PHRASENET
DETECCIÓN Y EXTRACCIÓN DE UNIDADES FRASEOLÓGICAS A PARTIR DE UN CORPUS TEXTUAL

Jose Luiz de Lucca
Universidad Politécnica de Valencia
jldlme@yahoo.com

Abstract
This article describes a new method to identify and extract phraseological units from textual corpora. There are different methods of classification of phraseological units, but we have to highlight the ones proposed by Corpas Pastor (1996). This author, starting from a wide conception of phraseology, classifies Spanish's phraseologisms in three different categories: collocations, locutions and phraseological enunciated units (fixed forms and routine formulas) from which we have to choose locutions and phraseological enunciated for our research.
The extraction is done sentence-by-sentence and the proposed architecture is based on statistics and relational algebra. The main characteristic of this architecture is the scarce use of linguistic resources, which are replaced by algorithms of searches and statistical methods. Four experiments are presented in this paper to show the extraction of phraseological units from textual corpora. This application demonstrates the useful architecture of the software design, comparing the results of this system with manual extraction. The advantages of this system are the huge database that can be analyzed
Keywords: information extraction, phraseological units, phraseology, textual corpora

I. INTRODUCCIÓN

Los sistemas de extracción de información (EI por su sigla en inglés) tienen como reto buscar y enlazar la información relevante muy concretamente, en colecciones o flujo de documentos, ignorando la extraña e irrelevante. Su reto es extraer información de datos no estructurados, transformando en informaciones estructuradas (base de datos).

IE es una tecnología de NLP cuya función es procesar textos no estructurados, localizar pedazos específicos de información, o hechos, en el texto, para llenar con éstos una base de datos. Su meta es extraer de los documentos los hechos salientes sobre los tipos pre-especificados de eventos, las entidades, o relaciones. Estos hechos se introducen entonces automáticamente en una base de datos, que puede usarse entonces en el proceso más adelante.

La información obtenida es presentada en un formato que pueda ser tratado posteriormente de forma automática. Estos sistemas son construidos para realizar una tarea específica, en función del tipo de información a extraer en cada caso. Un ejemplo podría ser un sistema de EI orientado a la extracción de las unidades fraseológicas que aparecen en textos literarios o científicos. Este sistema evidentemente precisaría ser alimentado por una base de datos relacional donde estarían almacenadas las unidades fraseológicas. Así, este sistema operaria de forma que automáticamente buscaría en el texto todas las unidades fraseológicas existentes en la base de datos, extrayendo la información correspondiente y incorporándola a otra base de datos o tesauro creado para tal efecto que haría de output.

Para extracción de informaciones no estructuradas de las páginas Web, es preciso construir una herramienta capaz de detectar y extraer expresiones, estructurando estas informaciones en una base de datos.

De esta forma decidimos hacer una investigación que ayude a colmar las lagunas existentes en el campo de la fraseología. La construcción de una herramienta informática destinada a la identificación de las Unidades Fraseológicas en el texto, cubrirá unos huecos que desde hace mucho complican la vida de estudiantes y traductores. Esta herramienta tiene nombre: PhraseNET – que empieza por incluir las unidades fraseológicas en una base de datos y termina con la detección y extracción de las unidades fraseológicas en un corpus textual.

Hay muchas investigaciones desde el procesamiento natural del lenguaje (NLP) para extracción de lss unidades fraseológicas: (Gaël Dias, Sylvie Guilloré, Jean C Bassano, José, 2000), Gregor Thurmair (2003), (Alegria, A. Gurrutxaga, P. Lizaso, X. Saralegi, S.Ugartetxea, R. Urizar 2004). Hay otras investigaciones basadas en corpus: (Forchini, Pierfranca, Murphy, Amanda; 2008), (Skut, W. & Brants T. 1998), (Oepen, S. et al. 1998), Keil, M. 1997:, Frantzi, K.T. & Ananiadou, S. 1996 y Ikehara, S. et al. 1996.

II. OBJETIVOS

La estrategia de trabajo que seguimos se sitúa en el ámbito de la fraseología, partiendo de la observación directa de los datos lingüísticos obtenidos de un corpus textual extraído de la Web. Esta aportación pretende explicitar los criterios y presentar una herramienta informática para detección y extracción automática de las unidades fraseológicas (UFS) desde la perspectiva del Procesamiento Natural del Lenguaje.

II.1 Los objetivos generales que se plantean en este estudio son:

I. Crear una herramienta informática que, conjuntamente con una base de datos sólida, pueda permitir la detección y extracción de las unidades fraseológicas en un corpus.

II. Desarrollo de un programa informático que permita detectar una unidad fraseológica en castellano con su equivalencia en portugués, que en estos momentos no se puede realizar con los traductores automáticos. Ello nos servirá para ver las equivalencias lingüísticas de dos lenguas y de ayuda a los traductores.

II.2. Los objetivos específicos de este estudio son:

I. Ayudar a identificar las unidades fraseológicas para que sirva de apoyo a estudiantes y traductores. No basta incluir las unidades fraseológicas en un diccionario electrónico junto con su equivalencia, es preciso que el sistema sepa cómo reconocerlas como tales en el corpus sea lo que sea su forma de aparición, distinguiéndolas de otras unidades sintagmáticas y, por fin, extraerlas automáticamente a partir de un corpus textual, así mismo se mostrarán ejemplos de las UFS en el texto.

II. Sistematizar el tratamiento de la información fraseológica (UFS) en una base de datos bilingüe que permita ver la variación que existe.

III. METODOLOGÍA

Nuestra investigación ha empezado por una selección de corpus formada por diccionarios monolingües, bilingües, de fraseología y tesis relativas a fraseología dónde empezamos a extraer las unidades fraseológicas más relevantes:

III.1. Diccionarios monolingües

REAL ACADEMIA ESPAÑOLA. *Diccionario de la Lengua Española.* Madrid: Espasa Calpe, 1995.

MOLINER, M. Diccionario de uso del español. Madrid: Gredos, 1996.

FERREIRA, A. B. H. *Dicionário Aurélio Eletrônico* – Século XXI. Rio de Janeiro: Nova Fronteira, 1999.

HOUAISS, A. *Dicionário Houaiss da língua portuguesa*. Rio de Janeiro: Objetiva, 2001.

III.2. Diccionarios bilingües

ORTEGA CAVERO, D. (1975). *Diccionario Portugués-Español*. Espanhol-Português, Barcelona: Ramon Sopena.

Flavian, E. y Fernández, G. E. (1994) *Minidicionário Español-Portugués, Portugués-Espanhol*, Editorial Atica, Sao Paulo.

III.3. Diccionarios fraseológicos

SBARBI, J. M. (1943). *Gran diccionario de refranes de la lengua española*. Buenos Aires: Joaquin Gil.

NUNES, Zeno Cardoso e NUNES, Rui Cardoso. (1982), *Dicionário de regionalismos do Rio Grande do Sul*. Porto Alegre, Martins Livreiro.

SECO, M.; ANDRÉS, O.; RAMOS, G. *Diccionario fraseológico documentado del español actual. Locuciones y modismos españoles*. Madrid: Aguilar, tercera reimpresión, 2005.
Tesis

III.4. Tesis

María Luisa Ortiz Alvarez. *Expressöes idiomáticas do porttuguês do Brasil e do espanhol de Cuba: Estudo contarstivo e implicações para o ensino de portugues como língua estrangeira*. UNICAMP, 2000

Myriam Jeannette Serey Leiva.
Lexicologia e lexicografia: A questäo das expressöes idiomáticas em espanhol variante chilena. USP, 2000

MOUZINHO FERRARO, RIita Giovana (2000). *Análisis contrastivo español/portugués de unidades fraseológicas*. Tese doctoral. Universidad de Cádiz.

El núcleo de este sistema de extracción automatizada de UFS es un algoritmo que toma un texto y obtiene una lista ordenada de todas las frases halladas tras su comparación con un tesauro (base del conocimiento o red conceptual). En cuanto al tratamiento lingüístico y el proceso de detección y extracción de unidades fraseológicas se distinguen **3** fases:

III.5. Diccionario de unidades fraseológicas

La construcción de un diccionario fue la próxima etapa. No basta con incluir las unidades fraseológicas en un diccionario junto con su equivalencia, es preciso que el sistema sepa cómo reconocerlas como tales en el corpus – cualquier que sea su forma de aparición. A continuación construimos la primera versión de PhraseNET para la Web. Realizamos solamente un experimento y lo abandonamos, pues hemos configurado la herramienta informática para funcionar en el disco duro. Hemos decidido que:

a. Teniendo en cuenta la taxonomía de las Unidades fraseológicas ofrecida por diversos autores decidimos que nuestro sistema de extracción de unidades fraseológicas se detendrá en detectar y extraer solamente las locuciones y los enunciados fraseológicos, siguiendo el modelo de unidades fraseológicas presentado por Corpas Pastor (1996:52), de cuya clasificación quitamos las colocaciones. Así nuestro sistema no detectará, ni extraerá las colocaciones, pero lo hará con las locuciones y los enunciados fraseológicos alineados en contexto (KWIC), de acuerdo con Sinclair (1991). Por supuesto, llamaremos las locuciones y enunciados fraseológicos de las unidades Fraseológicas.

b. La selección de las unidades fraseológicas incluidas en nuestra base de datos, se ha llevado a cabo: mediante el vaciado de las unidades fraseológicas existentes en nuestro corpus. La base de datos semasiológica (= estudio que parte del signo en busca de la determinación del concepto) bilingüe, aunque esté ordenada alfabéticamente, no es un diccionario, sino una base de datos de lexemas complejos, exceptuando las colocaciones, de unidades fraseológicas, sin marcaciones que indiquen niveles lingüísticos como, familiar, vulgar, jerga, etc., ausencia de marcaciones que indiquen connotaciones como, peyorativo, coloquial, eufemístico etc.

Espanol	Fuenteesp	Defineesp	Portugues	Fuenteptg	Defineptg
A enemigo que huye, puente de plata	TEPDC		Ao inimigo que foge, ponte de prata	TEPDC	Para se livrar de alguém que você não gosta, algumas veze
A lo hecho, pecho	DFDEA		Ao que está feito, [o] peito	MOLP	Enfrenta as conseqüências do que fizeste
A perro flaco, todo son pulgas	TEPDC		Para o cachorro magro, tudo é pulga	TEPDC	Se você está fraco, parece que só problemas estão à sua vol
Agua que no has de beber déjala correr	TEPDC		À água que não vais beber, deixa-a correr	TEPDC	Deixa aquilo que não precisas, deixe para o próximo, que a p
Al perro enfermo se le suben las pulgas	TEPDC		Ao pulgas sobem no cachorro fraco	TEPDC	Os fracos são os que são atacados
Al perro flaco se le suben las pulgas	TAZEPUF		Ao pulgas sobem no cachorro fraco	TAZEPUF	Os fracos são os que são atacados
apretar el zapato	TEPDC		saber onde o sapato aperta	TEPDC	ter pontos vulneráveis
Barriga llena, corazon contento	TAZEPUF		Barriga cheia, coração contente	TAZEPUF	Quando alguém come, fica melhor, contente
chuparse el dedo	DFDEA		ficar chupando dedo	DALP	ser ingênuo
con puntos y comas	DFDEA		com todas as vírgulas	DALP	com precisão, com todos os detalhes
cuidarse del que dirán	TEPDC		fugir das más línguas	TEPDC	precaver-se dos falsiários
Dios le da pan al que no tiene dientes	DRAE		Deus dá pão a quem não tem dentes	MOLP	Quem tem algo, por vezes não sabe o que fazer com ele ou n
Dios le da sombrero al que no tiene cabeza	TAZEPUF		Deus dá um chapéu a quem não tem cabeça. Dá	TAZEPUF	Quem tem algo, por vezes não sabe o que fazer com ele ou n
donde menos se piensa salta la liebre	TAZEPUF		as coisas acontecem quando menos se espera	TAZEPUF	as coisas acontecem quando menos se espera
echar abajo	DFDEA		derrubar a casa; pôr tudo a perder	MOLP	derrubar
echar flores	DFDEA		jogar flores	DALP	dizer galanteios
en absoluto	DFDEA		de jeito nenhum	DALP	ser muito toco
estar apoltado	TEPDC		ser agraviado	TEPDC	obter aprovação em exame
estar en pañales	TAZEPUF		estar de calças curtas	TAZEPUF	não estar preparado para fazer alguma coisa
gastar pólvora en chimengo	TAZEPUF		gastar pólvora em chimango	TAZEPUF	gastar energia à toa
llamar a la puerta	DFDEA		bater na porta	DALP	pedir ajuda
meter la mano hasta el codo	TEPDC		estar enterrado até o pescoço	TEPDC	estar muito comprometido
meterse en camisa de once varas	TAZEPUF		meter-se em camisa de onze varas	TAZEPUF	entrar em enrascada
no caber un alfiler	DFDEA		não caber um alfinete	MOLP	estar superlotado
no decir palabras	TAZEPUF		não dizer palavra	TAZEPUF	ficar calado
perro viejo	TEPDC		macaco velho	TEPDC	pessoa astuta, matreira
ponerse los cuernos	DFDEA		pôr chifre	DALP	ser infiel
ponerse rojo	DFDEA		ficar vermelho	DALP	envergonhar-se
quedarse para vestir santos	TEPDC		ficar para tia	TEPDC	ficar solteira
recién salido del cascarón	TEPDC		que acabou de sair das fraldas	TEPDC	inexperiente
sacar en limpio	DFDEA		tirar a limpo	DHLP	esclarecer

Figura 1: Diccionario de unidades fraseológicas

III.6. El Análisis vectorial

La etapa siguiente es la indexación de un documento del corpus y su emparejamiento con nuestra base de datos. A continuación, la construcción de una matriz de las unidades fraseológicas y del documento (un documento a la vez). Las filas de la matriz, vectores en términos algebraicos, son representadas por los términos de los documentos y las columnas por los términos de la base de datos., que se expresen en función de las apariciones (frecuencia) de cada término. En esta etapa el modelo toma en consideración las sentencias que sólo se emparejen parcialmente con las unidades fraseológicas extraídas, asignando pesos a los términos índice de los documentos y de las UFS almacenadas en la base de datos, así el emparejamiento se torna más preciso.

Cada documento añadido al sistema es, automáticamente, dividido en frases. Entendemos frases por aquellos segmentos de texto separados por punto final o punto de interrogación. Así tendremos un listado de frases, indexado.

2. Se efectúa la búsqueda de UFS significativas en las frases del texto, para lo cual, recurren a la denominada **base del conocimiento o red conceptual** y que contiene más de 1500 entradas. En particular, el sistema encuentra frases y las convierte en candidatas, ya estén en mayúsculas o minúsculas, para la indización tras un proceso morfológico.

Como esta etapa viene después del segmentador, el documento es representado por frases. Es decir:

$$F_1 = [f1,f1,f3...fn]$$

donde f_1 es uno de los elementos de esto conjunto. La frase f_1 será representada por un vector de pesos:

$$F_w = [w1,w2,w3...wn]$$

Cada término tuve un peso. Distintamente de la recuperación de información, aquí el peso asignado a un término nada tiene a ver con su frecuencia que compone un documento. El peso es asignado en función de cómo esta UF fue registrada en nuestro diccionario. En verdad la comparación de dará con cada frase y su respecto en la base de datos.

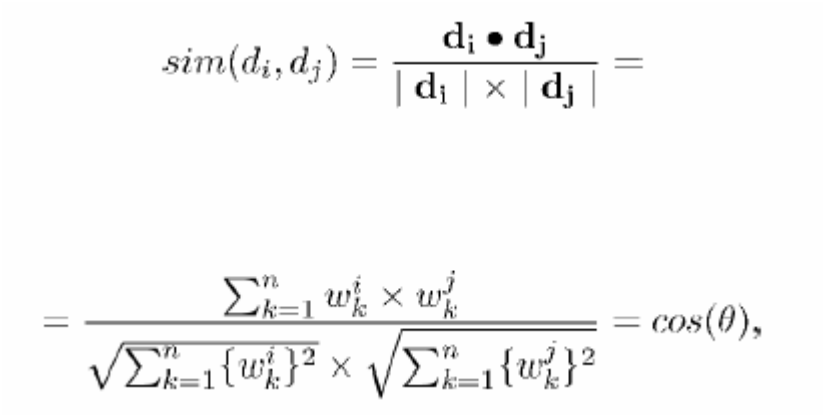
Así que "*pasar el Rubicón*" y "*pasar su propio Rubicón*", se queda así:

Tabla 1 Representación vectorial de una frase

Índice	**Peso**	**Término**
1	4	Pasar
2	1	Su
3	1	Propio
4	4	Rubicón

Tabla 2 Representación de los 2 documentos

	F1	**F2**	**F3**
Frase	4	4	0
Base de datos	44	4	2



$$sim(d_i, d_j) = \frac{\mathbf{d_i} \bullet \mathbf{d_j}}{|\ \mathbf{d_i}\ | \times |\ \mathbf{d_j}\ |} =$$

$$= \frac{\sum_{k=1}^{n} w_k^i \times w_k^j}{\sqrt{\sum_{k=1}^{n} \{w_k^i\}^2} \times \sqrt{\sum_{k=1}^{n} \{w_k^j\}^2}} = cos(\theta),$$

Figura 2 – Cálculo de la similitud

Para saber la similitud entre las frases y la base de datos, precisamos calcular la similitud de Sim (**F**i,**UF**j). Haciendo una adaptación del modelo y de (BAEZA-YATES; RIBEIRO-NETO, 1998), la cuenta a hacer es la siguiente: donde |Fi| es el módulo del vector Fi. Cos(θ) es el coseno del ángulo de los vectores que representan la frase y la UF (de la base de datos) **F**i,**UF**j. El valor del coseno de un ángulo varia en un intervalo de 0 y 1. Ese hecho, nos posibilita saber la distancia entre **F**i y **UF**i, donde 0 significará el más alto grado de similitud y 1 completa similitud. Por su vez, **W**1k indica el peso referente al término *tk* en el documento **F**i, como ya descrito. Ejemplificando:

$$\text{Sim}(F1,UF2) = \frac{4x4 + 4x4 + 0x4}{\text{Raíz cuadrada } (4(2)+4(2)+0(2) \text{ x Raíz cuadrada } (4(2)+4(2)+2(2)} = 32$$

RQ(32) x RQ (36) 32 =34,2 = 0,94

El cálculo indica que la frase **F**1 y la **UF**2 tienen más de 94% de similitud. Es decir, pasar el Rubicón, deja de ser una candidata a UF para se tornar una unidad fraseológica.

III.7. Truncamientos

En esta etapa se utiliza el método del truncamiento o método de extracción por medio de raíces y sufijos. Es una técnica para recuperar información en sistemas que utilizan el vocabulario natural para almacenar informaciones. La técnica de raíces y sufijos fue primeramente utilizada por Salton (1980).

Usamos la búsqueda por truncamiento: Los truncamientos suelen ser por la derecha y por la izquierda. Busca las unidades fraseológicas con base en los truncamientos almacenados en el lexicon, a partir de la raíz o del sufijo del mismo. La búsqueda distingue entre mayúsculas y minúsculas. No es posible tener un fichero de palabras vacías (stop words), como preposicones, artículos y pronombres. Aunque mejorase los tiempos de búsqueda en la base de datos, no se lo puede hacer, pues las unidades fraseológicas, en general, tienen palabras vacías.

En esta etapa se utiliza el método del truncamiento o método de extracción por medio de raíces y sufijos. Es una técnica para recuperar información en sistemas que utilizan el vocabulario natural para almacenar informaciones. La técnica de raíces y sufijos fue primeramente utilizada por Salton (1980).
Los truncamientos suelen ser por la derecha y por la izquierda. Busca las unidades fraseológicas con base en los truncamientos almacenados en el lexicon, a partir de la raíz o del sufijo del mismo. La búsqueda distingue entre mayúsculas y minúsculas.La base fundamental de este sistema son los algoritmos de busca y estadística parametrizada. ". La idea de usar un diccionario de raíces y sufijos no es nueva, aunque no lo sabíamos a la hora de construir nuestros algoritmos. Salton (1980).

El suceso del evento se logra en distintos sucesos. Variaciones morfológicas, Variaciones morfosintácticas, variaciones sintácticas, variaciones semánticas, etc.

IV. RESULTADOS

Hemos llevado a cabo cuatro experimentos, con el objetivo principal de obtener resultados sobre la calidad del algoritmo de detección y extracción de UFS. Para obtener el rendimiento esperado fue necesario primero experimentar en un corpus de reducida dimensión para poder verificar "manualmente" el silencio y el ruido de los resultados y comprobar cuáles son los problemas y faltas inherentes a la propia metodología experimental. En una base de datos muy grande suele ocurrir una cierta desorientación en la búsqueda automatizada de datos entre tanta información, lo que conduce a resultados no deseados (ruido o silencio documental según sea el caso). **Silencio documental** *es el conjunto de documentos almacenados en la base de datos que no han sido recuperados, cuando éste es interrogado.* El inconveniente se debe a que la búsqueda ha sido demasiado específica. **Ruido documental** *es el conjunto de documentos recuperados por la búsqueda que no son relevantes.* El inconveniente se debe a estrategia de búsqueda cuando esta ha definido demasiado genérica. En concreto, buscamos construir un sistema de busca automatizada muy refinado para tratar de solventar estas dificultades, pues la pertinencia de los resultados está en función de la calidad de las técnicas utilizadas en la búsqueda automatizada.

Nuestros experimentos están basados en tres documentos – todos documentos con textos breves. Los textos examinados fueron los siguientes: El Lazarillo de Tormes, de autor desconocido; El aderezo de esmeraldas de Gustavo Bécquer y Reliquias y Relatos: construcción del concepto de «Historia fenoménica», de Gustavo Bueno, publicado en "El Basilisco, 1ª época, nº 1, 1978, páginas 5-16".

Figura 3 - Interface

IV.1. Los experimentos

Para todos los experimentos elaboramos manualmente una lista de las UFS encontradas en el texto, indiferentemente se estaban o no en nuestro tesauro, es decir, una lista de todas las apariciones en el texto de alguna UF. Para reconocer las UFS usamos como referencia el diccionario fraseológico de Seco et al (2005), con rarísimas excepciones. Es decir, las UFS encontradas en los textos se refieren a UFS encontradas en Seco et al.(2005). Eventualmente pueden existir otras UFS en la obra, que no fueron registradas por nosotros y consecuentemente por Seco et al. (2005). Y finalmente, aunque algunos se parezcan con UF no les ha recortado por no haberlos encontrado en Seco et al. 2005), especialmente por estar con grafia antigua, como por ejemplo “y ansí”. Este chequeo fue hecho manualmente y por dos veces. La pantalla de salida de datos se puede ver en Figura 3.

Para el primero, elegimos como campo de pruebas el Lazarillo, suficientemente reducido para verificar de pronto sobre el texto la calidad de la información obtenida (Pamies & Pazos, 2003). Una fuente importante de ruido fueron los casos de inversión de morfemas de las UFs. En el primero experimento en que hemos trabajado con el texto de Lazarillo, se ha generado demasiado silencio, es decir las UFS que PhraseNET debería reconocer no las ha reconocido completamente, y también demasiado ruido, es decir las UFS que PhraseNET ha seleccionado como candidatos a UF y no debería haber reconocido. El silencio se puede comprobar con muchos ejemplos. En ese caso, fueron extraídas 12 UFS, siendo que 3 correctas. Entonces el índice de precisión fue de 25%. El ruido se puede ver por ocurrencias como esta: El coeficiente de cobertura, fue bajísimo (3/116), es decir, solamente 2,6%.

Segundo experimento. Gustavo A. Bécquer. Después de algunos arreglos en el lexicón, ponemos en marcha el segundo experimento. Un texto pequeño de Gustavo A. Bécquer (El aderezo de esmeraldas). El texto de Gustavo Bécquer fue convertido desde formato .PDF para .TXT. En esta conversión hubo algunos errores debido al propio conversor, pues fue hecha sin intervención humana. Fueron encontrados 4 types y 4 tokens.

En verdad, de acuerdo con la versión más actual, hay 22 tokens y 15 types. Por consiguiente, la estimativa actualizada de Precisión fue de 100%, aunque la cobertura fue de solamente 18% (4/22) a la época de este experimento.

Tercer experimento (El Lazarillo de Tormes). De lo expuesto en los dos primeros experimentos resultó evidente que era preciso cambiar la base de datos e algoritmo. La metodología ha continuado la misma, pues no depende del corpus. Realizamos así el experimento con el texto Lazarillo de Tormes, y el resultado, confirma los cambios profundos que se dieron en la extracción de las UFS. Nos preocupaba la metodología usada para detectar correctamente las UFS. La atención estaba enfocada en el lexicón (base de datos) y en el algoritmo. Había un problema metodológico en ambos. De la aplicación de cambios, tanto al lexicón como al algoritmo obtuvimos como resultado 116 tokens de las unidades fraseológicas y 56 types de las unidades fraseológicas encontrados. En verdad hay 151 tokens, es decir el coeficiente de cobertura fue de (76,8%) y la precisión 100%.
Cuando este experimento fue llevado a cabo, la base de datos de UFS aún carecía de una actualización. Esta actualización fue hecha cuando llevamos a cabo el 4º experimento.

Cuarto experimento (El Basilisco, 1ª época, nº 1, 1978, páginas 5-16) Corpus: Reliquias y Relatos: construcción del concepto de «Historia fenoménica», Gustavo Bueno.
El Basilisco, es una revista de filosofía, ciencias humanas, teoría de la ciencia y de la cultura, fue fundada en 1978 para facilitar «la publicación regular de trabajos cuyo común denominador fuera el estar concebidos desde una perspectiva filosófico-crítica (materialista)». Durante treinta años viene publicando artículos originales escritos en español. (http://www.filosofia.org/rev/bas/index.htm) De lo expuesto en el tercero experimento resultó evidente que era preciso ahora solamente añadir más registros a la base de datos, pues el algoritmo funcionaba bastante bien.

Tabla 3. Evaluación de PhraseNET.

Experimento	Texto	Área	Tipo	Precision	Recall	No. Palabras
1	Lazarillo	Literatura	Capítulo	25%	2,6%	20.159
2	G. Bécquer	Literatura	Libro	100%	18%	2.450
3	Lazarillo	Literatura	Capítulo	100%	76,8%	20.159
4	Rev. Basilisco	Filosofía	Artículo	85%	90%	10.540

Tabla 4. Descripción de la escala de evaluación de las unidades fraseológicas.

EVALUACIÓN	**DESCRIPCIÓN**	**EJEMPLOS**
Excelente	La UF candidata es semánticamente equivalente a encontrada en el texto	Al cabo de/ Al cabo de
Bueno	La UF candidata pertenece al segmento de texto encontrado, mas solamente parcialmente identificado	Al cabo de/ Cabo
Débil	La UF no es exactamente lo que se buscaba	Al cabo de/ Al cabo del
Mala	La UF candidata aunque tenga los mismos términos no está en la misma orden	Al cabo de/ Cabo de Almería

V. CONCLUSIONES

A lo largo de los cuatro experimentos (Tabla 3) podemos ver que los resultados han mejorado bastante, con un pequeño empeoramiento del índice de precisión en el cuarto experimento, lo que se puede asignar a unidades fraseológicas que no estaban en la base de datos de PhraseNET. Cuanto más UFS hubiere en el lexicón, más cerca estaremos de una evaluación excelente (Tabla 4). El nivel de precisión "excelente" (Tabla 4) ha sido prácticamente alcanzado a partir del segundo experimento, con las correcciones hechas en al algoritmo y en el lexicón, no habiendo más ocurrencias "débeis", "malas" o simplemente "buenas". PhraseNET se puede aplicar a varias lenguas. En este caso el español, pero esto depende solamente de crear un diccionario para el inglés o el francés o el portugués, así como fue creado para el español. En suma, la elección del corpus y la amplitud del lexicón tienen un papel clave en el funcionamiento de la metodología
Sin embargo, hay tipos de modificaciones que pueden no ser posibles de ser identificadas. Destacamos las siguientes limitaciones del sistema:

- Entre las modificaciones internas es posible detectar y extraer muchas de las modificaciones por adición, siempre que la periferia de la UF se mantenga.
- Es posible detectar y extraer las UF modificadas por reducción, desde que la periferia de la UF se mantenga.
- La desautomatización, según Mena Martinez (2003), necesitará de un algoritmo diferente para detectar y extraer las UFS que presenten tales modificaciones.
- Desconocemos un número relevante de los tipos de modificaciones arriba mencionadas que justifique cambios en los algoritmos, de tal forma, a tener en consideración estas modificaciones y excepciones. Los pocos casos conocidos y usados como ejemplos en la literatura no justifican cualquier añadidura a nuestra aplicación,

En este momento, lo que demandaría un coste mucho grande de dinero y tiempo.

VI. BIBLIOGRAFÍA

Alegria, A., Gurrutxaga, P., Lizaso, X., Saralegi, S., Ugartetxea., Urizar, R. (2003) "A Xml-Based Term Extraction Tool for Basque" . IN: LREC, Portugal

BAEZA-YATES, R.; RIBIERO-NETO, B. (1998). Modern Information Retrieval. 1. ed. New York: Addison-Wesley.

Corpas Pastor, G. (1996). Manual de fraseología española. (Biblioteca Románica Hispánica). Madrid: Gredos.

Forchini, Pierfranca y Murphy, Amanda 2008. N-grams in comparable specialized corpora: Perspectives on phraseology, translation, and pedagogy. Patterns, meaningful units and specialized discourses, Römer, Ute and Rainer Schulze (eds.), 351–367.

Frantzi, K. T., & Ananiadou, S. (1996). Extracting nested collocations. In: Proceedings of the 16th Conference on Computational Linguistics (pp. 41–46). COLING.

Gaël Dias, Sylvie Guilloré, Bassano, Jean-Claude, Pereira Lopes, José Gabriel, (2000) Normalisation of Association Measures for Multiword Lexical Unit Extraction. In "International Conference on Artificial and Computational Intelligence for Decision,

Control and Automation in Engineering and Industrial Applications", Monastir, Tunisia.

Ikehara, S. et al. (1996). A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. COLING-1996.

Keil, M. (1997). Wort fuer Wort -- Repraesentation und Verarbeitung verbaler Phraseologismen. [Sprache + Information.] Niemeyer, Tuebingen.

Mena Martínez, F. M. (2003). En torno al concepto de desautomatización fraseológica: aspectos básicos en Tonos. Revista electrónica de estudios filológicos 5,15 págs.

Oepen, Stephan ; Netter, Klaus ; Klein, Judith (1998). "TSNLP — test suites for natural language processing", in Linguistic Databases, J. Nerbonne (eds.), Stanford, CSLI Publications.

Salton, G. (1961-1976). The SMART system: Experiments in dynamic document processing. Encyclopedia of Library and Information Science, vol. 28, 1980, p. 1-28

Seco, M.; Andrés, O.; Ramos, G. (2005). Diccionario fraseológico dcumentado del español actual. Locuciones y modismos españoles. Madrid: Aguilar, tercera reimpresión.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Skut, W. & Brants, T. (1998). A Maximum-Entropy Partial Parser for Unrestricted Text. InProceedings of the Sixth Workshop on Very Large Corpora. Montreal, Canada

Thurmair, G. (2003). Making Term Extraction Tools Usable. Proc EAMT-CLAW Dublin.