 1) sugieren algunos beneficios derivados de la implementación de normas para recursos lingüísticos, como contar con una base estable para su representación y facilitar la reutilización de software y de datos que no están atados a formatos propietarios, siempre sujetos a cuestiones comerciales.

Según Moreno (2000) desde hace unas dos décadas los investigadores del área de la lexicografía computacional promueven la importancia del diseño de un conjunto de estándares para la constitución de recursos lingüísticos reutilizables e interoperables. Para tal fin, se han realizado proyectos para unificar la codificación de los lexicones computacionales y terminologías mediante la promulgación de normas, con la intención de que estas normas sean implementadas por parte de las organizaciones, grupos de investigación, empresas y profesionales del área en aras de favorecer el intercambio de información sin obstáculos ni pérdidas en la transmisión debido a incompatibilidad por usar tecnologías o protocolos disímiles. Entre estos proyectos podemos mencionar, entre otros, GENELEX[2], MULTEXT[3], EAGLES[4], ISLE[5] y SIMPLE[6].

Adquisición léxica y diccionarios MRD y MTD

Actualmente, en la era de la aldea global, de la información y de las comunicaciones digitales e inmediatas, al igual que los profesionales de otras áreas, los traductores, terminólogos, lexicógrafos e ingenieros del lenguaje deben manejar cantidades enormes de datos y textos provenientes de muy diversos ámbitos del conocimiento. En el caso concreto del traductor de textos especializados, en sus proyectos de traducción precisa encontrar los equivalentes adecuados para traducir

2 GENELEX, http://llc.oxfordjournals.org/cgi/content/abstract/9/1/47
3 MULTEXT, http://acl.ldc.upenn.edu/C/C94/C94-1097.pdf
4 EAGLES, http://www.ilc.cnr.it/EAGLES/browse.html
5 ISLE, http://portal.acm.org/citation.cfm?doid=1118062.1118075
6 SIMPLE, http://www.ub.es/gilcub/SIMPLE/simple.html

de manera óptima los textos encargados por sus clientes y para ello se sirve de diversas fuentes documentales y herramientas auxiliares como diccionarios (en papel o en formato electrónico), bases de datos terminológicas, glosarios, memorias de traducción, textos paralelos y motores de búsqueda en Internet. En el caso de los diccionarios, de forma creciente están disponibles en formato electrónico preferiblemente, por las claras ventajas que ello presenta para la recuperación más rápida y eficiente de la información deseada, como por ejemplo la facilidad de copiar y pegar los equivalentes en un procesador de textos o en un programa para la gestión de las memorias de traducción, en comparación con la forma tradicional de buscar los equivalentes en un diccionario voluminoso de papel.

No obstante, los diccionarios "tradicionales", así estén en formato electrónico para ser consultados en línea, no están codificados para el procesamiento computacional, pues fueron concebidos para ser leídos por humanos y no por máquinas, es decir que en su etapa inicial, estos recursos eran una transcripción fiel de la versión en papel, aunque con algunos valores agregados como la posibilidad de efectuar búsquedas más rápidas y exhaustivas, escuchar la pronunciación de la entrada mediante archivos de audio y acceder a sinónimos u otra información complementaria mediante hipervínculos. Sin embargo, a la hora de intentar alguna tarea de procesamiento, presentan desventajas para usarlos como un repositorio desde el cual extraer la información lingüística de tipo léxico, semántico, fonológico o morfosintáctico (Hanks, 2003, p. 56).

Uno de los aspectos de vital importancia para el procesamiento de lenguaje natural (PLN) es el de la adquisición léxica, puesto que el desempeño de cualquier sistema para procesar texto escrito o hablado depende del grado de "conocimiento" que el sistema tenga sobre los datos lingüísticos que está procesando (Grishman & Calzolari, 1997, p. 392). Según McCarthy (2006, p. 61) la adquisición léxica es la producción o enriquecimiento de un lexicón para emplearlo en un sistema para el procesamiento del lenguaje natural. El lexicón resultante es un recurso como un diccionario o tesauro en formato electrónico pero en un formato legible para la máquina y no para humanos, en el que se incluyen las formas, los significados, las colocaciones y la información estadística asociada, la cual no es de interés para un lector humano, pero que es vital para que el sistema computacional pueda efectuar operaciones como las de

desambiguación. Calzolari (1994, p. 267) afirma que es casi una tautología afirmar que un buen lexicón computacional es un componente esencial de cualquier aplicación lingüística dentro de las llamadas "industrias de la lengua", desde los sistemas para PLN hasta los proyectos lexicográficos. Dicho de otro modo, para que un sistema informático para el procesamiento del léxico desempeñe su labor de manera eficiente y efectiva, debe contar con un repertorio lo más completo posible de información léxica. Sin embargo, a la adquisición léxica se le considera un cuello de botella para el desarrollo de herramientas para el PLN, ya que la confección manual de un lexicón es costosa, requiere un amplio equipo de profesionales cualificados, que no siempre están disponibles, y eso sin mencionar que además es un proceso que toma mucho tiempo y es proclive al error y las inconsistencias, aunque bien puede decirse lo mismo de los diccionarios convencionales en formato papel. (Matsumoto. 2003, p. 396).

Hanks (2003, p. 56) afirma que un diccionario en formato electrónico pero que originalmente se concibió para la lectura por parte de humanos, tras una etapa de preparación y una buena dosis de paciencia e ingenio, puede ser una importante fuente de datos. En esa misma línea, Wilks, Fass, Guo, McDonald, Plate y Slator (1988) introducen la diferencia entre diccionarios en formato electrónico ("machine-readable dictionaries", MRDs) (Amsler, 1982) y diccionarios preparados para el procesamiento[7] ("machine-tractable dictionaries", MTDs), y presentan diversas estrategias para la conversión de MRD a MTD. En igual sentido apunta el trabajo de Litkowski (2006) y McCarthy (2006, p. 61), quien anota que existen diferencias significativas entre los requisitos de un lexicón destinado para un sistema informático y los contenidos de un diccionario o tesauro escrito para humanos. Para que un diccionario esté preparado para el procesamiento computacional, se han de separar los metadatos de la información lingüística y para ello se emplean lenguajes de marcado por etiquetas; inicialmente se usaba el SGML pero en la actualidad se emplea especialmente el lenguaje XML (eXtensible Markup Language) (Litkowski, 2006, p. 753). Estos recursos lingüísticos, ya como lexicones o diccionarios para el PLN propiamente dichos, incluyen, entre otros, la información sintáctica, morfológica, semántica, temática y ejemplos, en un código que la máquina puede procesar.

7 El autor no conoce una traducción en español para las siglas inglesas MRD y MTD, y la traducción "diccionario en formato electrónico" no refleja la diferencia entre ambas.

METODOLOGÍA

Para realizar una primera aproximación a la extracción de locuciones de manera semiautomática con base en textos del ámbito de la economía en español, se proponen estos objetivos específicos:

- Establecer patrones morfosintácticos frecuentes en la formación de locuciones en español.
- Realizar un análisis estadístico de las locuciones extraídas.

Según lo que se comentó en la sección de *unidades fraseológicas: locuciones* de este artículo, abordamos las locuciones desde una perspectiva estrecha, pues excluimos los refranes y proverbios y nos centramos en las unidades pluriverbales que presentan un comportamiento similar al de las palabras o los sintagmas que no permiten una lectura composicional. Los ejemplos se toman de locuciones que aparecen en textos de Economía en español, para lo cual como fuente de datos para la extracción de las locuciones, se usó el diccionario *Routledge Spanish Dictionary of Business, Commerce and Finance* (1998), obra que contiene 30.949 entradas. Luego, las unidades detectadas se buscan en Bwananet, la herramienta de explotación del Corpus Técnico del IULA[8], en la sección de Economía del corpus, para determinar la frecuencia de aparición de las mismas unidades. Igualmente se emplean los patrones morfosintácticos más frecuentes para detectar otras unidades de forma semiautomática usando el mismo corpus.

Constitución de la muestra

Con el fin de establecer algunos patrones morfosintácticos frecuentes en la formación de locuciones, se empezó por hacer una lista de las 773 locuciones incluidas en el apéndice de locuciones de Pavón (1999) que figuran en la *Gramática descriptiva de la lengua española* (Bosque & Demonte, 1999), correspondientes a 39 patrones.[9] Después, se tomaron como punto de partida los patrones morfosintácticos de las locuciones incluidas en dicha gramática, para compararlos, mediante tablas introducidas en una hoja de cálculo, con los lemas del diccionario que

8 Bwananet, Programa d'explotació del corpus tècnic (CT) de l'IULA de la UPF [http://bwananet.iula.upf.edu].

9 Véase el listado en http://atraducir.info/LocucionesGramatDescrEsp.xml

correspondían con los patrones tomados de dicha gramática. De esta forma, se obtuvo una muestra de 715 unidades candidatas a locuciones del diccionario Routledge, de las cuales se descartaron 37, que, si bien correspondían con la estructura de los patrones morfosintácticos, se encontró que eran unidades que no presentaban fijación, habían quedado truncadas y pertenecían a un sintagma nominal o se habían incluido de forma repetida, quedando un listado de 678 unidades.

Extracción

Para realizar el proceso de extraer locuciones de forma semiautomática mediante patrones morfosintácticos, se adaptó el código de un script de Perl[10] tomado del libro "Perl pour les linguistes" (Tanguy & Hathout, 2007), concebido para la extracción de patrones correspondientes al tipo *N de N*. Dicho *script* toma como entrada un texto etiquetado con el programa TreeTagger, que presenta los datos procesados en tres columnas, separados por tabuladores, donde la primera columna corresponde a la forma, la segunda a la categoría gramatical y la tercera al lema, como en este ejemplo:

diferencias	*NC*	*diferencia*
en	*PREP*	*en*
el	*ART*	*el*
ritmo	*NC*	*ritmo*
de	*PREP*	*de*
integración	*NC*	*integración*
de	*PREP*	*de*
los	*ART*	*el*
países	*NC*	*país*

El *script* se ejecuta así: perl Prep-N.pl < entrada.txt > salida.txt, es decir que tras ejecutarse, crea un archivo que lista las unidades extraídas, a razón de una por línea. Este *script* se adaptó para la extracción de candidatos a locuciones usando los patrones morfosintácticos más frecuentes en la formación de locuciones, a saber: Prep + N; Prep + N + Adj; N + Conj + N; Prep + Adj + N. Adicionalmente se adaptó el *script* para extraer unidades constituidas así: N_1 + Prep + N_1, es decir una unidad en la que ambos nombres son iguales.

10 Active State Perl [http://www.activestate.com/].

Empleando estos *scripts*, se extrajo un total de 50.936 ocurrencias de 7.389 unidades que corresponden al patrón Prep + N, de las cuales preliminarmente se descartaron 329 debido a errores en la codificación del corpus como caracteres extraños o por error del lematizador, como el caso de los números etiquetados con categoría gramatical nombre, de tal forma que quedó un total de 7.056. Luego, para el patrón Prep + N + Adj se extrajeron 8.626 ocurrencias de 5.795 unidades y se descartaron 102, quedando en total 5.693. Seguidamente se extrajeron 3.930 ocurrencias de las unidades formadas por el patrón N + Conj + N, que corresponden a 3.117 unidades, de las cuales se descartaron 33, quedando un total de 3.084.

Finalmente, se extrajeron 2.419 ocurrencias de las unidades formadas por el patrón Prep + Adj + N, que corresponden a 1.487 unidades, de las cuales se descartaron 12, de forma tal que al final quedaron 1.475 unidades. Para el patrón N_1 + Prep + N_1, se encontraron 54 ocurrencias, de las cuales se descartaron dos, quedando 39 unidades, como *año por año, casa por casa, caso por caso, mes a mes, paso a paso, pieza por pieza*.

La Figura 1 presenta el código de un script en Perl para la extracción de unidades de tipo Prep + Adj + N a partir de un texto etiquetado con TreeTagger.

```
#! /usr/bin/perl –w
use strict;
use locale;

my ( @forma, @categoria, @lema );

while ( my $line = <STDIN> ){
   chomp $line;

   my @t = split ( /\t/, $line );
   push ( @forma, $t[0] );
     push ( @lema, $t[0] );
   push ( @categoria, $t[1] );
}

for ( my $i = 0 ; $i <= ($#forma - 2) ; $i++ ) {
     if ( ($categoria[$i] eq "PREP") and ($categoria[$i+1] eq "ADJ")
                  and ($categoria[$i+2] eq "NC") ) {
                  print $forma[$i], " ", $forma[$i+1], " ", $forma[$i+2], "\n";
                  }
}
```

Figura 1. Script Prep-Adj-N.pl para la extracción de unidades de tipo Prep + Adj + N a partir de un texto etiquetado con TreeTagger.

ANÁLISIS DE RESULTADOS

En esta sección, se comenta el análisis efectuado con las locuciones y luego se comentan los patrones morfosintácticos más representativos de la muestra, y luego se propone un *script* en el lenguaje de programación Perl[11] para buscar esos patrones en un corpus etiquetado con TreeTagger[12].

Para establecer los patrones morfosintácticos más frecuentes en esta muestra de 678 unidades provenientes del diccionario Routledge, se empleó la herramienta de análisis estadístico R[13], de código abierto, junto con el programa Rattle[14] que se integra con R y facilita el trabajo para los usuarios que no son estadísticos.

La Figura 2 describe la distribución de estos patrones. Se encontró que el patrón Prep + N es el más frecuente, con 194 casos, lo cual representa el 28,5 %, seguido de Prep + N + Adj, con 103 casos, es decir el 15,1 %. En tercer lugar aparece N Conj N, con 110 ocurrencias, lo que representa un porcentaje de 16,2% de la muestra de 678 unidades. De estos patrones, encontramos que la suma de los tres primeros por su frecuencia de aparición representa el 59,8% de las unidades que conforman la muestra. En cuarto lugar figura el patrón Prep + D + N, seguido del patrón Prep + Adj + N, ambos con 56 casos, es decir que cada uno representa el 8,23% del total de la muestra. En sexto lugar aparece el patrón Prep + N + Prep + N con 46 casos, es decir, un 6,76% del total de la muestra.

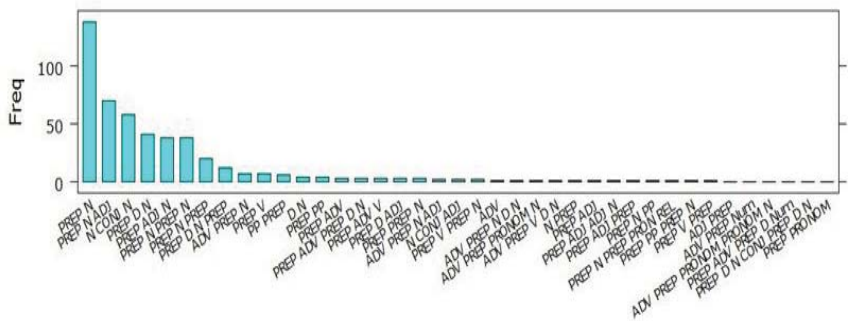

Figura 2. Distribución de los patrones morfosintácticos en la muestra de 678 unidades del diccionario Routledge.

11 http://www.perl.org/

12 TreeTagger, A language independent part-of-speech tagger. [http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger].

13 R, The R Project for Statistical Computing [http://www.r-project.org].

14 Rattle, the R Analytical Tool To Learn Easily. [http://rattle.togaware.com].

A continuación se presentan ejemplos de los 6 patrones con mayor frecuencia en la muestra de 678 unidades extraídas:

Tabla 1. Ejemplos tomados de los 6 patrones con mayor frecuencia en la muestra de 678 unidades extraídas.

Patrón	Ocurrencias	Porcentaje	Ejemplos
Prep + N	194	28,5	*en cuenta, de acción, de conformidad, en marcha, en línea, de moda, de derecho, como consecuencia, a distancia, a cambio*
Prep + N + Adj	103	15,1	*a nivel internacional, a escala mundial, por cuenta propia, en términos absolutos, de ámbito nacional, en términos relativos, a puerta cerrada, en condiciones normales*
N Conj N	110	11,6	*nombre y apellidos, bienes y servicios, investigación y desarrollo, compra y venta, daños y perjuicios, ingresos y gastos, seguridad e higiene, señoras y señores, oferta y demanda, usos y costumbres, ingresos y egresos, fusiones y adquisiciones*
Prep + D + N	56	8,23	*en el mercado, en el futuro, en la actualidad, al margen, en la práctica, hasta la fecha, sobre el valor, a la vista, en el día, en el poder, en el acto*
Prep + Adj + N	56	8,23	*a largo plazo, a corto plazo, en gran parte, de alto nivel, de alta calidad, en última instancia, de alto riesgo, en buen estado, en mayor medida, en cierta medida, de larga duración, de buena fe*
Prep + N + Prep + N	46	6,76	*sin ánimo de lucro, en condiciones de igualdad, en pie de igualdad, con ánimo de lucro, con visión de futuro, por orden de importancia, a precio de mercado, en fase de pruebas, de gobierno a gobierno, en orden de prioridad*

Estos resultados concuerdan básicamente con los trabajos presentados por Ruiz (2001, p. 50) y Martínez (2007). La primera autora presenta los patrones más frecuentes en la formación de locuciones en este orden: Prep + Núcleo; Prep + D + Núcleo; Prep + Núcleo + SAdj

(con D y sin D); Prep + Núcleo + SPrep; SPrep + Conj + Sprep. En el segundo trabajo, Martínez también presenta un estudio de los patrones más frecuentes con base en 336 locuciones adverbiales extraídas manualmente del *DRAE* y el *Diccionario de expresiones y locuciones del español*, específicamente las que inician con la preposición *a*, y también encuentra que el patrón más productivo es Prep + N, que, en el caso de su estudio, corresponde al 29,4% de la muestra, seguido de Prep + Det + N con 15,1% y Prep + N + Adj con 11,3%.

En la Tabla 2 tenemos los siguientes patrones, mucho menos representativos en términos porcentuales:

Tabla 2. Ejemplos de patrones con menor frecuencia.

Patrón	Ocurrencias	Porcentaje	Ejemplos
Prep + N + Prep	27	3,97	*a favor de, como medio de, como proporción de, como recompensa por, con beneficio de, con independencia de, de acuerdo con, de conformidad con*
Prep + D + N + Prep	19	2,79	*a la atención de, a la recepción de, a un precio de, al alcance de, al lado de, ante la ausencia de, bajo la protección de, como una función de, con la excepción de, con la exclusión de, durante el periodo de*
Prep + VInf	14	2,06	*para firmar, por resolver, sin aprobar, sin auditar, sin clasificar, sin confirmar, sin corregir, sin explotar, sin firmar, sin modificar*
PP + Prep	12	1,76	*acompañado por, asociado a, asociado con, atribuido a, basado en, debido a, dividido por, impulsado por, promulgado por, relacionado con, respaldado por, resuelto a*
Adv + Prep + N	11	1,62	*después de fecha, fuera de acta, fuera de almacén, fuera de circulación, fuera de cotización, fuera de funcionamiento, fuera de gálibo, fuera de plantación, fuera de temporada, justo a tiempo*

Luego, la Tabla 3 presenta los patrones menos frecuentes de la muestra, con menos de 10 ocurrencias cada uno, es decir que cada uno de ellos no representa más del 1,5% de la muestra:

Tabla 3. Patrones menos frecuentes de la muestra, con menos de 10 ocurrencias

Patrón	**Ejemplos**
Prep + Prep + N	*de bajo presupuesto, en pro de*
Prep + PP	*de contado, de hecho, de vuelta, por adelantado, por escrito*
Prep + D + [Adj|PP]	*a la inversa, al mejor, en lo sucesivo, según lo convenido, según lo previsto*
Prep + Adv	*de aquí, de nuevo, hacia abajo, hacia arriba*
Det + N	*los porqués, todo comprendido, todo incluido, todo riesgo*
Prep + V + Prep + N	*contra entrega de documentos, en busca de alquiler*

Seguidamente se consultó la frecuencia absoluta y relativa de estas 678 unidades extraídas del diccionario Routledge de economía en el Corpus Técnico del IULA[15] (Bach, Saurí, Vivaldi & Cabré, 1997), en el que se consultó la sección de economía, que consta de 1.091.314 palabras. Para agilizar el proceso, se empleó un módulo del programa Jaguar, para explotación estadística de corpus (Nazar, Vivaldi, & Cabré, 2008), que interroga el corpus para cada una de las unidades e imprime en un archivo de texto la frecuencia absoluta y relativa de cada unidad consultada. En la Tabla 4 se presentan las unidades con una frecuencia mayor a 20 ocurrencias en esta sección del Corpus Técnico del IULA.

Como puede observarse en la Figura 3, las locuciones formadas por Prep + N son las que mayor frecuencia presentan, seguido de los patrones Prep + D + N y Prep + D + N + Prep.

15 [http://bwananet.iula.upf.edu/].

Tabla 4. Unidades con una frecuencia mayor a 20 ocurrencias en la sección de economía del Corpus Técnico del IULA

Unidad	Ocurrencias	Unidad	Ocurrencias
por ejemplo	757	de derecho	68
en cuenta	328	en vez de	64
en el mercado	303	en favor de	56
en el caso de	290	en virtud de	56
de hecho	257	como resultado	53
debido a	256	como medio de	50
acerca de	221	en la actualidad	50
de acuerdo con	192	en vigor	46
en desarrollo	173	basado en	45
a largo plazo	133	de reserva	42
como consecuencia	133	en gran parte	40
a cambio	107	a favor de	37
en lugar de	107	en marcha	33
a cambio de	94	oferta y demanda	30
por medio de	94	con independencia de	28
de nuevo	83	por cuenta propia	28
a la vista	71	en la bolsa	27
bienes y servicios	71	en el curso de	26
en el futuro	71	a la larga	24
en el marco de	71	a priori	24
a corto plazo	69	en especie	21
por unidad	69	sobre el valor	21
		de lujo	20

En la muestra de 678 unidades recolectadas desde el diccionario Routledge, encontramos que las unidades del tipo Prep + N, con 194 casos, entre otras, están conformadas por unidades de tipo adverbial, en 51 casos por unidades introducidas por la preposición *en*, como: *en aduana, en almacén, en alquiler, en alza, en mora*; 49 de las unidades son introducidas por la preposición *sin*, como: *sin cargo, sin fondos, sin descuento, sin dividendo*; 22 de ellas con la preposición *por*, como: *por año, por cabeza, por defecto, por persona*; en 17 casos por unidades introducidas por la preposición *a*, también de tipo adverbial, como: *a cuenta, a granel, a mano*; y finalmente 13 unidades con la preposición *con*, como: *con descuento, con prima, con cupón.*

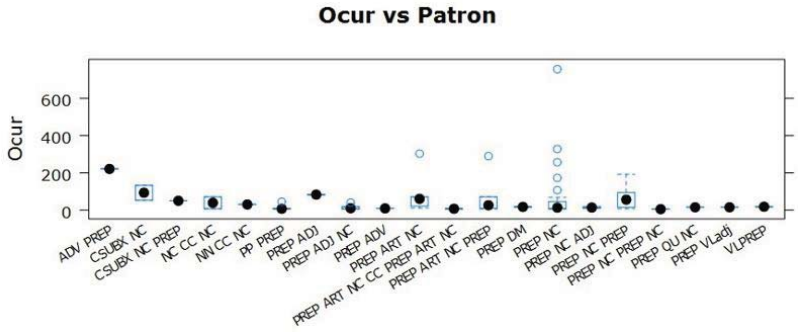

Figura 3. Ocurrencias de cada unidad y patrón morfosintáctico de las unidades consultadas en el Corpus Técnico del IULA.

Respecto al patrón formado por Prep + N + Adj, que también forma locuciones adverbiales, y que en la muestra presenta 103 casos, es introducido en su mayoría por preposiciones como estas: en 39 casos la preposición *de*, como *de categoría superior, de origen nacional, de ámbito nacional;* en 13 casos por la preposición *con*, e.g. *con efecto retroactivo, con valor agregado, con carácter extraoficial;* 7 casos con *a*, como *a puerta cerrada, a efectos contables, a título honorífico;* en 8 casos introduce la preposición *por*, como *por derecho propio, por servicios prestados, por consentimiento mutuo* y *en* 18 casos la preposición *en*, como *en cifras redondas, en números redondos, en plazos mensuales, en términos absolutos, en términos relativos.* En estos últimos ejemplos con la preposición *en*, notamos que, como lo afirma Ruiz (2001, pp. 59-60), las locuciones, como lexemas complejos, presentan las mismas características semánticas que el resto del vocabulario de la lengua, por lo que manifiestan relaciones semánticas, de sinonimia en este caso: *en cifras redondas, en números redondos*.

CONCLUSIONES

La parte empírica de este trabajo corrobora que los patrones morfosintácticos más frecuentes en la formación de locuciones en español corresponden a *Prep + N, Prep + N + Adj, N Conj N, Prep + D + N, Prep + Adj + N y Prep + N + Prep + N.* Hemos propuesto un código en Perl que permite detectar semiautomáticamente una gran cantidad de candidatos a locuciones, pero es necesario realizar un filtrado manual para descartar

las unidades que no presentan fijación o que el software lematizador Treetagger ha lematizado erróneamente.

Las cifras mencionadas en cuanto a las unidades extraídas mediante el uso de los *scripts* de Perl para cada uno de los patrones incluyen todas las coincidencias y por tanto se hace necesaria una validación posterior por parte de un humano para determinar la fijación de las unidades extraídas. Claramente, el uso de estos *scripts* es sólo un primer paso hacia la extracción de estas unidades candidatas a locuciones y para mejorar la extracción es necesario implementar otras medidas para mejorar la precisión, como añadir información estadística para mirar coocurrencias, o de restricciones temáticas o de selección, debido a la dificultad para establecer una distinción formal entre las locuciones y otras unidades producto de la sintaxis de la lengua. En el caso del patrón Prep + N es donde se presentan más falsos positivos en la extracción, por la longitud tan corta de la estructura, razón por la cual se eliminaron manualmente 329 unidades.

Por otro lado, de los patrones empleados en la extracción, el patrón Prep + N + Adj es el que permite detectar los mejores candidatos, de los cuales aquí presentamos algunos ejemplos: *a efectos estadísticos, a escala microscópica, a escala reducida, a nivel económico, a nivel empírico, a nivel global, a nivel regional, a nivel teórico.* Igualmente con el patrón Prep + Adj + N, se extrajeron ejemplos como éstos: *a buen recaudo de buena fe, de común acuerdo, de crucial interés.* En este patrón, merece atención la cantidad de unidades extraídas donde uno de estos tres adjetivos, a saber, {*mayor* | *grandes* | *diferentes}* es uno de sus constituyentes. En el primer caso, encontramos 143 ocurrencias de 76 unidades diferentes, como: *a mayor escala, con mayor celeridad, con mayor facilidad, con mayor frecuencia.* En 31 casos, estas unidades contaban con la preposición *con* entre sus constituyentes y en 24 casos aparecía la preposición *de*. En el caso de *grandes*, aparecían 118 ocurrencias de 81 unidades. Y en el caso de *diferentes*, se encontraron 98 ocurrencias de 73 unidades. Sin embargo, al comparar los resultados extraídos, el caso de *mayor*, era el que arrojaba más unidades que presentaban fijación, mientras que las unidades con *grandes* o *diferentes* eran en su mayoría unidades libres.

De cara a futuros trabajos, se podría implementar un método híbrido, mediante un filtro de tipo lingüístico y estadístico para descartar unidades que no presentan fijación. Otro aspecto posible sería corroborar

el valor terminológico que pueden tener las locuciones, usando para ello un listado de nombres, adjetivos y verbos que se tomen como palabras clave de un área temática y que en un discurso especializado actúan como términos. Tomando como ejemplo el área de Economía, entre las locuciones usadas en la muestra, "*a modo de subvención*" incluye un constituyente de ese tipo que permite delimitar el área temática mediante la inclusión de piezas léxicas con las que presenta alguna relación semántica o de coocurrencia léxica.

REFERENCIAS

Amsler, R. (1982). Computational lexicology: a research program. En AFIPS National Computer Conference (pp. 657-663). Houston: ACM.

Bach, C., Saurí, R., Vivaldi, J., & Cabré, M. (1997). El Corpus de l'IULA. IULA/INF017/97. Universitat Pompeu Fabra, Barcelona.

Bosque, I. & Demonte, V. (Dirs.) (1999). *Gramática descriptiva de la lengua española*. Vol. I, capítulo 9. Madrid: España Calpe.

Calzolari, N. (1994). Issues for Lexicon Building. En A. Zampolli, N. Calzolari & M. Palmer (Eds.), *Linguistica Computazionale Vol. IX-X. Current Issues in Computational Linguistics: In Honour of Don Walker* (pp. 267-281). Pisa: Giardini Editori e Stampatori.

Casares, J. (1992[1969]). *Introducción a la lexicografía moderna*. 3ª edición. Madrid: CSIC.

Corpas, G. (1997). *Manual de fraseología española*. Madrid: Gredos.

Cunningham, H. & Bontcheva, K. (2006). Computational Language Systems: Architectures. En K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (pp. 733-752). Londres: Elsevier Ltd.

Francopoulo, G., Declerck, T., Monachini, M. & Romary, L. (2006). The relevance of standards for research infrastructures. En *International Conference on Language Resources and Evaluation, LREC 2006*, Génova. Consultado el 14 de febrero de 2010 en http://www.tagmatica.fr/publications/LREC2006WS-RI-20AprilBis.pdf.

Godfrey, J. & Zampolli, A. (1997). Language Resources. En G. Battista & A. Zampolli (Eds.), *Survey of the State of the Art in Human Language Technology* (pp. 381-384). Cambridge: Cambridge University Press.

Grishman, R. & Calzolari, N. (1997). Lexicons. En G. Battista & A. Zampolli (Eds.), *Survey of the State of the Art in Human Language Technology* (pp. 392-395). Cambridge: Cambridge University Press.

Hanks, P. (2003). Lexicography. En R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 48-69). Oxford: The Oxford University Press.

Litkowski, K. (2006). Computational Lexicons and Dictionaries. En K. Brown. (Ed.), *Encyclopedia of Language and Linguistics* (pp. 753-759). Londres: Elsevier Ltd.

Martínez, J. (2007). Patrones e índice de frecuencia en algunas locuciones adverbiales. *Forma & Función,* 20, 59-78.

Martínez, J. & Jorgensen A. (2009): *Diccionario de expresiones y locuciones del español*. Madrid: Ediciones de la Torre.

Matsumoto, Y. (2003). Lexical Knowledge Acquisition. En R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 395-413). Oxford: The Oxford University Press.

McCarthy, D. (2006). Lexical Acquisition. En K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd ed.), Vol. 7 (pp. 61-68). Oxford: Elsevier Ltd.

Montero, S. (2003). Estructuración conceptual y formalización terminográfica de frasemas en el subdominio de la oncología. *Estudios de Lingüística Española* (ELiEs), 19. Consultado el 14 de febrero de 2010 en http://elies.rediris.es/elies19.

Moreno, A. (2000). Diseño e implementación de un lexicón computacional para lexicografía y traducción automática. *Estudios de Lingüística Española* (ELiEs), 9. Consultado el 14 de febrero de 2010 en http://elies.rediris.es/elies19/.

Nazar, R., Vivaldi, J. & Cabré, M. (2008). A suite to compile and analyze an LSP corpus. En *Actas de LREC 2008*. Consultado el 14 de febrero de 2010 en http://www.lrec-conf.org/proceedings/lrec2008/pdf/296_paper.pdf.

Navarro, C. (2002). La fraseología en el discurso político y económico de los medios de comunicación. Università di Verona. Centro Virtual Cervantes. Consultado el 14 de febrero de 2010 en http://dialnet.unirioja.es/servlet/oaiart?codigo=2356531.

Pavón, M. V. (1999). Clases de partículas: preposición, conjunción y adverbio. En I. Bosque & V. Demonte (Dirs.), *Gramática descriptiva de la lengua española*. Vol. I, capítulo 9 (pp. 565-655). Madrid: Espasa Calpe.

Penadés, I. (1999). *La enseñanza de las unidades fraseológicas*. Madrid: Arco/Libros.

Routledge Spanish Dictionary of Business, Commerce and Finance (1998). Versión 1.1. Londres: Routledge Software.

Ruiz, L. (2001). *Las locuciones en español actual*. Madrid: Arco/Libros.

Saussure, F. (1967 [1916]). *Cours de Linguistique générale*. Edición crítica de R. Engler. Wiesbaden: Otto Harrassowitz.

Sosinski, M. (2006). Fraseología comparada del polaco y del español: su tratamiento en los diccionarios bilingües. Universidad de Granada. Tesis doctoral inédita.

Tanguy, L. & Hathout, N. (2007). *Perl pour les linguistes*. París: Lavoisier.

Wilks Y., Fass, D., Guo, C., McDonald, J., Plate, T. & Slator, B. (1988). Machine tractable dictionaries as tools and resources for natural language processing. En *Proceedings of the 12th conference on Computational linguistics* (pp. 750-755). Morristown, Nueva Jersey: Association for Computational Linguistics. Consultado el 14 de febrero de 2010 en http://www.aclweb.org/anthology/C/C88/C88-2153.pdf.

Zuluaga, A. (1980). *Introducción al estudio de las expresiones fijas*. Bern, Frankfurt: Peter Lang Verlag.

SOBRE EL AUTOR

Pedro Patiño García
Es Magíster en Lingüística y Aplicaciones Tecnológicas por la Universitat Pompeu Fabra y Traductor Inglés-Francés-Español por la Universidad de Antioquia, donde es docente ocasional de tiempo completo e integrante del Grupo de Investigación Traducción y Nuevas Tecnologías. Sus áreas de interés son el procesamiento del lenguaje natural, la fraseología, terminología y lexicografía computacionales, la traducción y las nuevas tecnologías, y la traducción audiovisual.

Correo electrónico: ppatino@idiomas.udea.edu.co

Fecha de recepción: 14-02-2010
Fecha de aceptación: 30-04-2010