El tratamiento computacional de paremias: problemática y estado de la cuestión

Elena ALONSO PÉREZ-ÁVILA
Università di Roma *La Sapienza* (Italia)
ealonsop@gmail.com

Recibido: 10-03-2010
Aceptado: 18-04-2010

Resumen: El presente artículo recoge algunas aproximaciones a la clasificación y tratamiento automáticos de las paremias. En primer lugar, nos cuestionamos la pertinencia de utilizar los instrumentos y metodologías informáticos con este tipo de unidades pluriverbales, teniendo en cuenta sus particulares características formales y de contenido. Seguidamente, describimos algunas propuestas efectuadas por diversos autores para, finalmente, presentar brevemente nuestra propia aportación. Nuestra propuesta tiene como objetivo establecer un sistema de clasificación de paremias basado en la descripción que el lexicón computacional WordNet proporciona de la red semántica, de modo que se puedan ofrecer unos criterios de ordenación paremiográfica válidos para distintas lenguas. Asimismo, se añade información semántica a la paremia sustentada en los mismos principios psicolingüísticos que rigen el diseño del lexicón computacional, es decir, se definen las unidades léxicas no de manera absoluta, sino en relación con otras unidades. Concluimos destacando las ventajas obtenidas de todas estas propuestas y los beneficios de contar con bases de datos y corpus anotados de paremias con el fin de facilitar el tratamiento computacional de estas unidades.
Palabras clave: Paremiología. Refrán. Tratamiento automático. WordNet.

Titre : « Le traitement computationnel des parémies : problématique et état de la question »
Résumé : Cet article recueille quelques approches de classification et de traitement automatiques des proverbes. Premièrement, nous nous penchons sur la pertinence d'utiliser des outils et des méthodologies informatiques avec ce type d'unités pluriverbales étant donné, notamment, leurs caractéristiques formelles et leurs contenus. Deuxièmement, nous décrivons des propositions formulées par plusieurs auteurs dans différentes langues et ensuite, nous présenterons brièvement notre propre contribution. Notre proposition a pour but l'établissement d'un système de classification de proverbes fondé sur la description du réseau sémantique fournie par le lexique informatique WordNet, de telle sorte qu'on puisse offrir des critères d'ordonnance des proverbes qui s'appliquent aux différentes langues. En outre, ce système enrichit les proverbes avec de l'information sémantique, basée sur les mêmes principes psycholingüistiques qui régissent la conception du WordNet, c'est-à-dire que les unités lexiques se définissent non pas en termes absolus mais par rapport à d'autres unités. Nous concluons en soulignant les avantages tirés de toutes ces propositions et les bénéfices obtenus dans l'utilisation des bases de données et des corpus annotés de proverbes afin de faciliter le traitement automatique de ces unités.
Mots-clé : Parémiologie. Proverbe. Traitement automatique. WordNet.

Title : "Computer treatment of paremiae: problems and state of the art"
Abstract: The following paper includes various approaches to the automatic classification and processing of proverbs. In view of the particular characteristics of these multiword units in relation with their form and content, this paper considers whether it is worth to process them using computer tools and methodologies. After a description of a number of proposals made by several authors, the article leads to a brief presentation of our own contribution. Our proposal is aimed at establishing a classification system for proverbs that is structured along the same lines as the description of the semantic network provided by

the lexical database WordNet, in order to supply suitable criteria to classify proverbs in different languages. Moreover, it gives semantic information to each proverb based on the same psycholinguistic principle of the lexical database, which means that lexical units are not defined isolated but in relation with other units. As a conclusion, we highlight the lessons learned from all these proposals and the advantages of using databases and annotated corpora of proverbs in order to facilitate the automatic processing of these units.
Keywords: Paremiology. Proverb. Automatic processing. WordNet.

1. PROBLEMÁTICA DEL TRATAMIENTO AUTOMÁTICO DE PAREMIAS

Muchos autores han decidido prescindir de las paremias en sus propuestas de análisis y estructuración lingüística, especialmente desde que Casares (1960: 203) decidió excluirlas del ámbito de estudio de la Lexicografía para sostener que debían de ser objeto de estudio de la Antropología y el folclore. Defiende la misma línea Coseriu (1977), quien establece una clara distinción entre las paremias y el resto de unidades que conforman el "discurso repetido", y así descartó todas aquellas unidades que equivalen a un breve texto de su estructuración semántica.

Si el hecho de que las paremias constituyan enunciados autónomos o, lo que viene a ser lo mismo, que no puedan ser conmutables por una unidad léxica simple impulsa a estos autores a no considerar estas unidades como posibles objetos de estudio de la Lingüística, a más graves problemas se tendrán que enfrentar aquellos que se aproximen al tratamiento automático del lenguaje para conseguir que un agente computacional sea capaz de procesar paremias.

La imposibilidad de conmutación por un único concepto que abarque el significado completo de la paremia constituye un problema de difícil solución. A esta particularidad de las paremias, hay que añadir la frecuente idiomaticidad o no composicionalidad semántica. Para llevar a cabo el análisis semántico de una oración[1] en el ámbito del procesamiento del lenguaje natural se parte del presupuesto que el significado de una frase es equivalente a la suma de los significados de las partes que la componen (Jurafsky y Martin, 2000: 583). No obstante, estos autores señalan que las unidades fraseológicas desafían este principio, pues su significado no puede deducirse de la suma de sus partes. Aún así, presentan posibles soluciones a este problema, como la creación de constituyentes específicos (Jurafsky y Martin, 2000: 573), entre otras posibilidades, pero entre los ejemplos que proporcionan se encuentran únicamente ejemplos de unidades fraseológicas que pueden sustituirse por una lexía simple como: *tip of iceberg* como *beginning*.

Mayores dificultades presentan las paremias de cara al análisis semántico, especialmente porque no en todas está presente la característica de la idiomaticidad, ni se da en los mismos grados. En otras palabras, podemos hablar de un *continuum* en relación con la idiomaticidad relativa a las paremias que parte desde las completamente idiomáticas hasta llegar a aquellas cuyo significado sí equivale a la suma de los significados de las partes que la componen. Es importante recordar, en nuestra opinión, que la mayor parte de las paremias se sitúan en los puntos intermedios y que, dependiendo de los contextos de uso podrán considerarse más o menos idiomáticas. Se presenta, seguidamente, un ejemplo con sus correspondencias[2] italiana y gallega.

1. En dinero y santidad la mitad de la mitad.
2. *Quattrini e santità metà della metà*.
3. *En falando de diñeiro e santidade, a metade pola metade*.

[1] Jurafsky y Martin (2000: 545) describen el análisis semántico como "the process whereby meaning representations [...] are composed and assigned to linguistic inputs".

[2] Véase el artículo de Julia y Manuel Sevilla Muñoz (2000) para más información sobre el concepto de *correspondencia paremiológica*.

En este caso, el grado de idiomaticidad no es muy elevado, los términos mantienen el significado que tendrían fuera de la unidad y, sin embargo, la suma de los significados de dichos términos no equivale al significado global de la paremia, puesto que se trata de una oración sintácticamente incompleta: falta el verbo principal, que se encuentra implícito y que podría ser, entre otros, *creerse*, *considerar*, *intuir*. Entendemos que resulta difícil que un programa informático pueda inferir el verbo principal de una oración de este tipo; para hacerlo debería llevar a cabo un proceso cognitivo que es propio única y exclusivamente de los seres humanos. No obstante, se registran en el refranero numerosos ejemplos en los que no está presente la idiomaticidad en ningún grado:

4. Tres gustos da el dinero: ganarlo, gastarlo, y guardar para cuando se es viejo.

Y algunos ejemplos en los que se constata una idiomaticidad muy elevada:

5. Una golondrina no hace verano.
6. *Una rondine non fa primavera*.
7. *Unha andoriña soa non fai verán*.

No sólo la idiomaticidad puede constituir un obstáculo en el proceso de tratamiento automático de paremias. La oscuridad propia de las paremias se alcanza por medio de diversos recursos retóricos, como la metáfora y la metonimia, y mediante la presencia de arcaísmos y estructuras gramaticales anómalas –elipsis, alteración del orden lógico de los constituyentes, etc.–, o incluso palabras inventadas. En muchas ocasiones, las asociaciones de significados que se encuentran en la base de metáforas y metonimias se han perdido con el paso del tiempo y el cambio de referentes culturales, de manera que muchas de esas paremias resultan difíciles de interpretar hoy en día y provocan muchos problemas, especialmente en los hablantes no nativos.

Mostramos a continuación uno de esos ejemplos que, en nuestra opinión, puede presentar dificultades:

8. Al doblón no le dan garrote.

Se produce una desviación del sentido recto de los términos presentes en esta unidad fundada en la relación de contigüidad existente entre los conceptos empleados y los conceptos aludidos, en otras palabras, nos encontramos ante dos metonimias[3]. Hay que entender el significado de la moneda *doblón* como forma de aludir a la persona que lo posee o lo utiliza; *garrote*, en cambio, es un instrumento a través del que se ejerce la pena capital, por lo que podría encarnar una representación de la justicia. Se trata por tanto de metonimias del tipo instrumento por artífice. Así, un posible significado de esta paremia podría ser que las personas que poseen dinero no sufren la pena capital. No obstante, la generalidad que es propia de las paremias induce a que se produzcan cambios semánticos según el contexto en el que estas unidades van aplicadas y, de esta forma, se podrá usar también esta paremia para indicar que si se paga se puede comprar la justicia y ganarse los favores de los que en un primer momento pueden ser contrarios; o bien, que los ricos y millonarios, en general, no deben temer problemas con la justicia porque el dinero comporta poder en la sociedad. Toda esta información ha de ser estudiada en el contexto concreto de aparición de la paremia y es muy difícil que se pueda añadir a la base de datos de un

[3] La metonimia es uno de los recursos más frecuentes en el caudal paremiológico –en combinación con la metáfora. Otros ejemplos que se pueden citar son las paremias “A tus escudos son los saludos” y “Abre para todos tu boca, y para pocos tu bolsa”, entre otros muchos.

programa informático para que pueda realizar inferencias y entender el significado de esta unidad o utilizarla en la generación automática de textos en los momentos apropiados.

Incluimos a continuación más ejemplos en los que se registran algunas de las características que contribuyen a "extrañar" (Lázaro Carreter, 1981) u oscurecer el mensaje comunicado en la paremia.

9. *Buon pagatore, dell'altrui borsa è signore.*
10. El buen pagador, señor es de bolsa ajena.
11. *O bo pagador é herdeiro do alleo.*
12. *Quando la borsa mia fa* ndindí *ne chiami uno e ne vengono tre; ma se la borsa mia non fa* ndindí *non ne viene nessuno dei tre.*
13. Don sin din, olla sin tocín.
14. Dádivas quebrantan peñas.
15. *I soldi non puzzano mai.*

En estos ejemplos se puede observar: i) la presencia de arcaísmos y de estructuras sintácticas y léxicas anómalas como la elipsis de elementos necesarios para la correcta estructuración gramatical de la oración: en el ejemplo 13, elipsis del verbo; en los ejemplos 9 y 14; elipsis del artículo; ii) la aparición de términos inventados o deformados en los ejemplos 12 y 13 –para favorecer la rima y el ritmo con el fin de lograr la memorización de estas unidades–; iii) desviaciones de sentido por analogía –como la sinécdoque, la metáfora o la metonimia–, presentes en todos los ejemplos, y iv) personificación, en el ejemplo 15.

Habida cuenta de todas las dificultades que entraña la naturaleza intrínseca de las paremias, cabe preguntarse no sólo si será posible llevar a cabo un tratamiento automático de estas unidades, sino si merece la pena siquiera el intento. Lo cierto es que las paremias son, en primer lugar y tal y como señala Franceschi (1994), elementos lingüísticos que sirven para comunicar un mensaje, normalmente abstracto y complicado, por medio de una estructura *prefabricada* y aceptada en el seno de una comunidad. Las paremias, al igual que otros tipos de unidades léxicas pluriverbales, desempeñan una función en la comunicación cotidiana –no sirven únicamente como almacenes de la sabiduría popular, como se ha defendido durante mucho tiempo– y, dado su carácter convencional e idiomático, presentan una dificultad muy elevada para hablantes procedentes de otras culturas o comunidades lingüísticas. Pueden recogerse paremias en textos de todo tipo: en la literatura, en la oralidad, en textos periodísticos, etc., y a ellas se tendrán que enfrentar traductores, lingüistas, estudiantes de lenguas extranjeras, entre otras muchas personas.

Por todo ello, consideramos necesaria la modernización y mejora de los instrumentos y recursos de los que se dispone para acceder a estas unidades, tanto por lo que se refiere a la informatización del material paremiológico, como a la creación de más herramientas y recursos lingüísticos que almacenen e incrementen la cantidad de información relativa a las paremias: contextos de uso, origen, fuentes, correspondencias en otras lenguas, variantes, etc. Creemos que se pueden aprovechar las posibilidades que ofrecen los avances experimentados en el área de la Inteligencia Artificial, de la Lingüística de Corpus y, en general, de la Lingüística Computacional, para que se puedan incorporar estos elementos a la revolución de la sociedad de la información, en particular, de la web.

El problema radica en encontrar una metodología que permita el tratamiento automático de estas unidades teniendo en cuenta tanto sus características particulares como la heterogeneidad de elementos que se engloban dentro de lo "paremiológico", anteriormente se han presentado algunos ejemplos en los que estas características se manifiestan en diferente grado: desde las paremias de considerable oscuridad semántica, pasando por aquellas en cuyo origen se encuentra una metáfora, las paremias con arcaísmos, hasta aquellas totalmente transparentes.

2. SISTEMAS DE CLASIFICACIÓN PAREMIOGRÁFICA Y TRATAMIENTO AUTOMÁTICO DE PAREMIAS

En el presente apartado se recuerdan las propuestas de una serie de autores que se han planteado la posibilidad de aprovechar las nuevas tecnologías para proceder a una clasificación del material paremiográfico y facilitar el tratamiento computacional de paremias.

2.1. La clasificación informática de paremias de Holbek

Bengt Holbek (1969) fue uno de los primeros en considerar que los métodos tradicionales de ordenación alfabética o de clasificación bajo un tema o palabra clave presentaban muchos inconvenientes, entre ellos, las dificultades de consulta de la obra –para localizar una paremia, el usuario tendría que recorrer manualmente toda la obra– debidas a la rigidez de dichas clasificaciones. A juicio de este autor, el tratamiento automático de estas unidades podría agilizar en gran medida la búsqueda y el estudio en las obras paremiográficas.

Holbek (1969: 372) describe la metodología seguida en su propuesta de procesamiento de un corpus de 100 paremias danesas mediante la que ha codificado tres tipos distintos de información: a) el texto[4], b) ciertas características estilísticas y c) datos sobre el origen y las fuentes de esos textos. En cuanto a las dificultades que le surgieron trabajando con el primer tipo, el texto, cabe destacar la problemática de las variantes. El autor es consciente de que una mera variación ortotipográfica diacrónica o dialectal puede cambiar totalmente el texto a ojos del ordenador. La solución que propone el autor, para estos casos, consiste en codificar todas las versiones de un texto juntas, para que el ordenador las pueda identificar. Por otro lado, Holbek estaba muy interesado en que el ordenador fuese capaz de localizar la misma paremia en varios repertorios –fenómeno que el autor denomina *doubles* (Holbek, 1969: 373)–, incluso cuando el texto no era exactamente idéntico. Para lograrlo, Holbek propone que se tengan en cuenta sólo dos o tres términos clave que han de aparecer en las dos paremias que se están cotejando: "[...] to select in each proverb a few main words and to define identity as the occurrence of these main words in two or more proverbs".

Por otro lado, también elaboró una lista de características estilísticas [tipo b)] de modo que se pudiera cotejar cada paremia con el esquema y así proponer una clasificación según: 1) el tipo de metáfora que tuviera lugar –divide trece grupos distintos de metáforas en los que clasificar las paremias: de naturaleza inanimada, climatología, ropa, alimentos, unidades y valores económicos, etc.–; 2) aquellas paremias no metafóricas en las que estuviera presente algún concepto abstracto; 3) las peculiaridades gramaticales como el tiempo del verbo, entre otras; 4) el ritmo y la rima; 5) el paralelismo y antítesis de forma o contenido y, por último, 6) si se trata de un consejo positivo o negativo.

Asimismo, Holbek señala la necesidad de introducir información sobre el origen de los textos [tipo c)], de forma que se habrá de anotar, para cada texto, el paremiógrafo, el lugar, la fecha, el informante –en el caso de fuentes orales–, los manuscritos de referencia –para las fuentes escritas manuscritas– y las referencias bibliográficas. Creemos que la visión de Holbek es muy acertada, pues deja constancia de la importancia de realizar distintas tareas de anotación a fin de cubrir todas las necesidades que ha descrito este autor.

Otra limitación sobre la que llama la atención Holbek es la imposibilidad de procesar paremias procedentes de varias lenguas de forma simultánea. Ofrece como ejemplo una paremia danesa con su correspondencia literal noruega, pero señala que, hasta ese momento, ninguna máquina puede reconocer ese hecho. Sin embargo, añade que sería aconsejable que se utilizaran los mismos sistemas y programas en todas las lenguas, para así poder intercambiar la

[4] Holbek entiende por "texto" el significante de la paremia.

información en un futuro, cuando se incrementase el número de recursos multilingües disponibles.

Lo cierto es que la propuesta de Holbek, avanzada ya hace más de cuarenta años, sigue plenamente vigente y que las dudas que este autor se planteaba continúan constituyendo el núcleo del trabajo pendiente para poder llegar a un sistema informatizado de clasificación de paremias (Holbek, 1969: 375):

> There is, of course, no way of knowing what our successors will be looking for fifty years hence, but at least we can describe and analyze our texts in accordance with the full range of present-day knowledge and interests.

2.2. Kuusi y Lauhakangas: *International Type-System of proverbs*

También han sido los problemas inherentes a la clasificación paremiográfica en contextos multilingües los que han movido a uno de los paremiólogos más importantes de todos los tiempos, el finlandés Matti Kuusi, a proponer una tipología de clasificación que pueda servir para todas las lenguas, de forma simultánea, en su trabajo de 1972, "Towards an International Type-System of Proverbs".

La labor paremiológica de Kuusi ha sido fundamental para el desarrollo de la Paremiología como disciplina: en primer lugar, por su interés por solucionar el espinoso problema de la clasificación de paremias y, en segundo lugar, por el empeño demostrado para lograr que dicho sistema pueda ser utilizado para los repertorios paremiográficos de todas las lenguas, incluidas aquellas más alejadas desde el punto de vista occidental. Con su trabajo, ha identificado tipos paremiológicos universales, que se dan en todas las culturas, y otros que son específicos de culturas particulares.

Kuusi y su hija Outi Lauhakangas, en estrecha colaboración, diseñaron y crearon una base de datos paremiológica internacional que actualmente está disponible para su consulta en la red.[5] Lauhakangas indica que durante años Kuusi estuvo elaborando manualmente un índice de fichas –*pink cards*– con paremias pertenecientes a distintas colecciones paremiográficas en diversas lenguas. Estas *pink cards* estaban ordenadas alfabéticamente según el primer nombre –en finés–. Posteriormente, decidieron utilizar un sistema de tratamiento de bases de datos con el índice de fichas de Kuusi para, de esta manera, tratar de evitar un sistema de clasificación rígido. Lauhakangas pone de manifiesto algunas de las ventajas de utilizar un ambiente informatizado para clasificar paremias (2001: 16):

> The most important quality of the relational database system is that the multi-level character of the proverb correspondences can be described through a net of internal references, and that the user can make quite complicated inquiries of the corpus.

La propuesta de Kuusi (1972) se basa en dos enfoques fundamentales que la diferencian de otras clasificaciones temáticas: a) las oposiciones binarias y b) las imágenes y las estructuras formales. Dado que las paremias conforman formas fijas, cuyo origen puede remontarse a épocas muy antiguas, Kuusi considera fundamental el análisis de los esquemas proverbiales que se repiten, en su opinión, en las diferentes culturas; el finlandés denomina a estos esquemas "*fórmulas*", y los define utilizando la terminología de la gramática generativa: "estructura profunda" y "estructura superficial". Según Kuusi, estas fórmulas presentan una estructura superficial que, en la mayor parte de los casos, no se corresponde con la estructura profunda. Estas nociones de "estructura profunda" y "superficial" son muy útiles para distinguir las variantes e invariantes paremiológicas.

[5] http://lauhakan.home.cern.ch/lauhakan/int/cerpint.html [Consulta del 9/2/2010].

En cuanto a la clasificación temática que proponen Kuusi y Lauhakangas (2001: 26-61), se recogen distintas tipologías dependiendo del tipo de interpretación que se realice de la paremia: literal, metafórica, una lectura en clave moral, literaria, etc. Lauhakangas resume todas estas opciones en los dos criterios principales de clasificación que observan: según la estructura superficial o según la estructura profunda o, lo que es lo mismo, una clasificación onomasiológica[6] basada en el significante de la paremia o en su significado. En opinión de Lauhakangas, en el sistema que presentan han preferido reflejar la estructura profunda de la paremia "to see unifying ideas under variable forms" (2001: 28), es decir, se basan en el significado global de la paremia. Para no perder información en ningún nivel de interpretación la autora recomienda el uso de "cross-references" entre los distintos niveles, pues una misma unidad puede ser objeto de diversas clasificaciones y, a su vez, todo un tema –en la terminología de los autores: *type*– puede estar vinculado a otros temas de clasificación. El sistema de "tipos" que proponen estos autores, al tener alcance internacional, ha de ser lo suficientemente amplio y flexible como para poder ser utilizado por culturas muy distintas. La autora defiende el carácter eminentemente práctico del sistema que ella y su padre proponen, que está lejos de considerarse un trabajo cerrado, y que debe funcionar como una herramienta para poder continuar la investigación y no como una biblia de la Paremiología.

Lauhakangas reconoce que, en ciertas ocasiones, ella misma ponía en duda el rigor de este tipo de clasificación, viendo con desaprobación cómo se mezclaban criterios estructurales –distinciones basadas en oposiciones binarias– con clasificaciones temáticas de carácter tradicional. Sin embargo, reconoce que a la hora de examinar el material paremiológico real no puede separarse la estructura del significante del significado, por tanto, en su opinión, de igual modo debe procederse en una base de datos paremiológica.

La base de datos de Kuusi y Lauhakangas está dividida en 13 capítulos principales o, como los denominan los autores (2001: 32), *home districts*, a su vez divididos en 52 clases principales, entre las que se distribuyen los 325 subgrupos codificados. Gracias a las posibilidades de una base de datos relacional se pueden facilitar muchos tipos de búsqueda distintos; incluso es posible realizar una búsqueda de una palabra para ver en qué paremias aparece. Estas búsquedas ayudan a evitar la pérdida de información y, además, la autora reconoce que se facilita la creación de vínculos de "cross-references" entre tipos proverbiales cercanos.

2.3. Pamies, Bálmacz e Iñesta: *AUTOFRAS*

Hemos tenido conocimiento de más intentos de conectar bases de datos fraseológicas y paremiológicas a ontologías para proponer criterios de clasificación onomasiológica, como es el caso del programa de búsqueda nocional *AUTOFRAS*, cuyo objetivo consistía en "elaborar un sistema de búsqueda capaz de localizar fraseologismos y proverbios en un entorno multilingüe (10 lenguas) a partir de criterios conceptuales" (Pamies, Bálmacz e Iñesta, 1998: 112). Estos autores han utilizado ontologías creadas *ad hoc*, en las que los términos que funcionan como descriptores en la base de datos se han creado siguiendo unos criterios subjetivos. Como nos ha indicado el profesor Antonio Pamies, el proyecto todavía no se ha llevado a término. No obstante, el trabajo ideado por estos autores supone un gran avance en los intentos de automatización de unidades fraseológicas en un contexto multilingüe y, asimismo, ha arrojado luz sobre otros planteamientos, entre otros, nuestra aportación (Bordoni y Alonso Pérez-Ávila, 2009; Alonso Pérez-Ávila, 2009), especialmente en los momentos en los que no veíamos nuestra propuesta como viable.

[6] Puede consultarse la descripción que hemos realizado de los distintos tipos de clasificación de paremias utilizados en los repertorios paremiográficos en Alonso Pérez-Ávila (2008).

2.4. El tratamiento informático de paremias de Conenna

Es necesario destacar los trabajos realizados por Mirella Conenna (1998, 2004), que se enmarcan en el modelo de la *lexique-grammaire* de Gross (1975), dirigidos al tratamiento automático de paremias en el contexto de un diccionario digital de paremias francesas. No obstante, como la propia autora señala (1998: 200), su trabajo presenta un enfoque eminentemente sintáctico y deja de lado la aproximación semántica.
Los objetivos que se marca Conenna son los siguientes (2004: 91):

> La présence des proverbes dans les textes, y compris certains textes techniques, n'est pas négligeable, ce qui soulève le problème de leur reconnaissance et de leur traduction. C'est dans cette optique que nous avons construit, pour quelques types de proverbes français et italiens, des dictionnaires électroniques et des bibliothèques d'automates finis [...], tous gérés par le système INTEX [...].

Conenna ha creado una serie de autómatas finitos y de transductores capaces de reconocer una secuencia proverbial en una lengua –basándose en estructuras sintácticas precisas– y de generar otra secuencia en otra lengua reemplazando la original. En opinión de Conenna, este sistema puede ser también muy útil para identificar las recreaciones de paremias en textos actuales de diversa índole. Es especialmente destacable en este trabajo el hecho de que se considere la paremia como una unidad de traducción y que se tenga en cuenta el peligro de llevar a cabo una traducción literal de las paremias.

2.5. La base de datos de refranes meteorológicos y del calendario en la Romania, *BADARE*

En este somero repaso de autores que se han ocupado del estudio de posibles sistemas de clasificación paremiográfica automatizada no querríamos dejar de mencionar BADARE, la base de datos de refranes meteorológicos y del calendario en la Romania, fruto del proyecto dirigido por Enrique Gargallo Gil (2008). En esta base de datos se organizan los refranes meteorológicos y del calendario procedentes de distintas lenguas y dialectos romances; en otras palabras, en esta base de datos multilingüe tienen cabida tanto lenguas oficiales como el gallego o el francés, como los dialectos, entre otros, el occitano o el ladino dolomítico. Todos los refranes han sido extraídos de fuentes escritas publicadas en el último siglo y medio. Dicha base puede consultarse en la página web http://stel.ub.edu/badare/.

2.6. El *Refranero multilingüe*

El proyecto del *Refranero multilingüe*, dirigido por Julia Sevilla Muñoz y María I. Teresa Zurdo Ruiz-Ayúcar, se inició en 2005, si bien empezó a fraguarse en 1996, y desde entonces el número de paremias que lo componen ha ido aumentando así como las lenguas de trabajo.

Se trata de un refranero multilingüe único en el mundo, tanto por la el número y la combinación de lenguas abarcadas (alemán, catalán, francés, gallego, griego antiguo, griego moderno, inglés, italiano, portugués y vasco) como por la información que acompaña cada paremia. Además, la base de datos presenta la enorme ventaja de estar disponible de forma gratuita en la red, en la página web del Instituto Cervantes (cvc.cervantes.es/lengua/refranero/). Esta base no es un mero corpus multilingüe de refranes y otros tipos de paremias, sino que ofrece correspondencias paremiológicas en las distintas lenguas de trabajo y una serie de datos que se consideran indispensables para la comprensión del enunciado sentencioso. De este modo, cualquier usuario podrá conocer la correspondencia en otra lengua de una paremia, su

significado, posibles variantes, contextos de uso y otras paremias con las que está relacionada semánticamente de forma inmediata. El *Refranero multilingüe* se ha convertido ya en una herramienta fundamental para traductores, estudiantes de lenguas, profesores, lingüistas, antropólogos, entre otros muchos usuarios.

2.7. Propuesta de clasificación de paremias basada en la estructuración semántica del lexicón computacional WordNet

En nuestro caso, hemos querido centrar nuestro estudio en las posibilidades que pueden ofrecer ciertos instrumentos y recursos computacionales —especialmente tesauros, ontologías, lexicones computacionales y analizadores automáticos— en la presentación, organización y clasificación del material paremiológico presente en las obras paremiográficas ya existentes. Durante el proceso de estudio, observamos que la aplicación de dichos instrumentos no sólo favorece y amplía las posibilidades de distribución de unidades en los repertorios y la recuperación veloz de los datos consultados, sino que se incrementa en gran medida la información semántica disponible para cada paremia.

Para establecer los fundamentos de nuestra propuesta de clasificación de paremias (Alonso Pérez-Ávila, 2009) nos hemos servido de uno de los recursos informáticos para el tratamiento multilingüe de la información más utilizados en el ámbito del procesamiento semántico: el lexicón computacional WordNet (Miller *et al.*, 1990). Hemos empleado la red semántica de WordNet (WN), tejida fundamentalmente por medio de las relaciones de hiponimia y sinonimia, como andamiaje estructural que conforma la base para anotar semánticamente las paremias y clasificarlas. Dicha red está formada por *synsets* –grupos de sinónimos que sirven para representar los conceptos en WN– enlazados por medio de relaciones léxico-semánticas. Cada *synset* ocupa un lugar concreto en las jerarquías hiponímicas que integran la red, de modo que toda la información semántica contenida en los nodos superiores se hereda automáticamente en los inferiores; esta característica permite el desarrollo de varias aplicaciones informáticas. Asimismo, a cada *synset* se le ha añadido además otro tipo de información lexicográfica, como la categoría gramatical, una definición, un ejemplo, y más datos que pueden resultar muy útiles como el dominio al que pertenece (por ejemplo, “economía”, “deporte”, etc.) o información ontológica.

Una primera ventaja del uso de WordNet en las tareas de procesamiento del lenguaje radica en que todos los datos que proceden de los *wordnets* son procesables por programas informáticos de distinta naturaleza. Previamente hemos analizado algunos casos de autores que han intentado proponer un sistema de clasificación u ordenación informatizado, entre ellos, muchos han creado sus propios niveles y categorías (cfr. Pamies *et al*., 1999; Lauhakangas, 2001) para, una vez creados, informatizarlos. En nuestro caso, hemos preferido utilizar un recurso que ya ofrece toda esa información de forma que pueda ser procesada y, así, aprovechar todas las aplicaciones que se han creado en el seno de WN –como las etiquetas de dominio, entre otras muchas– a lo largo de muchos años y en distintos centros de trabajo.

Otro beneficio fundamental de servirse de un recurso como WN consiste en que existen versiones para numerosas lenguas –para nuestra propuesta hemos utilizado los *wordnets* español e italiano–, y así hemos logrado formular una metodología de anotación y clasificación paremiológica que puede aplicarse a lenguas distintas. Una clave numérica, el *offset*, identifica de forma unívoca a cada *synset* en la red semántica y se utiliza el mismo tipo de clave en todas las lenguas. En nuestra propuesta de anotación y clasificación de paremias, los *offsets* constituyen las etiquetas destinadas a la anotación semántica.

Con el fin de ejemplificar la metodología presentada, retomamos una de las paremias con las que se abrió este artículo y de la que se destacó su poca transparencia semántica: *Al doblón no le dan garrote*. Por medio de la puesta en práctica de nuestra propuesta se anotan los términos

doblón y *garrote* con las etiquetas correspondientes de WN –el código numérico incluido entre corchetes constituye el *offset*–, de forma que se vinculará esta paremia a dos dominios semánticos distintos. En nuestra opinión, la información que proporcionan las distintas aplicaciones de los *wordnets* podrá ayudar a esclarecer la oscuridad presente en esta paremia. El resultado de nuestra labor de anotación es el siguiente:

- "Doblón 1[7]", sustantivo, [09646553], hipónimo de "moneda 1, pieza 2", pertenece a la jerarquía léxica del dinero, y el *wordnet* español le adjunta la siguiente glosa: [Antigua moneda de oro equivalente en el siglo XVIII a unas 20 pesetas]. Además, está anotado con la siguiente etiqueta de dominio ofrecida por la aplicación de *WordNet Domains*: [*money numismatics*].
- "Garrote 2", sustantivo, [02742009], en el *wordnet* inglés se ofrece la siguiente definición: [*An instrument for execution by strangulation*]. Se trata de un hipónimo de "instrumento de ejecución 1"[02860305n]. A este *synset* le corresponde la etiqueta de dominio siguiente: [*law*].

Con esta información no se ha obtenido el significado íntegro de la paremia que, como señalamos anteriormente, varía dependiendo de cada contexto de aplicación: puede estar relacionado con el soborno, el poder del dinero, o las diferencias entre ricos y pobres ante la justicia, entre otras posibles explicaciones. Pero mediante este sistema un usuario obtiene valiosa información semántica sobre la paremia, facilitándose así la comprensión de palabras que puede desconocer por encontrarse en desuso –como *doblón* o *garrote*– y además se favorecerá su adscripción a dominios concretos de conocimiento. Este tipo de anotación semántica refleja que esta paremia se vincula tanto al dominio del dinero [*money*] como al de la justicia [*law*], y que forma parte de las jerarquías de las monedas y de los instrumentos de ejecución. Gracias a la vinculación de las paremias a distintos dominios, se incrementa el número de posibilidades de que el usuario pueda localizar la unidad que le interesa, sin las constricciones de las clasificaciones propuestas por los repertorios impresos, en los que una unidad se suele clasificar únicamente bajo una entrada. Además, la información semántica contenida en la anotación puede ser detectada por un programa informático.

En la presentación de las propuestas de autores como Lauhakangas, Pamies *et al*. o Conenna, hemos observado que, en general, se intenta buscar una forma de tratamiento de la paremia como unidad indivisible, como un todo, en la línea que defiende Čermák (2002: 44), quien afirma que deberían incluirse estas unidades *sentence-types* en las ontologías diseñadas en el ámbito de la Lingüística Computacional y la Inteligencia Artificial. Sin embargo, en nuestra aproximación al tratamiento de paremias hemos experimentado la dificultad de proponer principios objetivos y fiables para llevar a cabo una anotación semántica basada en la globalidad del significado de estas unidades. La razón de esta imposibilidad radica en que las paremias no pueden reducirse a un solo concepto pues constituyen enunciados autónomos y son el resultado de varias asociaciones de ideas y, por tanto, han de ser objeto de un proceso de anotación múltiple, como el que proponemos, en el que se dé cuenta de la complejidad de su significado.

Nuestra propuesta tiene como objetivo establecer un sistema de clasificación de paremias basado en la descripción que WordNet proporciona de la red semántica, de modo que se puedan ofrecer unos criterios de ordenación paremiográfica compartidos por distintas comunidades lingüísticas. Asimismo, se añade información semántica a la paremia sustentada en los mismos principios psicolingüísticos que rigen el diseño del lexicón computacional, es decir, se definen las unidades léxicas no de manera absoluta, sino en relación con otras unidades. No se trata de ofrecer un instrumento de detección automática de correspondencias paremiológicas en otros idiomas, ni siquiera se podrá asociar cada paremia a un único concepto (o *synset*) de la red semántica, pues las características propias de dichas unidades no lo permiten –las paremias

[7] El número que acompaña al término indica de qué acepción se trata. Por ejemplo, en este caso, nos encontramos ante la primera acepción de *doblón*.

nunca equivalen a unidades léxicas simples, sino que conforman enunciados autónomos con rango textual. No obstante, sí constituirá un instrumento que ofrezca información muy útil al traductor, de manera inmediata y eficaz, como instrumento para encontrar dichas correspondencias: por medio del *synset* se podrá saber de forma instantánea tanto a qué dominio conceptual está asociada la paremia, como en qué nivel de la jerarquía se encuentra, qué conceptos la incluyen y en qué conceptos se incluye esta a su vez.

CONCLUSIONES

Nos consta que cada vez son más los estudiosos y grupos de investigación interesados en proponer soluciones informáticas para el tratamiento de paremias, ya sea desde perspectivas puramente lingüísticas y traductológicas (Conenna, 1998, 2004) o desde un punto de vista lexicográfico (Pamies *et al.*, 1998). Nuestro propósito no ha sido el de realizar aquí un estudio exhaustivo de todos los trabajos en los que se hace uso de herramientas procedentes de las nuevas tecnologías para el tratamiento de paremias. Así, a estos autores se podrían añadir otros muchos, entre los que recordamos las fundamentales aportaciones de Čermák (1998), quien defiende la adopción del enfoque de la Lingüística de Corpus para analizar el uso de las paremias en función de su aparición en textos reales. Entre las nuevas propuestas que están apareciendo en los últimos años se puede mencionar a Hrisztova-Gotthardt (2008), que también ha realizado una base de datos multilingüe de paremias.

La conclusión que extraemos de los trabajos presentados en este artículo es que a pesar de todas las dificultades que las paremias entrañan para su tratamiento y procesamiento computacional, la Paremiología puede verse notablemente beneficiada de los avances que se están realizando en los últimos años en el ámbito de la Lingüística Computacional. En concreto, el abanico de posibilidades de consulta de distintas recopilaciones paremiográficas a disposición del usuario será muy amplio gracias al apoyo de diversos recursos informáticos.

REFERENCIAS BIBLIOGRÁFICAS

ALONSO PÉREZ-ÁVILA, E. (2008): "*Paremia guardada dos veces ganada*. Criterios de ordenación de paremias en los repertorios paremiográficos españoles e italianos", en A. Punzi e I. Tomassetti (ed.), *Critica del testo*, XI/1-2: 447-467.

ALONSO PÉREZ-ÁVILA, E. (2009): *Criterios de ordenación paremiográfica: estudio y propuesta de anotación para clasificar las paremias españolas e italianas pertenecientes al dominio del dinero*. Tesis Doctoral. Universidad de Turín.

BORDONI, L. y E. ALONSO PÉREZ-ÁVILA (2009): "Computational approaches for studying proverbs", en R. JB Soares y O. Lauhakangas (eds.): *The 2nd Interdiciplinary Colloquium on proverbs. [2. Colóquio Interdiciplinar sobre Provérbios.]*, Tavira, 98-108.

CASARES, J. (1950=1960): "Introducción a la lexicografía moderna", *Revista de Filología española*, Anejo LII: 185-205.

ČERMÁK, F. (1998): "Usage of proverbs. What the Czech National Corpus Shows", en Durčo, P. (ed.): *Europhras '97*. Bratislava: Akademia PZ: 37-59.

ČERMÁK, F. (2002): "Ontologies in today's computational linguistics", en Lewandowska-Tomaszczyk (ed.): *PALC 2001: Practical Applications in Language Corpora*. Peter Lang: 43-47.

CONDE TARRÍO, G. (2001): *Diccionario de refráns*. Vigo: Galaxia.

CONENNA, M. (1998): "Acerca del tratamiento informático de los proverbios", en Luque Durán, J.de D. y A. Pamies Bertrán (eds.): *Léxico y Fraseología*. Granada: Método: 197-204.

CONENNA, M. (2004), "Principes d'analyse automátique des proverbes", en Leclère, C., E. Laporte, M. Piot, M. Silberztein (eds.): *Lexique, syntaxe et lexique-grammaire/ syntax, lexis & lexicon-grammar, Linguisticae Investigationes Supplementa*, 24, 91-105.

COSERIU, E. (1977=1991): *Principios de semántica estructural*. Madrid: Gredos.

FERRO RUIBAL, X. (1987): *Refraneiro Galego Básico*. Vigo: Galaxia.

FRANCESCHI, T., CERVINI, C., GUARINO AMATO, M., MAZO, M. y L. MELIS (1998=2000): *Atlante Paremiologico Italiano: Questionario*. Alessandria: Edizioni dell'Orso.

FRANCESCHI, T. (1994): "Il proverbio e la Scuola Geoparemiologica Italiana", *Paremia*, 3: 27-36.

GARGALLO GIL, J. E., A. TORRES TORRES, V. FRANCO ARCHELERGUES (2008): "Líneas metogolóficas de Badare (Base de datos sobre el refranes del calendario y meteorológicos en la Romania)", en I. Olza Moreno, M. Casado Velarde, R. González Ruiz (eds,): *Actas del XXXVII Simposio Internacional de la Sociedad Española de Lingüística (SEL)*. Pamplona: Servicio de Publicaciones de la Universidad de Pamplona.

GROSS, M. (1975): *Méthodes en syntaxe*. Paris: Hermann.

HOLBEK, B. (1969): "Computer classification of proverbs", *Proverbium*, 14: 372-376.

HRISZTOVA-GOTTHARDT, H. (2008): "Multilingual Proverb Database", en Soares, R. J., O. Lauhakangas (eds.): *1º Colóquio Interdisciplinar sobre Provérbios – Actas*, Tavira, Tipografia Tavirense: 187-196.

KUUSI, M. (1972): "Towards an International Type-System of Proverbs", *Proverbium*, 19: 700-735.

JURAFSKY, D., J.H. MARTIN (2000): *Speech and Language Processing – An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall: 501-666.

LAUHAKANGAS, O. (2001): *The Matti Kuusi international type system of proverbs*. Helsinki: Academia scientiarum fennica.

LÁZARO CARRETER, F. (1981=1980): *Estudios de lingüística*. Barcelona: Crítica.

MARTÍNEZ KLEISER, L. (1953=1989): *Refranero general ideológico español*. Madrid: Hernando.

MILLER, G., R. BECKWITH, C. FELLBAUM, D. GROSS, K.J. MILLER (1990): "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography*, 3 (4): 235-244.

PAMIES, A., M. BÁLMACZ, E. Mª IÑESTA (1998): "Criterios para una fraseografía onomasiológica automatizable", en Luque Durán, J. de D., A. Pamies Bertrán (eds.): *Léxico y Fraseología*. Granada: Método: 111-124.

SEVILLA MUÑOZ, J., M. SEVILLA MUÑOZ (2000): "Técnicas de la 'traducción paremiológica' (Francés-Español)", *Proverbium*, 17, Universidad de Vermont: 369-386.

SEVILLA MUÑOZ, J.; ZURDO, M. I. T. (dir.) (2009): *Refranero multilingüe*. Madrid: Centro Virtual Cervantes. http://cvc.cervantes.es/lengua/refranero/ [Consulta del 2/3/2010].

Páginas web consultadas

http://stel.ub.edu/badare/ [Consulta del 20/02/2010]

http://lauhakan.home.cern.ch/lauhakan/int/cerpint.html [Consulta del 9/2/2010].