RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 54 (2), II Sem. 2016, pp. 85-112.

CL ISSN 0033 - 698X

UN CORPUS DE PARÁFRASIS EN ESPAÑOL: METODOLOGÍA, ELABORACIÓN Y ANÁLISIS[*]

A CORPUS SPANISH PARAPHRASE: METHODOLOGY, PROCESSING AND ANALYSIS

MARGARITA A. MOTA MONTOYA
Grupo de Ingeniería Lingüística (GIL)
Instituto de Ingeniería (IINGEN)
Universidad Nacional Autónoma de México (UNAM)
margaritamotamontoya@gmail.com

IRIA DA CUNHA
Departamento de Filologías Extranjeras y sus Lingüísticas
Facultad de Filología
Universidad Nacional de Educación a Distancia (UNED)
iriad@flog.uned.es

FERNANDA LÓPEZ-ESCOBEDO
Licenciatura en Ciencia Forense (LCF)
Facultad de Medicina (FM)
Universidad Nacional Autónoma de México (UNAM)
flopeze@unam.mx

RESUMEN

En este artículo se describe el proceso de elaboración de un corpus de paráfrasis para el español. Se describe la metodología empleada para llevarlo a cabo, haciendo hincapié en la especificación de los tipos de paráfrasis involucrados en cada nivel parafrástico y en los recursos lingüísticos utilizados. Una vez constituido el corpus, se realiza un análisis cuantitativo y cualitativo detallado de los fenómenos lingüísticos observados en el mismo.

Palabras clave: Paráfrasis, corpus, Procesamiento del Lenguaje Natural (PLN).

[*] Este trabajo fue posible gracias al apoyo del Consejo Nacional de Ciencias y Tecnología (CONACyT) dentro del proyecto *Detección y medición automática de similitud textual* con la clave CB2012/178248. De igual manera, esta investigación ha sido parcialmente financiada por un contrato de investigación Ramón y Cajal (RYC-2014-16935) y el proyecto de investigación APLE 2 (FFI2009-12188-C05-01) del Instituto Universitario de Lingüística Aplicada (IULA).

ABSTRACT

This work outlines the process of elaboration a paraphrase corpus for the Spanish language. It describes the methodology used for creating the corpus emphasising on one hand, the specification of the types of paraphrases involved in each level and, on the other, the linguistic resources used. Finally, a quantitative and qualitative analysis of the linguistic observed phenomena in the corpus is performed.

Keywords: Paraphrase, corpus, Natural Language Processing (NLP).

Recibido: 04.08.2016. Aceptado: 25.10.2016.

1. INTRODUCCIÓN

La paráfrasis ha sido un tema de gran interés en los últimos años en el ámbito del Procesamiento del Lenguaje Natural (PLN), ya que parafrasear automáticamente un texto es un recurso utilizado en muchas de sus aplicaciones, como la búsqueda de respuestas, el resumen automático y la traducción automática. Por este motivo, se han realizado muchas investigaciones en este sentido, como por ejemplo sistemas automáticos que generan paráfrasis, diferentes clasificaciones de paráfrasis y corpus de paráfrasis. Sin embargo, para el español son pocos los trabajos existentes hasta el momento.

El presente trabajo tiene como referente la investigación realizada por Castro, Sierra, Torres-Moreno y Da Cunha (2011), en la que se presentó un método de detección de similitud textual que se evaluó empleando un corpus que incluía 12 textos originales en español parafraseados a dos niveles. La paráfrasis de nivel bajo incluía únicamente variación léxica y la paráfrasis de nivel alto contenía variación léxica, sintáctica, de organización textual o discursiva, y fusión o separación de oraciones. Dado que el tamaño de ese corpus era reducido, en este trabajo se planteó ampliarlo para aumentar su representatividad y ponerlo a disposición de investigadores de diferentes áreas del PLN.

Así, el presente artículo tiene un doble objetivo. Por un lado, el primer objetivo es describir la metodología empleada para constituir el corpus, haciendo énfasis en la especificación de los tipos de paráfrasis involucrados en cada nivel parafrástico y en los recursos lingüísticos utilizados. Por otro, una vez constituido el corpus, el segundo objetivo es realizar un análisis cuantitativo y cualitativo detallado de los fenómenos lingüísticos observados en el mismo.

En el apartado 2 se realiza una revisión de los trabajos sobre la definición de paráfrasis, las diferentes clasificaciones y los diferentes corpus de paráfrasis. En el apartado 3 se expone la metodología utilizada para la elaboración del corpus de

paráfrasis en español. En el apartado 4 se analiza el corpus y se obtienen resultados. Finalmente, en el apartado 5 se plantean las conclusiones.

2. ESTADO DEL ARTE

2.1. Definiciones de paráfrasis

Se ha definido la paráfrasis como expresiones, formas lingüísticas o verbalizaciones alternativas que transmiten la misma información de una expresión original dentro de un idioma (Fujita, 2005; Bannard y Callison-Burch, 2005; Zhou, Lin, Munteanu y Hovy, 2006); o bien como la reescritura del contenido mientras se preserva el significado original (Burrows, Potthast y Stein, 2013).

Por su parte, Barrón-Cedeño, Vila, Martí y Rosso (2013) agregan que, si bien las paráfrasis transmiten la misma información o contenido, también se debe considerar a aquellas que transmiten aproximadamente el mismo significado o contenido equivalente, lo que Bhagat (2009) denomina como "cuasi-paráfrasis". Asimismo, Milićević (2007) señala que la paráfrasis es la relación que une dos frases o expresiones lingüísticas (sintagmas, periodos sintácticos) cuasi-sinonímicas. También señala que la relación de paráfrasis no sólo se da en dos pares de expresiones de una misma lengua (paráfrasis intralingüística), sino también entre pares de expresiones de lenguas diferentes (paráfrasis interlingüística). Estas definiciones se han dado en el área del PLN y resultan vagas o generales, ya que definir el concepto de paráfrasis no es el fin de las investigaciones que se realizan en este ámbito. Dichas investigaciones se centran en desarrollar técnicas para la generación y comprensión automática del lenguaje natural.

La paráfrasis también se ha estudiado desde la perspectiva lingüística, especialmente en el análisis del discurso. Bajo este enfoque, se ha definido como la reformulación en una situación discursiva con un texto fuente de partida del cual se produce un texto nuevo. Algunos autores denominan este fenómeno como paráfrasis reformulativa o reformulación parafrástica, que se lleva a cabo por razones discursivas como énfasis, corrección o clarificación, además de contribuir a la cohesión y desarrollo discursivo (Milićević, 2007 y Barbeito, 2013).

Después de la revisión de diferentes definiciones se puede concluir que para la elaboración de paráfrasis es necesario el uso de conocimiento lingüístico, ya que la paráfrasis es un fenómeno que involucra una amplia gama de mecanismos (morfológicos, léxicos, semánticos, sintácticos y discursivos) con la finalidad de mantener el mismo significado o significado equivalente entre diferentes expresiones lingüísticas (palabras, frases, oraciones, segmentos discursivos).

2.2. Clasificaciones de paráfrasis

Las definiciones que se han dado sobre el concepto de paráfrasis son generales. Respecto a las propuestas de clasificación sobre paráfrasis, existe una gran variedad de enfoques: desde clasificaciones muy exhaustivas hasta clasificaciones muy generales. Teniendo en cuenta el aspecto lingüístico en el que se centran, se pueden mencionar la de Dras (1999), que se basa en la perspectiva sintáctica; la de Bhagat (2009), que se centra en los mecanismos léxicos, y la de Fujita (2005), que se centra tanto en los mecanismos léxicos como en los sintácticos de la paráfrasis.

Algunas clasificaciones resultan menos complejas, pues únicamente enlistan tipos de paráfrasis que son útiles para un sistema o aplicación específica, o los tipos más comunes encontrados en un corpus. Tal es el caso de los trabajos de Barzilay, Mckeown y Elhadad (1999); Kozlowski, McCoy y Vijay-Shanker (2003); Rinaldi, Dowdall, Kaljurand, Hess y Mollá (2003); Dorr et al. (2004) y Boonthum (2005). Otras clasificaciones son más generales, con dos o tres tipos de paráfrasis, como la de Shimohata (2004).

Por su parte, tanto Vila, Martí y Rodríguez (2011, 2014) como Barrón-Cedeño et al. (2013) realizan clasificaciones que abarcan un gran número de fenómenos parafrásticos, con el objetivo de entender este fenómeno, pero sin ser exhaustivas. Estas clasificaciones no representan una simple lista de fenómenos parafrásticos, sino que están basadas en una reflexión lingüística. Como antecedente de estas clasificaciones que ofrecen una visión amplia e inclusiva de la paráfrasis, se encuentra Barrón-Cedeño, Vila y Rosso (2010). Además, mirando en dirección al presente trabajo, la importancia de esta última clasificación es que se realizó en español.

2.3. Corpus de paráfrasis

Sierra (2008: 445) define corpus lingüístico como "la recopilación de un conjunto de textos –escritos y/u orales– basada en determinados criterios con el objetivo de realizar análisis lingüísticos".

Este mismo autor señala que los corpus tienen un gran valor para las investigaciones lingüísticas, pero que "su importancia trasciende este ámbito y es materia de interés tanto para la lingüística teórica como aplicada, así como para las investigaciones y desarrollos en el PLN" (Sierra, 2008: 446).

Existen diversos corpus que incluyen paráfrasis, sin embargo, la mayoría se encuentra en inglés. Generalmente, los corpus de paráfrasis están compuestos por textos literarios o notas periodísticas de otras lenguas traducidos al inglés, o bien, notas periodísticas en inglés que narran el mismo evento. A continuación se hace referencia a los corpus de paráfrasis más relevantes y a los contextos en los que fueron creados.

Barzilay (2003) implementó y evaluó el sistema *MultiGen*, que identifica y sintetiza la información redundante para realizar resúmenes coherentes. Para esta investigación se utilizaron dos tipos de corpus. El primero contiene varias traducciones al inglés de textos literarios escritos por autores extranjeros. Dichas traducciones fueron realizadas por diferentes traductores. El segundo contiene artículos de periódicos sobre el mismo evento descrito por diferentes periodistas. Barzilay explica que se trata de un corpus de paráfrasis porque en el primero las traducciones preservan el significado de la fuente original, aunque pueden usar diferentes palabras y estructuras para transmitir el mismo significado; el cambio también se da por la creatividad del traductor. El segundo también es un corpus parafrástico dado que los artículos explican el mismo evento y coinciden en información, pero se diferencia del primero en que los periodistas seleccionan independientemente las formas lingüísticas para verbalizarlas.

Cohn, Callison-Burch y Lapata (2008) crearon un corpus de 900 pares de oraciones de paráfrasis alineados a nivel palabra o frase. Estos pares fueron compilados de tres fuentes diferentes: el corpus *Twenty Thousand Leagues Under the Sea* (Leagues), el corpus *Multiple-Translation Chinese* (MTC), y el corpus *Microsoft Research Paraphrase* (MSRP). El primero fue creado por Tagyoung Chung y contiene dos traducciones al inglés de la novela francesa *Veinte mil leguas de viaje submarino* escrita por Jules Verne. El segundo fue creado por Huang, Graff y Doddington (2002)[1] y contiene 105 notas periodísticas de tres fuentes en chino mandarín traducidas al inglés. El tercero fue creado por Dolan y Brockett (2005) y contiene 5.801 pares de oraciones en inglés de notas periodísticas. Cada par de oraciones fue analizado por personas, quienes consideraron que 3.900 pares (67%) eran paráfrasis, es decir, equivalentes semánticamente y 1.901 (33%) no lo eran.

El corpus paralelo monolingüe usado por Barzilay y Elhadad (2003) contiene 103 pares de descripciones de ciudades obtenidas de la Enciclopedia Británica y la *Britannica Elementary*. La *Britannica Elementary* contiene generalmente la información presentada en la Enciclopedia Británica; sin embargo, en numerosos casos la entrada de la *Britannica Elementary* contiene información adicional o más fechas. Este corpus fue anotado por dos hablantes nativos de inglés.

A su vez, Burrows, Potthast y Stein (2013) crearon el corpus *Webis Crowd Paraphrase*. Para la generación de paráfrasis se utilizó *Amazon Mechanical Turk* (AMT), un servicio comercial de externalización[2], que actúa como intermediario

[1] https://catalog.ldc.upenn.edu/LDC2002T01

[2] Las empresas o los desarrolladores que necesiten realizar tareas denominadas de inteligencia humana o "HIT" pueden acudir a *Amazon Mechanical Turk* (AMT) para acceder a miles de empleados bajo demanda, de calidad alta, a bajo costo y de todo el mundo. Esta opción es de gran utilidad ya que, a pesar de que la tecnología informática continúa mejorando, siguen existiendo cuestiones que los seres humanos pueden hacer de manera más eficaz que las computadoras, como la identificación de objetos en una foto o un vídeo, la desduplicación de datos, la transcripción de grabaciones de audio o la búsqueda de detalles en los datos. http://aws.amazon.com/es/mturk/

entre los anotadores y los solicitantes. Como textos originales utilizaron 4.067 fragmentos elegidos al azar de 7.000 libros descargados de *Project de Gutenberg*[3]. Para la creación del corpus fue necesario que los anotadores tuvieran fluidez al leer y escribir en inglés.

En el caso de corpus que están relacionados con el plagio se pueden mencionar varios casos. Potthast, Stein, Barrón-Cedeño y Rosso (2010) crearon PAN-PC-10, un corpus que contiene 700.000 casos de plagio. El 40% de los casos son copias exactas y el 60% involucra algún tipo de paráfrasis. El 94% de los casos de este corpus de paráfrasis fueron generados automáticamente y el 6% manualmente. Los casos de paráfrasis creados manualmente se recopilaron mediante AMT. Los anotadores debían tener fluidez en inglés, tanto para leer como para escribir. Se les solicitó que rescribieran el texto original con la instrucción de que esta versión rescrita debía tener el mismo significado que el original, pero debía incluir palabras y frases diferentes.

Clough y Stevenson (2011) elaboraron el corpus *Plagiarized Short Answers*, que consiste en 95 respuestas de entre 200 y 300 palabras a preguntas de ciencias de la computación, en las cuales el plagio tuvo que ser simulado. Como textos fuente se tomaron cinco artículos de Wikipedia. Para la creación del corpus participaron 19 anotadores que eran hablantes nativos del inglés y no nativos.

Castro et al. (2011) presentaron un método de detección de similitud textual basado en el discurso y la semántica. Dentro de esta investigación crearon un corpus que se compone de 12 textos en español, obtenidos de Wikipedia, de revistas científicas y de periódicos. Asimismo, los textos contienen tres temáticas: sushi, sexualidad y astronomía. Estos textos fueron parafraseados en nivel bajo, que consistía en variación solamente léxica, y nivel alto, que consistía en variación léxica, sintáctica, de organización textual o discursiva y fusión o separación de oraciones.

Barrón-Cedeño et al. (2013) crearon el corpus de paráfrasis *Paraphrase for Plagiarism* (P4P). P4P contiene una parte de los casos de plagio del corpus PAN-PC-10 anotados manualmente con base en la tipología parafrástica que crearon. Este corpus contiene 847 pares fuente-plagio en inglés.

3. METODOLOGÍA

3.1. Selección del corpus

Los textos que forman parte del corpus que se elaboró en esta investigación se extrajeron del *RST Spanish Treebank*[4] (Da Cunha, Torres-Moreno y Sierra, 2011),

[3] Es un proyecto cuya finalidad ha sido crear una biblioteca de libros electrónicos gratuitos. Este proyecto fue creado por Michael Hart en 1971.

[4] http://corpus.iingen.unam.mx/rst/

un corpus en español anotado con relaciones del discurso. Se conforma de textos especializados de múltiples ámbitos que incluyen tres niveles de especialización, siguiendo la clasificación de Cabré (1999): nivel alto, donde tanto el autor como el receptor del texto son especialistas del ámbito (por ejemplo, artículos científicos, actas de congresos, tesis doctorales, etc.); nivel medio, donde el autor del texto es un especialista del ámbito y el receptor es un estudiante o un aprendiz (por ejemplo, libros de texto, manuales, etc.); nivel bajo, donde el autor del texto es un especialista y el receptor es el público en general (por ejemplo, artículos y reportajes de divulgación, sitios web de asociaciones, etc.).

En el caso de nuestro corpus se tomaron 12 textos de tres dominios muy diferentes: matemáticas, psicología y sexualidad. En total se emplearon 36 textos de una longitud entre 27 y 193 palabras. Los textos de matemáticas son resúmenes de artículos científicos (*abstracts*) de las revistas *Miscelánea Matemática* (revista de divulgación de la Sociedad Matemática Mexicana) y *Divulgaciones Matemáticas* (revista de la Universidad de Zulia, Venezuela). Los textos de psicología también son *abstracts* de la *Revista Electrónica de Psicología de Iztacala.* Así, tanto los textos de matemáticas como los de psicología son textos especializados de nivel alto. Por el contrario, los textos del dominio de sexualidad son parte del *Periódico Mural*, un medio de difusión que el Departamento de Salud Pública de México ofrece a la comunidad de la Facultad de Medicina. Estos textos están destinados a estudiantes de dicha facultad y se emplean como complemento en la educación allí impartida. Por lo tanto, los textos de sexualidad son textos de especialización de nivel medio.

3.2. Marco teórico

En este trabajo se tuvieron en cuenta dos marcos teóricos: por un lado, la Teoría Comunicativa de la Terminología (TCT) de Cabré (1999) y, por otro, la *Rhetorical Structure Theory* (RST) de Mann y Thompson (1988).

Dado que este corpus está compuesto por textos especializados de nivel alto y de nivel medio, se tomó como marco teórico la TCT, que admite la variación lingüística utilizada para la paráfrasis baja (PB). La TCT concibe la terminología como una materia interdisciplinar e intenta explicarla dentro de una teoría del lenguaje, una teoría de la comunicación y en una teoría del conocimiento. Así, el objeto de estudio de la TCT son los términos, a los que define como unidades singulares y a la vez similares a otras unidades de comunicación dentro de un esquema global de la representación de la realidad, ya que considera que los términos forman parte del lenguaje natural y de la gramática de cada lengua. Por lo tanto, admite la variación conceptual y denominativa de los términos. La variación conceptual se refiere al carácter polisémico de los términos, puesto que algunos se pueden usar en diferentes ámbitos o en el ámbito especializado y en la comunicación

general. Respecto a la variación denominativa consiste en la sinonimia, es decir, formas alternativas de denominación del mismo concepto; sin embargo se señala que las relaciones de sinonimia pueden tener un valor similar o muy distinto de acuerdo con el contexto (Cabré, 1999; 2001).

Respecto a la RST, es una teoría de análisis discursivo mediante la cual es posible caracterizar la estructura jerárquica de un texto. Para realizar análisis con la RST se tienen en cuenta los segmentos discursivos (que pueden ser oraciones o partes de ellas), las relaciones discursivas o retóricas, y la estructura discursiva jerárquica del texto. Los segmentos discursivos también se denominan Unidades Discursivas Mínimas, en inglés *Elementary Discurse Units* (EDUs) (Marcu, 2000). Las características de las EDUs pueden variar dependiendo del analista; en nuestro caso se tomaron las que aparecen en Da Cunha e Iruskieta (2010). Por lo tanto, las EDUs deben incluir un verbo, ya sea en forma conjugada, en infinitivo o en gerundio. Asimismo, no se consideran EDUs a las oraciones de relativo, ni las de objeto directo o indirecto. Dependiendo la importancia que tenga una EDU dentro del texto y según su relación con otras EDUs, éstas pueden ser:

a) Núcleo: incluye información relevante para los propósitos del autor.
b) Satélite: incluye información adicional sobre el núcleo del que depende.

A su vez, las relaciones discursivas pueden ser de tipo Núcleo-Satélite (en las que el satélite depende del núcleo) o de tipo Multinuclear (si incluyen varios núcleos al mismo nivel). Ejemplos de relaciones Núcleo-Satélite son condición, reformulación, causa, resultado, elaboración, etc. y de relaciones Multinucleares son lista, secuencia o contraste.

3.3. Selección y entrenamiento de los anotadores

Para la creación del corpus se contó con tres anotadoras, todas pasantes de la licenciatura en Lengua y Literaturas Hispánicas de la Facultad de Filosofía y Letras de la UNAM. Se realizó una primera propuesta de clasificación de los fenómenos parafrásticos basada en investigaciones de otros autores, principalmente en Castro et al. (2011), Barrón-Cedeño et al. (2010) y Barrón-Cedeño et al. (2013), además de propuestas propias, y se generaron ejemplos para cada tipo o subtipo parafrástico. Asimismo, se señaló la combinación de algunos fenómenos parafrásticos.

La primera propuesta de clasificación fue modificada a lo largo de la creación del corpus, ya que algunos fenómenos no resultaban claros y confundían a las anotadoras. Por ejemplo, la definición de eliminación de contenido proposicional es vaga. Barrón-Cedeño et al. (2010: 11) la definen como "eliminación de una o

más piezas léxicas". Para esclarecer este fenómeno, se definió como eliminación de verbos y, como consecuencia, transformación de la oración. Sin embargo, la confusión permanecía, pues no se entendía este fenómeno y no se diferenciaba de otros.

El ejemplo ofrecido es:

a) Juan **hizo un intento** para dejar de fumar.
b) Juan **intentó** dejar de fumar (Barrón-Cedeño et al. 2010: 11).

Visto desde otra perspectiva, este ejemplo puede clasificarse dentro del fenómeno de cambio de sustantivo a verbo y, como consecuencia, la eliminación de palabras (*hizo un*). La principal duda es a qué se refieren Barrón-Cedeño et al. (2010) cuando hablan de contenido no proposicional. Raúl Rodríguez (2013: 9) define contenido proposicional como la información en un sentido fáctico, es decir, "el tipo de información que puede ser verdadera o falsa". Aunado a esto, Pérez Jiménez (1998: 262) menciona que "existen elementos que no contribuyen al significado proposicional del enunciado como los adverbios y adjetivos, específicamente los evaluativos". Lo anterior lo ejemplifica en:

a) Lamentablemente, ese joven delincuente robó mi coche.

La autora explica que los valores de verdad se determinan en la oración *ese joven delincuente robó mi coche*, mientras que el adverbio evaluativo *lamentablemente* no forma parte del contenido proposicional, ya que el enunciado es verdadero, si es verdad que el delincuente hurtó el coche, independientemente de que se juzgue un acto lamentable.

También causaban confusión los fenómenos de inserción y eliminación de adjuntos, así como la inserción y eliminación de especificadores. La confusión principal era ocasionada por la definición de adjuntos y especificadores. Adjuntos se definen como complementos no seleccionados, pero compatibles con las características semánticas de los núcleos. Los adjetivos y las oraciones de relativo se interpretan como adjuntos de los sustantivos, así como los adverbios son adjuntos de los verbos (RAE, 2009: §1.12f). La función de los especificadores, por su parte, es determinar, situar y cuantificar. Se componen de determinantes (artículos, adjetivos demostrativos, posesivos) y cuantificadores (numerales cardinales) (Bosque y Gutiérrez-Rexach, 2009).

Dado que estos cuatro tipos parafrásticos compartían clases de palabras, como el caso de los adjetivos, causaban duda sobre a cuál pertenecían. Además, este tipo de eliminaciones e inserciones no abarcaban varias clases de palabras que

se añadieron o se eliminaron, tales como preposiciones, abreviaturas, etc. Puesto que no eran útiles para la elaboración del corpus, se decidió renombrarlos como eliminación de palabras e inserción de palabras.

También se decidió fusionar varios fenómenos en uno ya que su separación no era útil ni productiva. El cambio de verboide a forma conjugada verbal y el cambio verbal transitivo/intransitivo se fusionaron solamente en el cambio de forma verbal. Asimismo, se redujo la fusión de yuxtapuestas y fusión de oraciones copulativas al tipo parafrástico fusión de oraciones. En la Tabla I se muestra la clasificación final para la anotación.

Tabla I. Propuesta de clasificación de fenómenos parafrásticos[5].

1. Cambios morfo-léxicos	1.1. Cambios morfológicos	1.1.1.	Cambio de flexión (CF)
		1.1.2.	Cambio de derivación (CD)
		1.1.3.	Cambio de composición/descomposición (CC/D)
	1.2. Cambios léxicos	1.2.1.	Sustitución palabra-definición (SP-D)
		1.2.2.	Sustitución por aproximación numérica (SAN)
		1.2.3.	Sustitución por una sigla o acrónimo (SSA)
		1.2.4.	Cambio de forma verbal (CFV)
		1.2.5.	Sustitución de un verbo por un conjunto de elementos equivalentes (SVCEE)
		1.2.6.	Inserción de palabras (IP)
		1.2.7.	Eliminación de palabras (EP)
		1.2.8.	Cambio de orden de palabras (COP)
2. Cambios semánticos		2.1.	Sustitución por sinónimos (SS)
		2.2.	Sustitución por hiperónimos (SHiper)
		2.3.	Sustitución por hipónimos (SHipo)
		2.4.	Sustitución por holónimos (SHol)
		2.5.	Sustitución por merónimos (SM)
		2.6.	Sustitución por antónimos (SA)
		2.7.	Sustitución acción-actante (SA-A)
		2.8.	Sustitución de acción por lugar prototípico (SALP)
		2.9.	Sustitución agente-instrumento (SA-I)
		2.10.	Diferentes formas para realizar el mismo contenido semántico (DFRMCS)

[5] Esta clasificación se basó en la revisión bibliográfica de Dras (1999: 59-75), Barzilay, Mckeown y Elhadad (1999: 553), Boonthum (2005: 2-4), Kozlowski, McCoy y Vijay-Shanker (2003: 3), Bhagat (2009: 30-45), Barrón-Cedeño, Vila y Rosso (2010: 9-12), Vila, Martí y Rodríguez (2011: 87-88), Barrón-Cedeño, Vila, Martí y Rosso (2013: 921-925) y propuestas propias.

Continuación Tabla I.

3. Cambios estructurales	3.1. Cambios sintácticos	3.1.1. Transformación de pasiva/activa (TP/A) 3.1.2. Repetición/elipsis (R/E) 3.1.3. Conmutación de negación (CN) 3.1.4. Inserción de oraciones de relativo (IOR) 3.1.5. Eliminación de oraciones de relativo (EOR)
	3.2. Cambios discursivos	3.2.1. Transformación de discurso directo/indirecto (TDD/I) 3.2.2. Fusión de oraciones (FO) 3.2.3. Segmentación de unidades discursivas (SUD) 3.2.4. Inserción de marcadores discursivos (IMD) 3.2.5. Eliminación de marcadores discursivos (EMD) 3.2.6. Inserción de segmentos discursivos (ISD) 3.2.7. Eliminación de segmentos discursivos (ESD) 3.2.8. Cambio de marcadores discursivos (CMD) 3.2.9. Cambio de orden de segmentos discursivos (COSD)

3.4. Diseño y gestión del procedimiento de anotación

Después de que todas las anotadoras interiorizaron los criterios de anotación, se le asignaron cuatro textos por ámbito, en total, 12 textos a cada una. El tiempo de anotación variaba según la cantidad de términos que contenía el texto, además de la disponibilidad de los recursos lexicográficos (diccionarios de lengua española, diccionario de antónimos y sinónimos), terminológicos (diccionarios especializados y bases de datos terminológicos) y textuales (textos de nivel alto, medio y alto de especialización). Asimismo, algunos textos eran más cortos que otros.

Los criterios empleados para realizar la anotación son similares a los criterios utilizados en la investigación de Castro et al. (2011). Se consideró que la paráfrasis baja (PB) consiste en la sustitución por sinónimos, hiperónimos, hipónimos, merónimos y holónimos. La paráfrasis alta (PA) consiste en la realización de los fenómenos de la PB, además de los fenómenos morfológicos, léxicos, semánticos, sintácticos y discursivos.

El proceso de anotación se realizó de la siguiente manera: el texto original (OR) se dividió en oraciones y éstas, de ser posible, en otros segmentos discursivos. Las anotadoras realizaron la PB y después la PA de cada uno de los segmentos discursivos de cada texto. Sin embargo, si el texto era muy especializado las anotadoras preferían realizar primero la PA, ya que había un mayor número de tipos parafrásticos para elegir, en comparación con los seis tipos parafrásticos de la PB. Posteriormente, realizaron el conteo de los fenómenos involucrados en el parafraseo. Finalmente, buscaron los textos de control (Pno), es decir, textos de las

mismas temáticas de los textos originales y con longitudes similares, pero que no eran paráfrasis[6].

En la creación del corpus se contabilizó el tiempo utilizado para elaborar el parafraseo, tanto de nivel bajo como de nivel alto, como puede observarse en la Tabla II.

Tabla II. Contabilización de horas del parafraseo.

Ámbitos	PB	PA	Total	Pno
Matemáticas	18:00	23:23	41:23	09:29
Psicología	18:42	23:44	42:26	09:15
Sexualidad	21:31	28:42	50:13	07:15
Total	58:13:00	75:49:00	134:02:00	25:59:00

En el ámbito de sexualidad, con un total de 50 horas con 13 minutos, fue en el que se empleó más tiempo para realizar la paráfrasis tanto de nivel bajo (21 horas con 31 minutos) como de nivel alto (28 horas con 42 minutos). La razón es que al tener más recursos lexicográficos, terminológicos y textuales, llevaba más tiempo decidir cuál era la mejor opción para parafrasear; además, buscar un término en las bases de datos terminológicos o dentro de los diversos recursos textuales requería más tiempo.

Por el contrario, la selección de los textos no parafrásticos de sexualidad para el corpus de contraste fue la que consumió menos tiempo, debido a la gran cantidad de textos existentes en este ámbito, como consecuencia del trabajo de difusión para la prevención de las enfermedades de transmisión sexual de las organizaciones o instituciones de salud pública.

Para el ámbito de psicología y para el de matemáticas se emplea un tiempo similar. Sin embargo, es importante destacar que se utilizaron menos tipos parafrásticos en este último, debido al poco conocimiento matemático de las anotadoras, a la falta de recursos lexicográficos y terminológicos, y al nivel de especialización alto de los artículos.

[6] El corpus de contraste permite comparar la similitud textual del texto original y los textos parafraseados (PB y PA) con la similitud textual del texto original con la no paráfrasis (Pno), para evaluar sistemas de PLN.

4. ANÁLISIS CUANTITATIVO Y CUALITATIVO

4.1. Análisis cuantitativo del corpus

En este trabajo, la ji cuadrada se utilizó para determinar la asociación que había entre anotadoras (A, B, C) y ámbitos (matemáticas, psicología y sexualidad). En la Tabla III se muestran los fenómenos parafrásticos permitidos en la PB en relación con las anotadoras. El resultado obtenido fue p-value= 0.00522, por lo tanto, se concluye que los tipos parafrásticos no son independientes de las anotadoras.

Tabla III. Relación de las anotadoras con los fenómenos parafrásticos en la PB.

Anotadoras (PB)	SS	SHiper	SHipo	SM	SHol
Anotadora A	302	1	3	0	0
Anotadora B	409	6	2	0	6
Anotadora C	293	6	5	3	0
Total	1004	13	10	3	6

En la Tabla IV se encuentra la relación de diferentes ámbitos con los fenómenos parafrásticos en la PB. El resultado obtenido fue p-value= 0.002526; por lo tanto, el uso de los tipos parafrásticos en la PB tampoco es independiente de los ámbitos.

Tabla IV. Relación de los ámbitos con los fenómenos parafrásticos en la PB.

Anotadoras (PB)	SS	SHiper	SHipo	SM	SHol
Matemáticas	217	0	0	0	1
Psicología	388	1	3	0	2
Sexualidad	399	12	7	3	3
Total	1004	13	10	3	6

En conclusión, la sustitución por sinónimos, la sustitución por hiperónimos, la sustitución por hipónimos, la sustitución por merónimos y la sustitución por holónimos en la PB dependen de las anotadoras y del ámbito, concretamente por la densidad terminológica del texto.

Lo importante, por lo tanto, será que se cuente con suficientes recursos lexicográficos, terminológicos y textuales de consulta.

En la Tabla V se muestra la relación de las anotadoras y los tipos parafrásticos exclusivos de la PA. El resultado obtenido fue p-value= 3.059e-14; por lo tanto, el uso de los tipos parafrásticos exclusivos de la PA no es independiente de las anotadoras. La elección de los tipos parafrásticos depende del estilo de las anotadoras. Se considera estilo como el conjunto de características en el modo de escribir de una persona que la distingue de las demás.

Tabla V. Relación de las anotadoras y los tipos parafrásticos exclusivos de la PA.

Anotadoras (PA)	CF	CD	DFRMCS	SP-D	CMD	CFV	SVCEE	SA	SAN	CC/D	SSA	TA/P	R/E
Anotadora A	7	21	6	47	13	11	1	2	2	0	1	2	11
Anotadora B	8	39	8	32	7	28	1	1	4	0	2	0	10
Anotadora C	4	17	8	37	4	22	3	0	2	1	1	0	21
Total	19	77	22	116	24	61	5	3	8	1	4	2	42
Anotadoras (PA)	CN	TDD/I	SUD	IP	IMD	ISD	IOR	EP	EMD	ESD	COP	CSD	CN
Anotadora A	0	0	7	32	27	4	15	14	0	0	13	1	0
Anotadora B	0	1	1	113	12	9	13	68	3	1	15	3	0
Anotadora C	1	0	16	116	22	22	18	37	8	0	27	8	1
Total	1	1	24	261	61	35	46	119	11	1	55	12	1

La anotadora A utilizó en más ocasiones la sustitución palabra-definición (47 ocurrencias), la inserción de marcadores discursivos (27 ocurrencias) y el cambio de marcadores discursivos (13 ocurrencias).

La anotadora B fue la que más utilizó los siguientes fenómenos: eliminación de palabras (68 ocurrencias), cambio de derivación (39 ocurrencias) y cambio de forma verbal (28 ocurrencias). Igualmente, esta anotadora fue la única que utilizó el fenómeno de transformación discurso directo/indirecto y la eliminación de segmentos discursivos.

La anotadora C, a su vez, prefirió usar más la inserción de palabras (116 ocurrencias), el cambio de orden de palabras (27 ocurrencias), inserción de segmentos discursivos (22 ocurrencias) y repetición/elipsis (21 ocurrencias). Además, esta anotadora fue la única que utilizó el cambio de composición/descomposición y el de conmutación de negación.

La Tabla VI muestra la relación de los ámbitos y tipos parafrásticos exclusivos de la PA. El resultado obtenido fue p-value= 1.026e-06; por lo tanto, se concluye que el uso de los tipos parafrásticos exclusivos de la PA no es independiente de los ámbitos.

Tabla VI. Relación de los ámbitos y tipos parafrásticos exclusivos de la PA.

Ámbitos (PA)	CF	CD	DFRMCS	SP-D	CMD	CFV	SVCEE	SA	SAN	CC/D	SSA	TA/P	R/E
Matemáticas	7	11	4	30	6	20	2	1	2	0	0	0	10
Psicología	9	41	6	20	12	23	1	2	3	0	2	1	18
Sexualidad	3	25	12	66	6	18	2	0	3	1	2	1	14
Total	19	77	22	116	24	61	5	3	8	1	4	2	42
Ámbitos (PA)	CN	TDD/I	SUD	IP	IMD	ISD	IOR	EP	EMD	ESD	COP	COSD	CN
Matemáticas	0	1	5	42	14	8	21	12	3	0	12	5	0
Psicología	0	0	13	103	15	10	8	60	3	1	21	6	0
Sexualidad	1	0	6	116	32	17	17	47	5	0	22	1	1
Total	1	1	24	261	61	35	46	119	11	1	55	12	1

El fenómeno más recurrente en el ámbito de **matemáticas** fue la inserción de oraciones de relativo, con 21 casos de las 46 apariciones de este tipo en todo el corpus. Las oraciones de relativo permitían agregar información sin modificar los términos, específicamente con las oraciones subordinadas adjetivas explicativas. En muchas ocasiones era difícil realizar la sustitución palabra-definición, que es uno de los tipos parafrásticos más usados en el corpus en el nivel parafrástico alto (PA). Esto se debe a la falta de recursos lexicográficos y terminológicos en este ámbito. La sustitución palabra-definición permitía mantener el mismo significado o significado equivalente en estos textos altamente especializados. Asimismo, eran difíciles de realizar también la eliminación e inserción de palabras, que son los tipos parafrásticos más usados y elementales. Se utilizaron en pocas ocasiones por la falta de conocimiento de este ámbito por parte de las anotadoras, puesto que no sabían si este tipo de modificaciones afectaría la cohesión y coherencia de los textos. Finalmente, el ámbito de matemáticas contenía textos altamente especializados, lo que dificultaba realizar otros fenómenos parafrásticos, pues son textos concisos y la terminología no se presta a la variación. En este ámbito se realizó el único caso de transformación de discurso directo/indirecto.

Los fenómenos más utilizados en el ámbito de **sexualidad** fueron inserción de palabras (116 casos), sustitución palabra-definición (66 casos), inserción de marcadores discursivos (32 casos), cambio de orden de palabras (22 casos), inserción de segmentos discursivos (17 casos) y diferentes formas para realizar el mismo contenido semántico (12 casos). La inserción de palabras, la inserción de segmentos discursivos, e incluso la sustitución de palabra-definición se relacionan con la inserción de marcadores discursivos, ya que al aumentar la información del texto es necesario relacionarla mediante los marcadores discursivos. Respecto a la sustitución palabra-definición, se debe al gran número de recursos terminológicos y textuales que existen en este ámbito.

El tipo denominado *diferentes formas para realizar el mismo contenido semántico* fue posible porque se contaba con más recursos textuales, los cuales eran en su mayoría de nivel de especialización bajo. Gracias a esto era fácil comprender los temas, lo que permitía mayores modificaciones léxicas, como este tipo parafrástico y el cambio de orden de palabras. En este ámbito se encuentra el único caso de cambio composición/descomposición y de conmutación de negación; este último tipo parafrástico es posible porque la negación puede manifestarse de maneras diversas: con determinantes y pronombres (nadie, ninguno, nada), adverbios (no, nunca, jamás, tampoco), conjunciones (ni, sino) y preposiciones (sin), por lo que es posible la alternancia negativa (no vino nadie > nadie vino) (RAE, 2009: §48.1c,d,i; 48.3a).

En el ámbito de **psicología** los fenómenos más utilizados fueron: eliminación de palabras (con 60 apariciones), cambio de derivación (con 41 apariciones), cambio de forma verbal (con 23 apariciones), repetición/elipsis (con 18 apariciones), segmentación de unidades discursivas (con 13 apariciones), cambio de marcadores discursivos (con 12 apariciones) y cambio de flexión (con 9 apariciones).

Los tipos *eliminación de palabras* y *cambios de derivación* están relacionados. Específicamente, esta relación surge cuando en el cambio de derivación se realiza el cambio de adjetivo a sustantivo. Se elimina el sustantivo que es modificado por el adjetivo; ya que dicho adjetivo se convierte en sustantivo en el texto parafraseado.

Asimismo, están vinculados el *cambio de forma verbal* y el *cambio de flexión*, siempre y cuando el cambio de forma verbal incluya cambio de número, ya que esta modificación se encuentra relacionada con el cambio de persona gramatical para mantener la concordancia.

También la segmentación de unidades discursivas y repetición/elipsis están relacionadas; la segmentación se da por la inserción de información y, en diversas ocasiones, fue necesaria la repetición del referente para darle cohesión al texto, además para dar énfasis al tema. El cambio de marcadores se dio principalmente entre marcadores de adición (*además* > *también*, *y* > *asimismo*, *también* > *igualmente*), de concesión (*aunque* > *a pesar de*, *aunque* > *aun cuando*). Aunque menos frecuentes, también se utilizaron los marcadores discursivos consecutivos (*luego* > *después*) y ordenadores (*finalmente* > *para terminar*). En este ámbito se encuentra el único caso de eliminación de segmentos discursivos. El hecho de que sólo exista un caso de este tipo parafrástico se debe a que es el más radical, pues afecta el significado de la expresión fuente.

En conclusión, los tipos parafrásticos exclusivos de la PA se relacionan con las anotadoras y con los ámbitos, es decir, si en el ámbito es posible efectuar los tipos parafrásticos y si las anotadoras los eligen, pueden llevarse a cabo.

En la Tabla VII se observa que, en general, los tipos parafrásticos que más se utilizaron en el corpus fueron la inserción de palabras, la eliminación de palabras,

la sustitución de palabra-definición, el cambio de derivación, el cambio de forma verbal, la inserción de marcadores discursivos, el cambio de orden de palabras, la inserción de oraciones de relativo y la repetición/elipsis.

Tabla VII. Porcentaje de aparición de los fenómenos parafrásticos en todo el corpus.

Fenómenos	Porcentaje de aparición
Inserción de palabras	25.81%
Eliminación de palabras	11.77%
Sustitución de palabra-definición	11.47%
Cambio de derivación	7.61%
Cambio de forma verbal	6.03%
Inserción de marcadores discursivos	6.03%
Cambio de orden de palabras	5.44%
Inserción de oraciones de relativo	4.54%
Repetición/elipsis	4.15%
Inserción de segmentos discursivos	3.46%
Cambio de marcadores discursivos	2.37%
Segmentación de unidades discursivas	2.37%
Diferentes formas para realizar el mismo contenido semántico	2.17%
Cambio de flexión	1.90%
Cambio de orden de segmentos discursivos	1.20%
Eliminación de marcadores discursivos	1.10%
Sustitución por aproximación numérica	0.80%
Sustitución de un verbo por un conjunto de elementos equivalentes	0.50%
Sustitución por una sigla o un acrónimo	0.40%
Sustitución por antónimos	0.30%
Transformación de activa/pasiva	0.20%
Cambio composición/descomposición	0.10%
Conmutación de negación	0.10%
Transformación de discurso directo/indirecto	0.10%
Eliminación de segmentos discursivos	0.10%

Entre los fenómenos menos frecuentes del corpus se encuentran la sustitución por aproximación numérica, la sustitución de un verbo por un conjunto de elementos equivalentes, la sustitución por una sigla o un acrónimo, y la transformación de activa/pasiva.

4.2. Análisis cualitativo

4.2.1. Inserción de palabras

Las categorías que más se insertaron fueron: conjunto de palabras[7] (135 casos), verbos (39 casos), adjetivos (22 casos), frases sustantivas (14 casos), frases preposicionales (14 casos) y preposiciones (11 casos). Aunque con menor frecuencia pero recurrentes: sustantivos (10 casos), adverbios (6 casos), pronombres (5 casos), frases adverbiales (1 caso), artículos (1 caso), términos (1 caso), frases adjetivas (1 caso) y siglas (1 caso).

En numerosas ocasiones se añadieron los verbos *haber*, *ser* y *existir*. Según Hernández (2002: 15), con estos verbos se construyen las oraciones existenciales, que "son básicas en las lenguas". Asimismo, menciona que las oraciones existenciales presentan entidades del discurso y tienen una función de señalamiento espacial, es decir, ubican entidades en un espacio físico o mental.

También se insertó frecuentemente el verbo *tener*, que es una de las maneras de expresar posesión en español. Aguilar (2007: 13) menciona que "la posesión es un concepto constante en el lenguaje", que consiste en "establecer una conexión entre dos entidades basadas en el reconocimiento de que entre ellas existe un vínculo o

[7] La categoría "conjunto de palabras" muestra las siguientes características:

I. No tiene un significado independiente, por lo que no lo consideramos enunciado, ni segmento discursivo. Tampoco se considera en el conjunto de palabras a las frases u oraciones.
II. Puede terminar en artículos (definidos o indefinidos), preposiciones (en su mayoría "de"), adjetivos, adverbios, pronombres relativos o nexos que permiten insertar el conjunto de palabras en la expresión que se parafrasea. Incluso puede terminar en verbo, tanto en su forma conjugada, como en su forma no personal (participio o infinitivo), que a su vez también facilita la inserción del conjunto de palabras.
III. El número de palabras que integran el denominado conjunto de palabras es variable; puede ser dos, tres, cuatro, etc.
IV. Las palabras que integran el conjunto, especialmente los sustantivos, son elegidos porque tienen relación con el ámbito (matemáticas, psicología y sexualidad), específicamente con el tema (matemáticas: método de inducción y de recursión, funciones multivaluadas, función zeta e hipótesis de Riemann, sucesiones baricéntricas, etc.; psicología: teoría del vínculo, diferenciación, celos y el tratamiento cognitivo conductual; sexualidad: ETS, VHB, gonorrea, enfermedad inflamatoria pélvica, sífilis, etapas de la sífilis, VIH).
V. Asimismo, el conjunto de palabras ayuda a la cohesión y coherencia, así como a mantener el significado, ya sea similar o equivalente, de la expresión fuente.

unión intrínseca" (Aguilar, 2007: 15). La autora señala, además, que este verbo expresa diferentes relaciones: parentesco, cualidades, espaciales, eventos, estados físicos y mentales, entre otras.

Retomando a Hernández (2002), la investigadora afirma que los verbos que expresan existencia, locación y posesión están relacionados en español, hecho relevante ya que en el corpus se eliminan este tipo de verbos y se insertan otros del mismo tipo; concretamente se muestra la relación entre los verbos *tener*, *ser* y *existir*.

Por lo que concierne a la inserción de pronombres relativos, ya sean solos o acompañados de verbos, se opta por convertir el contenido en oraciones subordinadas adjetivas especificativas o explicativas. Además, los relativos pueden funcionar también como conjunciones e introducir oraciones subordinadas sustantivas de objeto directo.

La mayoría de la inserción de frases preposicionales en el corpus cumple con la función de complemento adnominal o preposicional, "relación que se presenta cuando un sustantivo determina, aclara o precisa el significado de otro sustantivo" (Gili Gaya, 1983: §159).

Por lo que respecta a las preposiciones, se pueden insertar porque establecen relaciones entre diferentes clases de palabras: un verbo con un sustantivo o un sustantivo con otro sustantivo, entre otras. Esto permite que las palabras vayan determinándose y complementándose mutuamente para formar un conjunto comprensible (Seco, 1989). Las preposiciones que más se insertaron en nuestro corpus fueron *en*, expresando ubicación y *de*, una de las preposiciones más usadas en español, que señalaba regularmente la relación de pertenencia, es decir, que un elemento forma parte de un conjunto o clase.

4.2.2. Eliminación de palabras

Las categorías que más se eliminaron en este tipo parafrástico fueron: conjunto de palabras (29 casos), verbos (25 casos), sustantivos (14 casos), adjetivos (12 casos), frases preposicionales (10 casos), frases sustantivas (9 casos), preposiciones (9 casos), adverbios (4 casos), artículos (3 casos), pronombres (2 casos), términos (1 caso) y frases adverbiales (1 caso).

La eliminación de sustantivos se relaciona principalmente con el cambio de derivación. Al convertir un adjetivo a un sustantivo se elimina el sustantivo que calificaba el adjetivo. La eliminación de adjetivos y verbos también están relacionadas con el cambio de derivación.

La inserción y eliminación de palabras se encuentran relacionadas, en muchos casos, con la eliminación de una palabra porque se opta por la inserción de otra, sin que exista una relación sinonímica, aunque tampoco se realiza un cambio radical en el significado de la expresión fuente.

En el caso de la eliminación de adverbios, Alarcos (1973: 308-309) señala que los adverbios son una clase de palabra cuya función es la adjunción, es decir, son "segmentos de una oración cuya presencia o ausencia no afecta a la estructura esencial de la expresión y además gozan de cierta movilidad". Magaña (2007) concluye que esto se debe a que los adverbios son unidades cuya función es autónoma.

La eliminación de un pronombre relativo solo o con verbo se debe a que se opta por cambiar una oración subordinada adjetiva especificativa o explicativa por una sola oración. La eliminación de preposiciones se debe a varios factores; uno de ellos es que al eliminar o sustituir el verbo se elimina la preposición porque es el complemento del verbo del régimen preposicional (CVRP).

4.2.3. Sustitución de palabra-definición

La mayoría de las sustituciones en los diferentes ámbitos eran unidades terminológicas (105 casos), debido a que los textos eran especializados. También se realizó este tipo de sustitución, aunque en menor cantidad, en unidades léxicas pertenecientes a la comunicación general (11 casos).

4.2.4. Cambio de derivación

El cambio de derivación en el corpus se realizó de sustantivos a verbos, de sustantivos a adjetivos y de sustantivos a sustantivos, asimismo, cambio de adjetivos a sustantivos, de adjetivos a adverbios y de adjetivos a verbos. Los cambios más utilizados fueron de sustantivo a verbo (23 casos), de adjetivo a sustantivo (17 casos) y de verbo a sustantivo (14 casos).

El cambio de sustantivos a verbos se da principalmente a infinitivos, forma no personal del verbo. Este cambio se debe a la forma híbrida del infinitivo, el cual presenta propiedades nominales y verbales.

La relación del cambio de derivación con la sustitución por sinónimos se estableció porque en la mayoría de los casos se realizaba primero la PB, lo que ayudaba a la realización de la PA; además la sustitución por sinónimos fue el cambio más utilizado en la PB.

El cambio de sustantivo a verbo es posible también con una forma conjugada. En el corpus este cambio se hizo a presente, pretérito, pretérito perfecto compuesto y futuro; en la mayoría de los casos se añadió un *se* impersonal, debido que el sujeto no es importante en los textos especializados.

El cambio de verbos a sustantivos tiene relación con la inserción de palabras, especialmente con la inserción de verbos, para mantener el significado.

4.2.5. Cambio de forma verbal

Este tipo parafrástico consistió en el cambio de tiempo. En el corpus se utilizaron: presente, pretérito, futuro, antepresente y copretérito como se muestra en el Gráfico 1.

Gráfico 1. Porcentaje del cambio de tiempos verbales.

El cambio de tiempo verbal se da principalmente entre presente, pretérito y futuro, debido a que son los tiempos que se utilizan para escribir artículos científicos, ya que se narra qué se realizó y qué se hará como trabajo futuro.

También se incluyó el cambio de modo: indicativo y subjuntivo. Además, se consideró el cambio de persona; en el corpus se llevó a cabo concretamente entre la primera persona del plural, la tercera persona del singular y la tercera persona del plural como se muestra en el Gráfico 2. El cambio se realizó en estas personas debido a que son con las que se acostumbra redactar los textos científicos.

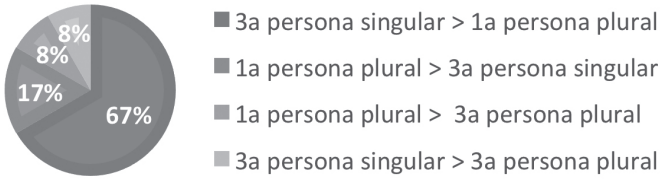

Gráfico 2. Porcentaje del cambio de persona gramatical.

Además, se incluyó el cambio de forma: perífrasis verbal y verboides. El cambio se realizó especialmente entre verboide y verbo conjugado (véase en el Gráfico 3), debido a que un verbo conjugado morfológicamente contiene mayor información tal como: persona, número, tiempo y modo. También se realiza en varias ocasiones de verbo simple a perífrasis, puesto que aporta cierto matiz o alteraciones expresivas al significado del verbo (modal [obligación, posibilidad, etc.] y aspectual [reiteración, duración, etc.]).

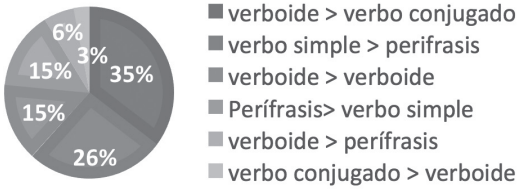

Gráfico 3. Porcentaje del cambio de forma verbal.

4.2.6. Cambio de orden de palabras

El cambio de orden de palabras se da principalmente entre sustantivos (19 casos) y/o términos (20 casos). Ocurre entre sustantivos porque son clases de palabras muy frecuentes que pueden desempeñar la función de sujeto, complemento directo, complemento indirecto, complemento agente y complemento de régimen de un verbo prepositivo. La capacidad de movilidad de esta clase de palabras se debe también a que tienen significado propio; además, el cambio de orden de palabras es posible si el contexto lo permite.

4.2.7. Inserción de oraciones de relativo

Las oraciones de relativo se pueden dividir en especificativas o restrictivas, y explicativas o incidentales. En el caso de la creación del corpus, las oraciones de relativo se utilizan para atribuir a un sustantivo, sobre todo a términos, una cualidad o característica compleja.

En el corpus se usaron, en la mayoría de las ocasiones, oraciones subordinadas adjetivas explicativas (38 casos), ya que como Porto Dapena (1997) menciona, desde el punto de vista semántico, las oraciones adjetivas explicativas añaden información secundaria al antecedente, es decir, no modifican el significado del antecedente o consecuente. Esto es importante ya que la paráfrasis consiste en mantener el mismo significado o significado equivalente.

También se insertaron oraciones adjetivas especificativas (8 casos), aunque en menos ocasiones. Este tipo de oraciones restringe la extensión del significado del antecedente, añadiendo una nueva información imprescindible que ayuda a la caracterización del antecedente, por lo que no se pueden eliminar.

4.2.8. Inserción de segmentos discursivos

En el corpus se usaron las relaciones discursivas de causa (4 casos), reformulación (5 casos) y concesión (1 caso); en ocasiones se insertó sólo el núcleo u oración

principal (3 casos), y segmentos discursivos más amplios de estructura multinuclear (22 casos). La relación de causa se define como la relación entre EDUs en la que "el núcleo es una acción o situación que encuentra su origen en lo que describe el contenido del satélite" (Castro, 2011: 33). La reformulación es la relación que se da cuando "el satélite contiene la misma información que su correspondiente núcleo pero expresada con otras palabras. En muchas ocasiones el satélite tiene mayor extensión al núcleo" (Castro, 2011: 39). La concesión es "la relación en la que el núcleo contiene una afirmación y el satélite aporta cierta información que pareciera negar la validez de lo que se afirma en el núcleo, pero realmente es complementaria" (Castro, 2011: 33-34).

En cuanto a la estructura multinuclear se puede definir como segmento discursivo que contienen más de un núcleo y sólo núcleos (Castro, 2011).

4.2.9. Eliminación de marcadores discursivos

Los marcadores discursivos son elementos marginales, ya que sólo contribuyen al procesamiento (introducir, concluir o finalizar un tema o idea, ordenar una secuencia, dar cuenta de la diferencia entre una idea y otra) de un texto (Portoles, 1998). Por eso se pueden eliminar sin perder excesiva información. En el corpus los tipos de marcadores que más se utilizaron fueron los conectores aditivos (4 casos) y los concesivos (3 casos), esto se debe a que se buscaba unir ideas y a la vez contraponerlas.

4.2.10. Sustitución por una sigla o acrónimo

Para que este fenómeno se pueda realizar es necesario que en el texto se mencionen instituciones o enfermedades, por ejemplo, que puedan expresarse también mediante siglas o acrónimos. En el corpus se utilizó en cuatro ocasiones este tipo parafrástico en el caso de Clínica Universitaria de la Salud Integral (CUSI), Terapia Cognitiva-Conductual (TCC) y en dos ocasiones para sustituir Enfermedad Inflamatoria Pélvica (EIP).

5. CONCLUSIONES

En este artículo se ha descrito la metodología del proceso de elaboración de un corpus de paráfrasis para el español, haciendo hincapié en la especificación de los tipos de paráfrasis involucrados en cada nivel parafrástico y en los recursos lingüísticos utilizados. También se ha realizado un análisis cuantitativo y cualitativo detallado de los fenómenos lingüísticos observados en el corpus constituido.

Realizar paráfrasis es una tarea compleja que requiere reflexión y conocimien-

tos lingüísticos. En esta investigación, el proceso de la creación del corpus se inició con la explicación de los tipos parafrásticos, ya que esto posibilitaba la realización de paráfrasis más complejas, especialmente en la PA.

Cepeda, López y Santoyo (2013) afirman que la imposibilidad de realizar paráfrasis se debe a la carencia o insuficiencia léxica, que en el caso de esta investigación se trató de cubrir con recursos lexicográficos, terminológicos y textuales. Además, explican que esta imposibilidad se debe a la descontextualización o falta de familiaridad respecto de algunos conceptos teóricos. Esto quedó confirmado en el ámbito de matemáticas, puesto que, si bien se realizó la paráfrasis, fue hecha con un número reducido de tipos parafrásticos. Contrariamente, en el ámbito de sexualidad había un gran número de información difundida por las organizaciones e instituciones de salud lo que permitió a las anotadoras una mejor comprensión de los textos y, por tanto, el uso de un número mayor de tipos parafrásticos.

Lo anterior quedó comprobado también en el análisis cuantitativo, pues fue visible que los fenómenos parafrásticos tanto en la PB como en la PA son dependientes del ámbito y de los anotadores. Esto se debe al nivel de especialización de los textos, de los recursos lexicográficos, terminológicos y textuales disponibles, además del estilo del anotador.

Debido a la combinación de los tipos parafrásticos, se concluye que la paráfrasis es un fenómeno lingüístico que involucra una amplia gama de mecanismos (morfológicos, léxicos, semánticos, sintácticos y discursivos) con la finalidad de mantener el mismo significado o significado equivalente entre diferentes expresiones lingüísticas (palabras, frases, oraciones, segmentos discursivos).

Sobre los tipos parafrásticos, además, los que más se utilizaron en todo el corpus fueron la inserción y eliminación de palabras. Tanto Bhagat (2009) como Barrón-Cedeño et al. (2013) unieron estos dos fenómenos en uno solo, ya que se encuentran relacionados entre sí. En este trabajo, por el contrario, se decidió separarlos debido a que cada uno contiene una gran cantidad de diferentes clases de palabras y porque no siempre se relacionan; en algunas ocasiones existía eliminación sin que hubiera inserción o viceversa.

Los fenómenos parafrásticos son difíciles de identificar porque implican conocer la expresión fuente, además de que el investigador debe ser consciente de los cambios realizados en el parafraseo, por lo que se decidió crear las paráfrasis y no sólo recopilarlas.

Este trabajo ha suplido la carencia existente de corpus en español que incluyan paráfrasis realizadas de manera manual. El corpus estará disponible de manera gratuita para la comunidad científica del ámbito del PLN. Precisamente, el trabajo futuro que se plantea es el desarrollo de aplicaciones de PLN tomando como base el corpus desarrollado, como por ejemplo detección automática de similitud textual.

REFERENCIAS

Alarcos Llorach, Emilio. (1973). *Estudios de gramática funcional del español*. Madrid: Gredos.

Aguilar, Nora. (2007). El verbo tener y las relaciones de posesión. Tesis de maestría en Lingüística Hispánica. México: UNAM.

Bannard, Colin y Callison-Burch, Chris. (2005). Paraphrasing with Bilingual Parallel Corpora. En *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 597-604.

Barbeito,Vanina Andrea. (2013). La aposición como estrategia parafrástica. *Boletín de Filología,* 48(1), 11-32.

Barrón-Cedeño, Alberto, Vila, Marta y Rosso, Paolo. (2010). Detección automática de plagio: de la copia exacta a la paráfrasis. En *Panorama actual de la lingüística forense en el ámbito legal y policial: Teoría y práctica. Jornadas (in) formativas de lingüística forense*. Madrid, España: Euphonia Ediciones SL.

Barrón-Cedeño, Alberto, Vila, Marta, Martí, Antonia y Rosso, Paolo. (2013). Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics*, 39(4).

Barzilay, Regina. (2003). Information Fusion for Multidocument Summarization: Paraphrasing and Generation. Tesis de doctorado en filosofía. New York: Universidad de Columbia.

Barzilay, Regina y Elhadad, Noemie. (2003). Sentence Alignment for Monolingual Comparable Corpora. En *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 25-32.

Barzilay, Regina, Mckeown, Kathleen y Elhadad, Michael. (1999). Information Fusion in the Context of Multi-Document. En *Proceedings of the 37th annual meeting of the Association for Computational Linguistic on Computational Linguistics*.

Bhagat, Rahul. (2009). Learning Paraphrases from Text. Tesis de doctorado en Ciencias de la computación. Faculty of the Graduate School University of Southern California.

Boonthum, Chutima. (2005). iSTART: Paraphrase Recognition. *Proceedings of the ACL 2004 workshop on Student research*, Association for Computational Linguistics.

Bosque, Ignacio y Gutiérrez-Rexach, Javier. (2009). *Fundamentos de sintaxis formal*. Madrid: Ediciones Akal.

Burrows, Steven, Potthast, Martin y Stein, Benno. (2013). Paraphrase Acquisition via Crowdsourcing and Machine Learning. *ACM Transactions on Intelligent Systems and Technology* (TIST), 4(3).

Cabré, María Teresa (1999). *La terminología: Representación y comunicación. Ele-*

mentos para una teoría de base comunicativa y otros artículos. Barcelona: Institut Universitari de Lingüística aplicada Universitat Pompeu Fabra.

_____ (2001). *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica*. Barcelona: Institut Universitari de lingüística aplicada Universitat Pompeu Fabra.

Castro, Brenda. (2011). Detección de similitud textual mediante criterios de discurso y semántica. Tesis de licenciatura en Lengua y literaturas hispánicas. México: UNAM.

Castro, Brenda, Sierra, Gerardo, Torres-Moreno, Juan-Manuel y Da Cunha, Iria. (2011). El discurso y la semántica como recursos para la detección de similitud textual. En *Proceedings of the III RST Meeting (8 th Brazilian Symposium in Information and Human Language Technology, STIL 2011)*. Cuiabá, Brasil: Brazilian Computer Society.

Cepeda, María Luisa, López, María del Refugio y Santoyo, Carlos. (2013). Relación entre la paráfrasis y el análisis de textos. En *Revista Electrónica de Investigación Educativa*,15(1).

Clough, Paul y Stevenson, Mark. (2011). Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45(1), 5-24.

Cohn, Trevor, Callison-Burch, Chris y Lapata, Mirella. (2008). Constructing Corpora for the Development and Evaluation of Paraphrase Systems. En *Computational Linguistics*, 34(4).

Da Cunha, Iria e Iruskieta, Mikel. (2010). Comparing rhetorical structures in different languages: the influence of translation strategies. En *Discourse Studies*, 12(5), 563-598.

Da Cunha, Iria, Torres-Moreno, Juan-Manuel y Sierra, Gerardo (2011). On the Development of the RST Spanish Treebank. En *Proceedings of the 5th Linguistic Annotation Workshop 49th Annual Meeting of the Association for Computational Linguistics* (ACL). Portland, Oregon, USA: Association for Computational Linguistics.

Dolan, William B. y Brockett, Chris. (2005). Automatically constructing a corpus of sentential paraphrases. En *Proceedings of the Third International Workshop on Paraphrasing*.

Dorr, Bonnie, Green, Rebecca, Levin, Lori, Rambow, Owen, Farwell, David, Habash, Nizar, Helmreich, Stephen, Hovy, Eduard, Miller, Keith J., Mitamura, Teruko, Reeder, Florence y Siddharthan, Advaith. (2004). Semantic Annotation and Lexico-Syntactic Paraphrase. En *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora, LREC*.

Dras, Mark. (1999). Tree Adjoining Grammar and the Reluctant Paraphrasing of Text. Tesis de doctorado en Filosofía. Australia: Macquarie University.

Fujita, Atsushi. (2005). Automatic Generation of Syntactically well-formed and Semantically Appropriate Paraphrases. Tesis de doctorado en Ingeniería. Nara

Institute of Science and Technology (NAIST).
Gili Gaya, Samuel. (1983). *Curso superior de sintaxis Española.* Barcelona: Vox.
Hernández, Axel. (2002). Las construcciones existenciales con el verbo haber en español: estructura y evolución. Tesis de maestría en Lingüística Hispánica. México: UNAM.
Huang, Graff y Doddington (2002). Multiple-Translation Chinese Corpus. [En línea] Disponible en: https://catalog.ldc.upenn.edu/LDC2002T01
Kozlowski, Raymond, McCoy, Kathleen y Vijay-Shanker, K. (2003). Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. En *Proceedings of the second international workshop on Paraphrasing-* Volume 16, *Association for Computational Linguistics*.
Magaña Juárez, Elsie. (2007). Adverbios temporales durativos: estudio diacrónico de una clase gramatical. Tesis de maestría en Lingüística hispánica. México: UNAM.
Mann, William C. y Thompson, Sandra A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk: Text-Interdisciplinary Journal for the Study of Discourse*, 8(3), 243-281.
Marcu, Daniel. (2000). The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics* 26(3).
Milićević, Jasmina. (2007). *La paraphrase: Modélisation de la paraphrase langagiére*. Alemania: Editions scientifiques internationales.
Pérez Jiménez, Isabel. (1998). Adverbios en -mente y adjetivos circunstanciales en la teoría de la relevancia". En *Interlingüística*, (9). Salamanca: Universidad de Salamanca.
Porto Dapena, José Álvaro. (1997). *Oraciones de relativo*. Madrid: Arcos/Libros.
Portoles, José. (1998). *Marcadores del discurso*. Madrid: Ariel.
Potthast, Martin, Stein, Benno, Barrón-Cedeño, Alberto y Rosso, Paolo. (2010). An Evaluation Framework for Plagiarism Detection. En *Proceedings of the 23rd International conference on computational linguistics: Posters,* 997-1005.
Real Academia Española y Asociación de Academias de la Lengua Española. (2009). *Nueva Gramática Básica de la Lengua Española*. México: Editorial Espasa-Calpe y Planeta.
Rinaldi, Fabio, Dowdall, James; Kaljurand, Kaarel, Hess, Michael y Mollá, Diego. (2003). Exploiting Paraphrases in a Question Answering System. En *Proceedings of the second internacional workshop on Paraphrasing-* Volume 16, *Association for Computational Linguistics*.
Rodríguez, Raúl. (2013). Significado y contexto. Tesis de doctorado en Filosofía. México: Universidad Nacional Autónoma de México.
Seco, Rafael. (1989) *Manual de gramática española*. Buenos Aires: Aguilar
Shimohata, Mitsuo. (2004). Acquiring Paraphrases from Corpora and Its Aplication to Machine Translation. Tesis de doctorado en Ingeniería. Nara, Japón:

Graduate School of Information Science, Nara Institute of Science and Technology.

Sierra, Gerardo. (2008). Diseño de corpus textuales para fines lingüísticos. En *Proceedings of the IX Encuentro Internacional de Lingüística en el Noroeste*, 2, 445-462.

Vila, Marta, Martí, Antonia y Rodríguez, Horacio. (2011). Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach. En *Revista Procesamiento del Lenguaje Natural*.

Vila, Marta, Martí, Antonia y Rodríguez, Horacio. (2014). Is this a paraphrase? What kind? Paraphrase Boundaries and Typology? En *Open Journal of Modern Linguistics*.

Zhou, Liang, Lin, Chin-Yew, Munteanu, Dragos Stefan y Hovy, Eduard. (2006). ParaEval: Using Paraphrases to Evaluate Summaries Automatically. En *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics,* 447-454.